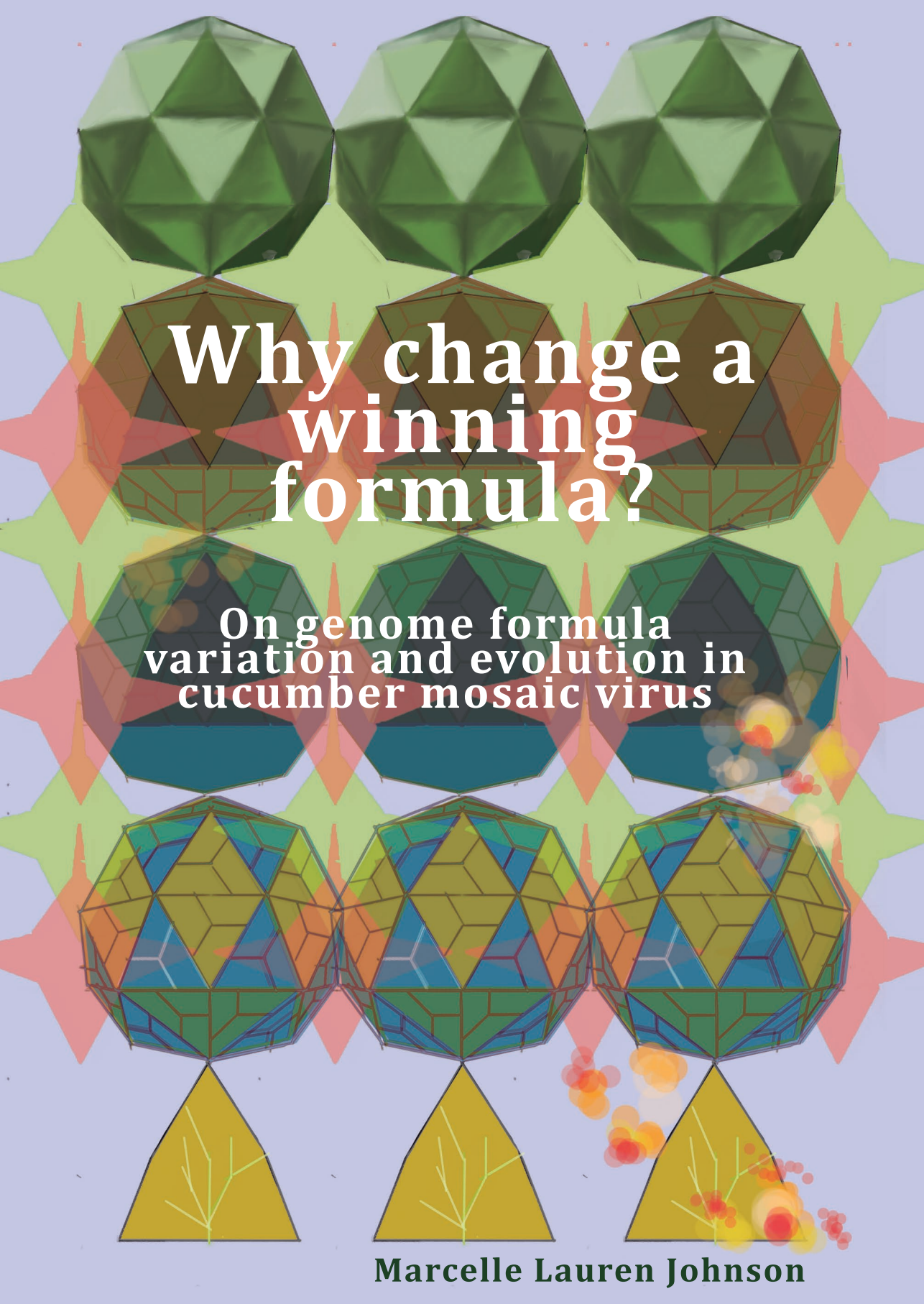


Why change a winning formula?

On genome formula variation and evolution in cucumber mosaic virus

Marcelle Lauren Johnson

Propositions

1. The genome formula space is not unimodal. (This thesis)
2. Genome formula variation is a liability. (This thesis)
3. A good scientist is also a good capitalist.
4. Academia would collapse without the invisible labour of women scientists.
5. Diversity and inclusion initiatives are Kafkaesque.
6. Provocative soundbites from non-experts fuel societal polarisation.
7. Common gardens are the path of most resistance.

Propositions belonging to the thesis, entitled

Why change a winning formula?

On genome formula variation and evolution in cucumber mosaic virus

Marcelle Lauren Johnson

Wageningen, 29 May 2024

Why change a winning formula?

On genome formula variation and evolution in cucumber mosaic virus

Marcelle Lauren Johnson

Thesis committee

Promotors
Prof. Dr R.A.A. van der Vlugt
Special Professor Ecological Plant Virology
Wageningen University & Research

Prof. Dr J.A.G.M. de Visser
Personal chair at the Laboratory of Genetics
Wageningen University & Research

Co-promotor
Dr M.P. Zwart
Senior Researcher, Department of Microbial Ecology
Netherlands Institute of Ecology (NIOO-KNAW), Wageningen

Other members
Dr Y. Michalakis, University of Montpellier, France
Prof. S.F. Elena, Spanish National Research Council (CSIC) and University of Valencia, Spain
Dr P. Dalcin-Martins, University of Amsterdam
Prof. M.E. Schranz, Wageningen University & Research

This research was conducted under the auspices of the Graduate School for Production Ecology and Resource Conservation (PE&RC)

Why change a winning formula?

On genome formula variation and evolution in cucumber mosaic virus

Marcelle Lauren Johnson

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr. C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
On Wednesday 29 May 2024
at 4 p.m. in the Omnia Auditorium.

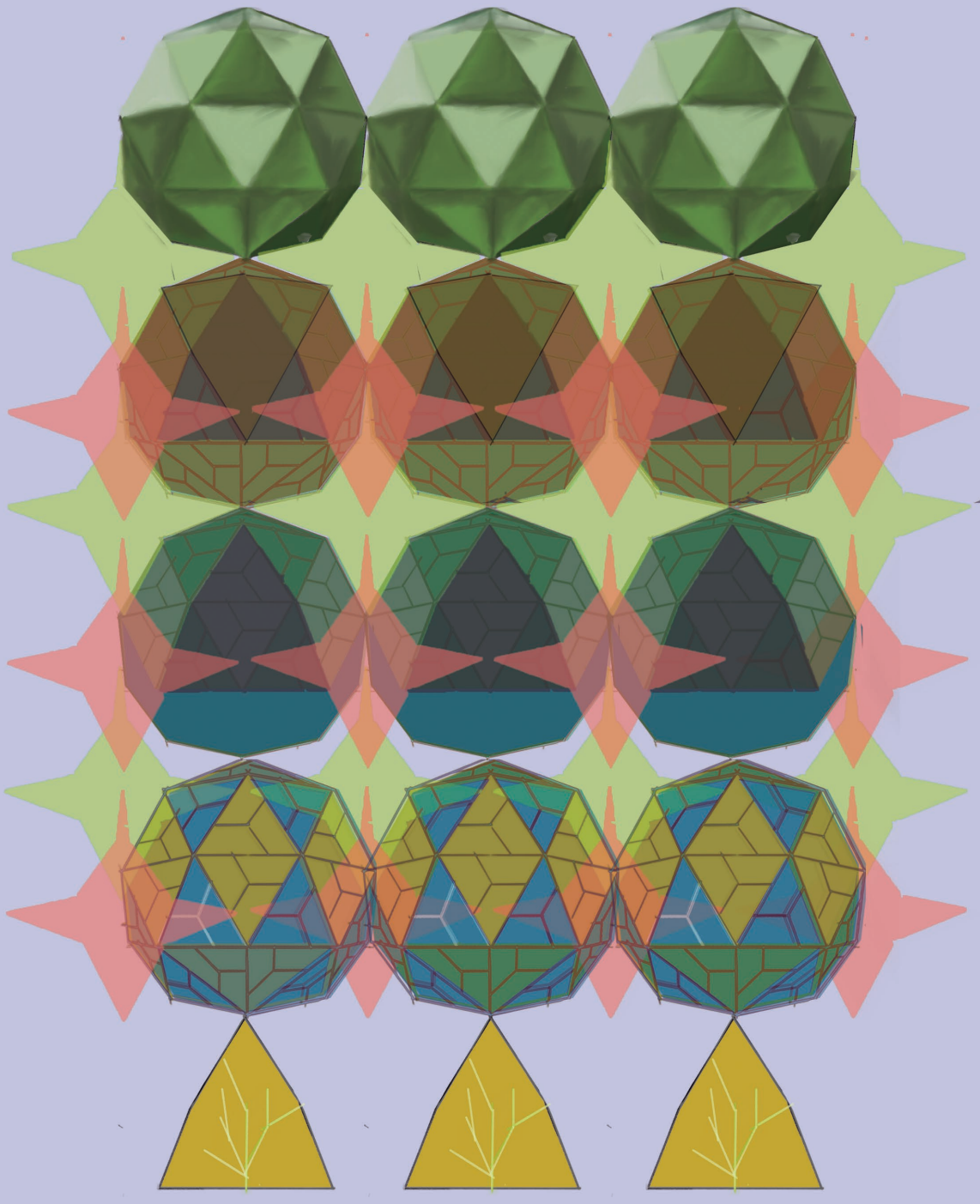
Marcelle Lauren Johnson
Why change a winning formula?
On genome formula variation and evolution in cucumber mosaic virus

236 pages

PhD thesis, Wageningen University, Wageningen, The Netherlands (2024)
With references, with a summary in English

DOI: https://doi.org/10.18174/654529

Table of Contents



General Introduction

Viruses are obligate cellular parasites infecting plant, fungi, animal and bacterial hosts. Viral genomes are highly diverse in terms of the type of nucleic acid (DNA/RNA), whether it is single-stranded (ss) or double stranded (ds), and the gene expression and replication strategy (Baltimore 1971). There are seven Baltimore Classes (BC I – VII): the ds- and ssDNA virus genomes (BC I – II), the positive (+) or negative sense (-) ssRNA and dsRNA virus genomes (BC III – V), and the reverse-transcribing viruses of ssRNA genome (BC VI) or DNA (BC VII) (Baltimore 1971; Koonin, Krupovic, and Agol 2021). The Baltimore classification scheme also describes the various modes of replication by DNA polymerase (BC I – II), RNA-dependent RNA polymerase (RdRp) (BC III – V), and reverse transcriptase (BC VI -VII) (Koonin, Krupovic, and Agol 2021). The diversity of virus genomes extends to the number of genome segments and their packaging strategy into virus particles. Viral genomes can be monopartite, being composed of a single genome segment packaged into a virus particle (Figure 1a). Segmented viruses (Figure 1b) have a varied number of genome segments which are collectively packaged and lastly there are the multipartite viruses; which have individual packaging of each genome segment (Figure 1c).

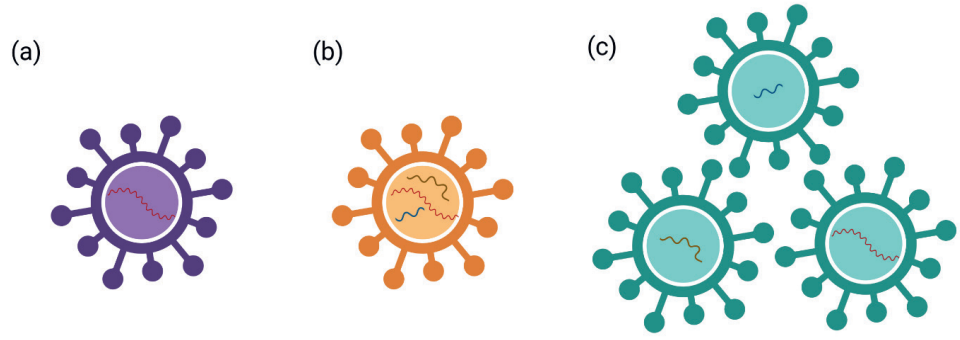


Figure 1. The virus genomes can be divided into several segments which differ in their packaging strategy in virus particles. *(a) Monopartite viruses have a single genome segment packaged into a virus particle (b) Segmented viruses have several genome segments that are co-packaged into one virus particle. (c) Multipartite viruses also have several segments. However, each genome segment is individually packaged and transmitted between hosts. Segmented and multipartite viruses may have genome segments of different lengths. The image presented here is from Chapter 2 of this thesis. Created with Biorender.*

Multicomponent viruses or covirus systems are an overarching term for viral systems which require the presence of more than a single virus particle for infection (Van Vloten-Doting and Jaspars 1977; Van Vloten-doting, Kruseman, and Jaspars 1968; Fulton 1980). This includes segmented viruses, multipartite viruses, accessory elements such as the selfish replicator-satellite viruses and defective-interfering particles, virus segments which are can only replicate by complementation with one another. The segmented and multipartite viruses require co-infection of a core of genomic segments for the initiation of viral replication and subsequent infection in hosts. Segmented virus co-packaging mechanisms ensure that all genomic components are transmitted together as one infectious unit, however, there may be

differences in the fidelity of co-packaging that result in virus particles in which an incomplete genomic set occurs (Nakatsu et al. 2016). In multipartite viruses, each genomic component is individually packaged and transmitted. Thus, the infectious unit is disconnected at each between- and within-host transmission event dependent on the translocation of these infectious units. Therefore, there is the physical decoupling of viral segments of a genome which imposes a potential cost to infection.

Historical overview and state-of-the art in multipartite virus research

There is a long history of research on multipartite viruses, including theoretical propositions on their existence and experimental evidence for their occurrence and dynamics within hosts. Here, I will briefly provide in chronological order an overview of a few key moments since the discovery of multipartite viruses and the contribution to understanding this group of viruses. A number of review papers by Michalakis and Blanc (2020), Lucia-Sanz et al. (2018), Lucia-Sanz and Manrubia (2017), and Sicard et al. (2016) detail progress in this research field since multipartite virus discovery. Historically, multipartite viruses were identified due to their deviation from the independent action hypothesis (IAH), which states that the dose of a parasite is directly related to the number of infectious sites formed (Druett 1952; Bald 1937). Prunus necrotic ringspot virus (PNRSV), an RNA plant virus was confirmed as a multipartite virus by analysing the dose-response relationship (Fulton 1962). In this work, increasing steepness of the dose-response relationship between the amount of the infectious agent and the response variable (local lesion formation) was used to infer the number of particles required to initiate an infection. Thus, viruses which are composed of more than one segment will have dose-response relationships with a steep gradient. (Bruening and Agrawal 1967) demonstrated with the bipartite cowpea mosaic virus (CPMV) that complementation between genome segments increases virus infectivity, and proposed infections of interacting genome segments do not occur in a 1:1 manner but may be composed of an unequal mixture. (Van Kammen 1967) identified not only differences in ratios of three genome segments of CPMV but also apparent stochastic variation in ratios isolated from local lesion infections in *Phaseolus vulgaris*. Genome segmentation was also identified for the bipartite tobacco rattle virus (TRV), where infections consisting of both short and long segments produced higher numbers of local lesions than infections which were enriched for each segment individually (Lister 1968). These experiments, therefore, suggested that there is complementation between segments. However, the contamination of virus particle purifications by each segment with the other makes it difficult to draw definitive conclusions on this work (Lister 1968). (Sänger 1968) was able to demonstrate that complementation between segments was required for successful infection. Later work confirmed that PNSRV has a tripartite genome with segments of unequal length (Loesch and Fulton 1975).

In the 1980s and 1990s, different proposals were made to explain the emergence and benefit of multipartite genomes for viruses, presenting different potential benefits for this strategy. Firstly, this included the notion that dividing the genome into smaller segments increases the number of segments without mutations for RNA virus genomes (Pressing and Reanney 1984) and leads to a reduction of the mutational load (Chao 1991). Secondly, Nee (1987) proposed that for a multipartite virus, the smaller genome segment sizes would allow for faster

replication of each segment, if the availability of the replication machinery is not the limiting factor. Thirdly, Chao (1988) proposed that segmented and multipartite viruses are able to increase genetic diversity in the virus population by the exchange of genome segments (i.e., re-assortment). Reassortment could bring together segments that are not affected by deleterious mutations, or those carrying beneficial mutations. Experimental support for the proposals of Nee (1987), Pressing and Reanney (1984) and Chao (1991) has to date not been demonstrated. Later in this chapter, I discuss and examine the experimental support for the hypotheses presented above, as well as newly proposed hypotheses on the benefit of multipartition.

An experimental study on foot and mouth disease virus (FMDV) showed the spontaneous development of a bipartite form of FMDV from the wild-type monopartite FMDV after serial passaging over 260 rounds at a high multiplicity of infection (MOI) (García-Arriaza et al. 2004). They were also able to show that the bipartite form of the FMDV had higher virus particle stability than FMDV in its' monopartite form, under the specific conditions used for passaging (Ojosnegros et al. 2011).

Results from Sicard et al. (2013) demonstrated that for the octapartite faba bean necrotic stunt virus (FBNSV), viral genome segments accumulate to unequal levels during infection in a host-specific manner, coined the "**genome formula**" (GF) and that it may reach an equilibrium called the "**setpoint genome formula**" (SGF) (Sicard et al. 2013). Sicard and colleagues (2013) put forward the hypothesis that the GF is an adaptive means of regulating viral gene expression in different host environments; based on the assumption that changing segment copy numbers has a direct effect on gene expression. These findings have led to a rejuvenation of experimental reports of unequal virus segment ratios in other systems, discussed later in this chapter.

Following the notable findings of Sicard et al. (2013), focus on multipartite viruses also shifted to finding other host species which might be infected by multipartite viruses. Keeping in mind that multipartite viruses had only been observed infecting plants, there was the confirmation of an insect-infecting multipartite virus, the bipartite ssDNA Bombyx mori bidensovirus (BmBDV) infecting silkworms (Hu et al. 2013). The virus had been discovered decades earlier (Bando et al. 1992; Seki and Iwashita 1983), but its multipartite nature had not been shown. Experiments with BmBDV showed that it also displayed the uneven accumulation of virus genome segments (Hu et al. 2016), providing support for the view that the GF may be a general feature for multipartite and segmented viruses. Additionally, a putative mosquito-infecting virus has been isolated, the guaico culex virus (GCXV) (Ladner et al. 2016). The segmented genome and steep dose-response suggest that his virus is multipartite, but only examination of the virus particles can rule out it is not a segmented virus with low fidelity packaging (Michalakis and Blanc 2020). Experiments with the tripartite +ssRNA alfalfa mosaic virus (AMV) showed that the GF stabilizes to an equilibrium which is associated with increased virus accumulation (Wu et al. 2017).

A key concept in segmented and multipartite virus research, is the need for complementation between segments within the same cell to initiate infection. Experiments on the localization of genome segments of FBNSV in petioles have shown that all segments are not present within the same cell (Sicard et al. 2019). Furthermore, the genome segments present in neighboring cells are able to complement one another. Tentatively the DNA-R segment encoding the

master replication protein (M-Rep) may be able to move from the cell where it is replicated to adjacent cells (although the specific mechanism is still unknown). This is a fascinating finding as one of the primary proposed costs for multipartition has been the requirement for all segments within the same cell, and this cost increases with the number of genome segments (Iranzo and Manrubia 2012). The sharing of gene products across cells may be a way to minimize the cost of multipartition within a host.

Evidence to support the role of the GF in virus gene expression regulation has also been found in FBNSV (Gallet et al. 2022). Briefly, Gallet et al. (2022) observed that GF differences between host species may act to buffer transcript levels. With a host-switching experiment, they show that the associated GF changes are related to changes in its' respective mRNA and that virus transcript levels between the initial and final host are similar to one another, whilst the genomic DNA GFs differ. Di Mattia et al. (Di Mattia et al. 2022) show that in FBNSV host plant infections can be initiated when one of the eight genome segments is missing, and that re-inoculation of the same plant can supplement the missing segment and the restoration of the full genome. Moreover, with an increased time interval between initial infection and re-inoculation, the likelihood of full genome restoration decreases.

Combined, these findings show that multipartite viruses are composed of genome segments which may vary in length. Each segment contains distinct gene functions, and complementation between genome segments is required for infection. Thus far, we've seen that genome segments may differ in their frequency ratio (the GF) during infection, which has been measured in multipartite plant viruses and one insect virus. It has been hypothesized that the GF may be adaptive, conferring advantages for virus gene expression in different host species; however, the evolutionary costs and benefits of the strategy remain unclear, especially as the benefits are not evident and the costs appear to be high. Next, I will discuss the proposed benefits of a multipartite virus genome and the evidence supporting this hypothesis.

Evolutionary benefits of multipartite virus genomes?

Multipartite viruses have an inherently more costly strategy for virus transmission, as a larger viral dose is required compared to monopartite viruses to ensure a full or core genome complement is present, enabling a new infection. This spurred research into the evolution of genome segmentation and its potential benefits. Several theoretical benefits have been put forward for the emergence and benefit of a multipartite virus genome, all of which may be beneficial to segmented viruses alike (Sicard et al. 2016). The benefits proposed in the past 40 years have been summarized in review papers (Sicard et al. (2016), Lucía-Sanz, Aguirre, and Manrubia (2018),). These proposals include: (1) faster replication of whole segmented genomes when polymerase is abundant (Nee 1987), (2) genome segmentation reduces the effect of error-prone RNA virus replication, as smaller template molecules have a lower error burden than longer template molecules (Pressing and Reanney 1984), therefore producing segment copies without errors, (3) the smaller template size of individual genome segments reduces the mutational load (Chao 1991), (4) viruses with segmented and multipartite genomes can rapidly exchange genetic information via reassortment (Chao 1988), a process which may bring beneficial mutations together and also remove deleterious mutations, (5)

increased virus particle stability due to smaller segment size (Ojosnegros et al. 2011), and (6) changing virus segment copy numbers may cause adaptive virus gene expression changes in new hosts (Sicard et al. 2013). The benefits (2) – (4) are reassortment benefits presented here separately within the context of a historical overview and the support at the time, and not a conceptual one as the arguments for (2) – (4) are largely the same, removing deleterious mutations increasing genetic diversity. In the following sections, I further discuss each proposed benefit and the supporting evidence.

Nee (1987) proposed that dividing the genome into smaller segments could drive the evolution of multipartite viruses. These segments would be incapable of autonomous replication but would replicate faster than full-length segments (for conditions when polymerase is not a limiting factor). If they occur readily and there are high levels of coinfection between these segments, these "defective" segments could take over the population, even if they have reduced spread between hosts. This argument has been recapitulated in much greater clarity and detail recently (Leeks et al. 2023). Under this argument, we would not expect to find the benefits of multipartition when considering fitness components that reflect higher levels of selection. Sakai et al. (1999) showed that in Sendai virus (SeV), the insertion of genes of various lengths, which increased overall genome length, were associated with a reduction in the replication speed, providing some evidence for faster replication by shorter segments. Although the arguments are promising, empirical tests of whether within-host competition drives the evolution of multipartition have not been reported.

Proposed benefits (2) – (4) are related to reassortment and address the ability of genome segments to remove deleterious mutations or increase genetic diversity of segmented and multipartite viruses. (Pressing and Reanney 1984), propose that RNA virus replication errors can be reduced (proposed benefit 2) by decreasing the template size, or the segment size for multipartite viruses. This would ensure that more progeny segments would be mutation free. At the time, known multipartite viruses had RNA genomes which were replicated using the RNA-dependent RNA polymerase (RdRP), which is error-prone as it lacks a proof-reading domain (Reanney 1982). The introduction of replication-related errors is not only dependent on the low fidelity of the RdRp, but it is also influenced by the replication mode: whether replication occurs via a single virus template – many copies from single template (stamping machine) or whether each progeny is used as a template for replication (geometric growth) (Martínez et al. 2011). The stamping machine approach will lead to fewer mutations than the geometric growth approach: although the stamping machine requires more replication events than geometric growth, any errors introduced are not amplified as new genome copies are not used for replication in that bout of replication.

For positive-sense RNA viruses, segments serve as the genome and template for translation of viral proteins. Virus replication happens by the RdRp, which causes a higher mutation rate than DNA polymerase (Sanjuán et al. 2010). The mutation rate has been measured for several virus species and estimated to be in the range of 10^{-3} - 10^{-6} mutations per nucleotide per cell infection ($\mu_{s/n/c}$), whilst in DNA viruses, the range is 10^{-8} - 10^{-6} $\mu_{s/n/c}$ (Sanjuán et al. 2010). These high mutation rates have inspired the quasispecies model of RNA virus evolution (Lauring and Andino 2010). High mutation rates result in a virus population that is characterised by a mutant spectrum (containing a large sample of the possible mutations within the mutational neighbourhood of the wild-type virus), and selection acts on the mutant population (Domingo, Sheldon, and Perales 2012). Chao (1991) suggested that reducing the

genome to shorter segments would decrease the target size for mutations because at the time, known RNA viruses had small genomes. Current knowledge suggests that a reduction in total genome size combined with shorter genome segments may reduce the mutational load. However, there is no evidence to support the claim that segmentation alone will reduce the mutation pressure on a viral genome.

Reassortment and recombination may provide benefits for both segmented and multipartite viruses, as a segmented genome increases the opportunity for genetic exchange and may increase genetic diversity (Chao 1988). In multipartite and segmented viruses, reassortment can occur via the exchange of whole genome segments from the same virus species, called "segment shuffling", or homologous recombination between virus segments within the same species or from unrelated virus species (Bujarski 2013; Varsani et al. 2018) . Reassortment and recombination occur at similar levels in multipartite viruses (Bujarski and Kaesberg 1986). Reassortment in the tripartite cucumber mosaic virus (CMV) is constrained to RNA3 and virus subgroups (Fraile et al. 1997), and reassortment and recombination occur at low frequency (Bonnet et al. 2005). Furthermore, even when recombinants and reassortants are readily available within a population, they are selected against and constrained by within-host virus local and systemic infection processes (Fernando Escriu, Fraile, and García-Arenal 2007). Reassorted strains of segmented tomato spotted wilt virus (TSWV) are linked to resistance-breaking (Qiu and Moyer 1999), and bluetongue virus (BTV) reassortment increases virus transmission by the insect vector (Sanders et al. 2022). In Influenza A virus (IAV), a mutation in the polymerase complex along with reassortment of the polymerase complex subunits have increased virus replication and pathogenicity (Mehle et al. 2012). These results indicate that the proposed benefit of reassortment between virus genotypes can be a source of genetic variation. However, not all segments may be easily exchanged, and there may be biases which favour reassortment between specific segments.

A benefit linked to dividing the genome up into smaller segments is increased virus particle stability. For plant viruses, this benefit would be manifest during vector-borne transmission, were viruses that rely on non-propagative transmission would have increased transmission because their virus particles remained intact and infectious for a longer period of time. To my knowledge, there is no experimental evidence for this hypothesis in plant viruses. For animal viruses, during serial passage of foot and mouth virus (FMDV) in cell culture at high multiplicity of infection (MOI), a bipartite version of FMDV spontaneously formed (García-Arriaza et al. 2004). Further studies on the bipartite FMDV showed that it had increased infectivity at the same dose and higher virus particle stability than the monopartite virus form of FMDV (Ojosnegros et al. 2011).

In experimental infections of FBNSV onto *Vicia faba* and *Medicago truncatula*, Sicard et al. (Sicard et al. 2013) demonstrated that the viral genome segments systematically shift to a host-specific ratio, which they called the **genome formula** (GF) (Sicard et al. 2013). The change in segment frequencies may be variable during the course of infection and in different host tissues, until stabilising at the equilibrium **setpoint genome formula** (SGF) (Sicard et al. 2013). Sicard et al. (2013) hypothesize that the change in segment stoichiometry may act as a mechanism for regulating virus gene expression, allowing for the virus to adapt to different host environments. They propose that segment copy number changes are directly proportional to gene expression-level changes for a given host. Furthermore, this can occur in a mutation-free manner, whereby changes in segment ratios occur faster than fixing beneficial point

mutations for virus gene expression. The GF has been measured in several multipartite plant viruses with different numbers of genome segments and nucleic acid types (Wu et al. 2017; Yu et al. 2019; Zhao et al. 2019; Boezen et al. 2023) and in a multipartite insect virus (Hu et al. 2016). The GF has also been measured for the segmented viruses, bluetongue virus (BTV) (ssDNA) (Moreau et al. 2020) and Rift Valley fever virus (RVFV) (Bermúdez-Méndez et al. 2022). In all of these cases, the different segments were never at the same frequency, and an unbalanced GF appeared to be a common phenomenon for multipartite viruses. Preliminary evidence for the adaptive role of the GF comes from experiments with FBNSV and AMV, where segment copies stabilise to the SGF associated with high virus accumulation (Wu et al. 2017; Sicard et al. 2013). Recently, Gallet et al. (2022) have shown that in a host-switching experiment virus transcript levels are less variable than the GF, suggesting that the GF may play a role in maintaining similar viral transcript levels in different host species. Thus, at least for an ssDNA virus, the GF may be one of the molecular tools in its' arsenal to regulate gene expression. However, many questions remain unanswered. One of the key observations from the studies of Sicard et al. (2013) and Wu et al. (2017)is the variability of the GF for a given virus across different host species. How could this genome segment copy number variation be beneficial for multipartite viruses? To answer this question, I will first explore copy number variation in other systems.

Hypothesis: The genome formula as a type of copy number variation in multipartite viruses

Genetic variation may be in the form of copy number variation (CNV), e.g. of genes or chromosomes, increasing or decreasing gene dosage and thereby gene expression (Freeman et al. 2006; Katju, Bergthorsson, and Chain 2013; Lauer and Gresham 2019). CNV may be adaptive at two time scales: changes in expression of certain genes may be adaptive at short time scales e.g. insecticide resistance of the planthopper *Nilaparvata lugens* to imidacloprid (Zimmer et al. 2018), *Candida albicans* antifungal resistance to fluconazole (Todd and Selmecki 2020) and *Escherichia coli* high temperature tolerance (41.5°C) (Riehle, Bennett, and Long 2001) , whereas changes in mutation supplies associated with CNV may be adaptive at longer time scales, e.g. for the evolution of antibiotic resistance (Sandegren and Andersson 2009),

In viruses, CNV in the form of gene duplications has been best described in the monopartite dsDNA vaccinia virus (VACV), a Poxvirus (Bayer, Brennan, and Geballe 2018). During infections of human cells, the host immune response activates the anti-viral Protein Kinase R (PKR) which is upregulated after detection of viral dsRNA (Weber et al. 2006), and initiates phosphorylation of the translation initiation factor eIF2α, which limits protein production and thereby reduces virus replication (Elde et al. 2012). VACV encodes K3L and E3L genes, which inhibit the activity of PKR (Davies et al. 1993). K3L has weak PKR inhibition activity and, using experiments with loss-of-function E3L HeLa cell mutants, increases the selection pressure for increased activity of K3L (Elde et al. 2012). Serial passage experiments showed an increase in virus accumulation over time and sequence analysis revealed amplifications of K3L which correlated with increased virus accumulation (Elde et al. 2012). K3L amplifications may have 15 - 16 K3L copies (Sasani et al. 2018; Elde et al. 2012). The K3L amplification is transient,

but also temporary expansions in the K3L gene increase the target size for beneficial mutations (Elde et al. 2012). The H47R amino acid substitution in K3L was observed in replicate populations and conferred increased virus replication and thus improved virus fitness (Elde et al. 2012). Selection favors genomes harboring H47R and may cause the amplified region to collapse back to single copy K3L (Elde et al. 2012; Sasani et al. 2018). Elde et al. (2012) describe the dynamic change of gene CNV in VACV as the "genomic accordion model". It describes a multi-step process,: (1) selective amplification of K3L, followed by (2) increased activity of K3L to counteract antiviral PKR, and higher mutation rate and (3) selection for genomes which contain the beneficial mutation (4) the fixation of a beneficial mutation which replaces the activity of the amplified region and (5) the reduction in the amplified region to mutant single copy (Elde et al. 2012). Thus, CNV dynamics in VACV follows a trajectory of expansion, fixation of beneficial mutations and contraction which bears a similarity to the CNV innovation-amplification-divergence (IAD) model proposed by (Näsvall et al. 2012)

In the IAD model, the ancestral gene or region of interest is amplified, when enhanced expression is beneficial under new environmental conditions, yielding increased functional activity and fitness (Näsvall et al. 2012). Thereafter, the amplified region is replaced by an allele with a beneficial mutation, which maintains the increased functional activity and benefits as a single copy, causing the concomitant collapse of the amplified wild-type sequences and fixation of the mutant single copy state (Näsvall et al. 2012). The IAD model follows a similar process as the genomic accordian model (Bayer, Brennan, and Geballe 2018). One question I wanted to investigate is whether the GF may allow for CNV-associated benefits of IAD, where changes in segment frequency may lead to increased gene dosage and mutation supply, possibly followed by decreases in segment frequency upon the occurrence of beneficial mutations that alleviate the need for high gene dosage.

Models predict rapid GF change under some conditions and that multipartite viruses can outcompete monopartite viruses in environments that demand specific but different levels of gene expression (Zwart and Elena 2020). By contrast, Gallet et al. (2022) showed the GF may play a role in maintaining similar mRNA levels in different hosts, suggesting that the GF may also play a role in stabilizing gene expression. The mechanism of control of the GF remains unclear, as it is still unclear how genomic GF changes are regulated and converted into mRNA levels which are maintained in different host species. Secondly, why the GF is dynamic in the first place remains unresolved, as the genomic accordion and IAD model have not been experimentally demonstrated. How general is an unbalanced GF within multipartite viruses? Next, I compare where the GF has been reported and discuss the possible role in plants and animals.

The genome formula of multipartite viruses

The proposed benefits and role of GF has been well-described in the model FBNSV (Sicard et al. 2013; Gallet et al. 2022) and here I will explore the GF in different classes of plant and animal viruses. Multipartite viruses are largely found infecting plant species (Lucía-Sanz and Manrubia 2017; Michalakis and Blanc 2020), with only *Bombyx mori* bidensovirus (BmBdV) confirmed infecting insects (Wang et al. 2007). A second animal multipartite virus has been reported infecting *Aedes albopictus* and *Culex quinquefasciatus* female mosquitoes, the

Guaico culex virus (GCXV) (Ladner et al. 2016). Ladner et al. (2016) used the dose-response relationship to infer that the GCXV is a multicomponent virus, however the distribution of genome segments over virus particles has not been measured and whether it is a true multipartite virus therefore remains open for discussion (Michalakis and Blanc 2020). Since the identification of the GF in FBNSV by Sicard and colleagues (2013), there has been an uptick in the reporting of GFs in other virus species.

Thus far, based on reported GF, the following trends become clear. Similar to FBNSV, the GF is both unbalanced and host-specific (Wu et al. 2017; Yu et al. 2019; Zhao et al. 2019). Several studies indicate that unbalanced GFs also occur in both plant and animal segmented viruses, as reviewed by Diefenbacher, Sun, and Brooke (2018) and Wichgers Schreur, Kormelink, and Kortekaas (2018). The GF has also been measured for a segmented ssDNA animal virus, bluetongue virus (BTV) (Moreau et al. 2020) and shown to be unbalanced and host-dependent. The varying, host-dependent distributions of genome segments over virus particles in Rift Valley fever virus (RVFV) also suggest the GF of this virus may be unbalanced (Bermúdez-Méndez et al. 2022), although GF values have not been reported yet. All of these observations suggest that unbalanced and host-dependent GFs may be a general feature of multipartite viruses, as well as some segmented viruses.

The GF has been reported for DNA and RNA viruses alike, indicating that this strategy is not limited to a specific nucleic acid type. However, it is likely that the differences in replication strategies may impact at which level the GF is active. It has been suggested that the FBNSV GF differences may stabilize transcript levels in the hosts *V. faba* and *M. truncatula* (Gallet et al. 2022). Similar to FBNSV, in BBTV there are also differences in the genomic DNA GF compared to viral mRNA transcripts (Yu et al. 2019). However, for BBTV it was found that the promoter sequence of each segment plays an important role in regulating transcript levels (Yu et al. 2019). How promotor sequence activity interacts with changes in the genomic GF is still unclear, but it highlights that there are likely to be multiple GF interactions involved in gene expression regulation. We anticipate that for positive sense RNA viruses, the GF gene dosage benefit is directly active upon virus entry and replication, as there is no distinction between genomic RNA and active RNA. However, there is no direct evidence to support this, although for AMV the packaging of RNAs into virus particles appears to modify the GF (Wu et al. 2017).

The use of high throughput sequencing and metagenomics has led to the discovery of the putative multipartite circular Rep-encoding ssDNA (CRESS DNA) viruses, based on phylogenetic analysis of the replication-associated protein and isolation of genome segments (Male et al. 2016; Kraberger et al. 2019). The putative multipartite viruses are tripartite Fusarium graminearum gemitripvirus 1 (FgGMTV1) (Li et al. 2020), Pacific flying fox faeces-associated multicomponent virus-1 (PfffaMCV-1) (Male et al. 2016), blackfly multi-component virus 1 (BfMCV-1) and blackfly multi-component virus 2 (BfMCV-2) (Kraberger et al. 2019). These results show that whilst most reported multipartite viruses are found in plants, their prevalence in fungi and insects has likely been underestimated, and with advances in virus detection techniques, their number is expected to rise.

Multipartite virus infection in animals

So far, only a small number of animal multipartite viruses have been discovered, but it is probable that additional ones will be recognized in the coming years. Here I will focus on BmBDV infecting silkworm (*Bombyx mori*) (Wang et al. 2007; Hu et al. 2013).

Bidnaviridae is a monophyletic family of bipartite linear ssDNA viruses with the exemplar species BmBDV (Adams and Carstens 2012). Virus particles contain either a single copy of the respective ssDNA genome segment or the complementary strand, and thus there may be four virus particle types (Bando et al. 1992). Segment VD1 encodes for 3 structural proteins and a type B DNA polymerase whilst segment VD2 encodes for a structural and non-structural protein (Li et al. 2015; Hayakawa et al. 2000). BmBDV infection occurs within the silkworm midgut, with accumulation in the posterior and columnar cells of the midgut (Seki and Iwashita 1983; Ito et al. 2016). Hu et al. 2016 determined the BmBDV GF in *B. mori* midgut and stool samples, as well as from purified occlusion bodies (OBs), where segment VD2 had a higher frequency than VD1. These reports are the first clear evidence of an unbalanced GF in an insect virus (Table 1). BmBDV replication within the host silkworm requires traversal of insect interior barriers, similar to the process of circulative propagative transmission in aphids.

Table 1. Genome formula (GF) of the animal multipartite virus Bombyx mori bidensovirus (BmBDV) is unbalanced in *Bombyx mori* larvae. GF values are reported as relative frequencies.

Source	Genome formula	Reference
Virus particles	(0.451:0.549)	(Hu et al. 2016)
Midgut	(0.366:0.634)	
Frass	(0.406:0.595)	
Occlusion body	(0.471:0.529) (0.395:0.605) (0.387:0.613)	(Gani et al. 2021)

Multipartition is a strategy largely found in plant-infecting viruses, suggesting that there may be benefits associated with plant hosts. The costs and benefits associated with multipartition may differ at the within and between-host levels, as indicated by experimental measures of accumulation and segment frequency differences in FBNSV hosts (Sicard et al. 2013), within vectors (Di Mattia et al. 2020) and computational approaches show that multipartition may have unique benefits for plant-infecting viruses (Valdano et al. 2019).

Factors which affect the genome formula

Factors which may influence changes in the GF have not been well-studied, but a few studies have shown that mixed virus infections and the presence of satellite viruses can perturb the GF. Satellite viruses are a class of selfish replicators often found accompanying monopartite, segmented and multipartite virus infections (Koonin et al. 2021). They are subviral agents of

short DNA or RNA segments which are dependent on a helper virus to complete replication (Murant and Mayo 1982). Their presence has been shown to decrease virus accumulation (Wrzesińska et al. 2018; Guyot et al. 2022; Saeed et al. 2007; Liao et al. 2007) and alter virus virulence (Palukaitis and Roossinck 1996). In the context of multipartite viruses, satellites could directly change the GF, modify virus accumulation and alter virus transmissibility.

Various studies have shown a range of different effects of satellites on multipartite viruses, and here I will consider some examples for CMV and BBTV. A satRNA co-infection with CMV decreased total virus accumulation and altered the GF of CMV in tomato and tobacco (Feng et al. 2012; Shen et al. 2015; Liao et al. 2007). CMV necrotic satRNA co-infection decreases total virus accumulation, but increases aphid transmission (Escriu, Perry, and García-Arenal 2000). CMV Y-sat infected plants' yellow phenotype preferentially attracts aphid vectors and facilitates virus and satellite transmission (Jayasinghe et al. 2021). BBTV infection with a novel alphasatellite alters the GF in plants and aphids and reduces both genomic DNA and mRNA transcript accumulation (Guyot et al. 2022). Furthermore, transmission experiments showed that BBTV alphasatellite was able to facilitate aphid transmission from monocot to dicot species (Guyot et al. 2022). In both these viruses, satellites could affect the GF, infection and transmission, although it remains unclear whether the GF changes mediated some of these effects. It is highly relevant to consider in more detail the interactions between satellites and the GF and their effect on virus and satellite fitness considered on different levels of selection.

There is mounting evidence that an unbalanced GF is common in multipartite viruses, as to date there are no reports of balanced GFs where it has been measured. Till now, studies have focused on ssDNA multipartite viruses with a large number of genome segments (>6); however, the overwhelming majority of multipartite viruses have ssRNA genomes with 2 – 5 genome segments and infect plants. This knowledge gap means that GF measurements and how they relate to viral fitness in RNA viruses cannot be easily interpreted, as the majority of multipartite viruses have compact genomes where each segment encodes one or two genes with essential functions for replication or transmission. We also note that in the situations where the GF has been measured, there is high GF variability between individual infected plants belonging to the same cultivar. The source of this variation is not understood: We also do not understand the roles of stochastic and deterministic forces that might act on the GF and how they cause both high GF variability between individuals and species-dependent GF equilibria. The GF has been experimentally measured for segmented viruses, highlighting that segment copy number changes can occur via mechanisms related to segment co-packaging or differences in packaging efficiency. To fill the identified gaps, I use the tripartite +ssRNA cucumber mosaic virus (CMV) as a model system. Below I provide a description of the CMV genome organisation, virus replication and its' suitability for investigating the GF.

Cucumber mosaic virus: A model for multipartite virus genome formula variability and evolution

CMV is a globally distributed virus reported to infect more than 1000 plant species (Roossinck 2001; Yoon, Palukaitis, and Choi 2019). CMV causes cucumber mosaic disease, first described in 1916 (Doolittle 1916). Symptoms include leaf yellowing, mosaic, leaf curling, stunting and necrotic lesions (Zitter and Murphy 2009). CMV is a +ssRNA virus with a tripartite

genome belonging to the genus *Cucumovirus* and the *Bromoviridae* family (Jacquemond 2012). The genome segments are designated RNA 1 – 3 with decreasing segment size, with the largest being RNA1 at 3.3kb, followed by RNA2 size of 3kb and the smallest, RNA3, with a segment size of 2.2kb (Jacquemond 2012) (Figure 2). RNA1 encodes for the 1a protein which contains methyltransferase and helicase motifs (Habili and Symons 1989; Gorbalenya et al. 1989; Rozanov, Koonin, and Gorbalenya 1992). RNA2 encodes for the 2a protein, RdRp, and the 2b protein, the viral suppressor of host RNAi which is expressed via sgRNA4A (Ding et al. 1994; Diaz-Pendon et al. 2007). Furthermore, the 2b protein has the primary function as the viral suppressor of host RNAi during infection and is important for determining cell and tissue type localisation and accumulation of CMV (Soards et al. 2002). RNA3 encodes the 3a movement protein (MP) and the 3b coat protein (CP), located downstream of the 3a ORF and expressed from sgRNA4 (Jacquemond 2012). RNA2 and RNA3 are bicistronic and express part of their genome segment by subgenomic RNA (sgRNA) from the minus strand of genomic RNA (Jacquemond 2012).

CMV infection starts upon virus entry into the host cell and by uncoating genomic RNA segments and initiates replication via a replication complex. The replication complex contains CMV proteins 1a (methyltransferase and helicase domains) and 2a (RNA-dependent RNA polymerase) which localise onto the tonoplast membrane of the vacuole (Cillo, Roberts, and Palukaitis 2002). Replication is an asymmetrical process whereby either the positive or negative-strand RNA is synthesized, a process mediated by the specific associations of 1a to the tonoplast and 2a to positive-strand synthesis (Seo et al. 2009). Replication begins with the complex of the viral 1a and 2a proteins (to form the replicase) and host factors for negative-strand synthesis (Seo et al. 2009). Phosphorylation of 2a leads to a reduction in the negative-strand synthesis and the switch of replicase activity to positive-strand synthesis, whilst 1a protein completes the capping of RNA (Seo et al. 2009).

During local movement the viral 3a movement protein (MP) is required for virus movement through plasmodesmata to adjacent cells. Briefly, MP increases the plasmodesmal size exclusion limits, allowing for the movement of viral ribonucleoprotein complexes locally (Wolf et al. 1989), by disrupting microtuble F-actin fibers (Ding et al. 1995; Su et al. 2010). The CP interacts with the MP to facilitate cell-cell and systemic movement (Nagano et al. 2001; Llamas, Moreno, and García-Arenal 2006). In the transition from local to systemic virus infection, viral RNA may be encapsidated or remain as a ribonucleoprotein complex and loaded in the phloem sieve elements (Blackman et al. 1998). For CMV, systemic infection occurs by movement along the photoassimilate pathway, via photosynthetic source and sink tissues (Leisner and Turgeon 1993).

Cucumber mosaic virus may be divided into three subgroups, IA, IB and II based on sequence alignments of RNA 1 – 3 from different virus isolates (Roossinck 2002; Roossinck, Zhang, and Hellwald 1999; Ohshima et al. 2016). Ohshima et al (2016) identify that for CMV within-subgroup reassortments of subgroup IA and IB are more common than between subgroup reassortments, although reassortment of subgroup II and IA do occur (Ohshima et al. 2016). CMV is a model system for studying plant virus evolution at the within-host and between-host levels (Roossinck 2001). The role of recombination and reassortment in CMV evolution has been the subject of several studies; reassortment and recombination are uncommon in natural populations of CMV in Spain (Fraile et al. 1997). (Bonnet et al. 2005) found that the majority of field isolates belonged to subgroup IA and IB with a low occurrence of reassortants and

recombinants in natural populations and that these had lower fitness than parent genotypes (Bonnet et al. 2005). CMV infection in *Arabidopsis thaliana* is associated with delayed onset of flowering and a developmental switch to seed production (Pagán, Alonso-Blanco, and García-Arenal 2008). During multi-host species infections, some host species may act as reservoirs for genetic exchange via recombination (Ouedraogo et al. 2019). The mutation frequency of CMV has been experimentally estimated in different host species and is in the range 0.60 – 25.0 mutations per 10^4 nucleotides (Ouedraogo and Roossinck 2019).

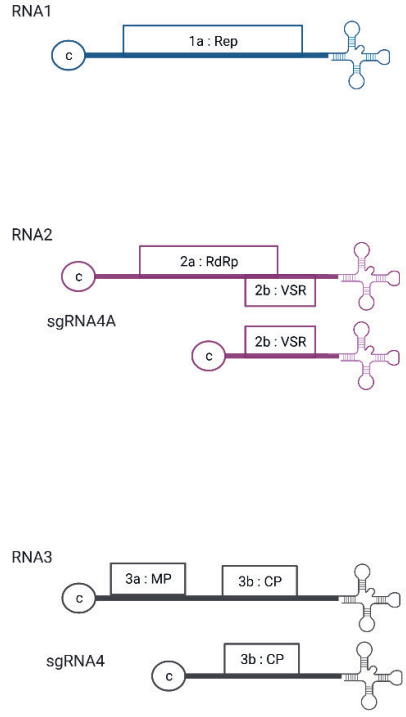


Figure 2. Cucumber mosaic virus has a genome composed of three +ssRNAs, RNA 1 – 3, with decreasing segment size. *It has two subgenomic (sg) RNAs, derived from RNA2 and RNA3, expressing the 2b protein viral suppressor of RNAi (VSR) and the coat protein (CP). There is 5' cap on each genome segment (c) and at the 3'end of all segments have a tRNA-like structure. Created with BioRender.com.*

Overview of thesis chapters

In this thesis, I explore the GF in a multipartite plant virus by investigating its role in short-term adaptation after a single infection cycle, how the GF changes over longer timescales in different hosts and the GF as a property in multipartite and segmented viruses.

In **Chapter 2**, we assess the state-of-the-art concerning the costs and proposed benefits of a multipartite virus genome organisation, reviewing key developments and using quantitative

approaches to address some key unanswered questions. We present an approach for estimating the cost of transmission for multipartite viruses. Also using computational modelling, we explore under which conditions the proposed benefit of virus gene product sharing across cells can minimise the within-host cost. We analyse how virus genome segmentation can increase host range and critically discuss the within and between-host benefits of the GF.

In **Chapter 3**, an approach for quantitatively analysing GF variation in multipartite and segmented viruses is developed. We introduce a new metric for analysing GF variability, the GF distance. We show how this metric can be used to compare GF observations, both for comparison between empirical observations and to theoretical predictions. We re-analyse published GF data to illustrate these approaches and, in so doing, uncover evidence that the GF is a transmissible property.

Chapter 4 presents a combined experimental and modelling approach to test GF variation during mechanical infection of *Chenopodium quinoa,* a local lesion host for CMV. Our data show that the GF is highly variable in local lesions and is influenced by the inoculum GF. The GF converges to two plateaus, associated with higher and lower virus titres. We discuss how stochastic population bottlenecks and directional GF change interact to affect GF variation and GF drift during infections.

Chapter 5 examines rapid adaptation of the GF and long-term evolutionary stability of the GF in *Arabidopsis thaliana*, *N. benthamiana* and *N. tabacum*. preliminary data indicates there may be a virus genotype GF for CMV. Serial passage experiments of CMV infection in the three hosts, A. thaliana, N. benthamiana and N. tabacum, show that for *A. thaliana*, there is a host-specific GF, whilst *N. benthamiana* and *N. tabacum* have similar GFs. The GF remains stable after the initial passage and at the final passage there is an increase in GF variation in *N. tabacum*. We observe a number of virus extinctions which are associated with low virus titre. Sequence analysis identified a repeated non-synonymous mutation in the RNA-dependent RNA polymerase which was associated with the accumulation of intermediate and low-frequency mutations in the populations where it was present.

In **Chapter 6,** I develop a computational model to quantitatively determine the cost of transmission for mono-, multipartite and segmented viruses. We extend this to include different classes of segmented viruses based on selective and non-selective packaging strategies.

In **Chapter 7**, I discuss the main findings from chapter 2 – 6 and discuss this within a broader context. I discuss the cost of transmission for monopartite, segmented and multipartite viruses and present a conceptual overview of genome packaging and the cost of transmission. I present and discuss the GF fitness landscape of CMV in four host species: *A. thaliana, C. quinoa, N. benthamiana* and *N. tabacum*. I find that CMV infections in *C. quinoa* have a broad plateau characterised by high virus titre, that A. thaliana has narrow peak associated with higher virus titre and that there is no relationship between GF variation and virus titre in *N. benthamiana* and *N. tabacum*.

References

Adams, M. J., and E. B. Carstens. 2012. “Ratification Vote on Taxonomic Proposals to the International Committee on Taxonomy of Viruses (2012).” *Archives of Virology* 157 (7): 1411–22. https://doi.org/10.1007/s00705-012-1299-6.

Bald, J. G. 1937. “The Use of Numbers of Infections For Comparing the Concentrations of Plant Virus Suspensions: Dilution Experiments with Purified Suspensions.” *The Annals of Applied Biology* 24 (1): 33–55. https://doi.org/10.1111/j.1744-7348.1937.tb05019.x.

Baltimore, D. 1971. “Expression of Animal Virus Genomes.” *Bacteriological Reviews* 35 (3): 235–41. https://www.ncbi.nlm.nih.gov/pubmed/4329869.

Bando, H., H. Choi, Y. Ito, M. Nakagaki, and S. Kawase. 1992. “Structural Analysis on the Single-Stranded Genomic DNAs of the Virus Newly Isolated from Silkworm: The DNA Molecules Share a Common Terminal Sequence.” *Archives of Virology* 124 (1–2): 187–93. https://doi.org/10.1007/BF01314637.

Bayer, Avraham, Greg Brennan, and Adam P. Geballe. 2018. “Adaptation by Copy Number Variation in Monopartite Viruses.” *Current Opinion in Virology*. Elsevier. https://doi.org/10.1016/j.coviro.2018.07.001.

Bermúdez-Méndez, Erick, Kirsten F. Bronsvoort, Mark P. Zwart, Sandra van de Water, Ingrid Cárdenas-Rey, Rianka P. M. Vloet, Constantianus J. M. Koenraadt, Gorben P. Pijlman, Jeroen Kortekaas, and Paul J. Wichgers Schreur. 2022. “Incomplete Bunyavirus Particles Can Cooperatively Support Virus Infection and Spread.” *PLoS Biology* 20 (11): e3001870. https://doi.org/10.1371/journal.pbio.3001870.

Blackman, Leila M., Petra Boevink, Simon Santa Cruz, Peter Palukaitis, and Karl J. Oparka. 1998. “The Movement Protein of Cucumber Mosaic Virus Traffics into Sieve Elements in Minor Veins of Nicotiana Clevelandii.” *The Plant Cell* 10: 525–37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC144016/pdf/100525.pdf.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064. https://doi.org/10.1016/j.virusres.2023.199064.

Bonnet, Julien, Aurora Fraile, Soledad Sacristán, José M. Malpica, and Fernando García-Arenal. 2005. “Role of Recombination in the Evolution of Natural Populations of Cucumber Mosaic Virus, a Tripartite RNA Plant Virus.” *Virology* 332 (1): 359–68. https://doi.org/10.1016/j.virol.2004.11.017.

Bruening, G., and H. O. Agrawal. 1967. “Infectivity of a Mixture of Cowpea Mosaic Virus Ribonucleoprotein Components.” *Virology* 32 (2): 306–20. https://doi.org/10.1016/0042-6822(67)90279-6.

Bujarski, Jozef J. 2013. “Genetic Recombination in Plant-Infecting Messenger-Sense RNA Viruses: Overview and Research Perspectives.” *Frontiers in Plant Science* 4 (March): 68. https://doi.org/10.3389/fpls.2013.00068.

Bujarski, Jozef J., and Paul Kaesberg. 1986. “Genetic Recombination between RNA Components of a Multipartite Plant Virus.” *Nature* 321 (6069): 528–31. https://doi.org/10.1038/321528a0.

Chao, Lin. 1988. “Evolution of Sex in RNA Viruses.” *Journal of Theoretical Biology* 133: 99–112. https://ac.els-cdn.com/S0022519388800274/1-s2.0-S0022519388800274-main.pdf?_tid=ab01d242-cd78-47f3-9277-1244aa87a42c&acdnat=1535462346_2e396edd562ebf6aec8f9c3f600eba53.

———. 1991. “Levels of Selection, Evolution of Sex in RNA Viruses, and the Origin of Life.” *Journal of Theoretical Biology* 153: 229–46. https://ac.els-cdn.com/S0022519305804242/1-s2.0-S0022519305804242-main.pdf?_tid=764a8861-57d4-4db1-9463-9da8d312d78b&acdnat=1535462473_cbb6158265d7444af7560be9cf10f0fd.

Cillo, Fabrizio, Ian M. Roberts, and Peter Palukaitis. 2002. “In Situ Localization and Tissue Distribution of the Replication-Associated Proteins of Cucumber Mosaic Virus in

Tobacco and Cucumber." *Journal of Virology* 76 (21): 10654–64. https://doi.org/10.1128/jvi.76.21.10654-10664.2002.
Davies, M. V., H. W. Chang, B. L. Jacobs, and R. J. Kaufman. 1993. "The E3L and K3L Vaccinia Virus Gene Products Stimulate Translation through Inhibition of the Double-Stranded RNA-Dependent Protein Kinase by Different Mechanisms." *Journal of Virology* 67 (3): 1688–92. https://doi.org/10.1128/JVI.67.3.1688-1692.1993.
Di Mattia, Jérémy, Babil Torralba, Michel Yvon, Jean-Louis Zeddam, Stéphane Blanc, and Yannis Michalakis. 2022. "Nonconcomitant Host-to-Host Transmission of Multipartite Virus Genome Segments May Lead to Complete Genome Reconstitution." *Proceedings of the National Academy of Sciences of the United States of America* 119 (32): e2201453119. https://doi.org/10.1073/pnas.2201453119.
Di Mattia, Jérémy, Marie-Stéphanie Vernerey, Michel Yvon, Elodie Pirolles, Mathilde Villegas, Yahya Gaafar, Heiko Ziebell, Yannis Michalakis, Jean-Louis Zeddam, and Stéphane Blanc. 2020. "Route of a Multipartite Nanovirus across the Body of Its Aphid Vector." *Journal of Virology* 94 (9). https://doi.org/10.1128/JVI.01998-19.
Diaz-Pendon, Juan A., Feng Li, Wan-Xiang Li, and Shou-Wei Ding. 2007. "Suppression of Antiviral Silencing by Cucumber Mosaic Virus 2b Protein in Arabidopsis Is Associated with Drastically Reduced Accumulation of Three Classes of Viral Small Interfering RNAs." *The Plant Cell* 19 (6): 2053–63. https://doi.org/10.1105/tpc.106.047449.
Diefenbacher, Meghan, Jiayi Sun, and Christopher B. Brooke. 2018. "The Parts Are Greater than the Whole: The Role of Semi-Infectious Particles in Influenza A Virus Biology." *Current Opinion in Virology* 33 (December): 42–46. https://doi.org/10.1016/j.coviro.2018.07.002.
Ding, B., Q. Li, L. Ngyuyen, P. Palukaitis, and W. J. Lucas. 1995. "Cucumber Mosaic Virus 3a Protein Potentiates Cell-to-Cell Trafficking of CMV RNA in Tobacco Plants." *Virology* 207: 345–53. https://ac.els-cdn.com/S0042682285710938/1-s2.0-S0042682285710938-main.pdf?_tid=e6bdce65-3d4b-4f36-928a-e7028c1b448d&acdnat=1547576562_0cbd7320370183e1b281912dac8716d7.
Ding, Shou-Wei, Beau J. Anderson, Helen R. Haase, and Robert H. Symons. 1994. "New Overlapping Gene Encoded by the Cucumber Mosaic Virus Genome." *Virology* 198 (February): 593–601. https://doi.org/10.1006/VIRO.1994.1071.
Domingo, Esteban, Julie Sheldon, and Celia Perales. 2012. "Viral Quasispecies Evolution." *Microbiology and Molecular Biology Reviews: MMBR* 76 (2): 159–216. https://doi.org/10.1128/MMBR.05023-11.
Doolittle, S. P. 1916. "A New Infectious Mosaic Disease of Cucumber." *Phytopathology*.
Druett, H. A. 1952. "Bacterial Invasion." *Nature* 170 (4320): 288–288. https://doi.org/10.1038/170288a0.
Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers, Jay Shendure, Adam P. Geballe, and Harmit S. Malik. 2012. "Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses." *Cell* 150 (4): 831–41. https://doi.org/10.1016/j.cell.2012.05.049.
Escriu, F., K. L. Perry, and F. García-Arenal. 2000. "Transmissibility of Cucumber Mosaic Virus by Aphis Gossypii Correlates with Viral Accumulation and Is Affected by the Presence of Its Satellite RNA." *Phytopathology* 90 (10): 1068–72. https://doi.org/10.1094/PHYTO.2000.90.10.1068.
Escriu, Fernando, Aurora Fraile, and Fernando García-Arenal. 2007. "Constraints to Genetic Exchange Support Gene Coadaptation in a Tripartite RNA Virus." *PLoS Pathogens* 3 (1): e8.
Feng, Junli, Leiyu Lai, Ruohong Lin, Chunzhi Jin, and Jishuang Chen. 2012. "Differential Effects of Cucumber Mosaic Virus Satellite RNAs in the Perturbation of MicroRNA-Regulated Gene Expression in Tomato." *Molecular Biology Reports* 39: 775–84. https://doi.org/10.1007/s11033-011-0798-y.
Fraile, Aurora, José Luis Alonso-prados, Miguel A. Aranda, Juan J. Bernal, José M. Malpica, and Fernando Garci. 1997. "Genetic Exchange by Recombination or Reassortment Is Infrequent in Natural Populations of a Tripartite RNA Plant Virus." *Journal of Virology*

71 (2): 934–40. https://www.ncbi.nlm.nih.gov/pubmed/8995610.
Freeman, Jennifer L., George H. Perry, Lars Feuk, Richard Redon, Steven A. McCarroll, David M. Altshuler, Hiroyuki Aburatani, et al. 2006. “Copy Number Variation: New Insights in Genome Diversity.” *Genome Research* 16 (8): 949–61. https://doi.org/10.1101/gr.3677206.
Fulton, R. W. 1980. “Biological Significance of Multicomponent Viruses.” *Annual Review of Phytopathology* 18: 131–46. https://doi.org/10.1146/annurev.py.18.090180.001023.
Fulton, Robert W. 1962. “The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus.” *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.
Gallet, Romain, Jérémy Di Mattia, Sébastien Ravel, Jean-Louis Zeddam, Renaud Vitalis, Yannis Michalakis, and Stéphane Blanc. 2022. “Gene Copy Number Variations at the Within-Host Population Level Modulate Gene Expression in a Multipartite Virus.” *Virus Evolution* 8 (2): veac058. https://doi.org/10.1093/ve/veac058.
Gani, Mudasir, Sergei Senger, Satish Lokanath, Pawan Saini, Kamlesh Bali, Rakesh Gupta, Vankadara Sivaprasad, Johannes A. Jehle, and Jörg T. Wennmann. 2021. “Patterns in Genotype Composition of Indian Isolates of the Bombyx Mori Nucleopolyhedrovirus and Bombyx Mori Bidensovirus.” *Viruses* 13 (5). https://doi.org/10.3390/v13050901.
García-Arriaza, Juan, Susanna C. Manrubia, Miguel Toja, Esteban Domingo, and Cristina Escarmís. 2004. “Evolutionary Transition toward Defective RNAs That Are Infectious by Complementation.” *Journal of Virology* 78 (21): 11678–85. https://doi.org/10.1128/JVI.78.21.11678-11685.2004.
Gorbalenya, A. E., E. V. Koonin, A. P. Donchenko, and V. M. Blinov. 1989. “Two Related Superfamilies of Putative Helicases Involved in Replication, Recombination, Repair and Expression of DNA and RNA Genomes.” *Nucleic Acids Research* 17 (12): 4713–30. https://doi.org/10.1093/nar/17.12.4713.
Guyot, Valentin, Rajendran Rajeswaran, Huong Cam Chu, Chockalingam Karthikeyan, Nathalie Laboureau, Serge Galzi, Lyna F. T. Mukwa, et al. 2022. “A Newly Emerging Alphasatellite Affects Banana Bunchy Top Virus Replication, Transcription, SiRNA Production and Transmission by Aphids.” *PLoS Pathogens* 18 (4): e1010448. https://doi.org/10.1371/journal.ppat.1010448.
Habili, N., and R. H. Symons. 1989. “Evolutionary Relationship between Luteoviruses and Other RNA Plant Viruses Based on Sequence Motifs in Their Putative RNA Polymerases and Nucleic Acid Helicases.” *Nucleic Acids Research* 17 (23): 9543–55. https://doi.org/10.1093/nar/17.23.9543.
Hayakawa, T., K. Kojima, K. Nonaka, M. Nakagaki, K. Sahara, S. i. Asano, T. Iizuka, and H. Bando. 2000. “Analysis of Proteins Encoded in the Bipartite Genome of a New Type of Parvo-like Virus Isolated from Silkworm - Structural Protein with DNA Polymerase Motif.” *Virus Research* 66 (1): 101–8. https://doi.org/10.1016/s0168-1702(99)00129-x.
Hu, Zhaoyang, Guohui Li, Guangtian Li, Qin Yao, and Keping Chen. 2013. “Bombyx Mori Bidensovirus: The Type Species of the New Genus Bidensovirus in the New Family Bidnaviridae.” *Chinese Science Bulletin = Kexue Tongbao* 58 (36): 4528–32. https://doi.org/10.1007/s11434-013-5876-1.
Hu, Zhaoyang, Xiaolong Zhang, Wei Liu, Qian Zhou, Qing Zhang, Guohui Li, and Qin Yao. 2016. “Genome Segments Accumulate with Different Frequencies in Bombyx Mori Bidensovirus.” *Journal of Basic Microbiology* 56 (12): 1338–43. https://doi.org/10.1002/jobm.201600120.
Iranzo, Jaime, and Susanna C. Manrubia. 2012. “Evolutionary Dynamics of Genome Segmentation in Multipartite Viruses.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1743): 3812–19. https://doi.org/10.1098/rspb.2012.1086.
Ito, Katsuhiko, Sachiko Shimura, Susumu Katsuma, Yasuhiro Tsuda, Jun Kobayashi, Hiroko Tabunoki, Takeshi Yokoyama, Toru Shimada, and Keiko Kadono-Okuda. 2016. “Gene Expression and Localization Analysis of Bombyx Mori Bidensovirus and Its

Putative Receptor in B. Mori Midgut." *Journal of Invertebrate Pathology* 136 (May): 50–56. https://doi.org/10.1016/j.jip.2016.03.005.
Jacquemond, Mireille. 2012. "Cucumber Mosaic Virus." Edited by Gad Loebenstein and Hervé Lecoq. *Advances in Virus Research* 84 (January): 439–504. https://doi.org/10.1016/B978-0-12-394314-9.00013-0.
Jayasinghe, Wikum H., Hangil Kim, Yusuke Nakada, and Chikara Masuta. 2021. "A Plant Virus Satellite RNA Directly Accelerates Wing Formation in Its Insect Vector for Spread." *Nature Communications* 12 (1): 1–10. https://doi.org/10.1038/s41467-021-27330-4.
Kammen, A. van. 1967. "Purification and Properties of the Components of Cowpea Mosaic Virus." *Virology* 31 (4): 633–42. https://doi.org/10.1016/0042-6822(67)90192-4.
Katju, Vaishali, Ulfar Bergthorsson, and Frederic Jj Chain. 2013. "Copy-Number Changes in Evolution: Rates, Fitness Effects and Adaptive Significance." *Frontiers in Genetics* 4: 273. https://doi.org/10.3389/fgene.2013.00273.
Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, and Jens H. Kuhn. 2021. "Viruses Defined by the Position of the Virosphere within the Replicator Space." *Microbiology and Molecular Biology Reviews: MMBR* 85 (4): e0019320. https://doi.org/10.1128/MMBR.00193-20.
Koonin, Eugene V., Mart Krupovic, and Vadim I. Agol. 2021. "The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution?" *Microbiology and Molecular Biology Reviews: MMBR* 85 (3): e0005321. https://doi.org/10.1128/MMBR.00053-21.
Kraberger, Simona, Kara Schmidlin, Rafaela S. Fontenele, Matthew Walters, and Arvind Varsani. 2019. "Unravelling the Single-Stranded DNA Virome of the New Zealand Blackfly." *Viruses* 11 (6). https://doi.org/10.3390/v11060532.
Ladner, Jason T., Michael R. Wiley, Brett Beitzel, Albert J. Auguste, Alan P. Dupuis, Michael E. Lindquist, Samuel D. Sibley, et al. 2016. "A Multicomponent Animal Virus Isolated from Mosquitoes." *Cell Host & Microbe* 20 (3): 357–67. https://doi.org/10.1016/j.chom.2016.07.011.
Lauer, Stephanie, and David Gresham. 2019. "An Evolving View of Copy Number Variants." *Current Genetics* 65 (6): 1287–95. https://doi.org/10.1007/s00294-019-00980-0.
Lauring, Adam S., and Raul Andino. 2010. "Quasispecies Theory and the Behavior of RNA Viruses." *PLoS Pathogens* 6 (7): e1001005. https://doi.org/10.1371/journal.ppat.1001005.
Leeks, Asher, Penny Grace Young, Paul Eugene Turner, Geoff Wild, and Stuart Andrew West. 2023. "Cheating Leads to the Evolution of Multipartite Viruses." *PLoS Biology* 21 (4): e3002092. https://doi.org/10.1371/journal.pbio.3002092.
Leisner, S. M., and R. Turgeon. 1993. "Movement of Virus and Photoassimilate in the Phloem: A Comparative Analysis." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 15 (11): 741–48. https://doi.org/10.1002/bies.950151107.
Li, Guohui, Qian Zhou, Zhaoyang Hu, Peng Wang, Qi Tang, Keping Chen, and Qin Yao. 2015. "Determination of the Proteins Encoded by BmBDV VD1-ORF4 and Their Interacting Proteins in BmBDV-Infected Midguts." *Current Microbiology* 70 (4): 623–29. https://doi.org/10.1007/s00284-014-0765-7.
Li, Pengfei, Shuangchao Wang, Lihang Zhang, Dewen Qiu, Xueping Zhou, and Lihua Guo. 2020. "A Tripartite SsDNA Mycovirus from a Plant Pathogenic Fungus Is Infectious as Cloned DNA and Purified Virions." *Science Advances* 6 (14): eaay9634. https://doi.org/10.1126/sciadv.aay9634.
Liao, Qiansheng, Liping Zhu, Zhiyou Du, Rong Zeng, Junli Feng, and Jishuang Chen. 2007. "Satellite RNA-Mediated Reduction of Cucumber Mosaic Virus Genomic RNAs Accumulation in Nicotiana Tabacum." *Acta Biochimica et Biophysica Sinica* 39 (3): 217–23. https://doi.org/10.1111/j.1745-7270.2007.00266.x.
Lister, R. M. 1968. "Functional Relationships between Virus-Specific Products of Infection by Viruses of the Tobacco Rattle Type." *The Journal of General Virology* 2 (1): 43–58.

https://doi.org/10.1099/0022-1317-2-1-43.
Llamas, Susana, Ignacio M. Moreno, and Fernando García-Arenal. 2006. “Analysis of the Viability of Coat-Protein Hybrids between Cucumber Mosaic Virus and Tomato Aspermy Virus.” *The Journal of General Virology* 87 (Pt 7): 2085–88. https://doi.org/10.1099/vir.0.81871-0.
Loesch, Loretta Sue, and Robert W. Fulton. 1975. “Prunus Necrotic Ringspot Virus as a Multicomponent System.” *Virology* 68: 71–78. https://ac.els-cdn.com/0042682275901488/1-s2.0-0042682275901488-main.pdf?_tid=a151fb9e-3dc5-4bcc-a7ad-6e7f0f9f854e&acdnat=1541768834_fab42d84f90a29a4d4c2954be508d8f1.
Lucía-Sanz, Adriana, Jacobo Aguirre, and Susanna Manrubia. 2018. “Theoretical Approaches to Disclosing the Emergence and Adaptive Advantages of Multipartite Viruses.” *Current Opinion in Virology* 33 (December): 89–95. https://doi.org/10.1016/j.coviro.2018.07.018.
Lucía-Sanz, Adriana, and Susanna Manrubia. 2017. “Multipartite Viruses: Adaptive Trick or Evolutionary Treat?” *Npj Systems Biology and Applications* 3 (1): 34. https://doi.org/10.1038/s41540-017-0035-y.
Male, Maketalena F., Simona Kraberger, Daisy Stainton, Viliami Kami, and Arvind Varsani. 2016. “Cycloviruses, Gemycircularviruses and Other Novel Replication-Associated Protein Encoding Circular Viruses in Pacific Flying Fox (Pteropus Tonganus) Faeces.” *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 39 (April): 279–92. https://doi.org/10.1016/j.meegid.2016.02.009.
Martínez, Fernando, Josep Sardanyés, Santiago F. Elena, and José-Antonio Daròs. 2011. “Dynamics of a Plant RNA Virus Intracellular Accumulation: Stamping 2 Machine versus Geometric Replication.” *Genetics* 188 (3): 637–46. https://doi.org/10.1534/genetics.111.129114.
Mehle, Andrew, Vivien G. Dugan, Jeffery K. Taubenberger, and Jennifer A. Doudna. 2012. “Reassortment and Mutation of the Avian Influenza Virus Polymerase PA Subunit Overcome Species Barriers.” *Journal of Virology* 86 (3): 1750–57. https://doi.org/10.1128/JVI.06203-11.
Michalakis, Yannis, and Stéphane Blanc. 2020. “The Curious Strategy of Multipartite Viruses.” *Annual Review of Virology* 7 (1): 203–18. https://doi.org/10.1146/annurev-virology-010220-063346.
Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. “The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner.” *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.
Murant, A. F., and M. A. Mayo. 1982. “Satellites of Plant Viruses.” *Annual Review of Phytopathology* 20 (1): 49–68. https://doi.org/10.1146/annurev.py.20.090182.000405.
Nagano, H., K. Mise, I. Furusawa, and T. Okuno. 2001. “Conversion in the Requirement of Coat Protein in Cell-to-Cell Movement Mediated by the Cucumber Mosaic Virus Movement Protein.” *Journal of Virology* 75 (17): 8045–53. https://doi.org/10.1128/jvi.75.17.8045-8053.2001.
Nakatsu, Sumiho, Hiroshi Sagara, Yuko Sakai-Tagawa, Norio Sugaya, Takeshi Noda, and Yoshihiro Kawaoka. 2016. “Complete and Incomplete Genome Packaging of Influenza A and B Viruses.” *MBio* 7 (5). https://doi.org/10.1128/mBio.01248-16.
Näsvall, Joakim, Lei Sun, John R. Roth, and Dan I. Andersson. 2012. “Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence.” *Science* 338 (6105): 384–87. https://doi.org/10.1126/science.1226521.
Nee, Scan. 1987. “The Evolution of Multicompartmental Genomes in Viruses.” *Journal of Molecular Evolution* 25: 277–81. https://link.springer.com/content/pdf/10.1007/BF02603110.pdf.
Ohshima, Kazusato, Kosuke Matsumoto, Ryosuke Yasaka, Mai Nishiyama, Kenta Soejima, Savas Korkmaz, Simon Y. W. Ho, Adrian J. Gibbs, and Minoru Takeshita. 2016.

"Temporal Analysis of Reassortment and Molecular Evolution of Cucumber Mosaic Virus: Extra Clues from Its Segmented Genome." *Virology* 487: 188–97. https://doi.org/10.1016/j.virol.2015.09.024.
Ojosnegros, S., J. García-Arriaza, C. Escarmís, S. C. Manrubia, and C. Perales. 2011. "Viral Genome Segmentation Can Result from a Trade-Off between Genetic Content and Particle Stability." *PLoS Genetics* 7 (3): 1001344. https://doi.org/10.1371/journal.pgen.1001344.
Ouedraogo, R. S., and M. J. Roossinck. 2019. "Chapter 18: Molecular Evolution." In *Cucumber Mosaic Virus*, edited by Peter Palukaitis and Fernando García-Arenal, 207–15. Virology. The American Phytopathological Society. https://doi.org/10.1094/9780890546109.022.
Ouedraogo, Rimnoma S., Justin S. Pita, Irenée P. Somda, Oumar Traore, and Marilyn J. Roossinck. 2019. "Impact of Cultivated Hosts on the Recombination of Cucumber Mosaic Virus." *Journal of Virology* 93 (7): e01770-18. https://doi.org/10.1128/JVI.01770-18.
Pagán, Israel, Carlos Alonso-Blanco, and Fernando García-Arenal. 2008. "Host Responses in Life-History Traits and Tolerance to Virus Infection in Arabidopsis Thaliana." *PLoS Pathogens* 4 (8): e1000124. https://doi.org/10.1371/journal.ppat.1000124.
Palukaitis, Peter, and Marilyn J. Roossinck. 1996. "Spontaneous Change of a Benign Satellite RNA of Cucumber Mosaic Virus to a Pathogenic Variant." *Nature Biotechnology* 14 (10): 1264–68. https://doi.org/10.1038/nbt1096-1264.
Pressing, J., and D. C. Reanney. 1984. "Divided Genomes and Intrinsic Noise." *Journal of Molecular Evolutionm* 20: 135–46. https://link.springer.com/content/pdf/10.1007%2FBF02257374.pdf.
Qiu, W., and J. W. Moyer. 1999. "Tomato Spotted Wilt Tospovirus Adapts to the TSWV N Gene-Derived Resistance by Genome Reassortment." *Phytopathology* 89 (7): 575–82. https://doi.org/10.1094/PHYTO.1999.89.7.575.
Reanney, D. C. 1982. "The Evolution of RNA Viruses." *Annual Review of Microbiology* 36: 47–73. https://doi.org/10.1146/annurev.mi.36.100182.000403.
Riehle, M. M., A. F. Bennett, and A. D. Long. 2001. "Genetic Architecture of Thermal Adaptation in Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 525–30. https://doi.org/10.1073/pnas.98.2.525.
Roossinck, M. J. 2001. "Cucumber Mosaic Virus, a Model for RNA Virus Evolution." *Molecular Plant Pathology* 2 (2): 59–63. https://doi.org/10.1046/j.1364-3703.2001.00058.x.
Roossinck, Marilyn J. 2002. "Evolutionary History of Cucumber Mosaic Virus Deduced by Phylogenetic Analyses." *Journal of Virology* 76 (7): 3382–87. https://doi.org/10.1128/JVI.76.7.3382-3387.2002.
Roossinck, Marilyn J., Lee Zhang, and Karl-Heinz Hellwald. 1999. "Rearrangements in the 5' Nontranslated Region and Phylogenetic Analyses of Cucumber Mosaic Virus RNA 3 Indicate Radial Evolution of Three Subgroups." *Journal of Virology* 73 (8): 6752–58. https://www.ncbi.nlm.nih.gov/pubmed/10400773.
Rozanov, M. N., E. V. Koonin, and A. E. Gorbalenya. 1992. "Conservation of the Putative Methyltransferase Domain: A Hallmark of the 'Sindbis-like' Supergroup of Positive-Strand RNA Viruses." *The Journal of General Virology* 73 (Pt 8) (8): 2129–34. https://doi.org/10.1099/0022-1317-73-8-2129.
Saeed, Muhammad, Yusuf Zafar, John W. Randles, and M. Ali Rezaian. 2007. "A Monopartite Begomovirus-Associated DNA Beta Satellite Substitutes for the DNA B of a Bipartite Begomovirus to Permit Systemic Infection." *The Journal of General Virology* 88 (Pt 10): 2881–89. https://doi.org/10.1099/vir.0.83049-0.
Sakai, Y., K. Kiyotani, M. Fukumura, M. Asakawa, A. Kato, T. Shioda, T. Yoshida, A. Tanaka, M. Hasegawa, and Y. Nagai. 1999. "Accommodation of Foreign Genes into the Sendai Virus Genome: Sizes of Inserted Genes and Viral Replication." *FEBS Letters* 456 (2): 221–26. https://doi.org/10.1016/s0014-5793(99)00960-6.
Sandegren, Linus, and Dan I. Andersson. 2009. "Bacterial Gene Amplification: Implications

for the Evolution of Antibiotic Resistance." *Nature Reviews. Microbiology* 7 (8): 578–88. https://doi.org/10.1038/nrmicro2174.
Sanders, Christopher, Eva Veronesi, Paulina Rajko-Nenow, Peter Paul Clement Mertens, Carrie Batten, Simon Gubbins, Simon Carpenter, and Karin Darpel. 2022. "Field-Reassortment of Bluetongue Virus Illustrates Plasticity of Virus Associated Phenotypic Traits in the Arthropod Vector and Mammalian Host In Vivo." *Journal of Virology* 96 (13): e0053122. https://doi.org/10.1128/jvi.00531-22.
Sänger, H. L. 1968. "Characteristics of Tobacco Rattle Virus." *Molecular & General Genetics: MGG* 101 (4): 346–67. https://doi.org/10.1007/BF00436232.
Sanjuán, Rafael, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. 2010. "Viral Mutation Rates." *Journal of Virology* 84 (19): 9733–48. https://doi.org/10.1128/JVI.00694-10.
Sasani, Thomas A., Kelsey R. Cone, Aaron R. Quinlan, and Nels C. Elde. 2018. "Long Read Sequencing Reveals Poxvirus Evolution through Rapid Homogenization of Gene Arrays." *ELife* 7 (August). https://doi.org/10.7554/eLife.35453.
Seki, H., and Y. Iwashita. 1983. "Histopathological Features and Pathogenicity of a Densonucleosis Virus of the Silkworm, Bombyx Mori, Isolated from Sericultural Farms in Yamanashi Prefecture." *The Journal of Sericultural Science Japan* 52: 400–405. https://doi.org/10.11416/kontyushigen1930.52.400.
Seo, Jang-Kyun, Sun-Jung Kwon, Hong-Soo Choi, and Kook-Hyung Kim. 2009. "Evidence for Alternate States of Cucumber Mosaic Virus Replicase Assembly in Positive- and Negative-Strand RNA Synthesis." *Virology* 383 (2): 248–60. https://doi.org/10.1016/j.virol.2008.10.033.
Shen, Wan-Xia, Phil Chi Khang Au, Bu-Jun Shi, Neil A. Smith, Elizabeth S. Dennis, Hui-Shan Guo, Chang-Yong Zhou, and Ming-Bo Wang. 2015. "Satellite RNAs Interfere with the Function of Viral RNA Silencing Suppressors." *Frontiers in Plant Science* 6 (April). https://doi.org/10.3389/fpls.2015.00281.
Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. "The Strange Lifestyle of Multipartite Viruses." Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819. https://doi.org/10.1371/journal.ppat.1005819.
Sicard, Anne, Elodie Pirolles, Romain Gallet, Marie-Stéphanie Vernerey, Michel Yvon, Michel Peterschmitt, Serafin Gutierrez, Yannis Michalakis, and Stéphane Blanc. 2019. "A Multicellular Way of Life for a Multipartite Virus." *ELife* 8 (March): e43599. https://doi.org/10.7554/eLife.43599.
Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. "Gene Copy Number Is Differentially Regulated in a Multipartite Virus." *Nature Communications* 4: 2248. https://doi.org/10.1038/ncomms3248.
Soards, Avril J., Alex M. Murphy, Peter Palukaitis, and John P. Carr. 2002. "Virulence and Differential Local and Systemic Spread of Cucumber Mosaic Virus in Tobacco Are Affected by the CMV 2b Protein." *Molecular Plant-Microbe Interactions: MPMI* 15 (7): 647–53. https://doi.org/10.1094/MPMI.2002.15.7.647.
Su, Shengzhong, Zhaohui Liu, Cheng Chen, Yan Zhang, Xu Wang, Lei Zhu, Long Miao, Xue-Chen Wang, and Ming Yuan. 2010. "Cucumber Mosaic Virus Movement Protein Severs Actin Filaments to Increase the Plasmodesmal Size Exclusion Limit in Tobacco." *The Plant Cell* 22 (4): 1373–87. https://doi.org/10.1105/tpc.108.064212.
Todd, Robert T., and Anna Selmecki. 2020. "Expandable and Reversible Copy Number Amplification Drives Rapid Adaptation to Antifungal Drugs." *ELife* 9 (July). https://doi.org/10.7554/eLife.58349.
Valdano, Eugenio, I. D. 1¤, Susanna Manrubia, Sergio Gó Mez Id, and Alex Arenas Id. 2019. "Endemicity and Prevalence of Multipartite Viruses under Heterogeneous Between-Host Transmission." *PLoS Computational Biology* 15 (3): e1006876. https://doi.org/10.1371/journal.pcbi.1006876.
Van Vloten-Doting, Lous, and E. M. J. Jaspars. 1977. "Plant Covirus Systems: Three-Component Systems." In *Comprehensive Virology 11: Regulation and Genetics Plant*

Viruses, edited by Heinz Fraenkel-Conrat and Robert R. Wagner, 1–53. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-2721-9_1.

Van Vloten-doting, Lous, J. Kruseman, and E. M. J. Jaspars. 1968. "The Biological Function and Mutual Dependence of Bottom Component and Top Component a of Alfalfa Mosaic Virus." *Virology* 34: 728–37. https://ac.els-cdn.com/0042682268900937/1-s2.0-0042682268900937-main.pdf?_tid=6e588fdd-cacf-40e0-a6b0-b9ce6e941bf2&acdnat=1541765186_017cbb62c1117c56b601710f65073eed.

Varsani, Arvind, Pierre Lefeuvre, Philippe Roumagnac, and Darren Martin. 2018. "Notes on Recombination and Reassortment in Multipartite/Segmented Viruses." *Current Opinion in Virology* 33 (December): 156–66. https://doi.org/10.1016/j.coviro.2018.08.013.

Wang, Yong Jie, Qin Yao, Ke Ping Chen, Yong Wang, Jian Lu, and Xu Han. 2007. "Characterization of the Genome Structure of Bombyx Mori Densovirus (China Isolate)." *Virus Genes* 35 (1): 103–8. https://doi.org/10.1007/s11262-006-0034-3.

Weber, Friedemann, Valentina Wagner, Simon B. Rasmussen, Rune Hartmann, and Søren R. Paludan. 2006. "Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses." *Journal of Virology* 80 (10): 5059–64. https://doi.org/10.1128/JVI.80.10.5059-5064.2006.

Wichgers Schreur, Paul J., Richard Kormelink, and Jeroen Kortekaas. 2018. "Genome Packaging of the Bunyavirales." *Current Opinion in Virology* 33 (December): 151–55. https://doi.org/10.1016/j.coviro.2018.08.011.

Wolf, S., C. M. Deom, R. N. Beachy, and W. J. Lucas. 1989. "Movement Protein of Tobacco Mosaic Virus Modifies Plasmodesmatal Size Exclusion Limit." *Science* 246 (4928): 377–79. https://doi.org/10.1126/science.246.4928.377.

Wrzesińska, Barbara, Lam Dai Vu, Kris Gevaert, Ive De Smet, and Aleksandra Obrępalska-Stęplowska. 2018. "Peanut Stunt Virus and Its Satellite RNA Trigger Changes in Phosphorylation in N. Benthamiana Infected Plants at the Early Stage of the Infection." *International Journal of Molecular Sciences* 19 (10). https://doi.org/10.3390/ijms19103223.

Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. "Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus." *Scientific Reports* 7 (1): 1–15. https://doi.org/10.1038/s41598-017-05335-8.

Yoon, J-Y, P. Palukaitis, and S-K Choi. 2019. "Chapter 1: Host Range." In *Cucumber Mosaic Virus*, 15–18. Virology. The American Phytopathological Society. https://doi.org/10.1094/9780890546109.004.

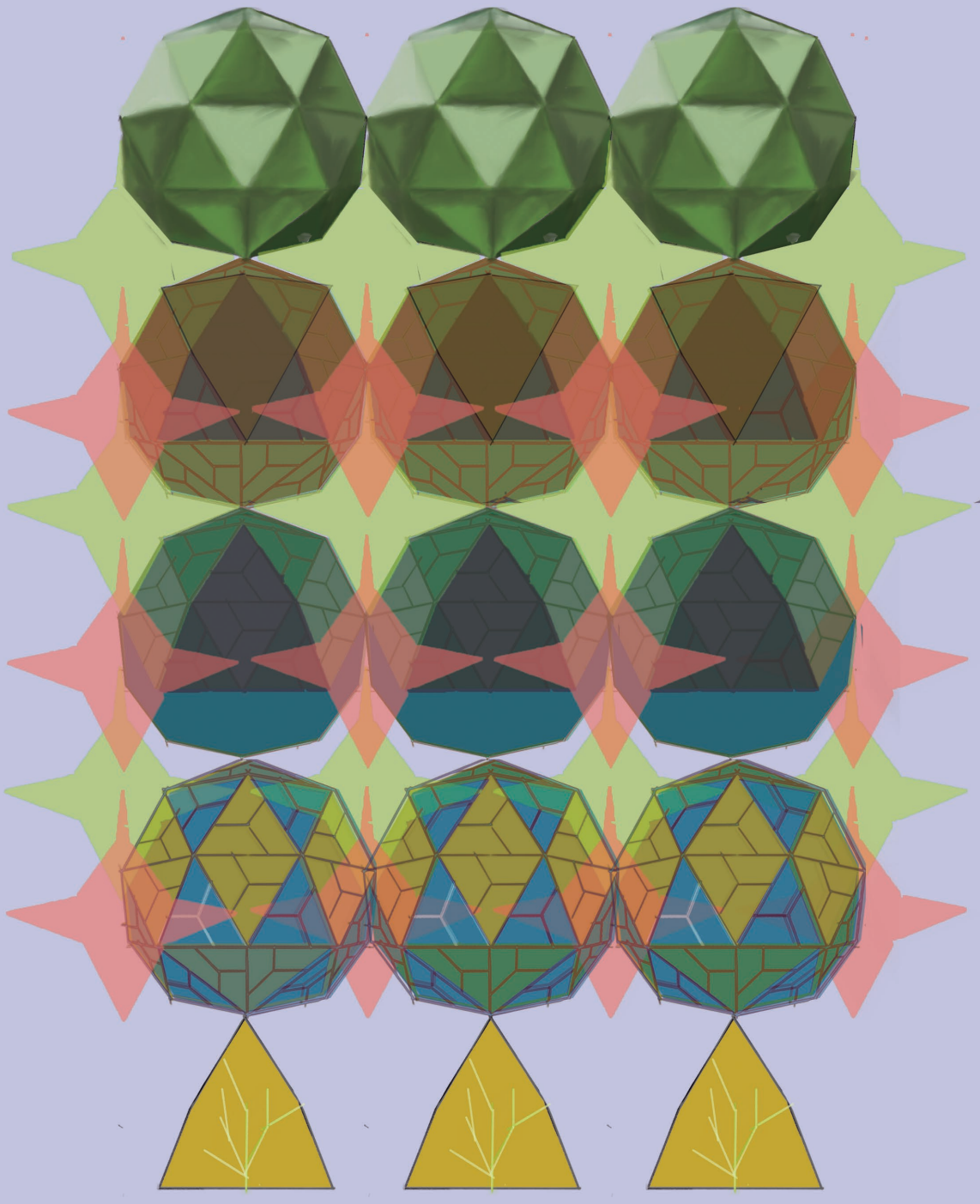
Yu, Nai-Tong, Hui-Min Xie, Yu-Liang Zhang, Jian-Hua Wang, Zhongguo Xiong, and Zhi-Xin Liu. 2019. "Independent Modulation of Individual Genomic Component Transcription and a Cis-Acting Element Related to High Transcriptional Activity in a Multipartite DNA Virus." *BMC Genomics* 20 (1): 573. https://doi.org/10.1186/s12864-019-5901-0.

Zhao, Wan, Qianshuo Wang, Zhongtian Xu, Renyi Liu, and Feng Cui. 2019. "Distinct Replication and Gene Expression Strategies of the Rice Stripe Virus in Vector Insects and Host Plants." *The Journal of General Virology* 100 (5): 877–88. https://doi.org/10.1099/jgv.0.001255.

Zimmer, Christoph T., William T. Garrood, Kumar Saurabh Singh, Emma Randall, Bettina Lueke, Oliver Gutbrod, Svend Matthiesen, et al. 2018. "Neofunctionalization of Duplicated P450 Genes Drives the Evolution of Insecticide Resistance in the Brown Planthopper." *Current Biology: CB* 28 (2): 268-274.e5. https://doi.org/10.1016/j.cub.2017.11.060.

Zitter, T. A., and J. F. Murphy. 2009. "Cucumber Mosaic Virus. The Plant Health Instructor." *American Phytopathological Society*. https://doi.org/10.1094/PHI-I-2009-0518-01.

Zwart, Mark P., and Santiago F. Elena. 2020. "Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments." *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa022.



A quantitative perspective on the evolutionary costs and benefits of multipartite virus genomes

Marcelle L. Johnson[1,2], Dieke Boezen[1,2], Alexey A. Grum-Grzhimaylo[1,3], René A. A. van der Vlugt[2], J. Arjan G.M. de Visser[4], Mark P. Zwart[1]

[1] Netherlands Institute of Ecology (NIOO-KNAW), P.O. BOX 50, 6700 AB, Wageningen, The Netherlands

[2] Laboratory of Virology, Wageningen University and Research, P.O. BOX 16, 6700 AA, Wageningen, The Netherlands

[3] Westerdijk Fungal Biodiversity Institute, Uppsalalaan 8, 3584 CT, Utrecht. The Netherlands

[4] Laboratory of Genetics, Wageningen University and Research, P.O. BOX 16, 6700 AA, Wageningen, The Netherlands

Submitted

Abstract

Multipartite viruses individually package the multiple segments that comprise their genome. Hence, their between-host transmission is dependent on spread and infection by multiple virus particles. This dependence implies a cost in the form of reduced transmission compared to viruses with only a single segment or those that package all constituent genome segments into a single virus particle (i.e. some segmented viruses). The notion of this cost to transmission has spurred a search for the possible benefits associated with a multipartite organisation. However, the exact costs of a multipartite organisation remain elusive, as only a few studies have considered the proposed mechanisms quantitatively. To evaluate the costs and benefits of multipartition we developed three quantitative modelling approaches. First, we present a stringent approach for measuring the cost to transmission and show its quantitative dependence on number of genome segments and dosage. Second, it was recently shown that a multipartite virus can share its gene products between cells, which could alleviate the cost to within-host spread. We show that this mechanism itself appears to be generally costly and confers benefits only under conditions that are incompatible with the putative benefits of multipartite viruses. Finally, we consider a possible benefit of multipartition, i.e. that rapid changes in gene expression through changes in the copy number of genome segments may help viruses adapt to their host environment. Since this may extend their host range, we compare the host range of monopartite, segmented and multipartite plant viruses using a genomics metadata approach that is less biased than traditional approaches. In support of this hypothesis, we show that multipartite and segmented viruses appear to have larger host ranges than monopartite viruses. Our synthesis highlights key outstanding questions about the evolutionary significance of multipartite viruses and the need to address these questions with quantitative approaches.

Introduction

Virus DNA or RNA genomes may be discretely organised into single or multiple segments which differ in their packaging into virus particles (Michalakis and Blanc 2020). Viruses with a single genome segment are termed monopartite viruses, whereas viruses that co-package all their genome segments into single viral particles or ribonucleoprotein complexes are termed segmented viruses. Alternatively, multipartite viruses have several genome segments that are each individually packaged into virus particles (Figure 1). The segmented and multipartite viruses require co-infection of multiple genome segments containing core viral functions for the initiation of viral replication and subsequent propagation in hosts. There is an inherent cost to transmission with multipartite viruses, as all segments are required to initiate an infection, there is a lower probability that all segments are sampled from an infection and reinoculated together in a new host upon transmission. This change in infection probability has been estimated by using dose-response assays and observing the change in the slope to identify the number of genome segments and the reduction in the number of primary infection sites (Lauffer and Price 1945; Fulton 1962). Michalakis and Blanc (2020) review support for the various costs of segment transmission at both within and between host level, such as (1) the co-occurrence of genomic segments with proteins or mRNA from other segment types within cells (Sicard et al. 2019), (2) possible avenues for sorted transmission by ribonucleoprotein complexes during cell-cell and long-distance movement, and (3) infections with essential virus segments, can be complemented with absent segments to reconstitute the full virus genome, as demonstrated with data from faba bean necrotic stunt virus (FBNSV) (Di Mattia et al. 2022). Furthermore, it was demonstrated that co-transmitted groups of the core FBNSV segments could acquire other FBNSV segments (Di Mattia et al.2022). These mechanisms are active during different stages of the transmission cycle and likely interdependently minimise the overall transmission cost of multipartition. Quantitative estimates for the cost to transmission and evidence for mechanisms which minimise cost acting in concert are scarce.

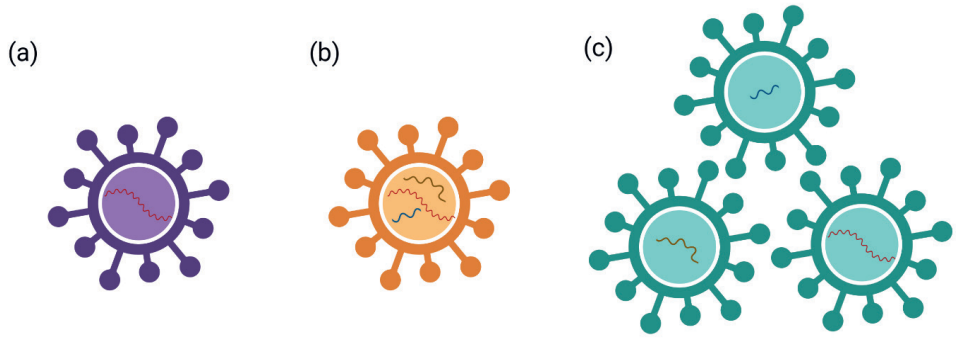


Figure 1. Conceptual overview of virus genome organisation. *Viral genomes may be organised into single or multiple genome segments which are packaged into capsids. Illustrated here: (a) monopartite viruses package their single segment into one particle, whilst (b) segmented viruses package their multiple segments into one particle and (c) multipartite viruses package their multiple segments individually. (b) and (c) may also be referred to as multicomponent systems. Created with Biorender*

Multipartite viruses represent ~30 - 40% of plant virus genera (Lucía-Sanz and Manrubia 2017; Michalakis and Blanc 2020), suggesting adaptive benefits that compensate for any hurdles to transmission imposed by their multipartite organisation. Sicard et al. (2016) provide a historical overview of the field and highlight that many of the proposed benefits associated with a multipartite genome organisation also would pertain to segmented viruses. For example, faster replication of short genome segments and reassortment between segments would also pertain to segmented viruses. There is one hypothesised benefit of genome segmentation which uniquely benefits multipartite viruses: the "**genome formula**" concept, where the frequency of multipartite genome segments can change rapidly in a host- and tissue-specific manner (Sicard et al. 2013; Wu et al. 2017). These changes could provide a means to rapidly adapt to a new host by altering gene expression (Sicard et al. 2013), suggesting that the genome formula increases the host range. Different genome formulae have been estimated for several plant virus species such as FBNSV (Sicard et al. 2013), alfalfa mosaic virus (AMV) (Wu et al. 2017), banana bunchy top virus (BBTV) (Yu et al. 2019), cucumber mosaic virus (CMV) (Boezen et al. 2023) and for a segmented animal-infecting virus; bluetongue virus (BTV) (Moreau et al. 2020). Measured genome formulae in different host species show an unequal stoichiometric ratio of segments, further providing support for its role in regulating gene expression (Sicard et al. 2013). The widespread nature of multipartite viruses infecting plants and the genome formula suggests that it may be a mechanism for increasing the host range of multipartite viruses.

In this study, our goal is to give an overview of recent theoretical and experimental work that has been done on the evolutionary costs and benefits of multipartite viruses. Where possible, we employ a quantitative perspective that allows testing theoretical predictions and considers the implications of the proposed mechanisms that underlie empirical observations. We will focus on whether there is evidence for a cost of multipartition when the sharing of viral gene products between cells (Sicard et al. 2019) is beneficial, and whether there is evidence that multipartite viruses have a broader host range than the host ranges of monopartite and segmented viruses.

The transmission cost of multipartition

Early research (Lauffer and Price 1945; Fulton 1962) not only suggested the existence of multipartite viruses but also noted that multipartition has a cost of reduced transmission. These studies considered the relationship between virus dose and host infection to make inferences on viral infection kinetics. The key experiments testing these predictions could be performed by exploiting the fact that in some hosts, plant viruses induced local lesions, i.e. readily visible necrotic responses on the inoculated leaves which can be quantified (Bald 1937; McKinney 1927). By preparing doses of virus inoculum in a dilution series and measuring the number of lesions formed, it is possible to infer the relationship between the number of virus particles needed to initiate infection (Parker 1938). Single-hit kinetics describe a direct relationship between the dose of an infectious unit and the number of plaques or lesions that form (Druett 1952). In multi-hit kinetics, there is a steeper gradient in the relationship between the dose of infectious units and the number of plaques or lesions formed due to the requirement for complementation (Fulton 1962; Lauffer and Price 1945). Certain viruses had responses which were inconsistent with predictions of single-hit kinetics and instead showed steep responses

characteristic of multi-hit kinetics (Lauffer and Price 1945) indicating the involvement of multiple infectious units. More definitive results were obtained by Fulton (1962), who considered the dose-response of the tri-segmented multipartite prunus necrotic ringspot virus (PNRSV) in detail and found direct evidence for complementation between virus particles. Fulton (1962) showed that partially UV-inactivated PNRSV particles increased the infectivity of the inoculum disproportionately, and from these strong synergistic interactions between the viral genetic component, he inferred that: “Two or more virus particles at one site might provide a complete complement of genetic units [...]” (Fulton 1962). Fulton (1962) demonstrated synergism between virus particles, but the implication of these observations is also that – all other things being equal – a multipartite virus has impaired transmission compared to a monopartite virus, an effect that is strongly dose-dependent.

Although these classic inferences of genome segmentation are based on very simple models of infection kinetics, dose-response is generally considered a reliable indicator of multipartition. Classic work with plant viruses considered local lesions as a response (Lauffer and Price 1945; Fulton 1962). In other studies, the response has been quantified by considering the proportion of infected hosts (Zwart, Daròs, and Elena 2011; Sánchez-Navarro, Zwart, and Elena 2013). When the empirical dose-response is considered across all different monopartite virus systems, these relationships invariably are compatible with predictions for monopartite viruses (van der Werf et al. 2011; Zwart and Elena 2015; Gutiérrez and Zwart 2018). In contrast, Ladner et al. (2016) found evidence that guaico culex virus (GCXV) might be a multipartite virus by analysing plaque formation in mosquito cell lines. From the steep dose responses observed, they predicted multiple virus particles are required for infection, which is congruent with the 3–5 segments detected using high throughput sequencing. It is important to stress that dose-response kinetics are not an infallible indicator of multipartite systems. First, Michalakis and Blanc (2020) remark that Ladner et al. (2016) do not conclusively show that GCXV is a multipartite virus, as the steep dose-response observed is also consonant with a segmented virus with low-fidelity packaging. A steep dose-response relation, therefore, can only suggest a multicomponent system and not confirm multipartition. Second, there are many reasons why a dose-response relation may differ from theoretical predictions. For example, differences in host susceptibility are predicted to lead to a more gradual response as the dose is increased (Regoes et al. 2003), and therefore, it is possible that a multipartite virus shows the predicted monopartite response in a host population with variable susceptibility (Zwart and Elena 2015). A historical footnote illustrates this point, as Fulton (1962) reported a single-hit dose response for the tri-segmented, multipartite CMV for reasons that remain unclear to us. Dose-response kinetics are a useful indicator of genome organisation, but the results need to be confirmed by the analysis of both the genome and virus particles.

All previous studies have focused on the steepness or shape of the dose-response relation. However, simple models make two clear predictions of the infection kinetics of multipartite viruses: as the number of genome segments increases, not only does the dose-response become steeper, but its position also shifts to the right, as it becomes less likely that the virus will successfully infect (Fig. 2a) (Gutiérrez and Zwart 2018). To date, the shifts in the position of the dose-response curve have not been included in analyses of virus infections. The combined effect of the shifted dose response and its altered shape ultimately determine the cost of multipartition to transmission. Therefore, dose-response analyses to date evaluate

whether the experimental results are compatible with a given number of segments, but they do not evaluate whether the cost of multipartition is in agreement with model predictions.

Comparing only the integral of dose-response of natural monopartite and multipartite viruses quantifying the transmission cost will be confounded by different segment numbers, genetic diversity and infectivity. A more satisfactory approach is to engineer a panel of hosts with different requirements for infection, for example by expressing viral genome segments so that they are no longer needed for infection. The same inoculum could then be used to challenge the different hosts, and the combined difference between the shift in the dose-response curve and dose-response integrals could be determined to estimate the cost of multipartition. In previous work, the dose-response was measured for tobacco plants constitutively expressing one or two genome segments of the tripartite alfalfa mosaic virus (AMV) (Taschner et al. 1991), but the authors only considered the shape of the dose-response (Sánchez-Navarro, Zwart, and Elena 2013). Here we have re-analysed these data, to consider both the shape and position of the response (Text Box 1). We found (Figure 2b-c, and Supplementary SI) that the cost of multipartition was larger than predicted by the null model of infection (Figure 2b), as the position of the dose-response shifted a greater distance than predicted. Alternative models that incorporated experimental variation in the infectivity of virus particles and differences in the susceptibility of host plants due to the expression of viral genome segments could better account for the experimental data (Fig. 2c). However, both of these models had similar levels of empirical support (Table S2), and our analysis, therefore, cannot identify which of these two mechanisms might underlie this larger-than-expected cost. Our analysis does provide further support for the idea that multipartition has a cost to transmission, confirming expectations based on the shape of the dose-response. This re-analysis of data confirms the cost of multipartition to transmission exists, and may even be larger than predicted by simple models of infection.

Text Box 1: Testing predictions for the between-host cost of multipartition

Sanchez-Navarro et al. (2013) performed dose-response experiments with a wild-type AMV inoculum in three tobacco plants: wild-type plants (Wt), plants expressing AMV RNA2 (P2) and plants expressing both AMV RNA1 and RNA2 (*P12*). A simple model of infection that assumes (1) independent action of the three RNA segments during viral invasion and (2) that all three segments need to be present for a productive infection was used to generate predictions (see supplementary text S1). We considered a tripartite virus with a balanced genome formula and tested a set of four models in which the probability of infection per virus particle type (i.e., containing a particular genome segment) was fixed (Model 1), or alternatively, it was dependent on the genome segment (Model 2), the host plant type (Model 3), or both on genome segment and host plant (Model 4). We fitted these models with a maximum likelihood approach (Table S1) and performed model selection with the Akaike information criterion (AIC; Table S2).

We found the highest support for Models 2 and 3, which had Akaike weights of 0.328 and 0.632. The Akaike weight indicates the relative likelihood of a model, and the sum of all weights for the set of all models included is one (Johnson and Omland 2004). Although Models 2 and 3 have considerably more support than Models 1 (Akaike weight = 0.000) and 4 (Akaike weight = 0.040), we cannot make a meaningful distinction between the best-supported models because they both have considerable empirical support. Although it incorporates both mechanisms introduced in Models 2 and 3, Model 4 is not supported because there is not an appreciable increase in fit over Models 2 or 3. The model selection results, therefore, suggest that there are differences in infectivity between virus particles, but we cannot say whether these depend on the virus particle type, host plant or both. When we plot the predicted dose-response, it is clear that Model 1 predicts the dose responses will be closer together than observed for the data (Figure 2b). Models 2–4 better fit the data and predict larger differences in the dose responses over the different types of host plants (Figure 2c). The analysis, therefore, suggests that the cost of multipartition is real and might be larger than predicted by simple models of infection kinetics. For a complete description of the results obtained and complete plots of all models, see supplementary materials (Supplementary S1).

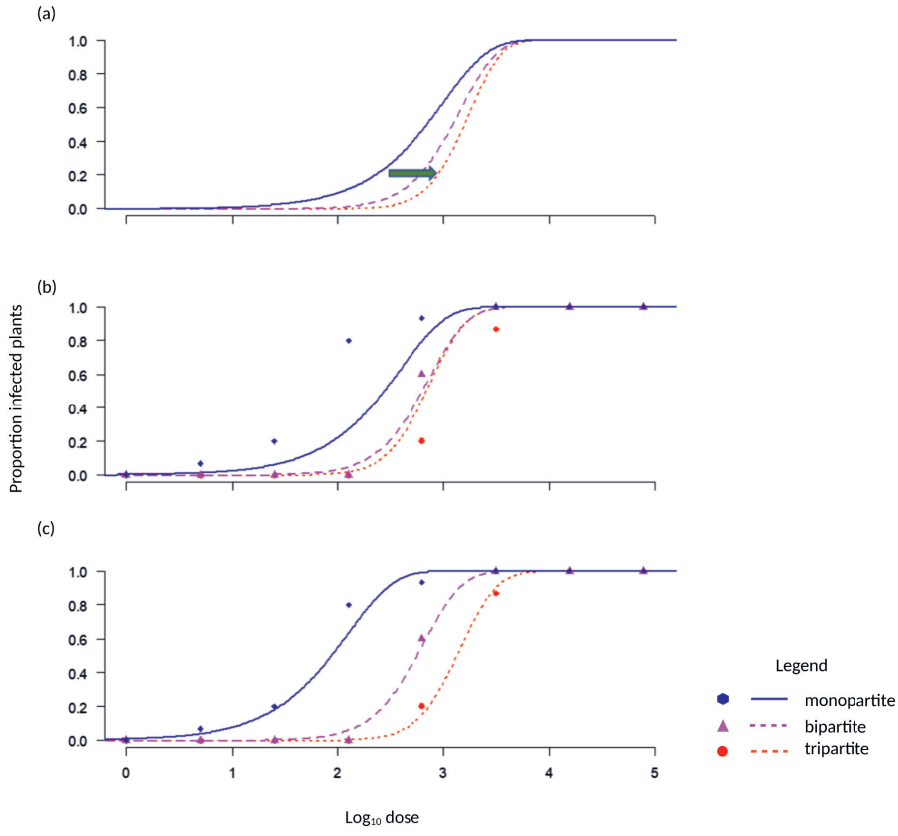


Figure 2. An overview of predictions and data for the cost of multipartition. *Lines indicate model predictions and points indicate empirical data. (a) Theoretical dose-response curves are shown for viruses with a different number of genome segments and balanced GF. As the number of segments increases, the dose-response becomes steeper and shifts to the right, as emphasised by the green arrow. (b) Fitted Model 1 and experimental data from Sanchez-Navarro et al (2013) are shown, where virus segments were made redundant by their constitutive expression in the host plant. For example, in the "monopartite" case, AMV RNA1 and RNA2 are expressed by the host plant and only RNA3 is required for infection. The dose-response predictions for the bipartite and tripartite viruses are different than in panel 1, both in terms of shape and position, because the AMV genome formula is not balanced. The model predicts that the dose-response for the different numbers of segments required for infection are closer together than observed. (c) Fitted model 3 and the same experimental data shown in (b) are shown. Models 2, 3, and 4 provide very similar predictions of dose-response, although Models 2 and 3 are better supported than 4 due to their lower complexity. We, therefore, only show the best-supported Model 3 here to illustrate all of these models do fit the data better.*

Mechanisms to minimise the cost to transmission of multipartite viruses

To remedy costs associated with a multipartite organisation, several mechanisms have been proposed that may help reduce these costs: (1) virus aggregation; the packaging of multiple virus particles in a single occlusion body, as seen for the alphabaculoviruses (Rohrmann 2019), by collective transmission of virus particles in the form of virus particle aggregates (Andreu-Moreno and Sanjuán 2018; Sanjuán and Thoulouze 2019), the between-host transmission of virus particles which adhere to another as a core genome (Gallet, Michalakis, and Blanc 2018), (2) as supramolecular viral RNA structures during within-host transmission (Gilmer, Ratti, and Michel 2018) and (3) the ability to share gene products across adjacent host cells, making replication possible despite the absence of a segment (Sicard et al. 2019). At present, there is little evidence showing that mechanisms of (1) – (2) pertain to multipartite viruses. However, there is evidence of gene product sharing in a multipartite virus (Sicard et al. 2019), and we thus focus on this proposed benefit.

The ability to share gene products across cells and tissues circumvents the requirement for all segments to be present within the same cell at the initial stages of infection. This mechanism has been demonstrated for FBNSV infection, in which gene products involved in key viral functions – replication, movement and coat protein synthesis – were found to accumulate in cells wherein the genome segments encoding these functions were absent (Sicard et al. 2019). This would lower the within-host costs associated with local and systemic movement and facilitate faster systemic spread, alleviating the costs of multipartition to within-host spread by relaxing the requirement for genome integrity. In the broadest terms, we expect a virus that does not share its gene products may infect fewer cells, whilst the replication of a virus that does share its products may be affected by changes in the presence of viral gene products or the lack of segments to replicate. However, there will also be costs associated with the sharing of gene products for the virus in the donor cell: some resources in an infected cell will need to be directed to the expression of gene products in recipient cells, at the expense of replication or virus particle production in the donor cell. This cost to the donor cell will depend on (1) the mechanism of gene product sharing between donor and recipient cells, as producing viral proteins will come at a higher cost than producing only messenger RNA, and (2) the intensity of product sharing. Both are currently unknown, and it is therefore unclear under what conditions viral gene product sharing is beneficial.

To explore when the sharing of viral gene products with other cells is beneficial, we extended a model of a bipartite virus incorporating variation in the genome formula (Zwart and Elena 2020) (Figure 3). In the original model, the intra-cellular genome formula and virus particle yield per cell are linked by the probability density function of the normal distribution, with mean μ and variance σ^2. Parameter σ^2, therefore, determines the magnitude of the decrease in virus particle yield as the genome formula departs from its optimal value μ, meaning that σ^2 determines how sensitive virus particle yield is to changes in the genome formula. In the original study, it was shown that multipartite viruses can outcompete their monopartite cognates when virus replication is sensitive to changes in gene expression (I.e., when $\sigma^2 \leq 0.1$). In our model here, we extended this model to include the production of both virus gene

products and virus particles in cells. A detailed description of the model is available in supplement S2.

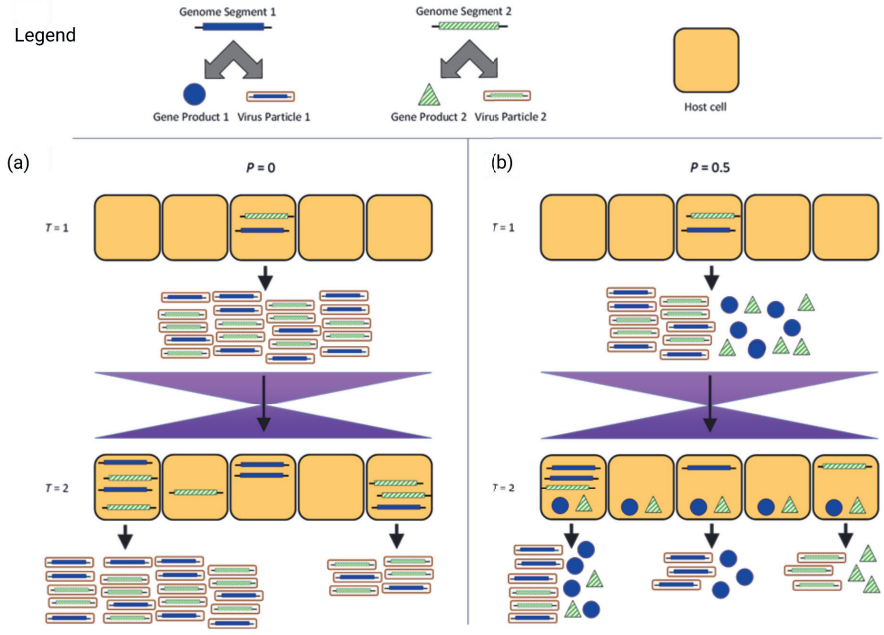


Figure 3. Overview of the virus gene-product sharing model*. Two virus segments are blue and green, as are the gene products they encode. The parameter ρ determines what proportion of the virus resources in each cell is dedicated to producing gene products that will be shared uniformly with all other cells. The model is agnostic about the mechanism of gene-product sharing, as this could occur through the movement of mRNA or trafficking of proteins. In both panels (a) and (b), we illustrate situations in which the production of virus resources is sensitive to changes in the proportion of gene products 1 and 2 ($\sigma^2 > 1$), and in the first round of infection (T = 1) a single cell is infected by virus particles in a 1:1 ratio for segments 1 and 2. We do not illustrate the gene products formed in a cell but not shared, but these are equivalent to the expression from a single genome segment present in a cell which supports replication. (a) Illustrates a scenario in the absence of gene-product sharing (ρ = 0). All viral resources are committed to virus particles, and hence infection does not proceed in those cells missing one of the two segments at T = 2. When the ratio of virus gene products deviates from 1:1 in a cell, the production of virus particles becomes lower. (b) A scenario when 50% of viral resources in each cell are committed to gene-product sharing (ρ = 0.5). Our model assumes gene products are shared equally with all cells that can be exposed to the virus in the next round of infection, allowing some level of segment replication in each cell in which a virus particle introduces a genome segment. However, when the ratio of gene products deviates from 1:1 in a cell, the total virus resources generated are again lower. When only one segment is present in a cell, only that segment can be replicated, and hence, all virus particles or gene products generated will be of a single type.*

From our model results, we can draw three main conclusions (Figure 4). First, for a large part of the parameter space we considered, our model predicts that viral gene product sharing is neutral or costly. As the fraction of gene products shared (ρ) becomes larger, the fitness cost becomes higher. However, at very low multiplicities of cellular infection (MOI), the cost disappears as the multipartite virus always performs very poorly. In our model, the gene products are not targeted to cells that are likely to be exposed to virus particles, which might explain its high cost. Second, there was a small parameter space in which virus gene product sharing was beneficial: for moderate levels of sharing ($\rho < 0.5$), when MOI was relatively low (MOI: ~ 3 ~ $10^{0.5}$), and when the virus is insensitive to the genome formula ($\sigma^2 = 10$). It is intuitive that there are benefits associated with gene product sharing under these conditions. Moderate levels of sharing reduce the opportunity cost to replication in the donor cell, and at low MOI, gene product sharing will be more beneficial because only a single segment will be present in many cells. Furthermore, low sensitivity to the genome formula allows virus replication in cells in which only a small amount of a missing segment has been shared. Third, the benefits of gene product sharing occur in a different parameter space than the benefits associated with adaptive changes in the genome formula. Genome formula variation is suggested to benefit virus gene expression regulation in novel environments (Sicard et al. 2013) and has been linked to transcriptional regulation in a multipartite DNA virus (Gallet et al. 2022). The benefits of changes in the genome formula (Sicard et al. 2013; Wu et al. 2017) are predicted to be greatest and outweigh the cost of multipartition when virus particle production is sensitive to changes in the genome formula ($\sigma^2 \leq 0.1$) (Zwart and Elena 2020). By contrast, we find that the benefits associated with gene product sharing only outweigh the costs when virus particle production is insensitive to the genome formula ($\sigma^2 = 10$). The model predicts that the genome formula and gene product-sharing mechanisms require different conditions to be beneficial, making it unlikely that a virus can exploit both of them concurrently.

We can, therefore, conclude that although gene product sharing has the potential to reduce the cost of multipartition at the within-host level, there are many conditions under which it imposes an additional cost. Given the specific set of conditions needed, we predict that the occurrence of gene product sharing to lower the cost of multipartition may not be common, unless (1) this sharing of gene products can be achieved at a very low cost, or (2) there are mechanisms that result in virus particles and gene products reaching the same susceptible cells. Finally, we note that a high level of infections supported by shared gene products may also result in a cost to between-host transmission. When replication without a complete genome is enabled by gene product sharing, the total of virus particles produced in these cells will represent incomplete genomes. Therefore, these particles can only contribute to the between-host transmission when they are complemented by the missing segments during virus acquisition or upon concomitant transmission of virus populations.

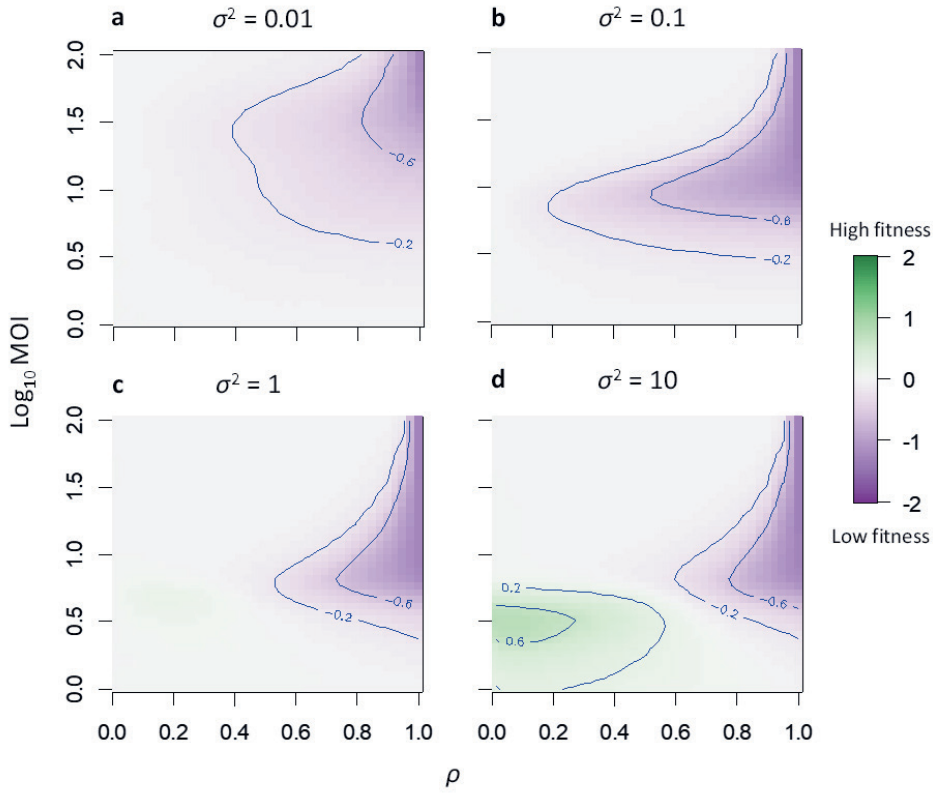


Figure 4. Model showing the effect of gene-product sharing of a single segment on viral fitness during local cell-cell movement. *The effects of viral-gene-product sharing on viral fitness on a bipartite virus. On the x-axis ρ is given, the proportion of viral gene products that spread to other cells. On the y-axis the log_{10} of the cellular multiplicity of infection (MOI) is given. The colours indicate viral fitness compared to a virus that does not share its gene products, as determined by total virus particle production during a simulation of multiple rounds of infection and as compared to a virus that does not share any gene products. When viral fitness is higher than the non-sharing reference the area is green, when viral fitness is worse the area is purple, and when viral fitness is similar the area is white. Contour lines have been included to highlight parameter space with high or low fitness. The value σ^2 indicates how sensitive virus particle production is to changes in the genome formula, with low values (i.e., $\sigma^2 = 0.01$) indicated high sensitivity and a quick drop in accumulation as the population moves away from the optimal value, and high values (i.e., $\sigma^2 = 10$) indicating low sensitivity. For each condition, 10^4 simulations were run with 5 cells per round of replication and 20 rounds of replication. See Supplementary S2 for a complete description of the model and model results for a broader range of conditions.*

Benefits of genome segmentation: evidence for extended host range

The discovery of multipartite viruses was linked to the effect of their genome and virus particle organisation on infection kinetics, and implied a high cost to virus transmission (Lauffer and Price 1945; Fulton 1962). The recognition of this predicted expected cost has fuelled research and a debate on whether there are benefits linked to this genome organisation that may alleviate these costs (Reanney 1982; Sicard et al. 2016; Lucía-Sanz and Manrubia 2017). However, to date the putative benefits of multipartition – if they exist – remain elusive. The benefits of a *de novo* multicomponent animal virus, in terms of higher virus particle stability, were convincingly demonstrated by (Ojosnegros et al. 2011). Similar verification for an authentic multipartite virus has not been reported to date.

Whereas previous work has suggested there is variation in the frequencies of segments (Hajimorad et al. 1991), Sicard and colleagues performed the first systematic study on the genome formula using FBNSV showing that the genome formula is unbalanced and converges on a host-dependent equilibrium, the so-called setpoint genome formula (Sicard et al. 2013). To date, for all multipartite viruses for which it has been measured, the genome formula is unbalanced. These viruses have different genome composition, including the positive single-stranded RNA (+ssRNA) AMV (Wu et al. 2017) and CMV (CMV)(Boezen et al. 2023), the negative single-stranded RNA (-ssRNA) rice stripe virus (RSV)(Zhao et al. 2019), the single-stranded DNA (ssDNA) BBTV (Yu et al. 2019) and the insect virus Bombyx mori bidensovirus (BmBdV) (Hu et al. 2016). Sicard et al. (2013) suggested that rapid changes in the genome formula could be adaptive, resulting in changes in gene expression in different host environments. The different multipartite viruses and their host species in which the genome formula has been measured suggest it has a role in facilitating gene expression regulation. Gene copy number is known to play a role in rapid evolutionary responses seen in monopartite viruses (Elde et al. 2012) and bacteria (Andersson and Hughes 2009). By contrast, later work on FBNSV showed that, in fact, the genome formula could also buffer changes in transcription rates in different host environments, perhaps contributing to stable gene expression (Gallet et al. 2022). Wu et al. (2017) also showed the genome formula is a negative frequency-dependent stable equilibrium for AMV, and showed the GF equilibrium is associated with higher virus accumulation. Moreau et al. (2020) demonstrated that the segmented BTV genome formula is also host-dependent. Finally, Zwart and Elena (2020) modelled the evolution of the genome formula and showed that, under some conditions, multipartite viruses can outcompete their monopartite cognates. This model predicts that the key conditions that determine how competitive multipartite viruses will be are the sensitivity of virus particle yield to the genome formula, whether the environment demands changes in the genome formula and the MOI. Taken together, this body of work suggests that the genome formula may play a role in the rapid adaptation of multipartite viruses to varied host environments.

If the genome formula facilitates virus adaptation to the host environment, we expect that it could extend the host range of viruses. This extension of host range would apply to multipartite viruses and selected segmented viruses. Segmented viruses are known to differ in the packaging fidelity of genome segments and can broadly be classified as high or low fidelity packagers (Nakatsu et al. 2016). High fidelity packagers ensure that a full genome

complement containing all virus segments are co-packaged into ribonucleoprotein complexes (Chou et al. 2012). Low fidelity packagers have imperfect segment co-packaging, where not all segment types are packaged together; there may be selective preferential co-packaging of segment types, disparate segment number and frequencies which influences infectivity (Diefenbacher, Sun, and Brooke 2018; Brooke 2014; Brooke et al. 2014). This packaging inefficiency allows low fidelity packagers to display genome-formula variation. Host range varies across viruses, with some exhibiting a narrow known host range limited to a single host species or even genotypes and variants within a species (e.g bacteriophages which strain-specifically infect host species) (de Jonge et al. 2019). Other viruses exhibit a broad host range, infecting many host species or taxa from different groups (e.g. Tomato spotted wilt virus (Parrella et al. 2003) or West Nile Virus which infects a range of animals from mosquitoes to birds to horses) (Marm Kilpatrick et al. 2006; Campbell et al. 2002). As only some segmented viruses are expected to be able to benefit from the genome formula, we would predict that the host range is the broadest for multipartite viruses, followed by segmented viruses, followed by monopartite viruses. Others have already noted that the number of genome segments is a predictor of host range; A study that measured host ranges with classical methods suggests that viruses with 3–4 genome segments have broadest host ranges (Moury et al. 2017).

Although this hypothesis is straightforward, in practice it is not easy to test. First, susceptibility of a host to a virus is not a binary characteristic: it depends on the conditions under which a host is exposed to a virus (Morris and Moury 2019), as many variables such as the extent of exposure (i.e., dose), host immune state, the presence of other viruses and microorganisms will affect whether infection occurs. Second, traditional host range testing will be biassed by the number of hosts tested for their susceptibility to a pathogen and any biases may be self-reinforcing. We therefore are interested in measuring the realised host range for a large number of viruses in a manner that is not biassed by the extent of host-range testing. The realised host range includes all susceptible host species in which a parasite can complete its life cycle and does not include the probability of encounter and other ecological barriers and would always constitute a much broader estimate of host range (Rohde 1994; Hutchinson 1957). The observed host range describes the proportion of susceptible host species in which a parasite completes its life cycle and includes the probability of exposure to the parasite as well as the influence of other ecological barriers that affect infection and represents the host range which can be appreciably measured (Rohde 1994; Hutchinson 1957). The observed host range will therefore be narrower than the realised host range. We therefore devised a less biassed approach for measuring the host range by using submissions from the NCBI Virus database which contains high-throughput sequence (HTS) metadata from DNA and RNA viruses of different virus genome organisations and notes the host from which it is derived (See supplementary S3 for a detailed description of methods). HTS data provides a better estimate of the observed host range, as a larger sample of host plants are assessed however it is difficult to distinguish if infections undergo the complete virus life cycle, i.e. virus replication and transmission, as there is no control on the timing of infection or if they may represent virus introductions onto non-hosts.

For each plant virus species, hosts were identified to genus level by combining data from the NCBI Virus database (Hatcher et al. 2017) and ICTV virus metadata resource (https://ictv.global/vmr). The genus *Begomovirus* contains both monopartite and multipartite members, and was therefore classified separately. For each virus species, we identified the number of unique host genera in which it had been identified. We performed our analysis on

the genus level, because we are interested in hosts that are highly diverging. We also wanted to know the number of observations for each virus species for our statistical analysis, so that we could weigh the observed host range by the total number of observations. Since each segment of a genome is initially counted as an observation, we corrected the number of observations for the multipartite and segmented viruses by dividing the number of segments observed by the mean number of segments for species that belong to a viral genus.

Our analysis suggests that the host range is significantly narrower for the monopartite viruses than for all other groups considered (Table 1), providing support for our hypothesis on differences between monopartite and multipartite viruses. There were no significant differences between any of the other groups, so there is no evidence that the host range is broader for multipartite than segmented viruses. However, despite the difference with multipartite viruses being insignificant, the largest host range was estimated for the segmented viruses. Our results do support the idea that genome segmentation is associated with a broader host range. Segmented viruses with high-fidelity packaging – which are presumably the majority of segmented viruses – cannot conserve changes in the genome formula during virus spread. Changes in the genome formula cannot therefore be the only reason for the differences in inferred host range.

Table 1. Estimate of host range for different groups of viruses

Group	N[a]	NLL[b]	Host range[c] [95% CI[d]]
Monopartite	869	1387.726	3.03 [2.63-3.23]
Multipartite	196	484.343	4.17 [3.33-6.25]
Segmented	67	216.413	6.67 [4.00-12.50]
Begomovirus	413	748.026	4.17 [3.85-5.00]

[a] The number of viruses per group. [b] Negative log likelihood of the calibrated model, using 10^5 permutations. [c] The estimated host range, given as $1/\theta$. [d] Confidence interval, as determined by 1000 bootstraps of the virus species included in the model calibration.

Discussion

In this paper we present quantitative approaches for measuring the cost to transmission of multipartite viruses, introduce a modelling approach to quantify how a within-host mechanism may reduce the cost to transmission and lastly estimate the benefit of genome segmentation for increasing the host range of multicomponent viruses. Multipartite viruses are a unique group of segmented viruses which package and transmit their genomic segments individually in virus particles. A multipartite genome organisation poses an inherent cost to infectivity; multiple virus particles are required during transmission between hosts, which limits the chances of a successful infection. Here we present a quantitative framework for estimating the cost of within- and between-host transmission for multipartite viruses, using both the change in position and shape of the dose-response relation. Using empirical data we show that the number of virus particle types required for infection affects the position of the dose

response curve, next to altering its shape. The change in dose-response position increases the estimated cost of transmission to a greater extent than predicted by our null model, suggesting that mechanisms which provide incremental reductions in the cost to transmission may have a disproportionate impact on increasing infection success. However, these experiments employed transmission by means of mechanical inoculation. Many plant viruses can be transmitted mechanically, and indeed in some cases like tobacco mosaic virus (TMV), this route of transmission is very effective (Sacristán et al. 2011). However, for most viruses vector-borne transmission is more important, with insect, nematode and fungal vectors acquiring, transporting and effectively inoculating virus particles (Whitfield, Falk, and Rotenberg 2015; Rochon et al. 2004; Bian et al. 2020; Brown, Robertson, and Trudgill 1995). For the most relevant route of transmission, the cost of multipartition therefore has not been measured. The cost of vector-borne transmission may vary enormously depending on the system and the context, and will depend on many factors including vector behaviour, vector-host interactions, and the time of transit between hosts. These factors conceivably increase or decrease the cost to transmission of multipartition. For example, a reduction in the dose of virus particles inoculated by the vector would increase the cost to transmission for a multipartite virus. By contrast, ecological settings that generate many opportunities for vector-borne transmission may alleviate this cost (Valdano et al. 2019). Tests of the cost to vector-borne transmission (Figure 5) would therefore be valuable, but they will need to be performed over a variety of conditions and virus-host systems to be truly informative. Interestingly, the same experimental approach (Sánchez-Navarro, Zwart, and Elena 2013) using transgenic plants (Taschner et al. 1991) could be used to quantify the cost of multipartition for aphid transmission.

One proposed mechanism for lowering the cost to transmission of multipartite viruses is by gene product sharing during the course of infection (Sicard et al. 2019). We modelled the effect of gene product sharing and showed that it can reduce the cost of multipartition. Our model predicts this reduction will occur (1) at low MOIs, (2) when the level of sharing of gene products is moderate, and (3) when there is low sensitivity of virus yield to the genome formula. Recapitulating these three conditions more intuitively: gene product sharing is likely to be helpful (1) when genome segments are missing in many cells, (2) investment is limited because only a little sharing is done, and (3) the return on investment is large because it only takes a little sharing to achieve moderate levels of infection in other cells.

Another mechanism which could alleviate the costs of multipartition at the early phase of infection, is the aggregation of transmission particles collectively containing all genome segments. Gilmer et al. (2018) hypothesise that viral RNA segments could form supramolecular complexes, allowing for the co-transmission of genome segments within a plant and thereby reducing the cost of multipartition during within-host transmission. Although RNA sequences suggest such structures may occur, this striking hypothesis has not been tested yet. Gallet et al. (2018) considered the possibility that virus particles might aggregate when estimating the size of FBNSV transmission bottlenecks for different genome segments, but the data did not allow them to draw conclusions. It is conceivable that gene product sharing may be paired with virus particle adhesion of 1 or more segments.

Zhang et al. (2019) suggested that multipartite viruses might be found mainly in plants, because these hosts are sedentary and this will facilitate the eventual transmission of a core genome containing segments which are responsible for viral reproduction and transmission.

We disagree with this perspective because most multipartite viruses depend on highly mobile vectors for their transmission, and hence their effective contact networks are not different to those of viruses that infect other organisms (Zwart et al.2021). In their model, Zhang et al. (2019) allow genome segments to accumulate over time in susceptible hosts until, eventually, a complete genome is present, at which point the host becomes infectious. Non-replicating genome segments will be degraded in the host environment, and it is therefore expected that their potential to contribute to transmission will be rapidly lost.

Di Mattia et al. (2022) considered whether non-concomitant transmission of multipartite segments is indeed possible, exploiting the fact that FBNSV does not require all genome segments for replication and transmission. The authors found that non-concomitant transmission of segments is possible, with infections missing single segments DNA-C DNA-N or DNA-U4 were subsequently fully complemented following aphid transmission. Reconstitution of the complete genome via sequential transmission by the same aphid cohort occurred with greater success than via parallel transmission with different aphid cohorts (Di Mattia et al. 2022). Vector behaviour samples a subset of genome segments and successive or parallel feeding events increase the probability of sampling a more complete infection, thereby increasing the transmission probability of complete genomes (Figure 5). However, neither of the segments tested by Di Mattia et al. (2022) are required to initiate FBNSV infection *in planta* and reconstitution of the full genome occurs in the presence of an already replicating FBNSV population within the host. This exciting work shows that segments that are not required for replication and transmission can be re-acquired readily, but on the other hand, all available evidence suggests that the core set of segments must be transmitted concomitantly. The work does show that there are no strong barriers to the coalescence of (groups of) genome segments, an important requirement for the rescue of incomplete sets of genomes. The study, therefore, sheds light on how accessory segments can be maintained and highlights that one important requirement for the reacquisition of core segments is met.

By using a less biased approach to estimate the host range of different genome organisations, we show that there is a trend of wider host range for segmented and multipartite viruses compared to monopartite viruses than expected. These results provide an indication that segmentation, either in the form of segmented or multipartite viruses may provide an advantage for extending the host range. Previous work has shown that tri-segmented viruses possess the largest host range, however this did not take into account the genome segmentation organisation (Moury et al. 2017). A suitable example for exploring if multipartition may extend the host range are the begomoviruses, which contain both monopartite and bipartite species. We show from our analysis that the host range of begomoviruses is similar to that of the multipartite viruses group as a whole. In our current approach, we have combined both the monopartite and bipartite begomoviruses, however analysing these separately may provide insights on how multipartite viruses host ranges can expand. They are particularly relevant as it is hypothesised for Begomoviruses that the second segment originated from a satellite virus (Briddon et al. 2010), and in a mixed infection of a monopartite and satellite begomovirus, the presence of a satellite virus increased virulence and vector transmission efficiency (Ouattara et al. 2022). Moury et al. (2017) showed in their analysis that virus host range is affected by the nucleic acid type and polarity, vertical transmissibility, horizontal transmission, and vector type (Moury et al. 2017). Our approach accounts for the number of virus genome segments and the total number of observations for a given host and provides a useful way to measure the host range of plant viruses.

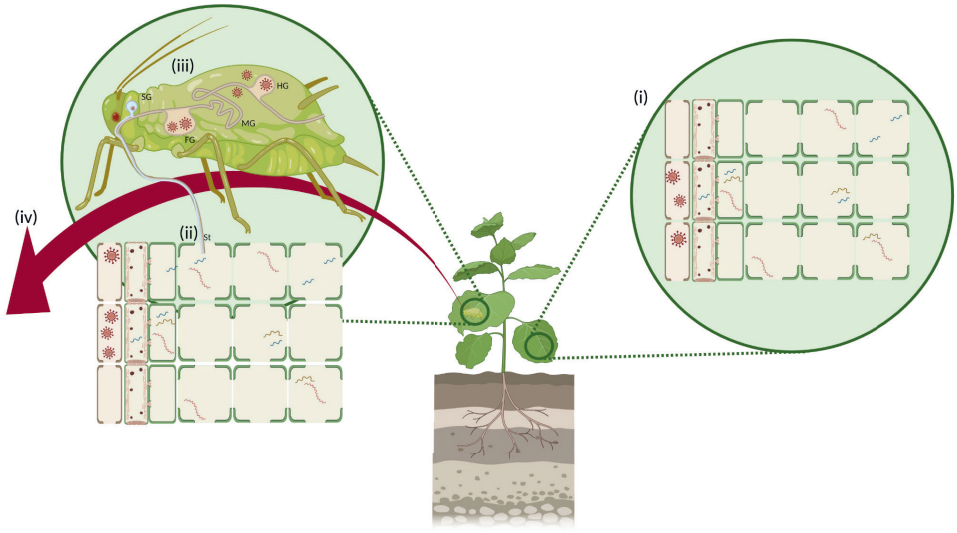


Figure 5. Overview of between and within-host costs of multipartition of plant viruses. *(i) Multipartite viral segments may be distributed over several cells, and therefore, many cells remain uninfected. (ii) Interplay between infectivity of individual segments, segment frequency, segment accumulation and total viral abundance during within-host transmission (iii) segment accumulation may differ during semi-persistent and persistent vector transmission. This suggests that there is a vector-specific GF in addition to a host-specific GF (iv) Vector transmission influences via sampling bias and vector GF the inoculum size and starting ratio of initial infections. Created with Biorender*

It can also be extended to include viruses from other kingdoms and could be integrated with data on transmission routes to identify conserved transmission pathways. Valdano et al. (2019) demonstrated in a compartmental modelling approach that multipartition can emerge when host contact networks are homogenous. A simplification to describe agricultural systems in which closely related species are connected not only spatially but also by vector behaviour. With this in mind and the approach described we can make the prediction that multipartite viruses will have host ranges that are skewed in favour of cultivated species. To test this, information from the plant virus transmission database (https://library.wur.nl/WebQuery/virus), a repository of virus transmission routes for plant viruses determined experimentally and the Virus-Host DB (https://www.genome.jp/virushostdb/note.html) a curated database for human and plant viruses infecting a small set of common cultivated crops can be combined (Mihara et al. 2016).

Our results show that segmented viruses have the broadest host range followed by multipartite viruses. Thus far, we have hypothetically attributed the extended host range to regulation of gene expression. However, enhanced recombination and reassortment of genome segment variants in segmented and multipartite viruses may also contribute to a broader range of host

species (Simon-Loriere and Holmes 2011). Reassortment and recombination in segmented and multipartite viruses may increase virus genetic diversity (Chao 1988). Both play an important role in virus host specificity and infections in novel host species (Subbarao, London, and Murphy 1993; Chen, Goldbach, and Prins 2002). For segmented influenza virus, reassortment is thought to increase virus host range (Ince et al. 2013; Imai et al. 2012). For segmented tomato spotted wilt virus, segment reassortment is associated with host resistance breaking (Qiu and Moyer 1999). Long-term analysis of natural multipartite plant virus infections has shown that reassortment and recombination are important factors in shaping virus evolution, that reassortment and recombination is low and differs between segments (Bonnet et al. 2005), genetic diversity is low (Nouri et al. 2014) and that for CMV, recombination between subgroup I and II isolate is generally low (Fraile et al. 1997).

However, low frequency RNA 3 recombinants are known to increase in frequency in different host species for types of subgroup IB and IA (Fraile et al. 1997), suggesting adaptive benefits from recombination. Ouedraogo et al. (2019) show that recombination and reassortment of a multipartite virus can be host-mediated, implying that focal cultivated plant species may act as sites for increasing genetic exchange and for emergence of new variants. This suggests that there is a complex ecological interaction of permissive and nonpermissive hosts for regulating recombination and reassortment rates in multipartite viruses on longer timescales, whereas changes in gene expression via the genome formula represents an adaptive mechanism at shorter timescale. Our focus on the genome formula hypothesis explaining the broad host range of segmented viruses was motivated by recent observations of genome formula dynamics. However, for a comprehensive understanding of genome segmentation in viruses, other adaptive consequences for adaptation and host range will need to be considered.

Supplementary S1: Empirical estimates of the between-host cost of multipartition

A multipartite genome and virus particle organization is often assumed to be costly. In classical studies, the shape of the dose local-lesion relationship was used to identify multipartite viruses. However, although multipartite viruses tend to have steep dose local-lesion relationships, this does not necessarily mean there is a cost to infection. It is concordant with infection model predictions - and the same models also predict a cost to infection – but it does not directly establish there is an infection cost. To do so, we would also need information on the position of the dose-response, ideally for viruses differing only in their number of genome segments.

To address this question of whether there really is a cost of multipartition, here we use an existing dataset (Sánchez-Navarro, Zwart, and Elena 2013). In this study, tobacco plants were challenged with a wild-type *Alfalfa mosaic virus* (AMV) strain, a tripartite positive sense RNA virus. Wild-type tobacco plants were used, but also transgenic plants constitutively expressing AMV RNA2 (*P2* plants) or AMV RNA1 and RNA2 (*P12* plants) (Taschner et al. 1991). In transgenic plants, viral replication is supported even when the plant-expressed AMV RNA is not present in the inoculum. Sánchez-Navarro et al. 2013 showed that the dose-response of AMV became more gradual in the *P2* and *P12* plants, corresponding to the predictions of simple infection models. However, one point that was not addressed in the previous study was whether the cost of infection predicted for multipartite viruses by simple infection models corresponds to model predictions, although this dataset can be used for this purpose.

From this study, we here consider only the data for infection in the inoculated leaf, and not for systemic infection. Expression of viral RNAs by the host plant appears to have a strong effect on movement, and AMV does not always cause systemic infection in wild-type plants, whereas it does for transgenic plants (Sánchez-Navarro et al. 2013). Therefore, differences in the rate of infection are probably further amplified in systemic tissues. We also used RT-qPCR estimates of the AMV setpoint genome formula (SGF), as measured for purified virions from infected tobacco plants (Wu et al. 2017), to establish the frequency of genome segments in the virus inoculum.

As a starting point for considering the cost of multipartition, we consider the previously described infection model for a multipartite virus with *j* segments:

$$I = \prod_{i=1}^{j} 1 - e^{-\rho_i n_i}$$

where I is the proportion of infected hosts, ρ is the probability of infection per virus particle, and n is the number of virus particles. The number of virus particles for each segment is known or assumed (i.e., a balanced SGF). If we then assume ρ_i is the same for all segments, the model only has a single free parameter. Conversely, if we estimate ρ_i for each segment the model will have *j* free parameters. For the transgenic plants used in our dataset (*P1*, *P12*), the expression of an AMV RNA by the plant results in $\rho = 1$. Therefore, whereas the full model for 3 virus particle types is needed to describe AMV infection of wild-type plants $I_{wt} =$

$(1-e^{-\rho_1 n_1})(1-e^{-\rho_2 n_2})(1-e^{-\rho_3 n_3})$, for *P2* plants: $I_{P2} = (1-e^{-\rho_1 n_1})(1-e^{-\rho_3 n_3})$ and for *P12* plants: $I_{P12} = (1-e^{-\rho_3 n_3})$.

Given the different slopes and positions of the responses in the different types of plants (Sánchez-Navarro, Zwart, and Elena 2013), it is uninformative to consider simpler models of infection. However, we can make different assumptions on whether ρ depends on the virus-particle type and host plant, leading to four different models with a different number of free parameters (Table S1). From first principles, we might expect that the probability of infection might be independent of the plant type, since we are modelling only the presence or absence of infection in the inoculated leaf. It is probably more difficult to make any predictions about whether the probability of infection depends on the virus-particle type, although previous work with this dataset suggests this is the case (Sánchez-Navarro, Zwart, and Elena 2013). On the one hand, there are differences in the morphology of the virus particles that depend on the genome segment encapsidated (Hull, Hills, and Markham 1969). On the other hand, all particles have the same capsid protein and similar physicochemical properties. One advantage of assuming that ρ is independent of virus-particle and plant types (i.e., Model 1), is that we only need values for the SGF to determine the cost of multipartition. In other words, the free parameter ρ can then shift the dose response in all three types of plants, but the shapes and relative positions of the responses are fixed. We consider Model 1 the null model, because of its parsimony (1 free parameter) and its property of fixing the cost of multipartition for a given SGF.

We then fitted all four models to the data using a stochastic hill-climbing algorithm to minimise the negative log likelihood (NLL). NLL was determined by assuming a binomial error structure such that for the kth plant type and lth dose:

$$L(a,b) = (a\ b\ ){I_{k,l}}^{b}\left(1-I_{k,l}\right)^{a-b}$$

The corresponding NLL was then summed over all doses and plant types (i.e., each model was fit to all the experimental data). The model fitting was done using a custom script in *R* 4.2.1 available at Zenodo (10.5281/zenodo.10652647). We used the Akaike Information Criterion (AIC) for model selection.

The experimental data and fitted models are shown in Figure S2, model parameter estimates are given in Table S1, and the model selection results are given in Table S2. There was little support for Model 1, our null model for the cost of multipartition (Table S2). For this model, the relative positions of the dose response curves for the different plant types are fixed. The difference between the dose response for the different plant types is larger for the experimental data than that predicted by the model (Figure S2). The results therefore suggest that the real cost of multipartition is even larger than that predicted by the null model. All the alternative models have similar fits, and hence the models with 3 free parameters (Models 2 and 3) enjoy greater support than the 6 free parameter model (Model 4). Model selection therefore cannot identify whether differences in the probability of infection between virus-particles types, or differences in the probability of infection between plant types better account for the experimental data.

The results suggest that the cost of multipartition exists, and that it might be greater than predictions by simple infection models. On the other hand, for the set of models we used here, we cannot identify a possible underlying mechanism explaining this greater cost, as the

support for different, more complex models is similar. The mechanism incorporated by Model 2 – differences in probability of infection for different virus-particle types – could be a general mechanism leading to such discrepancies in real world virus populations. On the other hand, plant-type-dependent differences in the probability of infection are likely to be artefacts of the expression of viral segments. These differences would then not reflect a true cost of multipartition, but only be artefacts of the experimental system. Finally, it is worth noting that these conclusions were not reached in the original study (Sánchez-Navarro et al. 2013) because of the focus on testing IAH-derived infection models, which focus on the steepness of a response rather than its relative position.

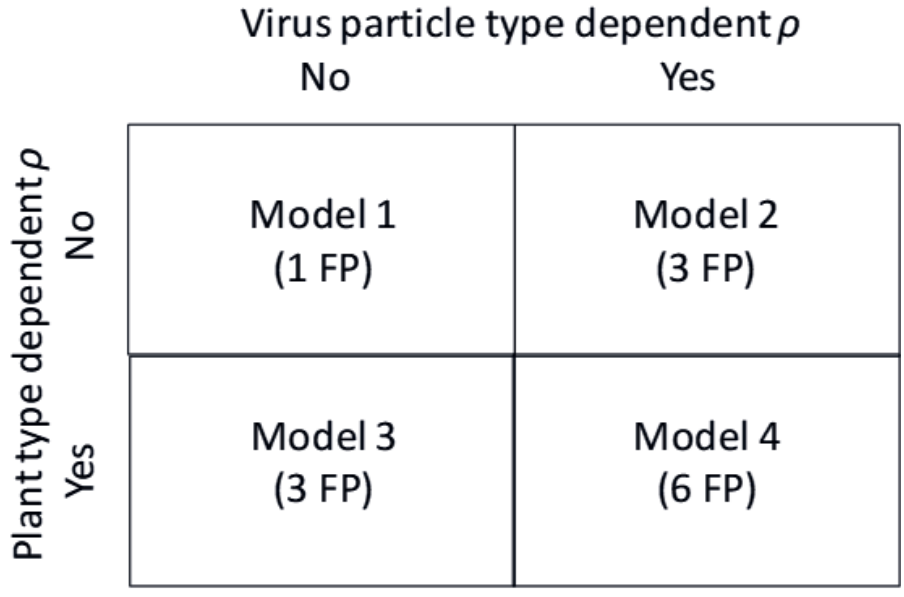


***Figure S1**. Overview of the different models fitted here. FP stands for free parameter, and ρ is the probability of infection per virus particle. Note the Model 4 only has 6 parameters (instead of 9) because when the transgenic plants express an AMV RNA, ρ is set to 1. Hence, for the P2 and P12 plants there are 2 and 1 free parameters, respectively.*

Table S1. Estimated model parameters

Model	Parameter estimates
1	$\rho = 7.82 \times 10^{-3}$
2	$\rho_1 = 6.85 \times 10^{-3}$, $\rho_2 = 1.42 \times 10^{-3}$, $\rho_3 = 2.32 \times 10^{-2}$
3	$\rho_{wt} = 4.00 \times 10^{-3}$, $\rho_{P2} = 8.81 \times 10^{-3}$, $\rho_{P12} = 2.47 \times 10^{-2}$
4	$\rho_{1,wt} = 1.15 \times 10^{-2}$, $\rho_{2,wt} = 4.75 \times 10^{-3}$, $\rho_{3,wt} = 2.00 \times 10^{-3}$, $\rho_{1,P2} = 1.09 \times 10^{-2}$, $\rho_{3,P2} = 6.71 \times 10^{-3}$, $\rho_{3,P12} = 2.47 \times 10^{-2}$

Table S2. Model Selection

Model	NLL	AIC	ΔAIC	AW
1	28.958	59.916	28.068	0.000
2	13.579	33.158	1.31	0.328
3	12.924	31.848	-	0.632
4	12.681	37.362	5.514	0.040

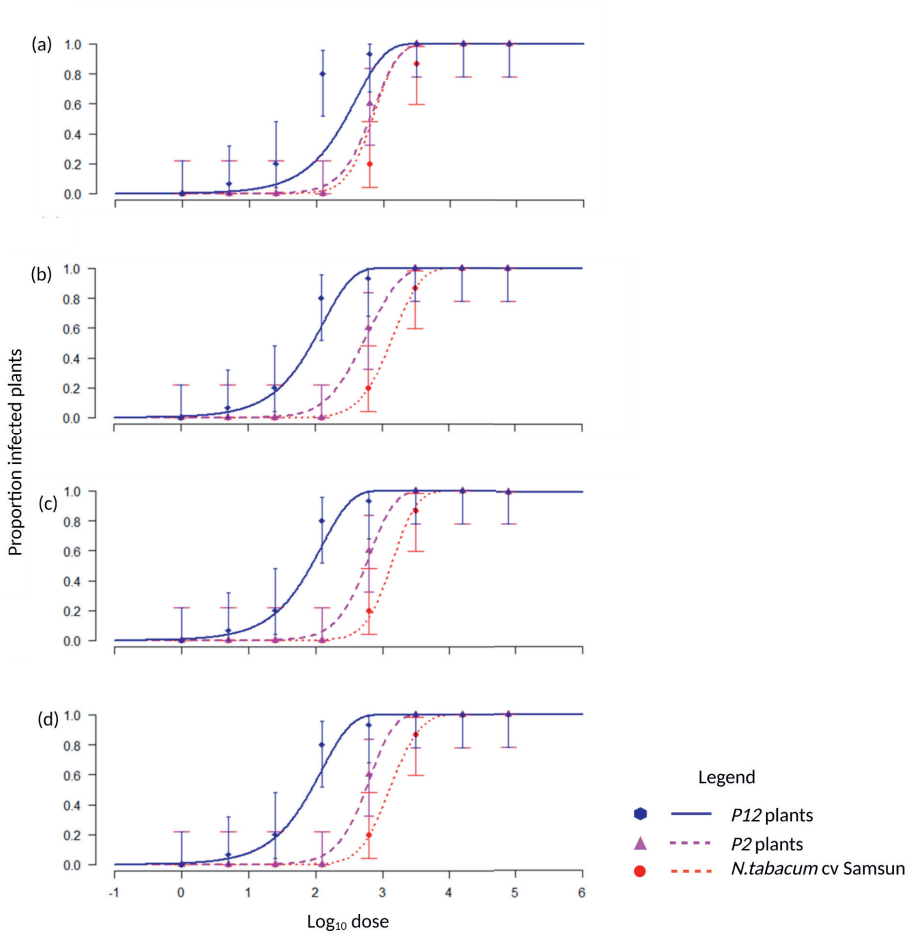


Figure S2. *Fitted dose response models (lines) and experimental data (symbols). The ordinate is the log transformed virus particle dose, and the abscissae is the proportion of infected plants. Error bars indicate the binomial 95% confidence interval. Panels a-d correspond to Models 1–4, respectively. The null model, Model 1, underestimates the differences between the responses for the different plants, and therefore underestimates the cost of multipartition. The fit for Models 2–4 is very similar, although Model 4 is over-parameterized and has little support.*

Supplementary S2: Gene product sharing is a mechanism to lower the within-host cost

To explore the effects of gene product sharing on virus fitness, and explore the relationship between the genome formula and gene product sharing, we adapt a simulation model of multipartite virus genome-formula evolution from Zwart and Elena (Zwart and Elena 2020). This model assumes a bipartite virus with genome segments 1 and 2. Each genome segment produces a unique gene product that is needed for a successful cellular infection. If both gene products are present in a cell, the virus genome segments (1, 2 or both) in that cell will be replicated, incorporated into virus particles and can infect new cells in subsequent rounds of infection. There is a fixed number of cells (c) in each generation, there are discreet rounds of cellular replication with this fixed number of cells, and there is no spatial structure or differences in susceptibility between cells. To this model, we add the possibility of a donor cell in passage t sharing a proportion of the gene products it generates across all cells in passage $t + 1$. A key parameter in this model is σ^2, a parameter which determines how changes in the ratio of the two viral gene products affect the virus's ability to exploit a cell, either by producing virus particles or assembling gene products that can be shared with other cells. Finally, this model is not concerned with direct competition between virus variants, and hence all model equations are simplified to describe a bipartite virus in isolation. We have considered the fitness of a single virus for simplicity, and because gene-product sharing is very vulnerable to exploitation by non-sharing strains, potentially leading to complex co-infection dynamics.. Below we provide a detailed description of the model.

At the start of each round of infection, cells are exposed to virus particles and gene products produced in the previous round of infection. In the original study, we fixed the cellular multiplicity of infection (λ), the total number of virus particles entering each cell. Here, we set a value for λ, but this number represents the maximum mean number of virus particles that infects a cell, and the realised mean can decrease due to a number of factors, as explained later on. There is a Poisson-distributed realization of the number of virus particles entering cells for each type (λ_1 and λ_2, i.e., virus particles containing genome segments 1 and 2, where $\lambda_1 + \lambda_2 = \lambda$).

Gene products are shared with units (β_1 and β_2) that are equivalent to the potential gene expression from an invading virus particle. However, we assume that these gene products are homogeneously distributed over all cells in each round of replication, such that the amount of shared gene product in each cell is e.g., $h_1 = \frac{\beta_1}{c}$. Therefore, virus gene product sharing is not subject to stochastic variation, as seen for the spread of virus particles.

The frequency of each genome segment in each cell then proceeds as before, e.g. $f_1 = \frac{k_1}{(k_1+k_2)}$, where f is the within-cell frequency of a genome segment. We assume there is no within-cell competition between genome segments, and therefore f_1 determines the relative frequency of the virus particle types produced by a cell. The ratio r of the virus gene products g for the two segments within a cell will determine the total virus resources of that cell: $r = \frac{g_1}{g_2} = \frac{(k_1+h_1)}{(k_2+h_2)}$. Under this model, a cell is infected and will produce virus resources if three conditions are

met: $k_1 + k_2 > 0$, $g_1 > 0$ and $g_2 > 0$. I.e., At least one genome segment must be present, whereas both gene products must be present, regardless of their origin (virus particles or gene-product sharing). The probability density function of the normal distribution is used to link viral proteins to virus resources generated, such that $\varphi(r) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\; exp\; \left[\frac{-(log_{10}r-\mu)^2}{2\sigma^2}\right]$. Here μ is the mean of the distribution, which is the value of r that results in the highest virus resources being generated, and σ^2 is its variance, the parameter which determines how quickly virus resources drop as r moves away from the optimum μ.

In a previous model, the function $\varphi(r)$ determine virus particle yield, but here we equate this with virus resources, because these resources generated can be dedicated to producing virus particles, producing viral gene products that are shared with cells in the next round of infection, or both. The proportion of these resources committed to sharing is ρ, with a range between 0 (only virus particles produced, no resources committed to gene-product sharing) to 1 (no virus particles produced, all resources committed to sharing). Moreover, the amount of virus particles produced is scaled to the maximum possible yield that would result in λ particles infecting cells in the next round of infection (i.e., $\varphi(1)$), such that the contribution of each cell to virus particle production (*vp*) for segment 1 is $vp_1 = \frac{(1-\rho)f_1\lambda*\varphi(r)}{\varphi(1)}$, the contribution of each cell to gene product (*gp*) sharing for segment 1 is $gp_1 = \frac{(\rho)f_1\lambda*\varphi(r)}{\varphi(1)}$. Unlike the previous model where MOI was fixed over rounds of passaging, suboptimal virus particle production over all cells will there lead to a reduction in MOI and possibly extinction of the virus. Note that the relative levels of production of the two gene products (gp_1 and gp_2) depends on the frequency of the two virus genome segments that entered the cell, analogous to virus particle production. We mean over all cells in this round of infection, where for cells that are uninfected per definition $vp_1 = vp_2 = 0$ and $gp_1 = gp_2 = 0$.

We performed 10^4 simulations per condition, as the individual simulations can give highly diverging results (e.g., due to the small number of cells c and low MOI λ). To measure the fitness of viruses in isolation, we calculated the aggregated virus accumulation over all cells in the simulation (c^*t_{final}), normalised by aggregated accumulation for a virus that does not share its gene products ($\rho = 0$), all other conditions being equal and using the mean values over all simulations. To explore model behaviour in a range of conditions, we varied model parameters λ, σ^2, c and ψ, which determines the decimal log transformed range in which μ can be sampled from a uniform distribution. I.e., when $\psi = 0$, μ is 1 and even gene expression always results in optimal exploitation of cells. When $\psi = 2$, μ is varied randomly resulting in values between 0.01 and 100. An overview of model parameters is given in Table S3.

Model code is available at Zenodo (10.5281/zenodo.10652647)

Table S3. For the model of the effects of gene product sharing on multipartite virus infection, for each model parameter we give the range of values used in the simulations, a brief explanation of the parameter, and miscellaneous comments for clarification.

Parameter	Value	Explanation and comments
λ	10^{0, 0.1, … 2}	Cellular multiplicity of infection; maximum value that can decrease under suboptimal conditions.
c	5, 100	Number of cells per round of infection; note that unlike the previous model, c is not adjust to maintain the same number of effectively infected cells per round of infection.
σ^2	0.01, 0.1, 1, 10	Variance of the normal probability function used to link virus gene products to the total production of viral resources
t_{final}	20	Number of rounds of infection
ρ	{0, 0.05, 0.1 … 1}	Proportion of virus resources dedicated to gene product sharing, the remaining fraction is used for virus particles.
ψ	0, 2	Determines the range of values from which values of μ can be drawn from a uniform distribution, from $0 - \psi$ to $0 + \psi$. When $\psi = 0$, μ is always 0.

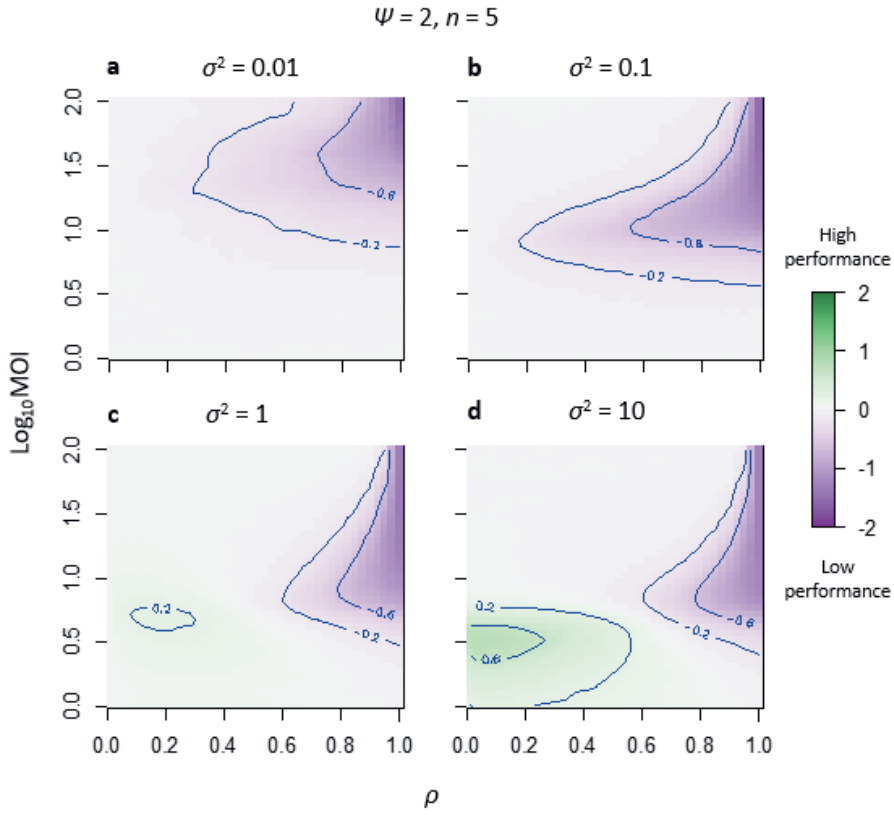


Figure S3. The effects of viral-gene-product sharing on performance. *On the x-axis ρ is given, the proportion of viral gene products that is exported to other cells. On the y-axis the log_{10} of the cellular multiplicity of infection (MOI) is given. The heat indicates viral performance, as compared to a virus that does not share any gene products. The performance metric $\theta_i=log_{10}(\alpha_{(\rho=i)}/\alpha_{(\rho=0)})$, where α is the sum of virus particles produced over all cells in the simulation. When viral performance is better than the non-sharing reference ($\theta_i>0$) the area is green, when viral performance is worse ($\theta_i<0$) the area is purple, and when viral performance is similar ($\theta_i\sim0$) the area is white. Contour lines have been included at the θ_i levels -0.6, -0.2, 0.2 and 0.6 to highlight parameter space with high or low performance. The value σ^2 indicates how sensitive virus particle production is to changes in the genome formula, with low values (i.e., $\sigma^2 = 0.01$) indicated high sensitivity and a quick drop in accumulation as the population moves away from the optimal value, and high values (i.e., $\sigma^2 = 10$) indicating low sensitivity. 10,000 independent simulations were run.*

In this case, in each generation the virus is passaged in a small number of cells (n = 5) and under high variation in the optimal GF between passages (ψ = 2). When virus accumulation is not sensitive to changes in the GF ($\sigma^2 \geq 1$), MOI is at intermediate values ($3\sim\log_{10}0.5$), and gene-product sharing is low to moderate (≤ 0.6), gene product sharing is predicted to offer an advantage.

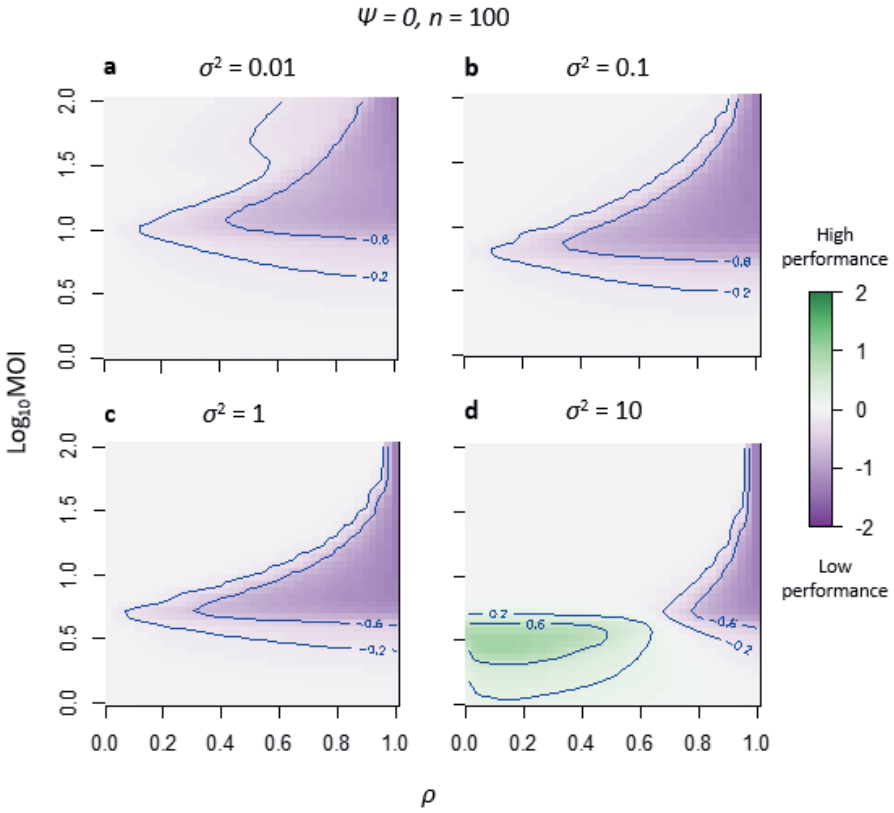


Figure S4. The effects of viral-gene-product sharing on performance*. On the x-axis ρ is given, the proportion of viral gene products that is exported to other cells. On the y-axis the log₁₀ of the cellular multiplicity of infection (MOI) is given. The heat indicates viral performance, as compared to a virus that does not share any gene products. The performance metric $\theta_i = log_{10}(\alpha_{(\rho=i)}/\alpha_{(\rho=0)})$, where α is the sum of virus particles produced over all cells in the simulation. When viral performance is better than the non-sharing reference ($\theta_i>0$) the area is green, when viral performance is worse ($\theta_i<0$) the area is purple, and when viral performance is similar ($\theta_i \sim 0$) the area is white. Contour lines have been included at the θ_i levels -0.6, -0.2, 0.2 and 0.6 to highlight parameter space with high or low performance. The value σ^2 indicates how sensitive virus particle production is to changes in the genome formula, with low values (i.e., $\sigma^2 = 0.01$) indicated high sensitivity and a quick drop in accumulation as the population moves away from the optimal value, and high values (i.e., $\sigma^2 = 10$) indicating low sensitivity. 1,000 independent simulations were run.*

In this case, in each generation the virus is passaged in a large number of cells (n = 100) and under no variation in the optimal GF between passages (ψ = 0). When virus accumulation is not sensitive to changes in the GF (σ^2 = 10), MOI is at intermediate values (3~$\log_{10}$0.5), and gene-product sharing is low to moderate (≤ 0.6), gene product sharing is predicted to offer an advantage, in a larger parameter space than when passaging in a smaller number of cells.

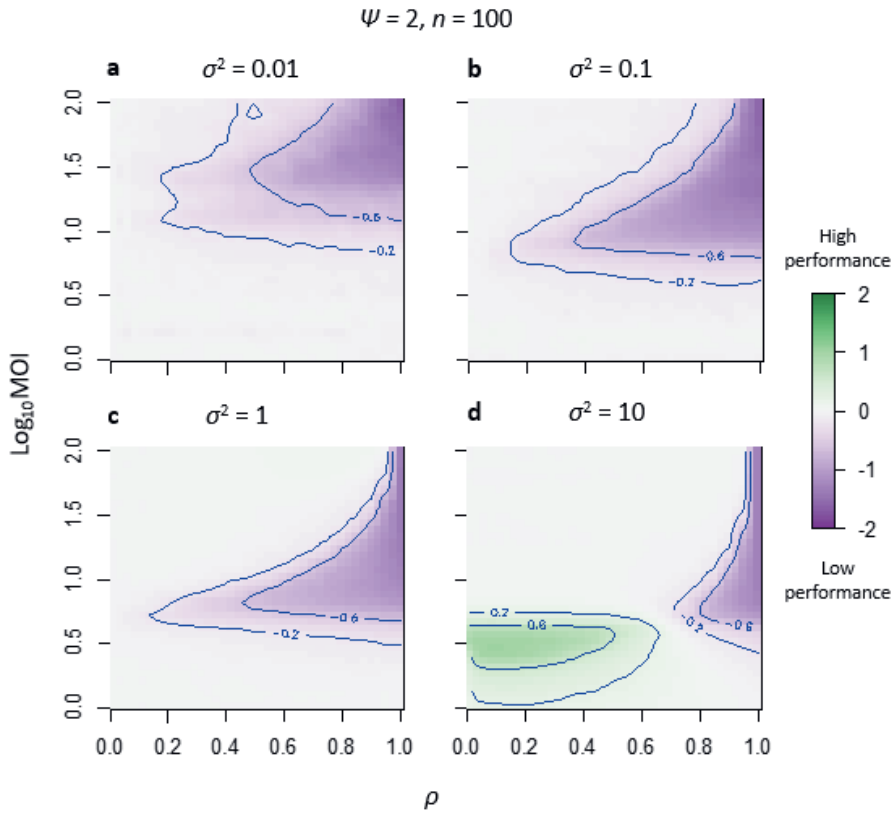


Figure S5: The effects of viral-gene-product sharing on performance. *On the x-axis ρ is given, the proportion of viral gene products that is exported to other cells. On the y-axis the log_{10} of the cellular multiplicity of infection (MOI) is given. The heat indicates viral performance, as compared to a virus that does not share any gene products. The performance metric $\theta_i=log_{10}(\alpha_{(\rho=i)}/\alpha_{(\rho=0)})$, where α is the sum of virus particles produced over all cells in the simulation. When viral performance is better than the non-sharing reference ($\theta_i>0$) the area is green, when viral performance is worse ($\theta_i<0$) the area is purple, and when viral performance is similar ($\theta_i\sim0$) the area is white. Contour lines have been included at the θ_i levels -0.6, -0.2, 0.2 and 0.6 to highlight parameter space with high or low performance. The value σ^2 indicates how sensitive virus particle production is to changes in the genome formula, with low values (i.e., $\sigma^2 = 0.01$) indicated high sensitivity and a quick drop in accumulation as the population moves away from the optimal value, and high values (i.e., $\sigma^2 = 10$) indicating low sensitivity. 1,000 independent simulations were run.*

In this case, in each generation the virus is passaged in a large number of cells (n = 100) and under high variation in the optimal GF between passages (ψ = 2). When virus accumulation is not sensitive to changes in the GF ($\sigma^2 = 10$), MOI is at intermediate values ($3\sim\log_{10}0.5$), and gene-product sharing is low to moderate (≤ 0.6), gene product sharing is predicted to offer an advantage, in a larger parameter space than when passaging in a smaller number of cells.

Supplementary S3: Global virus host range analysis

To test whether there are systematic differences in host range depending on genome organisation, we performed a metaanalysis to determine host ranges across plant viruses with monopartite, segmented, and multipartite genomes. We investigated the viral host range using virus sequence metadata from the NCBI Virus portal database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/, Accessed 17th April 2023). The NCBI virus portal database provides an overview of the metadata associated with high throughput sequencing derived viral (meta)genomes. The NCBI Virus portal (Hatcher et al. 2017) was downloaded as a dataset consisting of metadata of whole and partial virus genome segments (n = 11,031,767). The metadata included host species, viral genus, viral species and viral family. We used the ICTV Virus Metadata Resource (VMR; https://ictv.global/vmr) (Version MSL37 released 2nd December 2022) to add information on the Baltimore classification of virus species (Baltimore 1971) and information on which kingdom is infected by each virus species. These two datasets were merged in R using *tidyverse* package *dplyr* (R Core Team, n.d.; Wickham et al. 2019). We then selected only observations for which the viral genus was considered plant-infecting based on the ICTV VMR 'host source' associated with specific viral genera (i.e. host source == 'Plants'). We used only plant viruses for the host range analyses to ensure a fair comparison, given that most multipartite viruses infect plants. For the plant-infecting subset of virus-host observations, we assigned genome organisation to each observation by manual curation. The number of unique host genera detected per virus species were then tallied to give an estimate of the host range. The total number of observations per virus species were tallied, and corrected for the number of segments. These numbers were then used in a model to test for differences in host range between the three different modes of genome organisation: monopartite, multipartite and segmented.

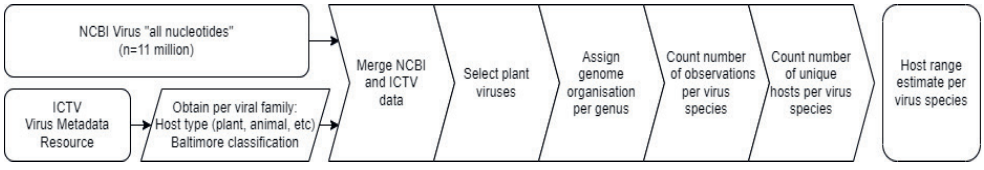


Figure S6. Schematic overview of virus host range analysis.

We developed a simple model to predict the genus-level host range (henceforth "host range") conditional upon the number of observations of that virus. The observed host range is the number of unique host genera observed per virus species. Our model makes two main assumptions. First, the effective host range within a group (i.e., monopartite, multipartite or segmented viruses) follows a zero-truncated geometric distribution over viruses, with a single free-parameter θ and a mean host range $1/\theta$. Effective host range is the number of hosts which the virus can infect under the real-world conditions for which the data were collected. We choose a geometric distribution based on the intuition that most viruses will infect one or few host genera, but a small fraction of viruses will infect a large number of hosts. Second, we assume that the virus is evenly distributed over all host species, to keep the model as simple as possible.

We can then generate a prediction for the observed host range conditional upon the number of observations of a virus based on an iterative approach. For each unique value *u* of the total number of observations of a virus, we first draw a value *v* of the host range from a geometric distribution with the rgeom function. Next we randomly sample *u* integers with replacement from the set {1, 2, … *v*} with the sample function, and then determine the number of unique values *w*, which is the realised host range for this iteration. For each value of *u* we perform *z* iterations, and use the Laplace's law of succession to estimate the pseudo-likelihood of an observed host range of *x* genera:

$$L(x|u) = -\frac{1+w}{2+z}.$$

Laplace's law of succession is used to avoid likelihood values of 0 during the model fitting procedure, while this means that the pseudo-likelihood values depend on *z*. To estimate *θ* for a set of observations, we used a grid search to find the value of *θ* that minimises the cumulative negative log likelihood (NLL). We bootstrapped virus species to estimate the 95% confidence intervals of *θ*. We first generated expectations of $L(x|u)$ for each value of *x, u* and *θ* in the dataset, and then used these expectations for model parameter estimation.

An overview of the model fitting settings and results is given in Table S4. Although we varied the range for free parameter *θ*, the step size was always set to 0.01. The number of permutations used to generate the model predictions of realised host range was varied over runs, but the number of bootstraps to determine the confidence interval was always set to 1000. After an initial run over the full range of *θ* with a low number of permutations, we restricted the range of *θ* but at the same time increased the number of permutations for more precise model parameter estimates. As we are working with pseudo-likelihoods and adjust predictions by Laplace's law of succession, the number of permutations affects both the model parameter estimation and the lowest NLL. Therefore, although we are optimising the model and obtaining more precise estimates of *θ*, in practice the NLL increases as the range of *θ* is restricted as more permutations are used in successive runs. One reason these effects occur can be that the very small or zero predicted likelihoods will affect the model parameter estimation more strongly, as with a larger number of iterations we gain more confidence that these values are indeed very small using this approach.

Table S4: Model fitting settings and parameter θ estimates

Group	Species[a]	Permutations	Range θ[b]	Estimated θ	95% CI[c]	NLL[d]
Monopartite	869	10^3	0.01 - 0.99	0.42	0.34 - 0.45	1287.012
		10^4	0.24 - 0.55	0.34	0.34 - 0.41	1350.753
		10^5	0.28 - 0.50	0.33	0.31 - 0.38	1387.726
Segmented	196	10^3	0.01 - 0.99	0.19	0.13 - 0.43	175.981
		10^4	0.03 - 0.53	0.18	0.11 - 0.29	196.808
		10^5	0.05 - 0.40	0.15	0.08 - 0.28	216.413
Multipartite	67	10^3	0.01 - 0.99	0.32	0.22 - 0.35	400.789
		10^4	0.12 - 0.45	0.25	0.20 - 0.33	446.624
		10^5	0.12 - 0.40	0.24	0.16 - 0.30	484.343
Mixed[e]	413	10^3	0.01 - 0.99	0.27	0.24 - 0.35	673.485
		10^4	0.14 - 0.45	0.24	0.19 - 0.29	718.048
		10^5	0.14 - 0.40	0.24	0.20 - 0.26	748.026

[a] The number of virus species included in the analysis. [b] The free model parameter to be estimated. [c] CI = Confidence interval [d] Negative log likelihood [e] Virus genera which included both monopartite and multipartite species are included here, while all species included due belong to the genus *Begomovirus*
[e] This group includes the entire dataset, including viroids and satellites.

References

Andersson, Dan I., and Diarmaid Hughes. 2009. “Gene Amplification and Adaptive Evolution in Bacteria.” *Annual Review of Genetics* 43: 167–95.

Andreu-Moreno, Iván, and Rafael Sanjuán. 2018. “Collective Infection of Cells by Viral Aggregates Promotes Early Viral Proliferation and Reveals a Cellular-Level Allee Effect.” *Current Biology: CB* 28 (20): 3212–19.e4.

Bald, J. G. 1937. “The Use of Numbers of Infections For Comparing the Concentrations of Plant Virus Suspensions: Dilution Experiments with Purified Suspensions.” *The Annals of Applied Biology* 24 (1): 33–55.

Bian, Ruiling, Ida Bagus Andika, Tianxing Pang, Ziqian Lian, Shuang Wei, Erbo Niu, Yunfeng Wu, Hideki Kondo, Xili Liu, and Liying Sun. 2020. “Facilitative and Synergistic Interactions between Fungal and Plant Viruses.” *Proceedings of the National Academy of Sciences of the United States of America*, January. https://doi.org/10.1073/pnas.1915996117.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064.

Bonnet, Julien, Aurora Fraile, Soledad Sacristán, José M. Malpica, and Fernando García-Arenal. 2005. “Role of Recombination in the Evolution of Natural Populations of Cucumber Mosaic Virus, a Tripartite RNA Plant Virus.” *Virology* 332 (1): 359–68.

Briddon, Rob W., Basavaprabhu L. Patil, Basavaraj Bagewadi, Muhammad Shah Nawaz-Ul-Rehman, and Claude M. Fauquet. 2010. “Distinct Evolutionary Histories of the DNA-A and DNA-B Components of Bipartite Begomoviruses.” *BMC Evolutionary Biology* 10: 97.

Brooke, Christopher B. 2014. “Biological Activities of ‘Noninfectious’ Influenza A Virus Particles.” *Future Virology* 9 (1): 41–51.

Brooke, Christopher B., William L. Ince, Jiajie Wei, Jack R. Bennink, and Jonathan W. Yewdell. 2014. “Influenza A Virus Nucleoprotein Selectively Decreases Neuraminidase Gene-Segment Packaging While Enhancing Viral Fitness and Transmissibility.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (47): 16854–59.

Brown, D. J., W. M. Robertson, and D. L. Trudgill. 1995. “Transmission of Viruses by Plant Nematodes.” *Annual Review of Phytopathology* 33: 223–49.

Campbell, Grant L., Anthony A. Marfin, Robert S. Lanciotti, and Duane J. Gubler. 2002. “West Nile Virus.” *The Lancet Infectious Diseases* 2 (9): 519–29.

Chao, Lin. 1988. “Evolution of Sex in RNA Viruses.” *Journal of Theoretical Biology* 133: 99–112.

Chen, Yuh-Kun, Rob Goldbach, and Marcel Prins. 2002. “Inter- and Intramolecular Recombinations in the Cucumber Mosaic Virus Genome Related to Adaptation to Alstroemeria.” *Journal of Virology* 76 (8): 4119–24.

Chou, Yi-Ying, Reza Vafabakhsh, Sultan Doğanay, Qinshan Gao, Taekjip Ha, and Peter Palese. 2012. “One Influenza Virus Particle Packages Eight Unique Viral RNAs as Shown by FISH Analysis.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (23): 9101–6.

Diefenbacher, Meghan, Jiayi Sun, and Christopher B. Brooke. 2018. “The Parts Are Greater than the Whole: The Role of Semi-Infectious Particles in Influenza A Virus Biology.” *Current Opinion in Virology* 33 (December): 42–46.

Di Mattia, Jérémy, Babil Torralba, Michel Yvon, Jean-Louis Zeddam, Stéphane Blanc, and Yannis Michalakis. 2022. “Nonconcomitant Host-to-Host Transmission of Multipartite Virus Genome Segments May Lead to Complete Genome Reconstitution.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (32): e2201453119.

Druett, H. A. 1952. “Bacterial Invasion.” *Nature* 170 (4320): 288–288.

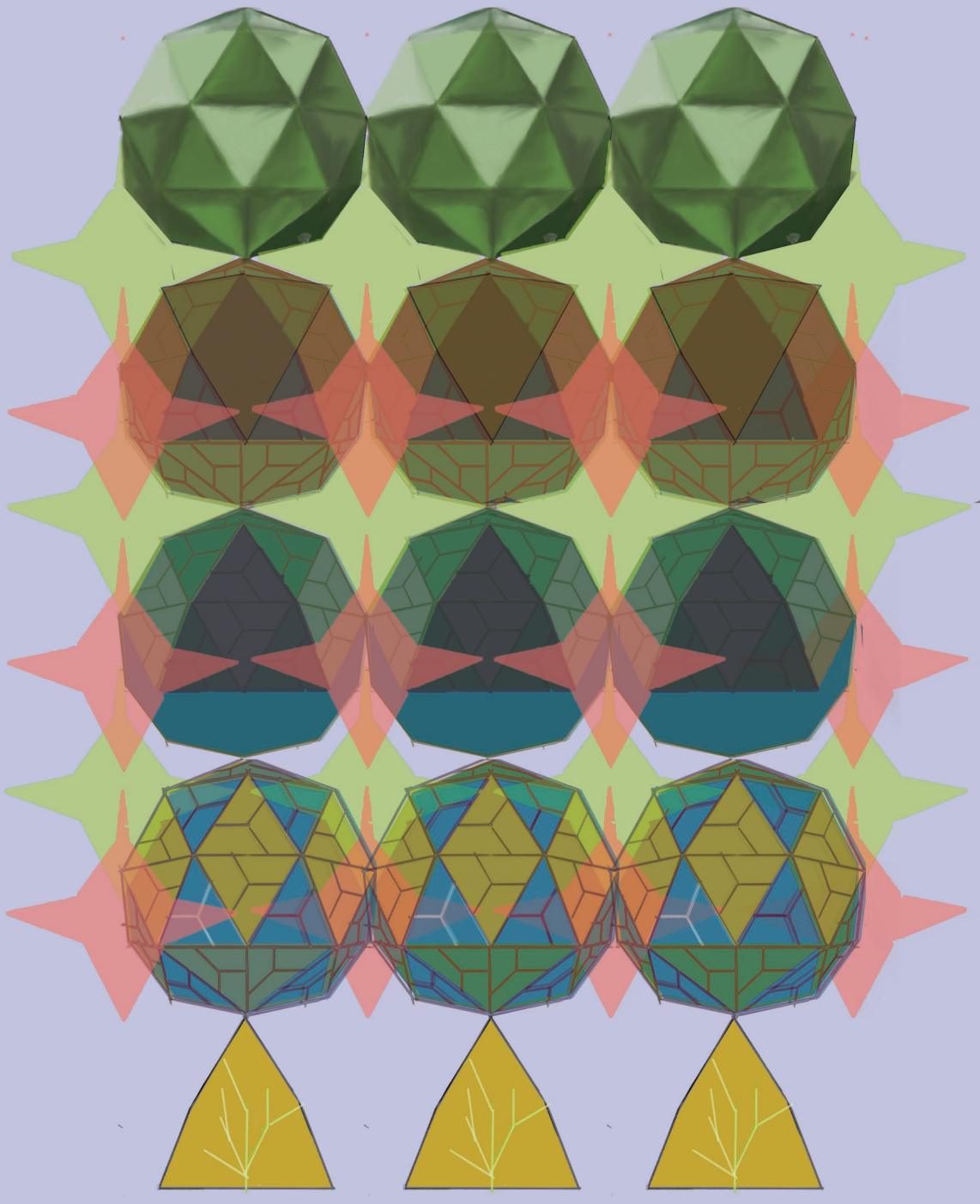
Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers,

Jay Shendure, Adam P. Geballe, and Harmit S. Malik. 2012. “Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses.” *Cell* 150 (4): 831–41.
Fraile, Aurora, José Luis Alonso-prados, Miguel A. Aranda, Juan J. Bernal, José M. Malpica, and Fernando Garci. 1997. “Genetic Exchange by Recombination or Reassortment Is Infrequent in Natural Populations of a Tripartite RNA Plant Virus.” *Journal of Virology* 71 (2): 934–40.
Fulton, Robert W. 1962. “The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus.” *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.
Gallet, R., Y. Michalakis, and S. Blanc. 2018. “Vector-Transmission of Plant Viruses and Constraints Imposed by Virus–vector Interactions.” *Current Opinion in Virology*. https://www.sciencedirect.com/science/article/pii/S1879625718300531.
Gallet, Romain, Jérémy Di Mattia, Sébastien Ravel, Jean-Louis Zeddam, Renaud Vitalis, Yannis Michalakis, and Stéphane Blanc. 2022. “Gene Copy Number Variations at the within-Host Population Level Modulate Gene Expression in a Multipartite Virus.” *Virus Evolution* 8 (2): veac058.
Gallet, Romain, Frédéric Fabre, Gaël Thébaud, Mircea T. Sofonea, Anne Sicard, Stéphane Blanc, and Yannis Michalakis. 2018. “Small Bottleneck Size in a Highly Multipartite Virus during a Complete Infection Cycle.” *Journal of Virology* 92 (14). https://doi.org/10.1128/JVI.00139-18.
Gilmer, David, Claudio Ratti, and Fabrice Michel. 2018. “Long-Distance Movement of Helical Multipartite Phytoviruses: Keep Connected or Die?” *Current Opinion in Virology* 33 (December): 120–28.
Gutiérrez, Serafín, and Mark P. Zwart. 2018. “Population Bottlenecks in Multicomponent Viruses: First Forays into the Uncharted Territory of Genome-Formula Drift.” *Current Opinion in Virology* 33 (December): 184–90.
Hajimorad, M. R., G. Kurath, J. W. Randles, and R. I. Francki. 1991. “Change in Phenotype and Encapsidated RNA Segments of an Isolate of Alfalfa Mosaic Virus: An Influence of Host Passage.” *The Journal of General Virology* 72 (Pt 12) (December): 2885–93.
Hatcher, Eneida L., Sergey A. Zhdanov, Yiming Bao, Olga Blinkova, Eric P. Nawrocki, Yuri Ostapchuck, Alejandro A. Schäffer, and J. Rodney Brister. 2017. “Virus Variation Resource - Improved Response to Emergent Viral Outbreaks.” *Nucleic Acids Research* 45 (D1): D482–90.
Hull, R., G. J. Hills, and R. Markham. 1969. “Studies on Alfalfa Mosaic Virus. II. The Structure of the Virus Components.” *Virology* 37 (3): 416–28.
Hutchinson, G. E. 1957. “Concluding Remarks. In: Cold Spring Harb Symp Quant Biol.”
Hu, Zhaoyang, Xiaolong Zhang, Wei Liu, Qian Zhou, Qing Zhang, Guohui Li, and Qin Yao. 2016. “Genome Segments Accumulate with Different Frequencies in Bombyx Mori Bidensovirus.” *Journal of Basic Microbiology* 56 (12): 1338–43.
Imai, Masaki, Tokiko Watanabe, Masato Hatta, Subash C. Das, Makoto Ozawa, Kyoko Shinya, Gongxun Zhong, et al. 2012. “Experimental Adaptation of an Influenza H5 HA Confers Respiratory Droplet Transmission to a Reassortant H5 HA/H1N1 Virus in Ferrets.” *Nature* 486 (7403): 420–28.
Ince, William L., Aissatou Gueye-Mbaye, Jack R. Bennink, and Jonathan W. Yewdell. 2013. “Reassortment Complements Spontaneous Mutation in Influenza A Virus NP and M1 Genes to Accelerate Adaptation to a New Host.” *Journal of Virology* 87 (8): 4330–38.
Johnson, Jerald B., and Kristian S. Omland. 2004. “Model Selection in Ecology and Evolution.” *Trends in Ecology & Evolution* 19 (2): 101–8.
Jonge, Patrick A. de, Franklin L. Nobrega, Stan J. J. Brouns, and Bas E. Dutilh. 2019. “Molecular and Evolutionary Determinants of Bacteriophage Host Range.” *Trends in Microbiology* 27 (1): 51–63.
Ladner, Jason T., Michael R. Wiley, Brett Beitzel, Albert J. Auguste, Alan P. Dupuis, Michael E. Lindquist, Samuel D. Sibley, et al. 2016. “A Multicomponent Animal Virus Isolated from Mosquitoes.” *Cell Host & Microbe* 20 (3): 357–67.

Lauffer, M. A., and W. C. Price. 1945. “Infection by Viruses.” *Archives of Biochemistry* 8 (December): 449–68.
Lucía-Sanz, Adriana, and Susanna Manrubia. 2017. “Multipartite Viruses: Adaptive Trick or Evolutionary Treat?” *Npj Systems Biology and Applications* 3 (1): 34.
Marm Kilpatrick, A., Peter Daszak, Matthew J. Jones, Peter P. Marra, and Laura D. Kramer. 2006. “Host Heterogeneity Dominates West Nile Virus Transmission.” *Proceedings of the Royal Society B: Biological Sciences* 273 (1599): 2327–33.
McKinney, H. H. 1927. “Quantitative and Purification Methods in Virus Studies.” *Journal of Agricultural Research* 35 (1): 13–38.
Michalakis, Yannis, and Stéphane Blanc. 2020. “The Curious Strategy of Multipartite Viruses.” *Annual Review of Virology* 7 (1): 203–18.
Mihara, Tomoko, Yosuke Nishimura, Yugo Shimizu, Hiroki Nishiyama, Genki Yoshikawa, Hideya Uehara, Pascal Hingamp, Susumu Goto, and Hiroyuki Ogata. 2016. “Linking Virus Genomes with Host Taxonomy.” *Viruses* 8 (3): 66.
Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. “The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner.” *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.
Morris, Cindy E., and Benoît Moury. 2019. “Revisiting the Concept of Host Range of Plant Pathogens.” *Annual Review of Phytopathology* 57 (August): 63–90.
Moury, Benoît, Frédéric Fabre, Eugénie Hébrard, and Rémy Froissart. 2017. “Determinants of Host Species Range in Plant Viruses.” *The Journal of General Virology* 98 (4): 862–73.
Nakatsu, Sumiho, Hiroshi Sagara, Yuko Sakai-Tagawa, Norio Sugaya, Takeshi Noda, and Yoshihiro Kawaoka. 2016. “Complete and Incomplete Genome Packaging of Influenza A and B Viruses.” *mBio* 7 (5). https://doi.org/10.1128/mBio.01248-16.
Nouri, Shahideh, Rafael Arevalo, Bryce W. Falk, and Russell L. Groves. 2014. “Genetic Structure and Molecular Variability of Cucumber Mosaic Virus Isolates in the United States.” *PloS One* 9 (5): 96582.
Ojosnegros, S., J. García-Arriaza, C. Escarmís, S. C. Manrubia, and C. Perales. 2011. “Viral Genome Segmentation Can Result from a Trade-Off between Genetic Content and Particle Stability.” *PLoS Genetics* 7 (3): 1001344.
Ouattara, Alassane, Fidèle Tiendrébéogo, Nathalie Becker, Cica Urbino, Gaël Thébaud, Murielle Hoareau, Agathe Allibert, et al. 2022. “Synergy between an Emerging Monopartite Begomovirus and a DNA-B Component.” *Scientific Reports* 12 (1): 695.
Gallet, Romain, Jérémy Di Mattia, Sébastien Ravel, Jean-Louis Zeddam, Renaud Vitalis, Yannis Michalakis, and Stéphane Blanc. 2022. “Gene Copy Number Variations at the Within-Host Population Level Modulate Gene Expression in a Multipartite Virus.” *Virus Evolution* 8 (2): veac058.
Parker, R. F. 1938. “Statistical Studies of the Nature of the Infectious Unit of Vaccine Virus.” *The Journal of Experimental Medicine* 67 (5): 725–38.
Parrella, G., P. Gognalons, K. Gebre-Selassiè, C. Vovlas, and G. Marchoux. 2003. “An Update of the Host Range of Tomato Spotted Wilt Virus.” *Journal of Plant Pathology: An International Journal of the Italian Phytopathological Society* 85 (4): 227–64.
Qiu, W., and J. W. Moyer. 1999. “Tomato Spotted Wilt Tospovirus Adapts to the TSWV N Gene-Derived Resistance by Genome Reassortment.” *Phytopathology* 89 (7): 575–82.
Reanney, D. C. 1982. “The Evolution of RNA Viruses.” *Annual Review of Microbiology* 36: 47–73.
Regoes, R. R., J. W. Hottinger, L. Sygnarski, and D. Ebert. 2003. “The Infection Rate of Daphnia Magna by Pasteuria Ramosa Conforms with the Mass-Action Principle.” *Epidemiology and Infection* 131 (2): 957–66.
Rochon, D ’ann, Kishore Kakani, Marjorie Robbins, and Ron Reade. 2004. “Molecular Aspects of Plant Virus Transmission by Olpidium and Plasmodiophorid Vectors.” *Annual Review of Phytopathology* 42: 211–41.
Rohde, K. 1994. “Niche Restriction in Parasites: Proximate and Ultimate Causes.”

Parasitology 109 Suppl: S69–84.
Rohrmann, George F. 2019. *Baculovirus Molecular Biology*. National Center for Biotechnology Information (US).
Sacristán, Soledad, Maira Díaz, Aurora Fraile, and Fernando García-Arenal. 2011. "Contact Transmission of Tobacco Mosaic Virus: A Quantitative Analysis of Parameters Relevant for Virus Evolution." *Journal of Virology* 85 (10): 4974–81.
Sánchez-Navarro, Jesús A., Mark P. Zwart, and Santiago F. Elena. 2013. "Effects of the Number of Genome Segments on Primary and Systemic Infections with a Multipartite Plant RNA Virus." *Journal of Virology* 87 (19): 10805–15.
Sanjuán, Rafael, and María-Isabel Thoulouze. 2019. "Why Viruses Sometimes Disperse in Groups†." *Virus Evolution* 5 (1). https://doi.org/10.1093/ve/vez014.
Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. "The Strange Lifestyle of Multipartite Viruses." Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819.
Sicard, Anne, Elodie Pirolles, Romain Gallet, Marie-Stéphanie Vernerey, Michel Yvon, Michel Peterschmitt, Serafin Gutierrez, Yannis Michalakis, and Stéphane Blanc. 2019. "A Multicellular Way of Life for a Multipartite Virus." *eLife* 8 (March): e43599.
Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. "Gene Copy Number Is Differentially Regulated in a Multipartite Virus." *Nature Communications* 4: 2248.
Simon-Loriere, Etienne, and Edward C. Holmes. 2011. "Why Do RNA Viruses Recombine?" *Nature Reviews. Microbiology* 9 (8): 617–26.
Subbarao, E. K., W. London, and B. R. Murphy. 1993. "A Single Amino Acid in the PB2 Gene of Influenza A Virus Is a Determinant of Host Range." *Journal of Virology* 67 (4): 1761–64.
Taschner, P. E., A. C. van der Kuyl, L. Neeleman, and J. F. Bol. 1991. "Replication of an Incomplete Alfalfa Mosaic Virus Genome in Plants Transformed with Viral Replicase Genes." *Virology* 181 (2): 445–50.
Valdano, Eugenio, Susanna Manrubia, Sergio Gómez, and Alex Arenas. 2019. "Endemicity and Prevalence of Multipartite Viruses under Heterogeneous between-Host Transmission." *PLoS Computational Biology* 15 (3): e1006876.
Werf, Wopke van der, Lia Hemerik, Just M. Vlak, and Mark P. Zwart. 2011. "Heterogeneous Host Susceptibility Enhances Prevalence of Mixed-Genotype Micro-Parasite Infections." *PLoS Computational Biology* 7 (6): e1002097.
Whitfield, Anna E., Bryce W. Falk, and Dorith Rotenberg. 2015. "Insect Vector-Mediated Transmission of Plant Viruses." *Virology*. https://doi.org/10.1016/j.virol.2015.03.026.
Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. "Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus." *Scientific Reports* 7 (1): 1–15.
Yu, Nai-Tong, Hui-Min Xie, Yu-Liang Zhang, Jian-Hua Wang, Zhongguo Xiong, and Zhi-Xin Liu. 2019. "Independent Modulation of Individual Genomic Component Transcription and a Cis-Acting Element Related to High Transcriptional Activity in a Multipartite DNA Virus." *BMC Genomics* 20 (1): 573.
Zhang, Yi-Jiao, Zhi-Xi Wu, Petter Holme, and Kai-Cheng Yang. 2019. "Advantage of Being Multicomponent and Spatial: Multipartite Viruses Colonize Structured Populations with Lower Thresholds." *Physical Review Letters* 123 (13): 138101.
Zhao, Wan, Qianshuo Wang, Zhongtian Xu, Renyi Liu, and Feng Cui. 2019. "Distinct Replication and Gene Expression Strategies of the Rice Stripe Virus in Vector Insects and Host Plants." *The Journal of General Virology* 100 (5): 877–88.
Zwart, Mark P., Stéphane Blanc, Marcelle Johnson, Susanna Manrubia, Yannis Michalakis, and Mircea T. Sofonea. 2021. "Unresolved Advantages of Multipartitism in Spatially Structured Environments." *Virus Evolution* 7 (1): veab004.
Zwart, Mark P., José Antonio Daròs, and Santiago F. Elena. 2011. "One Is Enough: In Vivo Effective Population Size Is Dose-Dependent for a Plant RNA Virus." *PLoS Pathogens* 7 (7): e1002122.

Zwart, Mark P., and Santiago F. Elena. 2015. “Testing the Independent Action Hypothesis of Plant Pathogen Mode of Action: A Simple and Powerful New Approach.” *Phytopathology*® 105 (1): 18–25.
———. 2020. “Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments†.” *Virus Evolution* 6 (1): veaa022.



Robust approaches to the quantitative analysis of genome formula variation in multipartite and segmented Viruses

Marcelle L. Johnson[1,2] and Mark P. Zwart[1]

[1] Netherlands Institute of Ecology (NIOO-KNAW), P.O. BOX 50, 6700 AB, Wageningen, The Netherlands

[2] Laboratory of Virology, Wageningen University and Research, P.O. BOX 16, 6700 AA, Wageningen, The Netherlands

Published in Viruses (2024), 16(2): 270
doi: https://doi.org/10.3390/v16020270

Abstract

When viruses have segmented genomes, the set of frequencies describing the abundance of segments is called the genome formula. The genome formula is often unbalanced and highly variable for both segmented and multipartite viruses. A growing number of studies are quantifying the genome formula to measure its effects on infection and to consider its ecological and evolutionary implications. Different approaches have been reported for analyzing genome formula data, including qualitative description, applying standard statistical tests such as ANOVA, and customized analyses. However, these approaches have different shortcomings, and test assumptions are often unmet, potentially leading to erroneous conclusions. Here, we address these challenges, leading to a threefold contribution. First, we propose a simple metric for analyzing genome formula variation: the genome formula distance. We describe the properties of this metric and provide a framework for understanding metric values. Second, we explain how this metric can be applied for different purposes, including testing for genome-formula differences and comparing observations to a reference genome formula value. Third, we re-analyze published data to illustrate the applications and weigh the evidence for previous conclusions. Our re-analysis of published datasets confirms many previous results but also provides evidence that the genome formula can be carried over from the inoculum to the virus population in a host. The simple procedures we propose contribute to the robust and accessible analysis of genome-formula data.

Introduction

Many viruses have segmented genomes: their complete hereditary material consists of multiple nucleic acid molecules. Packaging these genome segments into virus particles can result in various distributions of genome segments over virus particles (Sicard et al. 2016; Michalakis and Blanc 2020) (Figure 1). Segmented viruses package one copy of each genome segment into each virus particle (Figure 1b). This arrangement is thought to ensure genome integrity and maximize opportunities for virus transmission. By contrast, multipartite viruses package each genome segment into a separate virus particle (Figure 1c). This arrangement results in a dependence on multiple virus particles for successful virus transmission, and it is thought to make transmission less efficient and, thereby, impose a substantial cost to virus spread (Sánchez-Navarro, Zwart, and Elena 2013; Fulton 1962). Interestingly, some viruses blur the distinction between segmented and multipartite viruses. These viruses do not always package a full complement of genome segments into each virus particle (Wichgers Schreur and Kortekaas 2016; Yvon et al. 2023), resulting in transmission that depends partly on incomplete particles (Jacobs et al. 2019; Diefenbacher, Sun, and Brooke 2018; Bermúdez-Méndez et al. 2022) (Figure 1d). Whereas segmented viruses are most common among animal viruses, multipartite viruses abound among plant viruses (Michalakis and Blanc 2020; Lucía-Sanz and Manrubia 2017). However, there are many examples of segmented plant viruses (Lucía-Sanz and Manrubia 2017; Michalakis and Blanc 2020). At least one multipartite animal virus has been identified (Hu et al. 2016), and there are likely more cases (Michalakis and Blanc 2020; Ladner et al. 2016).

For some multipartite and segmented viruses, variation in the frequency of genome segments has been observed (Diefenbacher, Sun, and Brooke 2018; Sicard et al. 2013; Wu et al. 2017; Boezen, Vermeulen, et al. 2023; Moreau et al. 2020). The genome formula is the abundance of all virus genome segments, and it is typically described in one of two ways. If we take a bi-segmented virus with segments at equal abundance as an example, the genome formula can be expressed as a ratio 1:1 (segment1:segment2) or as a set of relative frequencies {0.5, 0.5} {segment1, segment2}. We use the latter convention throughout this paper. Current interest in the genome formula was sparked by the seminal work of Sicard and coworkers on faba bean necrotic stunt virus (FBNSV), a multipartite DNA virus with eight genome segments (Sicard et al. 2013). These authors showed that the genome formula converges on an unbalanced equilibrium when disrupted, and this equilibrium is host-species-dependent. Notably, the authors also observed considerable variation within and between plants in the genome formula, highlighting its stochastic nature. Later work confirmed similar findings for alfalfa mosaic virus (AMV), a multipartite plant RNA virus with three genome segments (Wu et al. 2017). From a historical perspective, it is interesting to note that previous observations already showed the variable nature of the genome formula for multipartite (Hajimorad et al. 1991) and segmented (Kormelink et al. 1992; Wichgers Schreur, Kormelink, and Kortekaas 2018) viruses, even if the implications may not have been acknowledged then. In the meantime, genome formula variation has also been shown for segmented animal viruses (Moreau et al. 2020; Diefenbacher, Sun, and Brooke 2018). Although studies on the genome formula have focused on full-length virus genome segments (Sicard et al. 2013; Wu et al. 2017; Boezen, Johnson, et al. 2023), other genetic elements are also relevant. For example, many RNA viruses produce sub-genomic RNAs, and for some viruses, these RNAs can be packaged into virus particles (Roossinck 2001). Parasitic genetic elements such as satellites are also known

to affect the genome formula (Mansourpour et al. 2021; Obrępalska-Stęplowska et al. 2015), and a full understanding will, therefore, require considering these elements. Given that genome formula variation appears to be a feature of many virus–host systems, what are the causes and consequences of this variation?

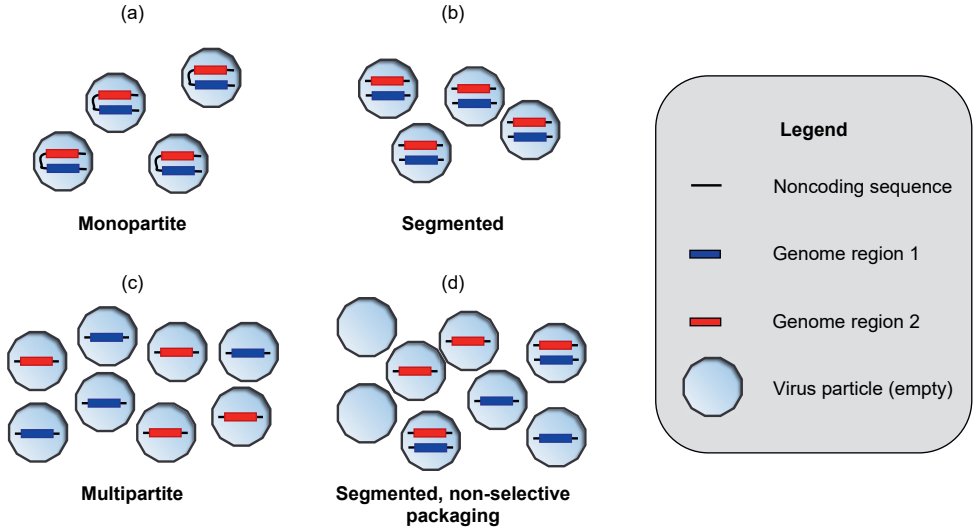


Figure 1. We provide a schematic illustration of the variation in the distribution of genome segments (nucleic acid molecules) over virus particles. *A legend is given on the far right. In each case shown, we assume the virus genome consists of the two identical coding genome regions, identified by blue and red fills, forming one or two segments. (**a**) Monopartite viruses have a single genome segment. Note that the two genome regions form a single molecule in the illustration. (**b**) Segmented viruses have multiple genome segments: two genome segments in this example. These viruses package a full complement of genome segments into each virus particle. (**c**) A multipartite virus with two genome segments is shown. Each segment is packaged individually into a virus particle. Infection will depend on the transmission of multiple virus particles, as both a blue and a red segment are needed. (**d**) A segmented virus with non-selective packaging is shown. The illustration is a hypothetical distribution based only on the observation that for some segmented viruses, many virus particles have an incomplete set of genome segments (Wichgers Schreur and Kortekaas 2016; Wichgers Schreur, Kormelink, and Kortekaas 2018). This organization is included to highlight that many distributions of genome segments over virus particles are possible, and that the genome formula of segmented viruses does not have to be balanced (i.e., not 1:1 ratio of genome segments).*

Both random and directional forces are likely to shape variation in the genome formula. Population bottlenecks are likely to result in stochastic variation in the genome formula. When the total number of segments entering a cell is small, the frequencies of the different segment types are likely to vary, a process known as genome formula drift (Gutiérrez and Zwart 2018). Sicard et al. (2013) suggested that variation in the genome formula is similar to copy number variation (CNV), possibly affecting gene expression and, thereby, enabling a rapid tuning of

gene expression (Sicard et al. 2013). Under this hypothesis, selection for a beneficial genome formula would also be a directional force (Zwart and Elena 2020). Other directional forces may include differences in the rates of replication or encapsidation for different segments (Sicard et al. 2016; Wu et al. 2017).

Many plant viruses that cause disease and economic losses in cultivated plants are multipartite or segmented viruses, including viruses with very broad host ranges (Rybicki 2015). For example, the multipartite viruses cucumber mosaic virus (CMV) and AMV have broad host ranges, as does the segmented tomato spotted wilt virus (TSWV) (Lamy-Besnier et al. 2021). Having three or four genome segments has been identified as a predictor for a large host range in plant viruses (Moury et al. 2017). As genome formula changes may enable these broad host ranges (Sicard et al. 2016), the genome formula may also have relevance for understanding virus emergence and disease outbreaks. There are no reports of genome formula variation in real-world virus populations; still, we speculate that the genome formula might have value as a tool for the monitoring of virus populations in crops and predicting disease outcomes. Finally, theory suggests that agro-ecosystems may also be conducive to the propagation of multipartite viruses due to many opportunities for transmission in dense monocultures (Valdano et al. 2019). For these reasons, studying the infection dynamics and genome formula variation of multipartite viruses in experiments and in agricultural ecosystems is a relevant topic within plant virus epidemiology.

Most studies quantify the genome formula with the same molecular method. For DNA viruses, quantitative polymerase chain reaction (qPCR) is used, whereas RNA viruses require reverse transcription—qPCR (RT-qPCR). In these assays, specific primers are used to amplify distinct template sequences on the different genome segments, and SYBR-Green-induced fluorescence is used to quantify amplicon copy numbers. For those viruses that generate subgenomic RNAs, primers are designed to amplify templates that only occur in the full-length RNA (Boezen, Johnson, et al. 2023). One study compared three other methods to RT-qPCR for the quantification of the CMV genome formula: RT—digital droplet PCR (RT-dPCR), Illumina short-read sequencing, and Oxford Nanopore Technologies (ONT) long-read sequencing. This study found that the methods give roughly similar results, although there are systematic differences (Boezen, Johnson, et al. 2023). Another study on FBNSV showed that rolling circle amplification (RCA), a common amplification step before sequencing for circular DNA viruses, may lead to discrepancies in the quantification of the genome formula compared to qPCR (Gallet et al. 2017).

Once the genome formula has been quantified, there are several different approaches for analyzing these data, driven in part by different research questions. For many studies, a key question is how to make rigorous genome formula comparisons for two or more groups. To show the breadth of approaches used to address this question, we provide a non-exhaustive overview (Table 1). When we consider the strengths and weaknesses of these approaches, we see that most approaches used have some crucial shortcomings (Table 1). In many cases, model assumptions are not met, or the procedure can only be applied to a bipartite virus or one specific genome segment. Ideally, we want a single method for comparing the complete genome formula with a limited set of model assumptions that can be met in practice.

While there are compelling hypotheses about the genome formula, exploring the causes and consequences of genome formula variation will require robust approaches. To date, studies have used a plethora of different approaches, ranging from simple qualitative comparisons to

Table 1. Approaches to comparing genome formula values for two or more groups.

Approach	Strengths	Weaknesses	Ref.
Analysis of variance (ANOVA) on the relative frequencies of individual genome segments	(i) Parsimony of the analysis	(i) Limited to the analysis of individual segments (ii) Model assumptions [1]	(Sicard et al. 2013)
Multivariate analysis of variance (MANOVA) on the relative frequency of all genome segments	(i) Single analysis of all segments (ii) Technical error included in the analysis	(i) Dependence between relative frequencies (ii) Model assumptions [1,2]	(Wu et al. 2017)
Model selection based on the ΔGF metric [3] for all genome segments	(i) Single analysis of all segments	(i) Assumptions for estimating the likelihoods and weighing of model parameters for model selection [4]	(Boezen, Johnson, et al. 2023)
T-tests on ratio of the log-tranformed RNA1:RNA2	(i) Parsimony (ii) Model assumptions met	(i) Only applicable to bipartite viruses (ii) Consider effects of a single factor	(Kennedy et al. 2023)
PERMANOVA on the genome formula distance metric [5]	(i) Parsimony (ii) Single analysis of all segments (iii) Model assumptions met	(i) If there are differences in spread, differences in centroid cannot be assessed	(Boezen, Vermeulen, et al. 2023)

[1] Normality of the residuals and equality of variance assumptions may not be met. For the comparison of single segments with ANOVA, the assumption of independence of observations is met. For comparison of multiple segments, the assumption is violated. [2] In addition to ANOVA assumptions, MANOVA assumes no multivariate outliers. [3] The cumulative distance between genome formula observations and a reference value (Sicard et al. 2013), which, in this case, is the mean value for the group under consideration. [4] To calculate the negative log likelihood for these data, residuals are assumed to be normally distributed. In addition, each group mean is weighed as a free parameter for model selection, whereas it follows directly from the data. [6] This metric is described in detail in in the results section Applications of the Genome Formula Distance Metric.

employing sophisticated statistical methods. This study is focused on these analysis methods and their effect on outcomes. Based on our previous experience with developing approaches for analyzing genome formula data, our hypothesis is that the method used can have a critical effect on study outcome. The result we work towards is having a robust, well-documented approach to analyzing genome formula data, which has been applied to various datasets,

illustrating its applications and demonstrating its relevance. Here, we propose a simple and robust approach to genome formula analysis that relies on the genome formula distance metric (Boezen, Vermeulen, et al. 2023). We document this method in detail as a resource for the analysis of genome-formula data. We provide a framework for interpreting our metric's values and explore how this approach can be applied to different problems. Finally, we re-analyze some previously published datasets to illustrate the benefits of this approach and as a validation of previous analyses.

Methods

All analyses were performed with R version 4.3.1 software for statistical computing (R Foundation for Statistical Computing 2023). Calculations of the genome formula distance were performed with the *vegdist* function, PERMANOVA was performed with the *adonis2* function, and PERMDISP2 was performed with the *betadisper* and *permutest* functions, which all pertain to the vegan Community Ecology Package version 2.6-4 (Oksanen et al. 2022).

All code for analysis and the data formatted for analysis are available as R markdown files at Zenodo (10.5281/zenodo.10355273). Access to the submission is currently restricted to avoid any confusion prior to the availability of the paper; please follow this link to gain access.

Results

The Genome Formula Distance Metric

Given the shortcomings of many methods for analyzing genome formula variation, we recently developed another approach, based on the genome formula distance metric (Boezen, Vermeulen, et al. 2023), in combination with permutation-based statistical approaches (Anderson 2017, 2001). Here, we build on this previous work by describing this metric in detail and considering some of its attributes, such as the range of values and its interpretation.

The Genome Formula Distance Metric

We consider the genome formula (*G*) as the set of relative frequencies for all virus genome segments. For a viral genome with *k* segments:

$$G = \{f_1, f_2, [\dots] f_k\} \quad (1)$$

Here, *f* is the relative mean frequency of a segment, such that for the j^{th} segment:

$$f_j = c_j / \sum_{i=1}^{k} c_i \quad (2)$$

Here, *c* is a measurement of accumulation for a specific segment, such as quantitative polymerase chain reaction (qPCR) measurements. Per definition, the sum of all *f* values is 1. When any measurement of segment accumulation *c* changes, it will affect the relative frequency of all other segments.

To compare two values of the genome formula, in a previous study, we proposed to consider the Euclidean distance between them (Boezen, Vermeulen, et al. 2023). We refer to this metric as the genome formula distance (*D*), such that for two genome formula observations *a* and *b*, the distance between them is as follows:

$$D_{a,b} = \sqrt{\sum_{i=1}^{k} \left(f_{a,i} - f_{b,i}\right)^2} \quad (3)$$

Intuitively, *D* is simply the length of the straight line connecting two points in an *n*-dimensional space (Figure 2). The multivariate genome formula data are, therefore, reduced to a single distance value, simplifying analysis and removing the dependence between measurements expressed as relative frequencies. Although we previously described this metric and applied it for comparing groups of genome formula observations, we did not consider the properties of this metric in detail. Therefore, before considering here how this metric can be applied to data for several different goals, we describe some properties of this metric and generate expectations based on first principles in detail.

Minimum and Maximum Values of the Genome Formula Distance Metric

Various properties of the metric *D* can be readily established. Its minimum value is $D_{a,b}^{min} = 0$, which is when two genome formula values coincide. Its maximum value is $D_{a,b}^{max} = \sqrt{2}$, as can be shown by induction (Figure 2). For a bipartite virus, the greatest possible *D* will be obtained when $G_a = \{1,0\}$ and $G_b = \{0,1\}$, when $D_{a,b} = \sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2}$. For tripartite and tetrapartite, the greatest distance occurs along the edges of the genome formula space. These edges represent the line connecting *G* values composed of the presence of only one segment, resulting in $D_{a,b} = \sqrt{2}$ (Figure 2). In real life, we do not expect to see such large values, as we do not expect to see replicating virus populations in which only a single segment is present. Although it is possible for some multipartite viruses to lose and reacquire a segment (Di Mattia et al. 2022), all or a number of core segments are often required for replication (Sánchez-Navarro, Zwart, and Elena 2013; Zwart et al. 2021). It is, therefore, interesting to consider what values of *D* can be expected under scenarios with a higher biological relevance.

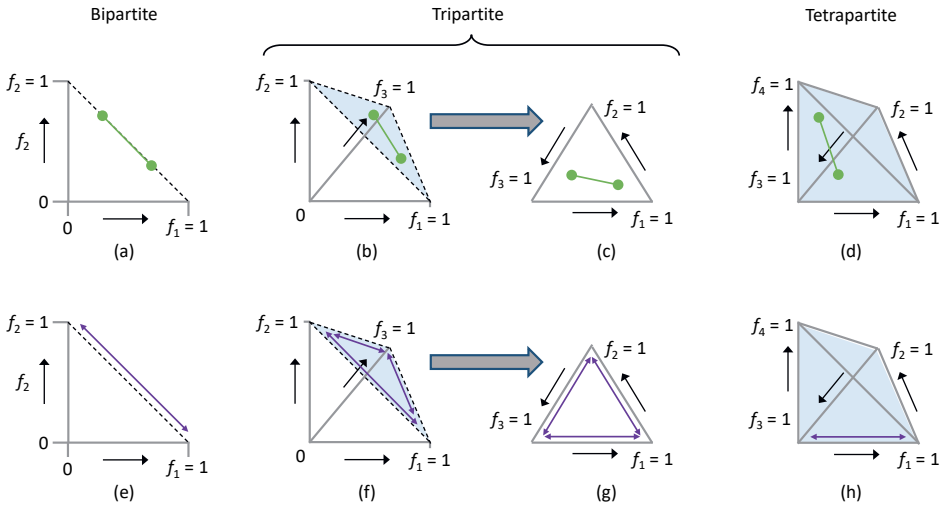


Figure 2. Here, we illustrate the genome formula distance metric (top panels, green lines) and its maximum possible distance for different numbers of genome segments (bottom panels, purple arrows). *Figure axes are genome segment frequencies (f) for 2 (panels a and b), 3 (panels b,c,f, and g), or 4 genome segments (panels d and h). (**a**) For a bipartite virus, we illustrate two possible genome formula values with green points and the distance between them with a line. Note that for the bipartite virus, all possible genome formula values fall on the dotted line connecting (1,0) and (0,1). (**b**) For a tripartite virus, we illustrate two possible genome formula values in three-dimensional genome formula space. As the sum of relative frequencies is 1, all possible genome formula values fall in the triangular plane illustrated by the dotted lines and light blue shading. (**c**) As all values fall in the same plane in panel b, genome formula values for a tri-segmented virus are often illustrated in only this plane, resulting in a ternary plot. (**d**) Two genome formula values and their distance are illustrated for a tetrapartite virus in a quarternary plot. All values in the tetrahedron represent possible genome formula values, as indicated by the light blue shading. (**e**) The maximum possible genome formula distance for a bipartite virus is simply the line connecting the points (1,0) and (0,1). (**f**) For the tripartite virus, the longest possible distance in the genome formula space is attained along its borders, resulting in an identical maximum genome formula distance to the bipartite virus. The light blue shading indicates the possible space for genome formula values. (**g**) The outcome described in panel f is clearer in the ternary plot of the genome formula space. (**h**) For a tetrapartite virus, there is no distance between two points in the genome formula space that is longer than the maximum distance for the bipartite and tripartite viruses. This maximum distance occurs at the edges of the genome formula space, as indicated by the light blue shading, connecting the vertices, which represent the presence of a single segment. To keep the panel clear, we only illustrate this for one edge for a tetrapartite virus, although there are six such edges.*

Distance Metric for Random Genome Formula Variation

To determine a plausible upper limit for the mean distance between two observations of the genome formula ($\overline{D}_{a,b}$), we assume that all genome segments must be present in the virus population, but that the level of accumulation is, otherwise, entirely random. For each segment, we, therefore, sample a value from a uniform distribution and then determine the mean pairwise distance $\overline{D}_{a,b}^{rand}$. The values of $\overline{D}_{a,b}^{rand}$ depend on the number of genome segments, with a maximum value of 0.391 for a tri-segmented virus (Table 2). If we find similar values for $\overline{D}_{a,b}$ for a real-world virus population, this result would suggest a genome formula shaped by random levels of accumulation for the different segments.

Table 2. Expected values of *D* for random genome formula variation ($\overline{D}_{a,b}^{rand}$) or the maximum genome formula drift introduced by a single bottleneck event ($\overline{D}_{a,b}^{drift}$).

Number of Genome Segments	$\overline{D}_{a,b}^{rand}$	$\overline{D}_{a,b}^{drift}$	λ [1]
2	0.3855	0.2877	5.37
3	0.3905	0.2801	7.08
4	0.3638	0.2629	9.12
5	0.3367	0.2494	10.47
6	0.3132	0.2341	12.30
7	0.2934	0.2189	14.12
8	0.2767	0.2060	15.85
9	0.2625	0.1929	18.20
10	0.2501	0.1847	19.50

[1] The bottleneck value corresponding to the maximum $\overline{D}_{a,b}^{drift}$ value.

Distance Metric for Maximum Genome Formula Drift

Whereas the strength of genetic drift decreases monotonically as effective population size increases, the strength of genome formula drift is maximized at an intermediate effective population size (Zwart and Elena 2020). Therefore, to determine the maximum level of genome formula drift that a single population bottleneck event can induce, we have to consider a range of bottleneck sizes. We assume that the total number of virus particles that initiates an infection follows a Poisson distribution with a mean value λ and consider the predicted genome formula distance over a broad range of λ values for different numbers of genome segments (Figure 3). The maximum genome formula distance values, $\overline{D}_{a,b}^{drift}$, are given in Table 2. As expected, these values are lower than those obtained for random genome formula variation ($\overline{D}_{a,b}^{rand}$), as the assumption of a Poisson-distributed number of founders constrains the variation in genome segment frequencies. If a population shows similar values of $\overline{D}_{a,b}$, this suggests that the genome formula variation observed is equivalent to the maximum variation that can be generated by a single bottlenecking event.

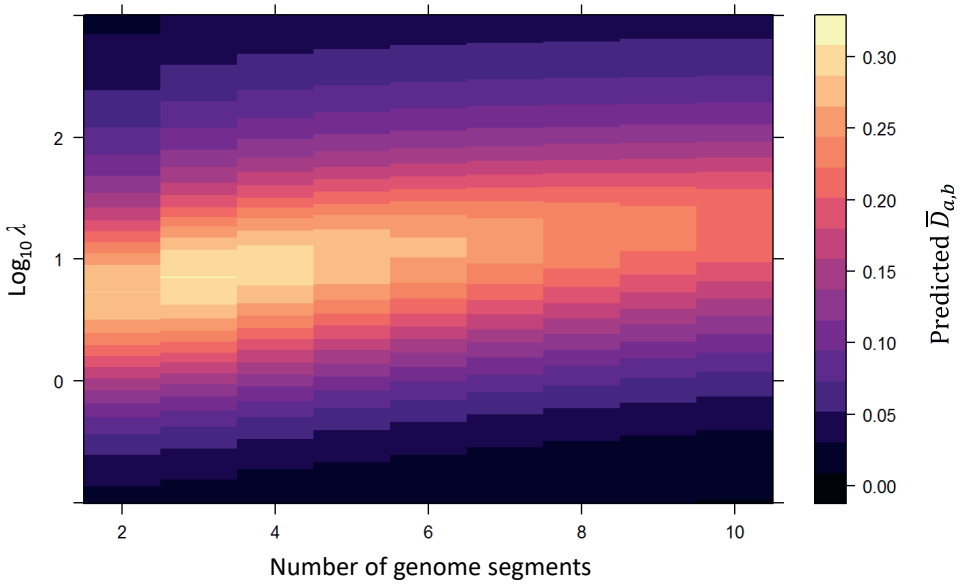


Figure 3. The effects of the number of segments and bottleneck size on the predicted genome formula *distance are illustrated. The x-axis indicates the number of virus genome segments, whereas the y-axis indicates the log-transformed number of infection founders (λ). For all combinations of these values, we predicted the mean genome formula distance $\overline{D}_{a,b}$, a value indicated by the heat according to the legend on the far right. We used these simulation results to determine the highest value of $\overline{D}_{a,b}$ for each number of genome segments, a value we term $\overline{D}_{a,b}^{drift}$. Note that the highest mean distance values occur at intermediate values of λ, as well as being associated with higher values of λ as the number of segments is increased.*

Applications of the Genome Formula Distance Metric

To illustrate how this metric can be applied to experimental data, we re-analyze datasets from several studies on plant multipartite viruses. We do not attempt to reproduce all analyses in these original studies here. Rather, we focus on a few cases to illustrate how an approach based on the genome formula distance can be used. Note that all the genome formula data re-analyzed throughout this study were obtained through qPCR or RT-qPCR. The only exception is the methods comparison by Boezen and coworkers (Boezen, Johnson, et al. 2023). Here, for that study, we also explicitly address the effect of different methods on genome formula quantification, as was performed in the original work.

Comparison of the Genome Formula to Theoretical Values

We defined clear expectations for the upper limit of the genome formula distance metric for the random accumulation of genome segments ($\overline{D}_{a,b}^{rand}$) or the maximum amount of genome formula drift generated by a single population bottleneck ($\overline{D}_{a,b}^{drift}$) (Table 2). First, we compare these theoretical predictions to observed values of genome formula distance ($\overline{D}_{a,b}$). We obtain these observed values by re-analyzing genome formula data reported in three experimental studies in which the genome formula was measured in single leaves or whole plants (Sicard et al. 2013; Wu et al. 2017; Boezen, Vermeulen, et al. 2023). For the tripartite RNA viruses AMV and CMV, we find that the observed values for the genome formula distance are below both of our reference values (Table 3), as expected for systems that appear to converge on an equilibrium value. Two out of three measurements for AMV are close to the value measured for CMV (~0.20), which is near to prediction for maximum genome formula drift ($\overline{D}_{a,b}^{drift} \sim 0.28$ for a tri-segmented virus). For the octapartite DNA virus FBNSV, we see a decrease in $\overline{D}_{a,b}$, indicating a reduction in variability over leaf levels (Table 3) as reported in the original study in Figure 3A (Sicard et al. 2013). The decrease in $D_{a,b}$ over leaf levels is highly significant (Kendall rank correlation: $\tau = 0.368$, $N = 77$, $p < 0.001$). When we compare values of $\overline{D}_{a,b}$ to model predictions, we find that it is higher than $\overline{D}_{a,b}^{rand}$ in the inoculated leaf (leaf level 1) but falls to and remains at levels below the $\overline{D}_{a,b}^{drift}$ predictions by leaf level 3 (Table 3).

Overall, these comparisons between model predictions and observed values of $\overline{D}_{a,b}$ underscore that there is considerable genome formula variation, suggesting that stochastic forces play an important role in shaping the genome formula. The differences in variability for the AMV estimates might reflect differences between the inoculated and systemic leaves but may reflect the relatively low number of replicates for each condition ($n = 6$). This variability stresses the need for high levels of replication for the representative estimates of these indexes. For FBNSV, the higher-than-expected genome formula variation in the inoculated leaf is striking. However, this phenomenon is probably related to the inoculation with *Agrobacterium*, as once the virus has systemically moved, it no longer surpasses model predictions of $\overline{D}_{a,b}^{rand}$.

Table 3. Observed values for the genome formula distance ($\bar{D}_{a,b}$) for two tripartite viruses.

Genome Segments	Model Predictions [1]		Ref	Experiment	*n*	$\bar{D}_{a,b}$ ± SD
	$\bar{D}_{a,b}^{rand}$	$\bar{D}_{a,b}^{drift}$				
3	0.391	0.280	(Wu et al. 2017)	AMV in *N. benthamiana*, inoculated	6	0.077 ± 0.015
				AMV in *N. benthamiana*, lower leaf	6	0.195 ± 0.029
				AMV in *N. benthamiana*, upper leaf	6	0.197 ± 0.124
			(Boezen, Vermeulen, et al. 2023)	CMV in *N. tabacum*, whole plant	9	0.207 ± 0.069
8	0.277	0.206	(Sicard et al. 2013)	FBNSV in *V. faba*, leaf level 1	9	0.352 ± 0.097
				FBNSV in *V. faba*, leaf level 2	8	0.275 ± 0.062
				FBNSV in *V. faba*, leaf level 3	13	0.198 ± 0.045
				FBNSV in *V. faba*, leaf level 4	15	0.175 ± 0.050
				FBNSV in *V. faba*, leaf level 5	16	0.198 ± 0.063
				FBNSV in *V. faba*, leaf level 6	16	0.178 ± 0.031

[1] Predictions of the mean genome formula distance under random accumulation ($\bar{D}_{a,b}^{rand}$) and the maximum genome formula drift introduced by a single bottleneck event ($\bar{D}_{a,b}^{drift}$) are given, depending on the number of genome segments, as given in Table 2.

Comparison of the Genome Formula for Different Groups

Boezen and coworkers first applied the genome formula distance metric to compare the genome formula for different treatments (Boezen, Vermeulen, et al. 2023). In this section, we first describe these previous results in detail, as they are important for understanding this approach and its limitations. This previous study explored the effects of mixed infection with other plant viruses on CMV's genome formula (Boezen, Vermeulen, et al. 2023). To compare the genome formula of CMV in different treatments, the authors calculated the genome formula distances and then performed PERMANOVA. PERMANOVA is a permutational multivariate analysis of variance, a non-parametric ANOVA widely applied in ecology (Anderson 2017, 2001). PERMANOVA is often applied to such analyses because of its robustness: the test makes fewer assumptions than parametric procedures. Note that if we

apply PERMANOVA to the genome formula distance as suggested here, we are performing a univariate analysis, for which PERMANOVA is also suitable. One interesting feature of PERMANOVA is that the procedure detects both differences in mean (or centroid for multivariate data) and spread. If we detect a significant difference, we must rule out a significant difference in spread before we can conclude that there are differences in the mean. The PERMDISP2 procedure tests whether there are significant differences in spread (Anderson 2017). When Boezen and coworkers applied this procedure, they found a significant difference between the PERMANOVA and PERMDISP2 procedures (Boezen, Vermeulen, et al. 2023). Therefore, in this case, the authors could only conclude that mixed infections had a significant effect on genome formula spread, surprisingly leading to a reduction in the spread compared to a CMV-only infection. Now that we have described this procedure and its application in previous work in detail, we consider how it can be applied to other datasets.

To further illustrate how PERMANOVA on the genome formula distance is useful, we re-analyzed data from four other experiments (see Appendix A for a detailed description). For the first dataset we consider here, the original study measured the genome formula of CMV with four different methods in three hosts (Boezen, Johnson, et al. 2023). The study found no effect of host species on the genome formula, and although the different methods gave similar results, there was a significant effect of method on the measured genome formula (Boezen, Johnson, et al. 2023). When we re-analyzed these genome formula data, we found largely similar results when comparing our new procedure to the model selection in the original study. The PERMANOVA-based procedure is more robust (Table 1) but still manages to identify some subtle species effects on the genome formula that were not detected by the original analysis (see Appendix A). The second dataset we considered was from a study that showed frequency-dependent selection results in an equilibrium for AMV's genome formula, and it showed that the genome formula of this RNA virus is host-species-dependent (Wu et al. 2017). A number of datasets are reported in this paper, and we choose to focus on one specific question for our re-analysis: are there differences in the genome formula in the inoculated leaf, for leaves inoculated with different genome formulae? Here, we did not find a significant effect (Appendix A). This result contradicts the result of the statistical test in the original study. However, all plant tissues were jointly analyzed in the original paper, whereas here, we focused exclusively on the inoculated leaf. From a biological perspective, it makes the most sense to look for an effect of the inoculum early in the infection process. In the final section of the results (Comparison of the Genome Formula to Reference), we explore a different approach to analyzing these AMV data that sheds more light on the underlying processes.

Next, we compared the genome formula distance for two sets of experiments on the octapartite FBNSV in a seminal study that reignited interest in these viruses (Sicard et al. 2013). The third dataset we re-analyzed considers the genome formula in different leaf levels (Sicard et al. 2013), the same dataset we used to determine the pairwise distance between genome formula measurements (Table 3). As we found large differences in genome formula variability (Table 3), we expect and indeed find that the PERMDISP2 result is significant (Appendix A). The results of the distance measurements and PERMANOVA are in good agreement. The original study used ANOVA to analyze the coefficient of variation for the genome formula in different leaf levels, also finding significant differences in variation between leaf levels (Sicard et al. 2013). Second, we considered the FBNSV genome formula in two plant species (Sicard et al. 2013), for which the authors analyzed the abundance of individual

segments. In agreement with the original analyses, we find highly significant differences in the genome formula distance between the two plant species, while the experiments in the same plant species render similar results (Appendix A).

These examples illustrate how readily our proposed approach can be used to analyze genome formula data. Our results are largely congruent with previous results in three out of four cases. However, there is a discrepancy for the data of Wu et al. on AMV infection (Wu et al. 2017), for which we analyzed a subset of the data using a different approach. This discrepancy illustrates that the approach and methods used matter for the results obtained.

Comparison of the Genome Formula to Reference

We can also use the genome formula distance metric to compare observations of the genome formula to a reference. The reference genome formula used will depend on the question being addressed. We provide some examples to illustrate a range of reference values and a purpose for the comparison, to show the breadth of potential applications. These possible reference values include the following: (i) the mean genome formula for a group of observations (which, in effect, also occurs for PERMANOVA); (ii) the genome formula used in the inoculum for an experiment, to test whether it is maintained; (iii) a balanced genome formula (i.e., 1:1:1), to quantify the imbalance in the genome formula (see examples using another metric (Sicard et al. 2013; Moreau et al. 2020); or (iv) theoretical predictions of the genome formula, to fit models to data and test these predictions. One example from previous work is worthy of mention because the authors used what is effectively the same metric we are proposing: Wu and coworkers used the genome formula distance metric to consider whether there was higher virus accumulation as virus populations approached the mean genome formula value (Wu et al. 2017). A rank correlation was used to test for an association between genome formula distance and accumulation, and the results were significant. Now that we have given some examples of purposes for which reference values can be used in combination with our metric, next we consider one application in detail.

We previously considered whether there were significant differences for the AMV genome formula measured in inoculated leaves (Wu et al. 2017) when the inoculum genome formula is considered for the treatment (see Comparison of the Genome Formula for Different Groups and Appendix A). However, in this instance, one could ask a more specific question: is the genome formula measured in the inoculated leaf more similar to the genome formula of the inoculum than expected by chance? To address this question, we first calculate the mean genome formula distance for each AMV observation to its corresponding inoculum (Wu et al. 2017). Next, we resampled the data by randomly assigning observations to inocula and calculated the mean genome formula distance for a large number of resampled datasets (10^4). We can then compare the observed outcome to the predicted range of genome formula distances for the resampled data to determine its likelihood. This analysis clearly shows that the observed genome formula distance is less than that predicted for the resampled data, showing that there is a clear effect of the inoculum on the genome formula measured in the inoculated leaf (Figure 4, Table 4). The genome formula distance is much smaller than the predicted value for randomized data, showing that the inoculum has a clear effect on the genome formula.

Table 4. Re-analysis of the AMV genome formula data (Wu et al. 2017) with a resampling approach.

Tissue	Genome Formula Distance to Inoculum		Ranking [3]
	Observed [1]	Predicted [2]	
Inoculated leaf	0.400 ± 0.242	0.556 [0.434–0.652]	5
Middle leaf	0.484 ± 0.261	0.494 [0.410–0.568]	3683
Upper leaf	0.530 ± 0.237	0.503 [0.418–0.576]	7919
Rest of plant	0.445 ± 0.245	0.486 [0.421–0.538]	533

[1] The observed value of the mean genome formula distance to the inoculum in the corresponding tissue, with its standard deviation. [2] The predicted value of the mean genome formula distance based on randomized datasets, with its 99% confidence interval. [3] The number of randomized datasets for which the mean genome formula distance was smaller than the observed value, out of 10^4 resampled datasets in total. Ranks < 250 or > 9750 fall outside of the 95% confidence interval, while ranks <50 or >9950 fall outside of the 99% confidence interval.

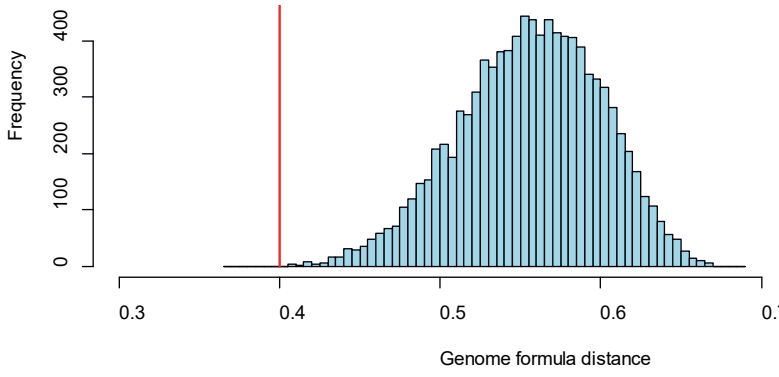


Figure 4. Resampling approach to testing for an effect of inoculum on the genome formula measured in the inoculated leaf. *The blue bars in the histogram indicate the frequency of predicted mean genome formula distance for 10^4 resampled datasets, in which observations in the inoculated leaf were randomly assigned to an inoculum. The red line indicates the genome formula distance for the actual data.*

This result appears to contradict the PERMANOVA test results on the same data, in which there was not a significant treatment effect. However, these two procedures address different questions and test different null hypotheses. Rather than considering whether there is an effect of treatment on the mean, here, we are asking whether means are closer to a reference corresponding to each treatment. The resampling test we have used in this section incorporates more information from the experimental setup, resulting in a specific null hypothesis that can be more readily rejected.

Finally, we can perform the same resampling procedure for other tissues analyzed in the same experiment, in which case we do not see an effect in any other tissue (Table 4 and Appendix B). Therefore, the effect of the inoculum on the genome formula appears to be transient, as this effect is absent in systemically infected tissues. In summary, by reanalyzing these data, we do find strong evidence for an effect of the inoculum: in the inoculated leaf alone, the genome formula is closer to the inoculum genome formula than would be expected by chance.

Discussion

In the past decade, there has been considerable interest in the genome formula of both multipartite and segmented viruses (Sicard et al. 2016; Michalakis and Blanc 2020; Sicard et al. 2013; Diefenbacher, Sun, and Brooke 2018; Wu et al. 2017; Zwart and Elena 2020; Di Mattia et al. 2022; Leeks et al. 2023). However, different studies have applied different analysis methods, many of which have serious shortcomings. To address this challenge and provide examples, here, we present some simple and robust approaches to analyzing genome formula data. Our approach is based on the genome formula distance metric, the Euclidean distance between two genome formula values. We demonstrated the properties of this metric and showed how it can be applied to different analyses. By reanalyzing previously published datasets, we showed that in some cases, the approach used matters for the outcome, in support of our expectation. The genome formula distance is amenable to formal analysis by simple and robust approaches such as PERMANOVA, using existing software packages such as the vegan package for community ecology in R (Oksanen et al. 2022).

We argue that permutational analyses based on the genome formula distance are superior to other approaches used to analyze genome formulae, primarily because the assumptions of the statistical test are met with this procedure. Many of the procedures used previously by others and ourselves do not meet these assumptions, with one common violation being the assumption of independence when relative frequencies are analyzed as independent measurements. The procedures we propose here avoid this problem by reducing relative frequencies to a single distance measurement. Ultimately, the main benefit of the procedures we are proposing is greater robustness and, consequently, validity, irrespective of test performance. Nevertheless, in two cases, this procedure found differences where other procedures did not find any, suggesting that the statistical power of these procedures is not lower.

Most of our reanalysis yielded similar results to the original study. For the work of Wu and coworkers (Wu et al. 2017), our initial re-analysis of the inoculated leaf contradicts the study's results, whereas our subsequent re-sampling analysis determined a clear effect of the inoculum on the genome formula in the inoculated leaf. By inference, there are, therefore, some differences between plants due to the inoculum, in agreement with the studies' conclusions. The different test results for the PERMANOVA and re-sampling based approaches are logically compatible given the different null hypotheses being evaluated, and they illustrate the importance of carefully considering which hypothesis to test. Ultimately, the results convincingly show a clear legacy of the inoculum genome formula in the inoculated leaf. What could explain this outcome? It cannot be categorically ruled out that the in vitro synthesized inoculum has an effect, although this is highly unlikely given the instability of RNA

under ambient conditions. The most likely explanation is, therefore, that insufficient generations of virus replication occurred for a frequency-dependent selection to alter the genome formula. Major changes in the genome formula might also be more likely to occur upon systemic movements of multipartite viruses, especially if these are associated with low multiplicities of cellular infection (MOI) that are predicted to facilitate rapid changes when using a theoretical model (Zwart and Elena 2020). What is exciting about this new result is that it shows that the genome formula can be transmissible, as this is an essential ingredient for its hypothesized role in virus adaptation to changing host environments (Sicard et al. 2016, 2013; Zwart and Elena 2020).

Alternative Metrics for Analyzing Genome Formula Data

In their landmark study on the FBNSV genome formula, Sicard et al. and coworkers (Sicard et al. 2013) proposed *ΔGF* as a metric, which is expressed in general terms as follows:

$$\Delta GF_{a,b} = \sum_{i=1}^{k} \left| f_{a,i} - f_{b,i} \right| \Big/ 2 \qquad (4)$$

This metric has been used for quantifying the imbalance in the genome formula (e.g., comparing empirical values to a balanced genome formula) (Sicard et al. 2013; Moreau et al. 2020). Given that we advocate reducing multivariate data to a single distance measurement and then using permutational statistics, *ΔGF* also could be used instead of the genome formula distance *D* and often yield similar results. We chose the genome formula distance metric mainly because it provides the simplest and most intuitive representation of the distance between two data points in an *n*-dimensional space, i.e., a straight line. Another advantage may be that squaring differences will more heavily weigh larger distances. Ultimately, both approaches are reasonable, and the effect on the results of analysis may often be small. To facilitate the interpretation of analyses based on the *ΔGF* metric, we also calculated expected values of genome formula variation for a random accumulation of segments and under maximum genome formula drift (Appendix C).

Caveats

The approaches we propose have some important benefits, but it is important to keep in mind some limitations. First, when samples have significant differences in genome formula spread (i.e., as indicated by the PERMDISP2 procedure), no firm conclusions can be reached on differences in mean using PERMANOVA. Significant differences in spread between treatments can also be interesting in their own right. For example, Boezen and co-workers used this procedure to show that mixed infections restricted genome formula variation (Boezen, Vermeulen, et al. 2023). However, if there is not a framework to interpret whether differences in spread are relevant, this outcome may not be very informative. Second, in some cases multipartite viruses can lose or gain genome segments that are not essential for replication (Di Mattia et al. 2022). The approaches we propose can handle such data, as segments can have a relative frequency of zero. However, when segments are missing altogether, we suggest considering other approaches for analysis. For example, essential

FBNSV segments (e.g., R and S) are typically present at low frequencies ($f < 0.05$). Their complete absence would have a minimal effect on the hypothetical GF distance, but result in virus populations incapable of replication. Third, methods used for the quantification of the genome formula can have an effect on the results, as shown previously (Boezen, Johnson, et al. 2023) and confirmed by our re-analysis here (Appendix A). The analysis of results obtained with different methods clearly should be avoided. However, as the genome formula quantification method could induce different amounts of technical variation, a comparison of indexes like genome formula distance ($\bar{D}_{a,b}$) obtained with different methods should also be avoided.

Concluding Remarks

Genome formula data can have a large number of dimensions, complicating their visualization, analysis, and, ultimately, the interpretation of results. The visualization of these data can be aided with the use of ternary plots or radar charts, whereas, here, we explore new approaches to the analysis. We show that the genome formula distance metric can be used for a number of different purposes, ranging from comparisons between experimental treatments to comparing data and theoretical expectations. One major advantage of these approaches is their simplicity and reliance on well-established statistical tests, such as PERMANOVA. However, other developments suggest future directions for analyzing these kinds of datasets. First, ecological communities, such as microbiomes, often have high species richness. Advanced approaches for analyzing the relative frequency of taxonomic units (Warton et al. 2015) could serve as inspiration for how to refine methods for genome formula analysis. Second, machine learning and deep learning algorithms (Pichler and Hartig 2023) may prove to be valuable for analyzing genome formula data, as these tools may identify trends that are difficult to visualize and may not be identified by testing hypotheses specified a priori.

Appendix A: Results for the Comparison of the Genome Formula for Different Groups

In this appendix, we describe in detail the results summarized in the results section, Comparison of the genome formula for different groups. To illustrate how this procedure can be used to address different questions, here, we consider some examples of comparisons of the genome formula for different groups.

First, we consider our previous work, which measured the genome formula with four different methods in three hosts (Boezen, Johnson, et al. 2023). Model selection suggested that only the method used had a significant effect on the genome formula. To re-analyze these data, we ran a PERMANOVA on the genome formula distance, including host and method as factors. We found significant effects for the method ($F_{1,44}$ = 12.174, p < 0.0001) and host species ($F_{1,44}$ = 9.746, p = 0.001) on the genome formula. The PERMDISP2 procedure does not show significant effects ($F_{11,36}$ = 2.073, p = 0.051). This reanalysis, therefore, confirms a clear effect of quantification method on the genome formula. However, there was also an effect of host in the new analysis, and differences in spread (PERMDISP2) were nearly significant.

We, therefore, looked in more detail at the results by performing one-way PERMANOVA for each host and method separately, as well as the corresponding PERMDISP2 tests (Table A1). These analyses revealed a significant effect of method on the mean in *C. quinoa* only, suggesting the effects of quantification method are strongest in this host. By contrast, a significant effect of the host was found only for one method (RT-dPCR), showing the methods do not agree on a host-species effect. Overall, this new analysis, therefore, confirms that there are biases in genome-formula quantification methods, while suggesting these effects manifest in one host species. As the methods do not agree on a host-species effect on the genome formula, we cannot draw clear conclusions on this effect. However, three out of four methods suggest that there is not a clear effect, suggesting that for this panel of host species, CMV does not show differences in the genome formula. Results from the original (Boezen, Johnson, et al. 2023) and new analysis are, therefore, congruent.

Second, we re-analyzed data from another study that measured the AMV genome formula (Wu et al. 2017). This study showed striking effects of host species on the genome formula while arguing that the genome formula converges on a host-species dependent equilibrium. Here, we considered the data showing convergence on an equilibrium in more detail. In the original study, the ratio of AMV RNAs was varied in the inoculum, and the genome formula was then measured in different tissues in inoculated plants. Here, we compared the genome formula in inoculated leaves. This simplifies the analysis and allowed us to consider the condition in which the genome formula is most likely to have carried over from the inoculum. The genome formula will most likely carry over to the inoculated leaf as the virus has not moved systematically, incurring additional bottleneck events and opportunities for directional forces to act on the genome formula (i.e., selection). We found an insignificant effect of the inoculum on the genome formula distance with PERMANOVA ($F_{1,17}$ = 0.991, p = 0.344) and PERMDISP2 ($F_{6,12}$ = 0.520, p = 0.812). Both the mean and spread of the genome formula, therefore, appear to be similar across plants treated with a different inoculum genome formula.

Next, we reanalyzed data from work on FBNSV by Sicard and coworkers (Sicard et al. 2013). There are two datasets of interest in this work. The genome formula was measured in different leaf levels, showing a drop in genome formula variability with leaf level as described in Figure 3a in the original study (Sicard et al. 2013), and as confirmed by our re-analysis here (see Results section Comparison of the Genome Formula to Theoretical Values and Table 3). When we reanalyzed these data to look for differences in the genome formula distance between leaf levels, we obtained a significant result for both PERMANOVA ($F_{1,75}$ = 4.472, p = 0.002) and PERMDISP2 ($F_{5,71}$ = 3.241, p = 0.010). These results confirm the differences in genome formula variation, while we cannot draw conclusions on whether the mean genome formula changes over leaf levels.

Finally, we compared a second dataset presented by Sicard and coworkers (Sicard et al. 2013). Here, the authors compared FBNSV genome formula measurements in different hosts, as shown in Figure 2b in the original study (Sicard et al. 2013). For simplicity, we restricted our analysis to plants inoculated with viruliferous aphids and excluded the (aggregated) data from agro-inoculated plants. First, we analyzed each experiment as a separate treatment to look for overall effects and found a highly significant result for PERMANOVA ($F_{1,71}$ = 40.946, p < 0.0001) and an insignificant result for PERMDISP2 ($F_{4,68}$ = 2.082, p = 0.088). Therefore, as there are no significant differences in spread as indicated by the PERMDISP2 results, we can conclude there is a significant difference in the mean. Next, we performed pairwise comparisons between experiments to establish which differ significantly (Table A2). Here, we found no significant differences for the PERMDISP2 procedure, whilst all the results from the two different hosts were significantly different for PERMANOVA. This result demonstrates that differences between experiments are due to a host species' effect on the genome formula.

Table A1. PERMANOVA and PERMDISP2 test results for genome formula observations in three hosts using four quantification methods, analyzed separately per host and method.

Data Included in Analysis	*PERMANOVA*		*PERMDISP2*	
	F (d.f.)	*P*	*F* (d.f.)	*P*
C. quinoa, all methods	9.523 (1,14)	0.007 **	2.293 (3,12)	0.069
N. tabacum, all methods	3.105 (1,14)	0.072	2.144 (3,12)	0.148
N. benthamiana, all methods	2.342 (1,14)	0.126	0.622 (3,12)	0.598
RT-qPCR, all host species	1.723 (1,10)	0.208	1.900 (2,9)	0.205
RT-dPCR, all host species	7.187 (1,10)	0.007 **	0.671 (2,9)	0.538
Illumina, all host species	3.242 (1,10)	0.101	12.65 (2,9)	<0.001 ***
Nanopore, all host species	3.632 (1,10)	0.072	2.988 (2,9)	0.105

** Significant at p < 0.01, *** Significant at p < 0.001.

Table A2. PERMANOVA and PERMDISP2 test results for the pairwise comparison of the FBNSV genome formula distance for five experiments in two host species (*Vicia faba* and *Medicago truncatula*). Cells below the diagonal give the PERMANOVA result, while cells above the diagnonal give the PERMDISP2 results. A Holm-Bonferroni correction was made to the threshold for significance, and all statistically significant results are marked (*). All statistically significant results were below a threshold value of 0.001, after Holm-Bonferroni correction.

		Experiment				
		V. faba 1	*V. faba* 2	*V. faba* 3	*M. truncatula* 1	*M. truncatula* 2
Experiment	*V. faba* 1		$F_{1,14} = 0.593$ $p = 0.483$	$F_{1,40} = 3.525$ $p = 0.062$	$F_{1,21} = 0.185$ $p = 0.679$	$F_{1,X} = 0.260$ $p = 0.712$
	V. faba 2	$F_{1,14} = 4.397$ $p = 0.011$		$F_{1,36} = 1.124$ $p = 0.297$	$F_{1,16} = 2.130$ $p = 0.170$	$F_{1,17} < 0.001$ $p = 0.985$
	V. faba 3	$F_{1,40} = 1.659$ $p = 0.164$	$F_{1,36} = 3.735$ $p = 0.013$		$F_{1,42} = 5.631$ $p = 0.021$	$F_{1,43} = 1.558$ $p = 0.227$
	M. truncatula 1	$F_{1,21} = 73.68$ $p < 0.0001$ *	$F_{1,16} = 52.959$ $p < 0.0001$ *	$F_{1,42} = 44.458$ $p < 0.0001$ *		$F_{1,24} = 0.679$ $p = 0.518$
	M. truncatula 2	$F_{1,X} = 40.926$ $p < 0.0001$ *	$F_{1,17} = 28.968$ $p = 0.0001$ *	$F_{1,43} = 35.289$ $p < 0.0001$ *	$F_{1,24} = 2.006$ $p = 0.116$	

Appendix B: Results for the Comparison of the Genome Formula to A Reference

Figure A1 provides the results for the resampling of genome formula distance values, as compared to the inoculum value, for other tissues in plants infected with AMV as described in results section Comparison of the Genome Formula to reference (see also Figure 4 and Table 4).

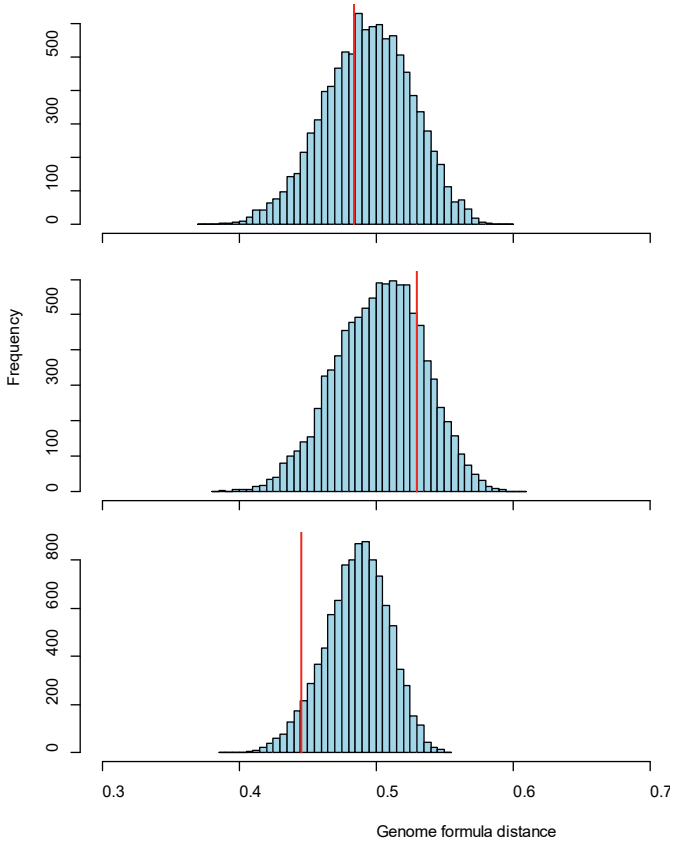


Figure A1. Resampling approach for testing for an effect of inoculum on the AMV genome formula measured in different tissues*. The blue bars in the histogram indicate the frequency of predicted mean genome formula distance for 10^4 resampled datasets, in which observations in the inoculated leaf were randomly assigned to an inoculum. The red line indicates the genome formula distance for the actual data, which in all cases falls well within the 99% confidence interval of the distribution predicted by resampling (see Table 4). (**a**) Results for the middle leaf of the plant are shown. (**b**) Results for the upper leaf are shown. (**c**) Results for the rest of the plant tissues are shown.*

Appendix C: Predicted Properties of the ΔGF Metric

For the genome formula distance ($D_{a,b}$), we predicted the variation under random accumulation of segments ($\bar{D}_{a,b}^{rand}$, Results section Distance metric for Random Genome Formula Variation) and the maximum variation under genome formula drift by a single bottleneck event ($\bar{D}_{a,b}^{drift}$, Results section Distance metric for Maximum Genome Formula Drift). These same predictions can be made for the ΔGF metric (Table A3), to help provide some context for observed values of the mean pairwise ΔGF ($\Delta GF_{a,b}$). Compared to $D_{a,b}$, there are differences in the absolute values and for random accumulation. The trend is also different, as it increases with the number of segments whereas $\bar{D}_{a,b}^{rand}$ decreases.

Table A3. **Expected values of $\Delta GF_{a,b}$ for random genome formula variation ($\overline{\Delta GF}_{a,b}^{rand}$) or the maximum genome formula drift introduced by a single bottleneck event ($\overline{\Delta GF}_{a,b}^{drift}$).**

Number of Genome Segments	$\overline{\Delta GF}_{a,b}^{rand}$	$\overline{\Delta GF}_{a,b}^{drift}$	λ [1]
2	0.2726	0.2034	5.37
3	0.3046	0.1981	7.08
4	0.3157	0.1850	9.33
5	0.3211	0.1742	10.96
6	0.3241	0.1585	13.49
7	0.3260	0.1493	15.14
8	0.3274	0.1411	16.98
9	0.3285	0.1324	19.05
10	0.3291	0.1236	21.88

[1] The bottleneck value corresponding to the maximum $\overline{\Delta GF}_{a,b}^{drift}$ value.

References

Anderson, Marti J. 2001. “A New Method for Non-parametric Multivariate Analysis of Variance.” *Austral Ecology* 26 (1): 32–46.

———. 2017. “Permutational Multivariate Analysis of Variance (PERMANOVA).” In *Wiley StatsRef: Statistics Reference Online*, 1–15. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat07841.

Bermúdez-Méndez, Erick, Kirsten F. Bronsvoort, Mark P. Zwart, Sandra van de Water, Ingrid Cárdenas-Rey, Rianka P. M. Vloet, Constantianus J. M. Koenraadt, Gorben P. Pijlman, Jeroen Kortekaas, and Paul J. Wichgers Schreur. 2022. “Incomplete Bunyavirus Particles Can Cooperatively Support Virus Infection and Spread.” *PLoS Biology* 20 (11): e3001870.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064.

Boezen, Dieke, Maritta Vermeulen, Marcelle Johnson, Rene Van Der Vlugt, Carolyn Malmstrom, and Mark Zwart. 2023. “Mixed Viral Infection Constrains the Genome Formula of Multipartite Cucumber Mosaic Virus.” *Frontiers in Virology* 3. https://doi.org/10.3389/fviro.2023.1225818.

Di Mattia, Jérémy, Babil Torralba, Michel Yvon, Jean-Louis Zeddam, Stéphane Blanc, and Yannis Michalakis. 2022. “Nonconcomitant Host-to-Host Transmission of Multipartite Virus Genome Segments May Lead to Complete Genome Reconstitution.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (32): e2201453119.

Diefenbacher, Meghan, Jiayi Sun, and Christopher B. Brooke. 2018. “The Parts Are Greater than the Whole: The Role of Semi-Infectious Particles in Influenza A Virus Biology.” *Current Opinion in Virology* 33 (December): 42–46.

Fulton, Robert W. 1962. “The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus.” *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.

Gallet, Romain, Frédéric Fabre, Yannis Michalakis, Stéphane Blanc, and Anne E. Simon. 2017. “The Number of Target Molecules of the Amplification Step Limits Accuracy and Sensitivity in Ultradeep-Sequencing Viral Population Studies.” *Journal of Virology* 91: 561–78.

Gutiérrez, Serafín, and Mark P. Zwart. 2018. “Population Bottlenecks in Multicomponent Viruses: First Forays into the Uncharted Territory of Genome-Formula Drift.” *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2018.09.001.

Hajimorad, M. R., G. Kurath, J. W. Randles, and R. I. Francki. 1991. “Change in Phenotype and Encapsidated RNA Segments of an Isolate of Alfalfa Mosaic Virus: An Influence of Host Passage.” *The Journal of General Virology* 72 (Pt 12) (December): 2885–93.

Hu, Zhaoyang, Xiaolong Zhang, Wei Liu, Qian Zhou, Qing Zhang, Guohui Li, and Qin Yao. 2016. “Genome Segments Accumulate with Different Frequencies in Bombyx Mori Bidensovirus.” *Journal of Basic Microbiology* 56 (12): 1338–43.

Jacobs, Nathan T., Nina O. Onuoha, Alice Antia, John Steel, Rustom Antia, and Anice C. Lowen. 2019. “Incomplete Influenza A Virus Genomes Occur Frequently but Are

Readily Complemented during Localized Viral Spread." *Nature Communications* 10 (1): 3526.

Kennedy, George G., William Sharpee, Alana L. Jacobson, Mary Wambugu, Benard Mware, and Linda Hanley-Bowdoin. 2023. "Genome Segment Ratios Change during Whitefly Transmission of Two Bipartite Cassava Mosaic Begomoviruses." *Scientific Reports* 13 (1): 10059.

Kormelink, R., P. de Haan, D. Peters, and R. Goldbach. 1992. "Viral RNA Synthesis in Tomato Spotted Wilt Virus-Infected Nicotiana Rustica Plants." *The Journal of General Virology* 73 (Pt 3) (March): 687–93.

Ladner, Jason T., Michael R. Wiley, Brett Beitzel, Albert J. Auguste, Alan P. Dupuis, Michael E. Lindquist, Samuel D. Sibley, et al. 2016. "A Multicomponent Animal Virus Isolated from Mosquitoes." *Cell Host & Microbe* 20 (3): 357–67.

Lamy-Besnier, Quentin, Bryan Brancotte, Hervé Ménager, and Laurent Debarbieux. 2021. "Viral Host Range Database, an Online Tool for Recording, Analyzing and Disseminating Virus-Host Interactions." *Bioinformatics* 37 (17): 2798–2801.

Leeks, Asher, Penny Grace Young, Paul Eugene Turner, Geoff Wild, and Stuart Andrew West. 2023. "Cheating Leads to the Evolution of Multipartite Viruses." *PLoS Biology* 21 (4): e3002092.

Lucía-Sanz, Adriana, and Susanna Manrubia. 2017. "Multipartite Viruses: Adaptive Trick or Evolutionary Treat?" *Npj Systems Biology and Applications* 3 (1): 34.

Mansourpour, Mahsa, Romain Gallet, Alireza Abbasi, Stephane Blanc, Akbar Dizadji, and Jean-Louis Zeddam. 2021. "Effects of an Alphasatellite on Life Cycle of the Nanovirus Faba Bean Necrotic Yellows Virus." *Journal of Virology*, November, JVI0138821.

Michalakis, Yannis, and Stéphane Blanc. 2020. "The Curious Strategy of Multipartite Viruses." *Annual Review of Virology* 7 (1): 203–18.

Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. "The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner." *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.

Moury, Benoît, Frédéric Fabre, Eugénie Hébrard, and Rémy Froissart. 2017. "Determinants of Host Species Range in Plant Viruses." *The Journal of General Virology* 98 (4): 862–73.

Obrępalska-Stęplowska, Aleksandra, Jenny Renaut, Sebastien Planchon, Arnika Przybylska, Przemysław Wieczorek, Jakub Barylski, and Peter Palukaitis. 2015. "Effect of Temperature on the Pathogenesis, Accumulation of Viral and Satellite RNAs and on Plant Proteome in Peanut Stunt Virus and Satellite RNA-Infected Plants." *Frontiers in Plant Science* 6 (October). https://doi.org/10.3389/fpls.2015.00903.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'hara, et al. 2022. *Vegan: Community Ecology Package* (version 2.6.4). Vegan: Community Ecology Package. https://cran.r-project.org/web/packages/vegan/index.html.

Pichler, Maximilian, and Florian Hartig. 2023. "Machine Learning and Deep Learning—A Review for Ecologists." *Methods in Ecology and Evolution / British Ecological Society* 14 (4): 994–1016.

R Foundation for Statistical Computing. 2023. *R: A Language and Environment for Statistical Computing* (version 4.3.1). Vienna, Austria. https://www.r-project.org/.

Roossinck, M. J. 2001. “Cucumber Mosaic Virus, a Model for RNA Virus Evolution.” *Molecular Plant Pathology* 2 (2): 59–63.

Rybicki, Edward P. 2015. “A Top Ten List for Economically Important Plant Viruses.” *Archives of Virology* 160 (1): 17–20.

Sánchez-Navarro, Jesús A., Mark P. Zwart, and Santiago F. Elena. 2013. “Effects of the Number of Genome Segments on Primary and Systemic Infections with a Multipartite Plant RNA Virus.” *Journal of Virology* 87 (19): 10805–15.

Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. “The Strange Lifestyle of Multipartite Viruses.” Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819.

Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. “Gene Copy Number Is Differentially Regulated in a Multipartite Virus.” *Nature Communications* 4: 2248.

Valdano, Eugenio, Susanna Manrubia, Sergio Gómez, and Alex Arenas. 2019. “Endemicity and Prevalence of Multipartite Viruses under Heterogeneous Between-Host Transmission.” *PLoS Computational Biology* 15 (3): e1006876.

Warton, David I., F. Guillaume Blanchet, Robert B. O’Hara, Otso Ovaskainen, Sara Taskinen, Steven C. Walker, and Francis K. C. Hui. 2015. “So Many Variables: Joint Modeling in Community Ecology.” *Trends in Ecology & Evolution* 30 (12): 766–79.

Wichgers Schreur, Paul J., Richard Kormelink, and Jeroen Kortekaas. 2018. “Genome Packaging of the Bunyavirales.” *Current Opinion in Virology* 33 (December): 151–55.

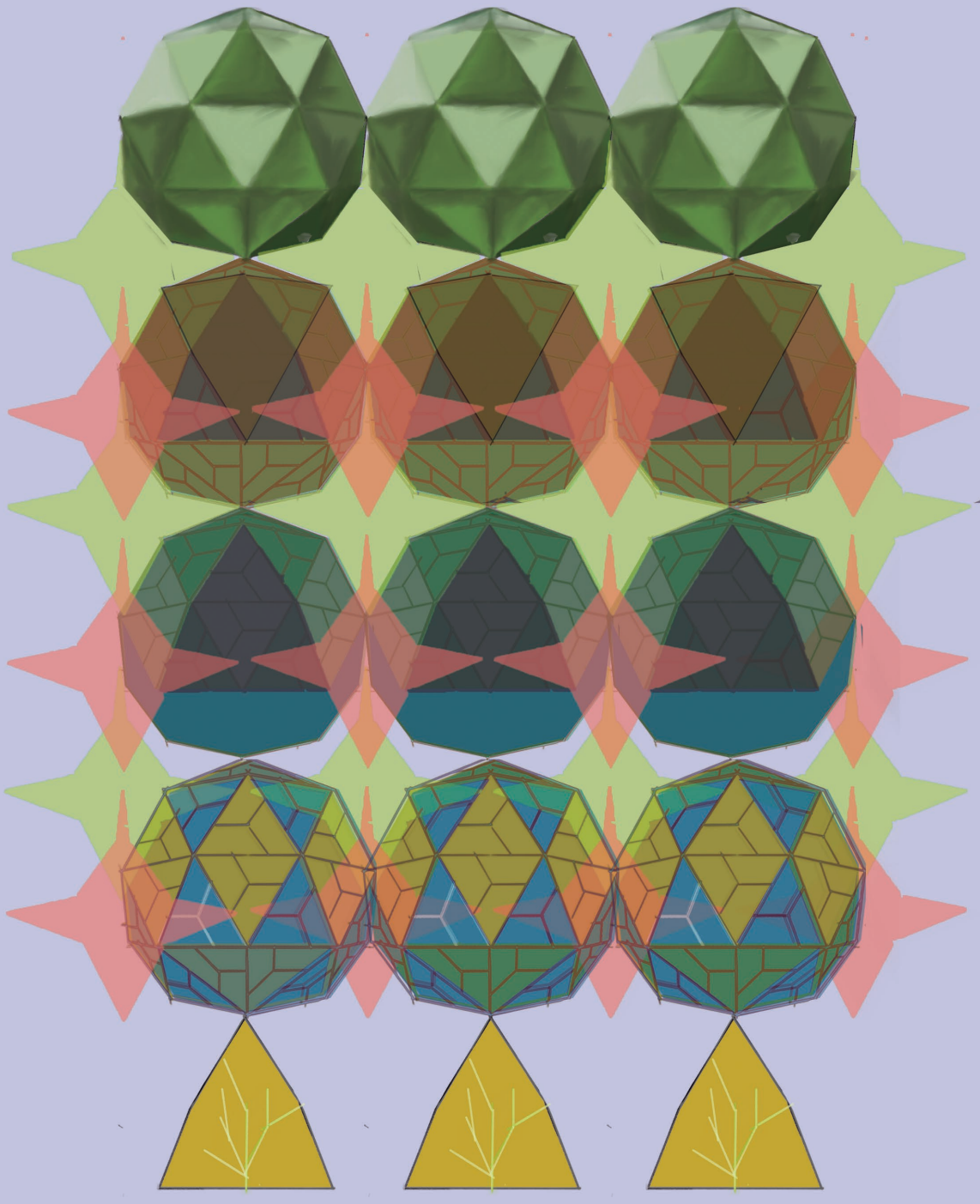
Wichgers Schreur, Paul J., and Jeroen Kortekaas. 2016. “Single-Molecule FISH Reveals Non-Selective Packaging of Rift Valley Fever Virus Genome Segments.” *PLoS Pathogens* 12 (8): e1005800.

Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. “Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus.” *Scientific Reports* 7 (1): 1–15.

Yvon, Michel, Thomas L. German, Diane E. Ullman, Ranjit Dasgupta, Maxwell H. Parker, Sulley Ben-Mahmoud, Eric Verdin, et al. 2023. “The Genome of a Bunyavirus Cannot Be Defined at the Level of the Viral Particle but Only at the Scale of the Viral Population.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (48): e2309412120.

Zwart, Mark P., Stéphane Blanc, Marcelle Johnson, Susanna Manrubia, Yannis Michalakis, and Mircea T. Sofonea. 2021. “Unresolved Advantages of Multipartitism in Spatially Structured Environments.” *Virus Evolution* 7 (1): veab004.

Zwart, Mark P., and Santiago F. Elena. 2020. “Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments†.” *Virus Evolution* 6 (1): veaa022.



Genome formula variation in local infections of a multipartite virus

Marcelle L. Johnson[1,2], Elisa Antonia Miranda Vasquez[1,2], René A.A. van der Vlugt[2], J. Arjan G.M. de Visser[3], Mark P. Zwart[1]

[1] Netherlands Institute of Ecology (NIOO-KNAW), P.O. Box 50, 6700 AB, Wageningen, The Netherlands

[2] Laboratory of Virology, Wageningen University and Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

[3] Laboratory of Genetics, Wageningen University and Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

Abstract

Multipartite viruses have segmented genomes that are individually packaged and transmitted. This transmission of individual segments permits considerable variation in the genome formula (GF), i.e. the relative frequencies of virus genome segments. However, the amount of GF variation and its biological significance have not been studied in detail. In this study, we investigate GF variation in local lesions, i.e. infections of a limited number of adjacent plant cells. Using the tripartite RNA cucumber mosaic virus (CMV), we quantify the GF in individual local lesions occurring in *Chenopodium quinoa*, leading to three main insights. First, the GF is highly variable between local lesions, and the observed variation was similar to model predictions in which stochastic processes drive GF variation. Second, the GF in local lesions depends on the inoculum GF, highlighting that despite its variability, it is a transmissible property. Third, the measured GF values were multimodal, and we identified a cluster of GF values associated with lower virus accumulation. Our results demonstrate the importance of stochastic forces in shaping the CMV GF, while also showing that GF variation is related to differences in virus fitness in *C. quinoa*. While the GF may be beneficial for rapidly tuning viral gene expression, our results also highlight the risks associated with GF variation.

Introduction

Multipartite viruses have segmented genomes, packaging and transmitting each segment independently (Sicard et al. 2016). The number of segments can range from two to nine segments per species (Lucía-Sanz, Aguirre, and Manrubia 2018). Each segment is mono- or bicistronic, encoding viral components including the replication machinery, coat protein, movement protein and suppressors of RNA interference. The requirement for transmission by multiple virus particles will often be associated with a cost, as genome integrity can be compromised (Sicard et al. 2016, Sanchez-Navarro et al. 2013). In contrast, the independent transmission of segments increases opportunities for genetic exchange via segment reassortment (Varsani et al. 2018), although in practice, reassortment is rare in the few studies that consider it (Ohshima et al. 2016; Fraile et al. 1997). Another proposed benefit associated with multipartition stems from the unbalanced accumulation of genome segments typically seen during infection. The frequencies of all genome segments are termed the genome formula (GF) (Sicard et al. 2013). The GF converges to a host-specific equilibrium termed the setpoint genome formula (SGF) in all multipartite viruses for which this property has been studied (Michalakis and Blanc 2020), due to frequency-dependent selection (Sicard et al. 2013; Wu et al. 2017). If virus gene copy number affects viral fitness, the GF may affect the stoichiometry of virus gene products and virus replication (Sicard et al. 2013, 2016). It has also been suggested that for faba bean necrotic stunt virus (FBNSV), differences in the GF may stabilise gene expression, as transcript levels are more stable than genome segments (Gallet et al. 2022). However, despite its rapid convergence on the SGF, the GF shows considerable variation between individual plants (Sicard et al. 2013; Wu et al. 2017). The GF may confer benefits to multipartite viruses due to its propensity for rapid change, but its variability could also be disadvantageous for multipartite viruses.

Both stochastic and directional forces are likely to act on the GF. The main driver of stochastic changes in the GF is likely to be narrow population bottlenecks, which occur for viruses as they spread between cells, organs and hosts (Miyashita et al. 2015; Betancourt et al. 2008). These population bottlenecks will result in genetic drift: stochastic changes in allele frequencies. If genome segments spread independently, for multipartite viruses there will also be genome formula drift, which is a stochastic variation in the frequencies of the different segments (Gutiérrez and Zwart 2018). Whereas genetic drift increases as the bottleneck narrows, GF drift is predicted to be strongest at intermediate-size bottlenecks due to the requirement for a complete set of genome segments (Zwart and Elena 2020). The main driver of directional change in the GF will likely be selection for efficient replication and within-host spread, as the GF might affect viral gene expression. Selection could act on the GF, pulling it towards advantageous values for a given environment. The relative strengths of these stochastic and directional forces will shape the observed GF variation between replicate populations.

One way to change the strength of stochastic forces acting on a virus population is to change the size of the host cell population in which the virus is replicating. Infections in a small number of host cells may (1) follow from the initial infection of a single cell, providing a narrow population bottleneck, and (2) provide a limited number of virus generations in which selection can act to change the GF. Due to virus-host interactions and the resulting patterns of within-host spread, for plant viruses the number of infected cells can be minimal under some

conditions, affording opportunities to study the GF under conditions that lead to high levels of variation between different infection loci. Spread of plant viruses within hosts follows two main routes: (1) local cell-to-cell movement, in which the virus moves to adjacent cells, often followed by (2) long-range movement leading to systemic infection, where the virus spreads through the vasculature. In some instances, strong host defences restrict infection to a limited number of cells, resulting in the formation of local lesions. In these cases, plant defence is characterised by the onset of the generalised pathogen defence, the hypersensitive response (HR) (Lam, Kato, and Lawton 2001). The HR is characterized by localized programmed cell death (PCD), a host-mediated response to infection in which cells at or near the initial infection site undergo apoptosis and form visible necrotic spots (i.e., local lesions) to limit virus movement (Coll, Epple, and Dangl 2011). The activation of salicylic and jasmonate pathways and the presence of reactive oxygen species of these cascades initiates PCD (Soosaar, Burch-Smith, and Dinesh-Kumar 2005; Coll, Epple, and Dangl 2011). Opportunities for cell-to-cell movement via plasmodesmata are severely limited, impeding the virus from reaching the vascular tissue and often preventing systemic infection (Ross 1961; Jacob, Hige, and Dangl 2023). In some cases, virus spread can occur outside the HR necrotic zone, as observed for tobacco mosaic virus (TMV) (Wright et al. 2000) and Potato virus X (Lukan et al. 2018). Local lesions caused by plant virus infection were first described for TMV in several *Nicotiana* species and recognized as the sites of primary infection and necrosis (Holmes 1929). *Chenopodium quinoa* is a local lesion host for many plant viruses (Cooper 2001). As local lesions are readily visible, they have been exploited to isolate single virus genotypes by removing and propagating the virus from an individual lesion. Therefore, local lesions could also be used to study GF variation during virus replication in a small number of cells, preceded by a population bottleneck.

To investigate GF variability in a multipartite virus, we studied CMV local lesions in *C. quinoa*. We measured the GF in local lesions ten days post inoculation (dpi) in three experiments. We quantified GF variation by analysing differences in the mean and spread of observed GF values, quantified the influence of inoculum on the observed GFs and compared empirical GF measurements to model predictions for GF variation. Our results illustrate the high variability of the GF while also highlighting that this variation has implications for virus fitness.

Methods and Materials

Infection of CMV in *C. quinoa*

C. quinoa plants were germinated and grown in gamma sterilised potting soil till three weeks old in growth chamber conditions (21°C, 16/8hrs light/dark cycle) and after that transferred to a climate-controlled greenhouse (22/20°C, 16/8hrs light/dark cycle). Plants were inoculated with CMV subgroup I isolate, CMV-i17f (Jacquemond and Lot 1981) obtained from the plant virus collection at Wageningen Plant Research (www.primediagnostics.com). Three dose-response experiments were conducted with varying doses of CMV-i17f from infected *N. tabacum* and *N. benthamiana* as source inoculum (Table 1). Inoculum stocks were prepared by weighing frozen leaf tissue and homogenised with a pestle and mortar in phosphate inoculation buffer (Roenhorst 2014) to a concentration of 0.25g.ml^{-1} and after that diluted.

Serial dilutions from stocks were prepared, spanning the range 1:5 - 1:1600 (Table 1) over all three experiments. 50ul of inoculum per dilution was applied per leaf onto each of 3 - 6 carborundum-dusted leaves per plant (Table 1), mock plants were inoculated with a phosphate inoculation buffer and after that plants were rinsed with demineralised water. Plants were phenotyped every 2 - 3 days till ten days post infection (dpi) when local lesions were formed. Whole leaves were scanned with an EPSON XPS scanner (800 dpi resolution), lesions were counted manually, and individual lesions were isolated as an Eppendorf disc and immediately placed in liquid nitrogen and stored at -80°C till further analysis.

Table 1. Local lesions were isolated from three dose-response experiments from infections of CMV-i17f derived from frozen *N. tabacum,* and *N. benthamiana* tissue prepared as stocks of 0.25 g.ml^{-1}.

Experiment	Inoculum	Number of inoculated *C. quinoa* plants	Dilution series
1	Mock	2	Stock, 1:25, 1:50. 1:100, 1:400, 1:1600
	N. benthamiana	9	
	N. tabacum	9	
2	Mock	1	Stock, 1:2, 1:4, 1:8, 1:16
	N. tabacum	18	
3	Mock	1	Stock, 1:2, 1:4, 1:8, 1:16
	N. benthamiana	8	

Estimating the GF and virus accumulation in local lesions

RNA extraction and CMV detection

Inocula and *C. quinoa* local lesions were used in RNA extractions (Qiagen RNeasy Plant Mini-Kit) according to manufacturer's instructions and treated with on-column DNASE I (Qiagen). The concentration and quality of RNA were measured on NanoDrop One (ThermoFisher Scientific). 250ng of RNA was converted to cDNA using random hexamers in the iScript cDNA synthesis kit (BioRad) according to the manufacturer's instructions. All *C. quinoa* lesions were tested for CMV infection by RT-PCR targeting RNA1 and RNA3 (Table 2). Reactions were carried out in 25ul volume containing GoTaq G2 DNA polymerase (5U/ul), 5x GoTaq Buffer Green, 10mM dNTP, 25mM MgCl2, 10uM of forward and reverse primers and 2ul of 100x diluted cDNA and nuclease free H_2O. Cycling conditions were as follows: 94°C for 5 min;

94°C for 30 sec, 47°C (RNA1) and 51°C (RNA3) for 30 sec, 72°C for 1 min (33 cycles); 72°C extension for 7 min; followed by 12°C hold. Amplicons were analysed on a 2% agarose gel.

Table 2. RT-PCR to determine the presence of CMV in *C. quinoa* local lesions.

Target	Primer ID	Primer sequence	Amplicon length	Annealing temperature (°C)	Position	Ref
CMV RNA1 (1a)	cmv_rna1a	CYCTGTAAAAYW ACCCTTTG	410	47	38 - 57	This paper
CMV RNA1 (1a)	cmv_rna1as	RTGTGTGACSCA ACTTCC			434 - 451	
CMV RNA3 (CP)	cmv_2f	GCATTCTAGATGG ACAAATCTGAATC	650	51	1248 - 1273	(Vishnoi, Kumar, and Raj 2013)
CMV RNA3 (CP)	cmv_2r	GCATGGTACCTCA AACTGGGAGCAC			1899 - 1923	

RT-qPCR to determine the GF of CMV in local lesions.

To quantify RNA segments, a qPCR was performed with primers targeting RNA1, RNA2 and RNA3 of CMV-i17f (Table 3). A SYBR green (iQ SYBR green, BioRad) qPCR assay was designed to target the three viral RNAs in simplex, with three technical replicates per target. A reaction mix of 8ul consisting of 2x iQ SYBR Green (BioRad) mastermix, 10uM forward and reverse primers, 3ul of template cDNA and nuclease-free H_2O. Cycling conditions were 95°C (3min), 40 cycles: 95°C for 10s, 60°C for 3s and a melt curve at 5°C increments from 65 - 95°C. The $2^{-\Delta\Delta Ct}$ method (Rao et al. 2013) was used to determine the GF relative to RNA1. The GF is calculated as the mean value of the RNA segment relative to the sum of all RNA segments, RNA1 + RNA2 + RNA3 = 1

Virus accumulation was analysed as the mean RT-qPCR cycle quantification (Cq) value of the three genomic CMV RNAs. Cq values are inversely related to template concentration, and these lower Cq values correspond to higher template concentrations and, thus, higher total virus accumulation. Thus, we do not estimate absolute virus accumulation but can robustly infer relative differences in virus accumulation.

Table 3. RT-qPCR for the quantification of GF of CMV in local infections, CMV primers are relative to the reference isolate CMV-Fny accessions (NC_002034; NC_002035; NC_001440).

Target	Primer ID	Primer sequence	Amplicon length	Annealing temperature (°C)	Position	Ref
CMV RNA1 (1a)	cmvrna1a_1f	GCACAACCCGTGAGTGAGG	83	60°C	1706 - 1724	(Boezen, Johnson, et al. 2023)
	cmvrna1a_1r	TCCCTTCCACAAACATCAGCAG			1767 - 1788	
CMV RNA2 (2a)	cmvrna2a_1f	GGTGTTGTTGATAATGCGACTCTG	94	60°C	789 - 812	
	cmvrna2a_1r	CGATGGTTGGCGTTGGACAT			863 - 882	
CMV RNA3 (CP)	cmvrna3a_1f	ACCATGATCTTCCCGCTTTGG	91	60°C	511 - 531	
	cmvrna3a_1r	ACGACAGCAAAACACCGCTT			582 - 601	

Statistical analysis and modelling

All statistical analysis and modelling were performed in R 4.3.1 (R Core Team 2001). GFs are visualised in ternary plots using the ggtern package version 3.4.2 (Hamilton and Ferry 2018).

Genome formula distance (*D*)

For quantitative analysis, we consider the GF as the set of relative frequencies (f) for the complete set of genome segments in the virus genome $\{f_1, f_2, f_3 \ldots, f_j\}$, where for CMV $j = 3$. To analyse genome formula data, we used the genome formula distance (D) as previously described (Boezen, Vermeulen, et al. 2023). This metric is based on the Euclidean distance between two genome formula values, a and b, such that:

$$D_{a,b} = \sqrt{\sum_{i=1}^{j=3} \left(f_{a,j} - f_{b,j}\right)^2}$$

Note that this metric is used for different purposes throughout the analysis. It can be used to determine the pairwise distance between observations (e.g. for PERMANOVA, see (Anderson 2017), to determine the distance to a mean value, or to determine the distance to another value of interest (i.e., the inoculum GF). A detailed description of the genome formula distance metric and many of the test procedures we use here is described in Chapter 3 of this thesis.

Comparison of the GF for experimental treatments

We estimate the pairwise Euclidean distance (*D*) between GF observations using the vegdist function from the VEGAN package version 2.6.4 (Oksanen et al. 2022). A permutational analysis of variance (PERMANOVA) (Anderson 2017) was then performed for these univariate *D* values using the adonis2 function (Oksanen et al. 2022). In addition, the PERMDISP2 test was performed to test for differences in spread, using the betadisper() function. For pairwise comparisons between groups with PERMANOVA and PERMDISP2, a Holm-Bonferroni correction for multiple comparisons was made (Holm 1979).

Test for effect of inoculum on the GF in local lesions

To test for an effect of the inoculum GF on that observed in local lesions, we used a resampling approach as described in Chapter 2 of this thesis. We first determined the mean Euclidean distance between the local lesion GFs and the corresponding inoculum GF. We then randomly assigned each local lesion to an inoculum and recalculated the mean GF distance. The number of local lesions randomly assigned to each inoculum was the same as for the empirical dataset. We repeated this process 10^4 times to obtain the predicted distribution of GF distances for the randomised data. Finally, we could compare the observed GF distance value to the ranked predicted values, and thereby determine the likelihood of the null hypothesis that the inoculum GF did not have an effect on the GF observed in lesions.

Test for multimodality of the GF and determining GF clusters

We used the folding test of unimodality to test whether empirical GF data are multimodal (Siffer et al. 2018). This nonparametric test identifies whether multivariate data have multiple modes by folding the data along a pivot point and testing whether the reductions in variance are significant. The folding ratio is estimated by the difference in variance at the fold point relative to the initial variance. The outputs are the folding statistic (Φ), where $\Phi \geq 1$ indicates a unimodal distribution and $\Phi \leq 1$ indicates a multimodal distribution, and a p-value. Contrary to our other analyses, the complete multivariate GF data (i.e., set of relative frequencies for all genome segments) were analysed, as we cannot work with *D* here. Rfolding package version 1.0 was used.

When the folding test established that GF data were multimodal, we wanted to assign individual local lesions to clusters of similar GF values. For parsimony and based on a visual inspection of the GF data, we assumed there were 2 clusters for this procedure. We established a range of GF values, from which we would determine the two sets of mean GF values and assign individual experimental GF measurements to each. First, we calculated the GF distance *D* to each experimental data point for all these test GF values to be evaluated. Then, for each possible combination of two sets of test GF values, for each GF observation, we considered which test GF value resulted in the lower *D* and assigned the observation to this cluster. Finally, we identified the two test GF values that minimise the sum of *D* for all GF observations, effectively performing a grid search. We then know the two mean GF values and the assignment to clusters of each GF observation. To limit the computational resources needed, we initially performed the grid searches using large (ΔGF = 0.05) step sizes over all possible GF values (i.e., a frequency of 0 to 1 for each of three segments, but only those combinations of frequencies that sum to 1). Based on the results obtained, we then narrowed the range of GF values to be evaluated and ran a grid size with smaller step sizes (ΔGF = 0.01) to obtain a more precise solution.

Results and Discussion

Local lesion genome formulae are diverse and widely distributed

Pilot dose-response experiments of CMV in *C. quinoa* showed that the number of local lesions formed is highly variable within a single experiment. This variation was used as an indicator for the amount of infected tissue to be used as inoculum, without saturating the leaf, so that individual local lesions could be isolated. Previous studies on FBNSV (Sicard et al. 2013) and AMV (Wu et al. 2017) have used infectious clones to alter segment ratios in the inoculum. However, we chose to work with a natural virus isolate and investigate if the natural variation in GF, induced by host species or between-plant variation, had an effect on GF variation in local lesions. The virus isolate we used for these experiments was deep sequenced and shown to contain minimal genetic variation (see Chapter 5), and hence our starting material can be considered largely isogenic. We performed separate three experiments infecting *C. quinoa* with CMV from either frozen 14 days post-infection (dpi) infected *N. benthamiana* or *N. tabacum* leaf tissue (Table 1). Individual local lesions were visually identified, excised and the GF in each lesion was determined (Figure 1, Table 4). We did not observe local lesions upon inoculation with higher dilutions of the virus stock. For experiment 1, a single lesion was observed after inoculation with *N. benthamiana* inoculum and was not included in later analyses. Overall, we assayed 69 lesions from the three experiments.

Table 4. Local lesions assayed in the three experiments. *For experiment 1 *N. benthamiana* inoculum and lesion were excluded from the analysis. Inoculum stock was prepared from 14 dpi CMV-i17F infected *N. tabacum* and *N. benthamiana* plant material and prepared in dilutions from 1:2 – 1:1600. Local lesions were formed for all three experiments on *C. quinoa* leaves infected with 0.25 g.ml^{-1} (stock) and lesions formed for experiment three inoculations from 0.125 g.ml^{-1} and 0.0625 g.ml^{-1}.

Experiment	Inoculum Host	Inoculum concentration (g.ml^{-1})	Number of lesions formed
1	*N. tabacum*	0.25 (Stock)	52
2	*N. tabacum*	0.25 (Stock)	4
3	*N. benthamiana*	0.25 (Stock)	5
		0.125 (1:2)	6
		0.0625 (1:4)	1

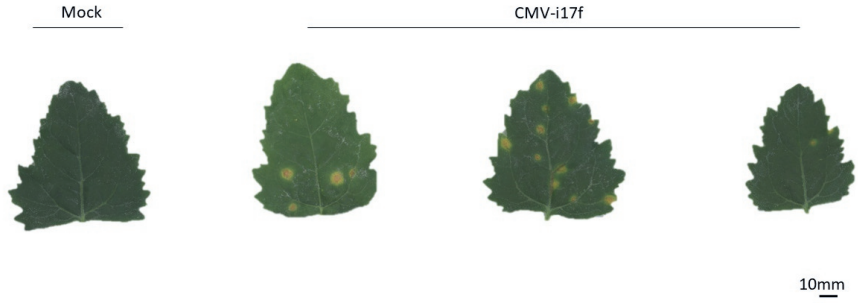


Figure 1. Local lesions on leaves of C. quinoa plant. *Mock-inoculated and leaves inoculated with 0.25 g ml^{-1} of CMV-i17f from N. tabacum: leaves b, d, f. Local lesions are sites of immune hypersensitive response characterised by cell necrosis and the formation of lesions at the sites of primary infection.*

We quantified the GF in the inocula and local lesions from these three experiments (Figure 3). GF values for the inocula used in the three experiments were appreciably different (Table S1): relative frequencies of CMV RNA2 were quite similar, whereas they were more variable for RNA1 and RNA3 (Figure 3). These observations are congruent with previous reports on the CMV-I17F GF in *N. tabacum* and *N. benthamiana,* in which the main variation occurs in RNAs 1 and 3 (Boezen, Johnson, et al. 2023). The inocula of experiments 1 and 3 fall at the two extremes of the range of GF values we have typically observed (Boezen, Johnson, et al. 2023; Boezen, Vermeulen, et al. 2023); Johnson et al. Unpublished data)

We found considerable variation in the GF in local lesions in *C. quinoa*, both within and between the experiments (Figure 3). Two trends in the data are remarkable. First, there is a high level of GF variation across lesions within experiments. For experiment 1, it appears as if there may be two clusters of GF observations, one close to the inoculum GF and a second closer to the GF values observed in experiments 2 and 3 as reported here, and in other studies (Boezen, Johnson, et al. 2023). Second, despite the variation within experiments, there appear to be differences between the experiments. The mean GF was calculated for each experiment (Table S2), and we observe that experiment 1 has higher RNA1 and balanced RNA2 and RNA3, similar to what is observed for the inoculum. In experiments 2 and 3, there are higher levels of RNA3, and this also appears similar to the GFs of the respective inocula (Figure 2, Table S1). Taken together, these results suggest that the observed local lesion GFs reflect their respective inoculum GF. As noted, this effect appears weakest in Experiment 1, where only one of two apparent GF clusters is near the inoculum GF.

Based on previous reports (Sicard et al. 2013), we had not expected to see an effect from inoculum GF and our experimental design is not well-suited to analyse these effects because each inoculum was used in a single experiment, making it difficult to distinguish between treatment (i.e., inoculum) vs. block (i.e., experiment) effects. However, given that we observe GF variation within and between experiments, analysing the experimental data may give us preliminary insights into inoculum-driven GF variation. In the following two sections, we explore whether statistical support exists for the trends we have noted.

Inoculum influences local lesion genome formula variation

We used PERMANOVA to test formally if there is an effect of the experiment on the local lesion GF. We find that there are significant differences in the GF between experiments (PERMANOVA: $F_{2,65} = 5.6128$, $P = 0.0018$). A significant PERMANOVA result can indicate differences in mean, spread or both (Anderson 2001), so we tested whether there were significant differences in spread (Anderson 2006, 2017). The spread between experiments was not significantly different (PERMDISP2 test: $F_{2,65} = 1.8026$, $P = 0.1793$). As this result confirms the GF means between experiments are different, we computed pairwise comparisons to identify which experiments differ significantly. GF means for experiments 1 and 3 are significantly different: $F_{1,62} = 10.484$, $P = 0.0018$, whilst comparisons of experiments 1-2 and 2-3 are not significantly different ($F_{1,54} = 3.7801$, $P = 0.07099$ and $F_{1,14} = 0.903$, $P = 0.386$, respectively). Although the local lesion GFs from experiment 2 are situated roughly between those of experiments 1 and 3 (Figure 3), the small number of lesions will also lower statistical power in these comparisons. Combined, these results show that there are significant differences in local lesion GFs between experiments, suggesting that the local lesion GFs may be influenced by the inoculum. However, each inoculum was tested in a separate experiment, and consequently, we do not consider significant differences between experiments as conclusive evidence.

To consider if the inoculum GF influences the GF in the local lesions, we therefore used a different procedure. We tested whether local lesion GFs were more similar to their respective inoculum GFs than expected by chance in a combined analysis on data from all three experiments that employs a resampling approach (see Methods Section and Chapter 3). We

randomly assigned each local lesion GF observation to an inoculum and measured the GF distance *(D)* between the inoculum GF and respective local lesion GFs (Figure S1). Using 10,000 resampled datasets, we determined that the predicted GF distance had a mean of 0.301, with a 99% confidence interval of 0.268 to 0.334. When comparing the experimental results to the resampled data, we note that only 2 of 10,000 resampled datasets had a lower predicted GF distance value than the observed GF distance of 0.255. These results indicate a strong effect of the inoculum GF on the observed GF in local lesions, because the GF observations in each experiment were closer to the inoculum GF than expected by chance. Therefore, this result suggests that the inoculum GF is an important factor driving the differences between experiments, irrespective of whether there may also be other variation between experiments.

The current results are unexpected as previous reports using infectious clones of FBNSV and AMV focused on frequency-dependent selection towards an equilibrium GF where the inoculum GF is not maintained during systemic infection (Sicard et al. 2013, Wu et al. 2017). However, for AMV there does appear to be an inoculum effect on the GF in the inoculated leaf, based on a re-analysis (Chapter 3 of this thesis) of the original data (Wu et al. 2017). Taken together these results suggest that although GF drift will affect multipartite virus populations subjected to bottlenecks, the GF is transmissible. This conclusion is important because a degree of transmissibility will facilitate adaptive change in the GF (Sicard et al. 2016, Zwart and Elena, 2020).

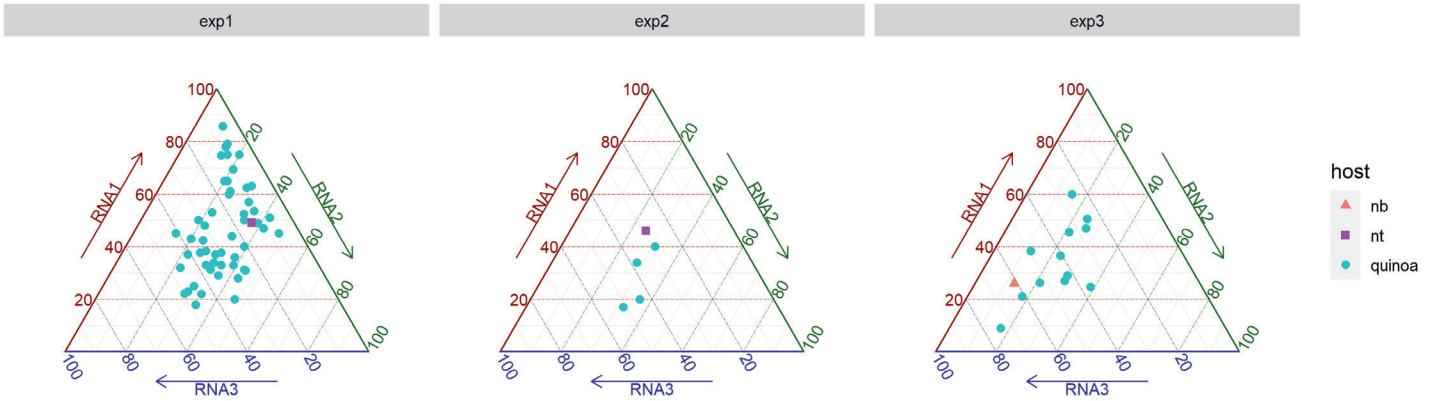


Figure 2. Genome formula variation for CMV in C. quinoa local lesion infections. *C. quinoa plants were inoculated with CMV preparations in three experiments (see Table 4). Experiment 1 was inoculated with CMV from N. tabacum (n = 52 lesions), Experiment 2 inoculum from N. tabacum (n = 4 lesions) and experiment 3 inoculum from N. benthamiana (n = 12 lesions) and monitored for local lesion formation till 10 days post inoculation. Inoculum source and local lesions are differentiated by colour and shape; blue circles indicate the GF for the individual local lesions formed on C. quinoa in experiments 1-3, purple squares indicate N. tabacum inoculum in experiment 1 and 2 and the red triangle indicates the N. benthamiana inoculum in experiment 3.*

Local lesions segregate into two GF clusters with differential virus accumulation.

The large, discontinuous GF spread observed across three experiments prompted us to investigate if the GF data are multimodal. We used the folding test for unimodality (Siffer et al. 2018), a nonparametric test which determines if multivariate data points belong to single (i.e., a unimodal distribution, the null hypothesis) or multiple groups (i.e., a multimodal distribution, the alternative hypothesis). If the data are determined to be multimodel, the the number or identity of groups is not determined. The output is a folding statistic (Φ) which describes the unimodality score (Φ≥1 reflects an unimodal distribution and Φ≤1 a multimodal distribution) and p-value for the significance of the test. We excluded experiment 2 in this analysis due to the small sample size. We find that the data from experiment 1 is multimodal with high statistical significance (Folding test: $\Phi = 0.034$, $P = 0.0001$), whereas the data from experiment 3 is multimodal but only with marginal statistical significance ($\Phi = 0.018$, $P = 0.034$). Thus, there is clear evidence for multimodality in experiment 1 and possibly for experiment 3.

We identify the GF clusters by calculating the GF distance (*D*) to the mean and assigning observations to groups to minimise the sum of *D* for all values (see methods section and Chapter 3 for full description of the GF distance metric). Here, we define clusters as a post hoc, statistically supported grouping of GF data points. In experiment 1, cluster 1 ($n = 30$) is characterised by a more balanced GF than cluster 2 ($n = 22$), which has higher RNA1 and lower RNA3 levels. We also see that the inoculum GF appears closer to observations from cluster 2 (Figure 4). This pattern suggests that the GF remains close to the inoculum in some local lesion viral populations while transitioning to a more balanced GF in others. In experiment 3, we identified two less well-supported clusters than those identified in experiment 1. Eight of 12 lesions from experiment 3 are found in cluster 1, characterised by higher RNA3 and RNA2, whilst the smaller cluster 2 ($n = 4$) has higher RNA1 and low RNA2 levels. Given the results of the folding test of unimodality and the visualisation of the predicted clusters, we think there is strong evidence for clustering in experiment 1 and weak evidence in experiment 3.

Given the different GF clusters in experiment 1 and possibly in experiment 3, we tested whether these clusters may be linked to differences in virus accumulation. As we do not have absolute estimates of virus accumulation, we calculated the mean RT-qPCR cycle quantification (Cq) value over the three viral RNAs and used this as a proxy of virus accumulation. Cq values are inversely related to virus accumulation, thus lower CQ values indicate higher virus accumulation. As we used a validated ΔΔCq-based method (Rao et al. 2013), our inferences on relative accumulation from these data are robust. There are significant differences in mean Cq and thus virus accumulation between clusters in experiment 1 (Mann-Whitney U test: $W = 57$, $P = 4.479 \times 10^{-7}$), but not between clusters in experiment 3 ($W = 8$, $P = 0.2027$). In experiment 1, cluster 1 has a higher virus accumulation than cluster 2, as these clusters have lower mean Cq values and there is an inverse relationship between RT-qPCR and Cq value. Although the differences in accumulation are not statistically significant, there is a similar pattern in experiment 3. For one of the the inocula used here (experiment 1), which is characterized by low levels of RNA3, a considerable number of populations are “trapped” in the low accumulation GF space.

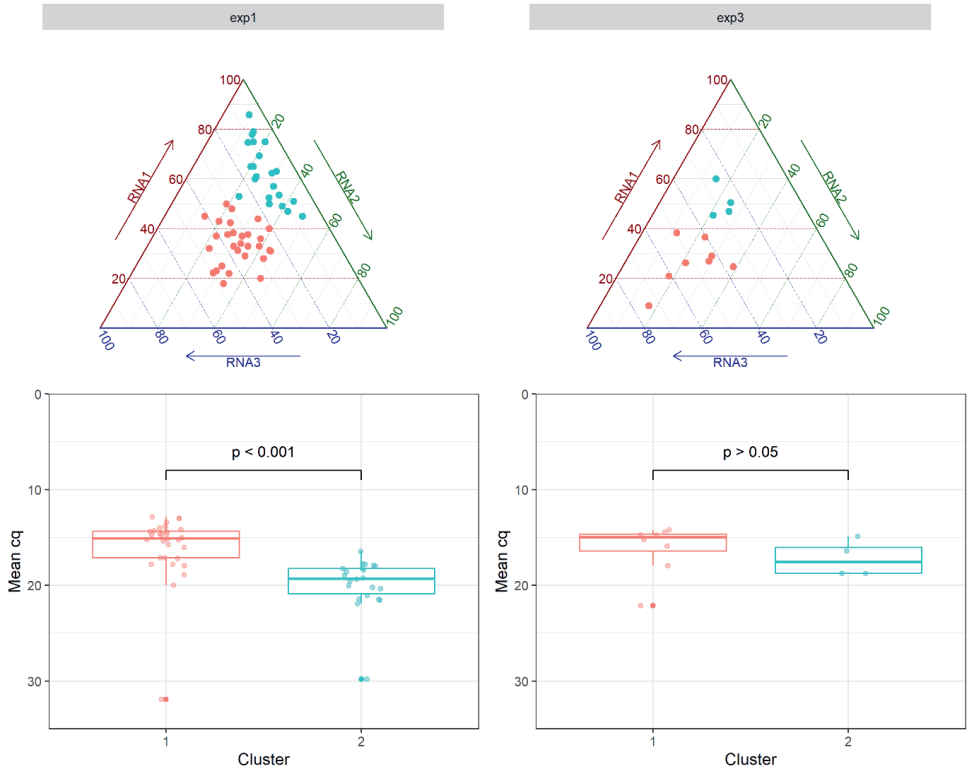


Figure 3. Genome formulae of CMV from local lesions cluster into distinct groups associated with differential virus accumulation*. In experiment 1, there is significant support for two GF clusters (p < 0.001): a central balanced GF cluster 1 (n = 30, indicated in red) and a high RNA1 and low RNA3 cluster 2 (n = 22, indicated in blue). Experiment 3 has weak clustering (p > 0.05), with a majority cluster 1 (n = 8) with high RNA3 and low RNA2 levels (indicated in red). Cluster 2 (n = 4) has higher RNA3 levels and balanced RNA2 and RNA1 levels (indicated in blue). Mean virus accumulation for RNAs 1-3 per local lesion is plotted in the lower graph. On the y-axis is the mean RT-qPCR cycle quantification (Cq) for RNA1-3 per local lesion and on the x-axis are GF clusters identified for experiments 1 and 3. We use Cq values as a proxy for virus accumulation; lower Cq values indicate higher virus accumulation.*

Comparison of the observed genome formula variation to model predictions

We generated predictions of the GF distance *D* under different theoretical scenarios, to have some reference from comparison with our GF observations. We describe our approach and predictions in detail in Chapter 3 of this thesis, and briefly summarise them here. The range of *D* is between zero and $\sqrt{2}$. However, in practice we will not reach the maximum value of *D* because it requires virus populations consisting of a single but distinct segment (e.g., GF

values {f_1,f_1, f_3} of {1, 0, 0} vs {0,1,0}). First, a more relevant upper value is $\overline{D_{a,b}^{rand}}$, the mean pairwise distance between populations when accumulation of each segment is non-zero but otherwise random. For a tri-segmented multipartite virus, this renders a prediction $\overline{D_{a,b}^{rand}}$ ~ 0.3905 (Chapter 3 of this thesis). Second, we want to know how much GF variation is generated by a single population bottleneck, e.g., the bottleneck which occurs at the start of infection in each local lesion. This prediction is more complex because GF variation will be maximised for an intermediate bottleneck size, as the requirement for a complete genome filters out much of the GF variation when the population bottleneck is narrow (Zwart & Elena, 2020).Therefore, $\overline{D_{a,b}^{drift}}$ is the mean pairwise distance between populations experiencing maximal GF drift For a tri-segmented multipartite virus, a bottleneck size of ~7 virus particles renders the maximum predicted value $\overline{D_{a,b}^{drift}}$ ~ 0.2801 (Chapter 3 of this thesis). We can compare these two model predictions of GF variation to the observed value of *D* for our experimental data (Table 5). We determined the GF variation in local lesions of experiments 1 and 3 only, as the number of data points in experiment 2 was too small to be representative. We find that for both experiments, the GF variation is near the Poisson-based model predictions: ($\underline{D}$ $\pm$ *SD)* 0.298$\pm$ 0.062 for experiment 1, and ($\underline{D}$ $\pm$ *SD)* 0.275$\pm$0.073 for experiment 3. Therefore, the observed GF variation is similar to the maximum GF drift that can be generated by a single bottleneck event, $\overline{D_{a,b}^{drift}}$. This result suggests that the GF variation seen in local lesions is appreciable and could be accounted for entirely by the expected GF drift caused only by the bottleneck in the initial infection. However, recall that we have predicted the distance for maximum GF drift, as many bottleneck sizes are predicted to have less drift. Therefore, it is plausible that the GF drift caused by the initial bottleneck was less, and subsequent stochastic processes during replication and cell-to-cell spread elevated the levels of GF variation.

Table 5. GF variation in local lesion infections in *C. quinoa*. The GF distance (*D*) from empirical observations is presented for experiment 1 and experiment 3 with 95% confidence interval (CI). Model predictions for GF variation under two conditions: $\overline{D_{a,b}^{rand}}$ random accumulation of segments and $\overline{D_{a,b}^{drift}}$ when variation is at maximum GF drift.

Model predictions		Experimental data		
$\overline{D_{a,b}^{rand}}$	$\overline{D_{a,b}^{drift}}$	Experiment	Number of lesions	*D* (95% CI)
0.3905	0.2801	1	52	0.2981(0.2813 - 0.3149)
		3	12	0.2754 (0.2338 – 0.3170)

Concluding Remarks

We investigated the variation of the GF in local lesions of *C. quinoa,* showing that the GF is variable over individual local lesions and occupies a large GF space (Figure 3). Local lesions are the consequence of the interaction between the virus and the host's hypersensitive response (HR), resulting in programmed cell death (PCD). During this phase, virus infection

is localised to a small number of cells, and we speculate that the observed variation in the GF at this stage of infection may reflect changes in segment frequencies which increase the likelihood of the virus to escape cells affected by PCD. The lesions with the lowest levels of accumulation also had the lowest levels of RNA3, which encodes for *3a* (the movement protein, MP) and *3b* (the capsid protein, CP). Therefore, we speculate that the GF in these populations may have limited the availability of the MP and thereby the capacity for rapid cell-to-cell movement. Although hypotheses on the adaptive role of the genome formula may explain our observations, alternative explanations that reverse causality cannot be discarded. For example, reduced cell-to-cell movement introduces random variation in local GF , and the GF may revert to a balanced value in populations which undergo more cell-to-cell movement even if GF changes have no intrinsic adaptive value. More work will be needed to show GF changes are adaptive.

We have observed that there are two clusters in the CMV GF space for local lesions, which differ in their virus accumulation. Previous work suggested that as the GF approaches its equilibrium value there is higher virus accumulation (Sicard et al. 2013, Wu et al. 2017), but all studies reported a GF space with a single equilibrium value, I.e., the setpoint genome formula (SGF) (Sicard et al. 2013). Although the SGF may be host dependent (Sicard et al. 2013, Wu et al. 2017), for a given set of environmental conditions the GF converged on a single equilibrium value in these studies. Our results here are the first indication that there may be multiple equilibria in GF space, even in one particular host environment. Previous research with local lesions has shown a GF skewed to high RNA1 and low RNA2 levels (Boezen, Johnson, et al. 2023), for CMV infection of *C. quinoa*. This position in GF space is analogous to the GF centroid of cluster 1, as identified in this study for experiment 1. In the study of Boezen, Johnson, et al. (2023), whole leaves were collected and these GF measurements therefore reflect the mean GF over all local lesions in a leaf, weighed by each lesion's level of viral accumulation. These results are, therefore, congruent with our results here, as they will be representative for those local lesions with high accumulation. We have demonstrated considerable variation in the GF. However, when the inoculum GF is close to the balanced, high accumulation equilibrium, the resulting populations usually do not end up in the GF space associated with low accumulation. Two local lesions from experiment 3 have relatively low accumulation and are close to the GF space for cluster 2 in experiment 1, but the majority of local lesions in experiments 2 and 3 have a more balanced GF. Therefore, the low accumulation GF space appears to be most accessible when the inoculum is in this GF space. We can tentatively conclude that the inoculum GF derived from another host with a different equilibrium, may lead to unfavourable GFs in a new host. Without the initial GF displacement from the inoculum, we do not observe that GF drift displaces populations from a high accumulation inoculum GF to the low accumulation GF space. Therefore, a specific set of circumstances may be needed to trap populations in a low-accumulation region in GF space, including the inoculum GF and limited virus replication and spread. These requirements may explain why such observations have not been made before.

Supplementary S1: Genome formula variation in local lesions

Table S1. Genome formulae of CMV inocula from three experiments. (mean ± standard deviation*). *Standard deviation is derived from target specific qPCR reactions per RNA (RNA 1- 3), performed with three technical replicates, *n* = 1. Inocula are derived from 14 days post-infection (dpi) systemic infection in *N. tabacum* and *N. benthamiana*.

Experiment	Inoculum Host	RNA1	RNA2	RNA3
1	*N. tabacum*	0.49 ± 0.39*	0.37 ± 0.02*	0.14 ± 0.07*
2	*N. tabacum*	0.46 ± 0.02*	0.25 ± 0.03*	0.29 ± 0.03*
3	*N. benthamiana*	0.26 ± 0.16*	0.13 ± 0.09*	0.61 ± 0.11*

Table S2. Genome formulae of CMV from local lesion infections of *C. quinoa* in three experiments. (mean ± standard deviation), samples at ten days post infection (dpi). Standard deviation is from biological replicates of local lesions per experiment. Experiment 1: *n* = 52 lesions, Experiment 2: *n* = 4 lesions, Experiment 3: *n* = 12 lesions. The GF is calculated as the mean value of the RNA segment relative to the sum of all RNA segments, RNA1 + RNA2 + RNA3 = 1

Experiment	Local lesion Host	RNA1	RNA2	RNA3
1	*C.quinoa*	0.46 ± 0.17	0.29 ± 0.09	0.25 ± 0.14
2	*C.quinoa*	0.28 ± 0.11	0.32 ± 0.03	0.40 ± 0.09
3	*C.quinoa*	0.34 ± 0.14	0.23 ± 0.07	0.43 ± 0.15

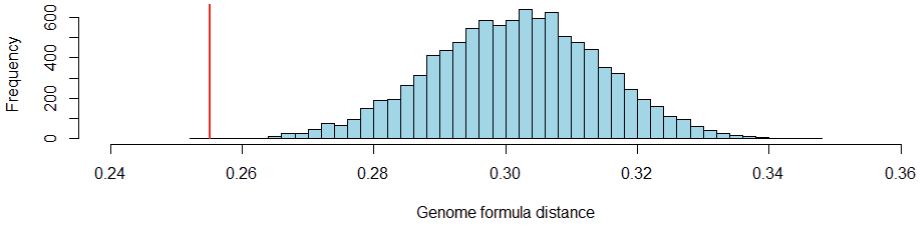


Figure S1. Effect of the inoculum on the genome formula in local lesions. *A resampling approach was used to determine whether the observed genome formula (GF) in local lesions was closer to the inoculum GF than expected by chance. The bars in the histogram indicate the predicted distribution of the GF distance (D) from each lesion to the GF in the inocula of the three experiments, whereas the red line indicates the observed GF distance (D).*

References

Anderson, Marti J. 2001. “A New Method for Non-parametric Multivariate Analysis of Variance.” *Austral Ecology* 26 (1): 32–46.

———. 2006. “Distance-Based Tests for Homogeneity of Multivariate Dispersions.” *Biometrics* 62 (1): 245–53.

———. 2017. “Permutational Multivariate Analysis of Variance (PERMANOVA).” In *Wiley StatsRef: Statistics Reference Online*, 1–15. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat07841.

Betancourt, Mónica, Alberto Fereres, Aurora Fraile, and F. Garcia-Arenal. 2008. “Estimation of the Effective Number of Founders That Initiate an Infection after Aphid Transmission of a Multipartite Plant Virus.” *Journal of Virology* 82 (24): 12416–21.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064.

Boezen, Dieke, Maritta Vermeulen, Marcelle Johnson, Rene Van Der Vlugt, Carolyn Malmstrom, and Mark Zwart. 2023. “Mixed Viral Infection Constrains the Genome Formula of Multipartite Cucumber Mosaic Virus.” *Frontiers in Virology* 3. https://doi.org/10.3389/fviro.2023.1225818.

Coll, N. S., P. Epple, and J. L. Dangl. 2011. “Programmed Cell Death in the Plant Immune System.” *Cell Death and Differentiation* 18 (8): 1247–56.

Cooper, B. 2001. “Collateral Gene Expression Changes Induced by Distinct Plant Viruses during the Hypersensitive Resistance Reaction in Chenopodium Amaranticolor.” *The Plant Journal: For Cell and Molecular Biology* 26 (3): 339–49.

Fraile, Aurora, José Luis Alonso-prados, Miguel A. Aranda, Juan J. Bernal, José M. Malpica, and Fernando Garci. 1997. “Genetic Exchange by Recombination or Reassortment Is Infrequent in Natural Populations of a Tripartite RNA Plant Virus.” *Journal of Virology* 71 (2): 934–40.

Gallet, Romain, Jérémy Di Mattia, Sébastien Ravel, Jean-Louis Zeddam, Renaud Vitalis, Yannis Michalakis, and Stéphane Blanc. 2022. “Gene Copy Number Variations at the within-Host Population Level Modulate Gene Expression in a Multipartite Virus.” *Virus Evolution* 8 (2): veac058.

Gutiérrez, Serafín, and Mark P. Zwart. 2018. “Population Bottlenecks in Multicomponent Viruses: First Forays into the Uncharted Territory of Genome-Formula Drift.” *Current Opinion in Virology* 33 (December): 184–90.

Hamilton, Nicholas E., and Michael Ferry. 2018. “Ggtern: Ternary Diagrams Using ggplot2.” *Journal of Statistical Software* 87 (December): 1–17.

Holmes, Francis O. 1929. “Local Lesions in Tobacco Mosaic.” *Botanical Gazette* 87 (1): 39–55.

Holm, Sture. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics, Theory and Applications* 6 (2): 65–70.

Jacob, Pierre, Junko Hige, and Jeffery L. Dangl. 2023. “Is Localized Acquired Resistance the Mechanism for Effector-Triggered Disease Resistance in Plants?” *Nature Plants* 9 (8): 1184–90.

Jacquemond, Mireille, and Herve Lot. 1981. “L’ARN Satellite Du Virus de La Mosaïque Du Concombre I. - Comparaison de L’aptitude à Induire La Nécrose de La Tomate d’ARN Satellites Isolés de Plusieurs Souches Du Virus.” *Agronomie* 1 (10): 927–32.

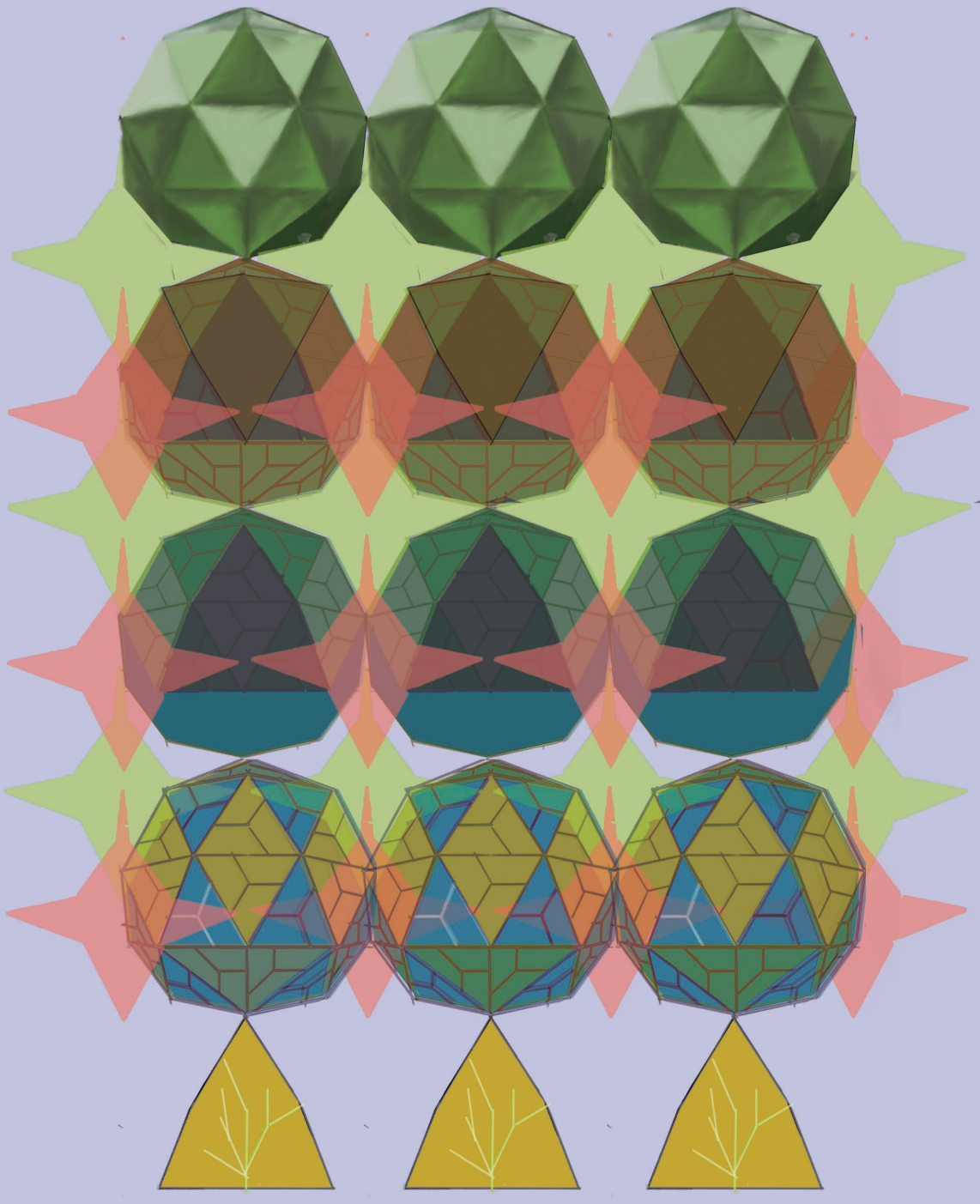
Lam, Eric, Naohiro Kato, and Michael Lawton. 2001. “Programmed Cell Death, Mitochondria and the Plant Hypersensitive Response.” *Nature* 411 (6839): 848–53.

Lucía-Sanz, Adriana, Jacobo Aguirre, and Susanna Manrubia. 2018. “Theoretical Approaches to Disclosing the Emergence and Adaptive Advantages of Multipartite Viruses.” *Current Opinion in Virology* 33 (December): 89–95.

Lukan, Tjaša, Špela Baebler, Maruša Pompe-Novak, Katja Guček, Maja Zagorščak, Anna Coll, and Kristina Gruden. 2018. “Cell Death Is Not Sufficient for the Restriction of

Potato Virus Y Spread in Hypersensitive Response-Conferred Resistance in Potato." *Frontiers in Plant Science* 9 (February): 168.
Michalakis, Yannis, and Stéphane Blanc. 2020. "The Curious Strategy of Multipartite Viruses." *Annual Review of Virology* 7 (1): 203–18.
Miyashita, Shuhei, Kazuhiro Ishibashi, Hirohisa Kishino, and Masayuki Ishikawa. 2015. "Viruses Roll the Dice: The Stochastic Behavior of Viral Genome Molecules Accelerates Viral Adaptation at the Cell and Tissue Levels." *PLoS Biology* 13 (3): e1002094.
Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. "The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner." *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.
Ohshima, Kazusato, Kosuke Matsumoto, Ryosuke Yasaka, Mai Nishiyama, Kenta Soejima, Savas Korkmaz, Simon Y. W. Ho, Adrian J. Gibbs, and Minoru Takeshita. 2016. "Temporal Analysis of Reassortment and Molecular Evolution of Cucumber Mosaic Virus: Extra Clues from Its Segmented Genome." *Virology* 487: 188–97.
Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'hara, et al. 2022. *Vegan: Community Ecology Package* (version 2.6.4). Vegan: Community Ecology Package. https://cran.r-project.org/web/packages/vegan/index.html.
Rao, Xiayu, Xuelin Huang, Zhicheng Zhou, and Xin Lin. 2013. "An Improvement of the 2ˆ(-Delta Delta CT) Method for Quantitative Real-Time Polymerase Chain Reaction Data Analysis." *Biostatistics, Bioinformatics and Biomathematics* 3 (3): 71–85.
R Core Team. 2001. "R: A Language and Environment for Statistical Computing." https://www.r-project.org/.
Roenhorst, Annelien. 2014. "Protocol for Mechanical Inoculation of Test Plants." The Netherlands: National Plant Protection Organization.
Ross, A. F. 1961. "Systemic Acquired Resistance Induced by Localized Virus Infections in Plants." *Virology* 14 (July): 340–58.
Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. "The Strange Lifestyle of Multipartite Viruses." . *PLoS Pathogens* 12 (11): e1005819.
Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. "Gene Copy Number Is Differentially Regulated in a Multipartite Virus." *Nature Communications* 4: 2248.
Siffer, Alban, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. 2018. "Are Your Data Gathered?" In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2210–18. KDD '18. New York, NY, USA: Association for Computing Machinery.
Soosaar, Jennifer L. M., Tessa M. Burch-Smith, and Savithramma P. Dinesh-Kumar. 2005. "Mechanisms of Plant Resistance to Viruses." *Nature Reviews. Microbiology* 3 (10): 789–98.
Varsani, Arvind, Pierre Lefeuvre, Philippe Roumagnac, and Darren Martin. 2018. "Notes on Recombination and Reassortment in Multipartite/segmented Viruses." *Current Opinion in Virology* 33 (December): 156–66.
Vishnoi, Radha, Susheel Kumar, and Shri Krishna Raj. 2013. "Molecular Characterization of a Cucumber Mosaic Virus Isolate Associated with Mosaic Disease of Banana in India." *Phytoparasitica; Israel Journal of Plant Protection Sciences* 41 (5): 545–55.
Wright, K. M., G. H. Duncan, K. S. Pradel, F. Carr, S. Wood, K. J. Oparka, and S. S. Cruz. 2000. "Analysis of the N Gene Hypersensitive Response Induced by a Fluorescently Tagged Tobacco Mosaic Virus." *Plant Physiology* 123 (4): 1375–86.
Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. "Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus." *Scientific Reports* 7 (1): 1–15.
Zwart, Mark P., and Santiago F. Elena. 2020. "Modeling Multipartite Virus Evolution: The

Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments†." *Virus Evolution* 6 (1): veaa022.



Host dependence and evolutionary stability of the genome formula in a multipartite RNA virus

Marcelle L. Johnson[1,2], Bo Hartman1, Rosalinde Keijzer[1], J. Arjan G.M. de Visser[3], Rene A.A. van der Vlugt[2], Mark P. Zwart[1]

[1] Netherlands Institute of Ecology (NIOO-KNAW), P.O. BOX 50, 6700 AB, Wageningen, The Netherlands

[2] Laboratory of Virology, Wageningen University and Research, P.O. BOX 16, 6700 AA, Wageningen, The Netherlands

[3] Laboratory of Genetics, Wageningen University and Research, P.O. BOX 16, 6700 AA, Wageningen, The Netherlands

Abstract

Multipartite viruses have segmented genomes and individual segments are separately packaged and transmitted. Genome segment copies have differential host-specific frequencies, the “genome formula” (GF). The GF undergoes rapid changes, hypothesized to play a role in adapting to diverse environments by regulating viral gene expression. In this study, we investigate the role of the GF in the local adaptation of a tripartite virus, cucumber mosaic virus, to three host species: *Nicotiana benthamiana*, *Nicotiana tabacum* and *Arabidopsis thaliana*. We combine serial passaging *in planta* with next generation sequencing to determine the genome formula, assess viral fitness by measuring virus titre, and elucidate the dynamics of GF variation. We find that the GF of cucumber mosaic virus is highly variable, in agreement with previous results. Despite this variation, we show that the GF is host-specific, as we observed significant GF changes in *A. thaliana.* For each species, passaging could not be completed for some replicate populations, and these populations were considered extinct. Although the GF showed considerable random variation over passages, we only saw a systematic shift in one species, *N. tabacum*. Point mutations were observed in multiple populations, with the most common loci with mutations being RNA 2a, the viral RNA-dependent RNA polymerase, and the untranslated regions. Most extinctions were associated with these repeated mutations, although some extinctions also appear linked to changes in the GF. Our study highlights the stability of the GF after initial host-dependent changes, as well as the potential risks associated with random GF variation.

Introduction

Multipartite viruses have segmented genomes in which individual segments are packaged within a virus capsid and transmitted (Michalakis and Blanc 2020). In the multipartite faba bean necrotic stunt virus (FBNSV), viral segments accumulate to a stable host-specific stoichiometric ratio, the "genome formula" (GF) (Sicard et al. 2013). The GF has been experimentally measured for the plant viruses alfalfa mosaic virus (AMV) (Wu et al. 2017), cucumber mosaic virus (CMV) (Boezen, Johnson, et al. 2023), banana bunchy top virus (BBTV) (Yu et al. 2019) and rice stripe virus (RSV) (Zhao et al. 2019). In addition, the GF has been measured for the animal multipartite virus Bombyx mori bidensovirus (BmBDV) (Hu et al. 2016) and for the segmented bluetongue virus (BTV) (Moreau et al. 2020). FBNSV has a genome composed of monocistronic segments, where host-specific GF segment copy differences in *Vicia faba* and *Medicago truncatula* have been proposed to improve viral fitness within these hosts by changing virus gene expression (Sicard et al. 2013). While the segments converge reproducibly on a stable host species-dependent equilibrium (Sicard et al. 2013; Wu et al. 2017; Moreau et al. 2020; Mware 2016), some studies noted the considerable GF variation within and between plants (Sicard et al. 2013; Wu et al. 2017). This variation would facilitate rapid invasion in novel hosts and replication within a specific cell or tissue type (Zwart and Elena 2020). Sicard et al. (2013) suggested that, viruses selectively regulate copy numbers of segments during infection to counter host immune responses and regulate viral gene expression during replication and transmission. A large body of research suggests that genome segment copy number variation (CNV) is a source of genetic diversity within populations and facilitates rapid adaptation. For, example, adaptation by gene CNV is known for increasing bacterial antibiotic resistance (Sandegren and Andersson 2009), insecticide resistance of the planthopper *Nilaparvata lugens* to imidacloprid (Zimmer et al. 2018) and *Eschericia coli* tolerance to high temperature (41.5°C) (Riehle, Bennett, and Long 2001). Here we explore whether CNV may also play a role in adaptation by multipartite viruses.

In viral systems, CNV has extensively been studied for the monopartite dsDNA human vaccinia virus (VACV) (Elde et al. 2012), a monopartite dsDNA virus that infects mammals, including humans. These researchers started with a viral strain missing the host immune suppresion gene *E3L*, resulting in poor replication in human cells. Upon passaging in human cells, another host immune suppression gene, *K3L*, is amplified to comprise many copies in the virus genome, increasing K3L protein levels and thereby anti-host phosphorylation activity to suppress host-immune effectivity (Elde et al. 2012). *K3L* gene amplification preceded a nonsynonymous mutation within *K3L*, which provided higher fitness gain than gene duplication (Elde et al. 2012). Finally, the many copies of the original gene are lost and only the mutated copy of the gene is retained. This amplification, mutation and subsequent collapse of a gene to single copy is dubbed the "genomic accordion" response (Elde et al. 2012; Cone et al. 2017; Näsvall et al. 2012). These results show that copy number variation can have immediate fitness benefits and transiently increases mutation supply, aiding the exploration of sequence space and the generation of beneficial innovations which could be exploited to increase viral fitness (Bayer, Brennan, and Geballe 2018). However, the interactions between different classes of mutations during evolution can be more complex. A hallmark study, found that under some conditions, CNV and adaptive point mutations are mutually exclusive in *Escherichia coli* populations during experimental evolution of bacterial growth on galactose substrate (Isabella Tomanek and Guet 2022). The *GALK* gene encodes galactokinase, which

enables *E.coli* growth for using galactose metabolism (Tomanek et al. 2020). The results from this study suggested that the high mutation rates of CNV could prevent the fixation of point mutations with a large effect on fitness, thereby limiting adaptation. Combined, these studies indicate that different interactions between CNV and point mutations can occur, and that these interactions can be shaped by environmental conditions.

CNV is relevant to rapid evolutionary adaptation in many organisms, including poxviruses. The variable GF of multipartite viruses will lead to high levels of CNV, albeit without linkage between gene copies. Although a number of studies have considered the effects of the GF on infection, its evolutionary implications have not been considered experimentally. Here, we use experimental evolution to study the role of the GF during the evolution of a three segmented, multipartite RNA virus, cucumber mosaic virus (CMV). CMV is a suitable model system for this purpose because it is an RNA virus with a high mutation rate (Ouedraogo and Roossinck 2019), it has a broad host range and many aspects of its biology have been studied (M. J. Roossinck 2001). CMV isolates have been collected across the world and sequences of all three RNAs have been analysed to determine the population structure; finding that the virus has a diverse genetic background and can be divided into two subgroups, subgroup I (SI) and II (SII) based on phylogenetic analysis (Ohshima et al. 2016; Marilyn J. Roossinck 2002). Studies on CMV populations in the US (Nouri et al. 2014) and Spain (Fraile et al. 1997) show that reassortment and recombination between subgroups is uncommon and may be selected against. CMV is a model for studying plant virus evolution with multiple studies exploring virus evolution in greenhouse and field settings (M. J. Roossinck 2001). Sacristan et al. (Sacristán et al. 2005) investigated host adaptation by experimental evolution of six CMV genotypes isolated from *Cucumis sativus*, *Phaseolus vulgaris* and *Solanum lycopersicum,* by comparing virus adaptation in the original and heterologous hosts. Virus accumulation in the original host was highest for all isolates and did not increase during successive passaging, suggesting that CMV may have a limited capacity to adapt to these hosts on the short term. As for other multipartite viruses, no studies have considered the role of the GF during the evolution of CMV.

In the current study, we investigate the evolutionary dynamics of the GF using CMV. First, we considered whether virus genotype had an effect on the GF. If such effects occur, we could expect to find shifts in the GF in evolving virus populations due to the fixation of mutations. Second, we studied CMV's evolution in three host species, *Nicotiana tabacum*, *Nicotiana benthamiana* and *Arabidopsis thaliana,* over multiple rounds of passaging. We hypothesized that GF dynamics may serve a similar purpose as the “genomic accordions” seen in VACV: we expect to see rapid changes in the GF in some host species, possibly followed by further changes in the GF when beneficial mutations become fixed in evolving populations. To our surprise, we found no evidence of adaptation and many virus populations went extinct. Instead, we focused on what caused these extinctions, including whether GF changes play a role.

Methods and Materials

Plants and inoculation

Nicotiana benthamiana, *N. tabacum* cv White Burley and *A. thaliana* col-0 were grown in standard greenhouse conditions (22/20°C day/night, 60% RH, 16/8hrs light/dark cycle. Plants were mechanically inoculated 3 weeks post germination with a single dose of CMV and grown till 14 dpi upon harvesting (Roenhorst 2014). Upper leaves of symptomatic plants were collected and stored at -80C.

CMV isolates and amplification of virus

CMV isolates were obtained from the plant virus collection of Wageningen Plant Research (www.primediagnostics.com, Wageningen University and Research), Wageningen, The Netherlands. A panel of isolates were selected from the existing collection based on the following criteria: (1) Confirmed CMV infection by DAS-ELISA, (2) CMV isolates from both subgroup I and II, host species and year of collection (Figure 1, Supplementary 1 Table S1). CMV isolates from subgroup I and II (Supplementary 1, Table S1) were amplified in *N. tabacum*.

Serial passaging of CMV-i17F in three host species

Subgroup I isolate CMV-i17F (Jacquemond and Lot 1981) was serially passaged in three hosts: *N. tabacum* cv White Burley, *N. benthamiana* and *A. thaliana* Col-0 (Figure 3). The experiment consisted of six biological replicate lines per host and 3 - 6 technical replicate plants per line. When lines went extinct, the inoculations were repeated in a second set of 6 plants. In each passage a single, positive infected plant was randomly selected and used as inoculum for the next passage round, inoculum concentration of 0.0002 $g.ml^{-1}$.

RNA extraction

Plant samples were homogenised with a handheld homogeniser (BIOREBA) in liquid nitrogen to a fine powder and stored at -80℃. 100mg of plant tissue was used in total RNA extraction using the Qiagen RNeasy Plant Mini Kit with on-column DNAse treatment following manufacturer's instructions (Qiagen). RNA concentration was quantified by NanodropOne spectrophotometer and stored at -80℃.

cDNA synthesis and qPCR of CMV-i17F

Powdered leaf samples were used in an immunochromotography assay (Agristrip, BIOREBA) to confirm infection. Positive samples were used for downstream analysis. To quantify the GF

in replicate lines, a qPCR targeting CMV-i17f RNA1, RNA2 and RNA3 (Table 1) was performed essentially as follows.

250ng of RNA was converted to cDNA using the iScript cDNA synthesis kit (Bio-rad) using random hexamers according to manufacturer's instructions; the cDNA concentration was quantified with NanodropOne spectrophotometer and stored at -20℃. A qPCR reaction mix consisted of 8ul per well with 4ul 2x iQ SYBR Green (BioRad) mastermix, 0.16 ul each of 10uM forward and reverse primers, 3ul of template cDNA and 0.68ul nuclease-free H_2O. Cycling conditions were 95°C (3min), 40 cycles: 95°C for 10s, 60°C for 3s and a melt curve at 5°C increments from 65 - 95°C. The $2^{-\Delta\Delta Ct}$ method (Rao et al. 2013) was used to determine the GF relative to RNA1.

Table 1. RT-qPCR for the quantification of GF of CMV in local and systemic infections, CMV primers are relative the reference isolate CMV-Fny accessions (NC_002034; NC_002035; NC_001440)

Target	Primer ID	Primer sequence	Amplicon length	Annealing temperature	Position	Reference
CMV RNA1 (1a)	cmvrna1a_1f	GCACAACCCGTGAGTGAGG	83	60°C	1706 - 1724	(Boezen, Johnson, et al. 2023)
	cmvrna1a_1r	TCCCTTCCACAAACATCAGCAG			1767 1788	
CMV RNA2 (2a)	cmvrna2a_1f	GGTGTTGTTGATAATGCGACTCTG	94	60°C	789 - 812	
	cmvrna2a_1r	CGATGGTTGGCGTTGGACAT			863 - 882	
CMV RNA3 (CP)	cmvrna3a_1f	ACCATGATCTTCCCGCTTTGG	91	60°C	511 - 531	
	cmvrna3a_1r	ACGACAGCAAAACACCGCTT			582 - 601	

Sequencing and bioinformatic analysis

For genomically characterising CMV isolates, total RNA isolated from infected plants was sent for library preparation and RNA sequencing using Illumina Miseq by BaseClear B.V. in Leiden, The Netherlands. CLC Genomics Workbench v 22.0.1 (Qiagen) was used to analyse sequence data. Briefly, reads were quality-trimmed to a limit of 0.05 of quality score, adapter trimming and mapped to the subgroup reference sequence for subgroup I CMV-Fny accession (RNA1;RNA2;RNA3: NC_002034; NC_002035; NC_001440) or subgroup II CMV-Ls accession (RNA1;RNA2;RNA3: AF416899; AF416900; AF127976) to generate the consensus sequence per CMV isolate. The mean coverage per position per isolate was exported as tsv file and used to generate the genome formula.

CMV samples from the final passage (p5) or last passage of positive infection samples, and an ancestral CMV-infected *N. tabacum* sample were sequenced. 100mg of frozen powdered leaf material was used in total RNA extraction using the RNeasy Plant Mini Kit (Qiagen) with

on-column DNase digestion (Qiagen) following manufacturer's instructions. RNA concentration was quantified by the NanodropOne spectrophotometer and Qubit. Total RNA was ribodepleted before commencing with double-stranded cDNA (ds-cDNA) synthesis (Liefting, Waite, and Thompson 2021). The FastSelect -rRNA plant kit (Qiagen) protocol was used to remove ribosomal RNA followed by Maxima H Minus Double-Stranded cDNA Synthesis kit (ThermoScientific). ds-cDNA concentration was determined by Qubit and ~2 - 27 ng of cDNA per sample was sent for Nextera DNA XT Library preparation and sequencing on NovaSeq platform for paired-end 100-bp sequencing at the Cologne Centre for Genomics (CCG), Cologne, Germany. All sequence data have been deposited with links to BioProject accession number PRJNA1076493 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Statistical analysis of the genome formula

Statistical analysis was performed in R 4.3.1 (R Foundation for Statistical Computing 2023). GFs were visualised in ternary plots using the ggtern package version 3.4.2 (Hamilton and Ferry 2018).

To quantitatively analyse the GF, we consider the GF as a set of relative frequencies (f) for all genome segments of a virus $\{f_1, f_2, f_3 \ldots, f_j\}$, where for CMV $j = 3$. We analyse the GF using a previously described metric, the genome formula distance (D) (Boezen, Vermeulen, et al. 2023). The use and different test for which D can be used are further described in Chapter 3 of this thesis. D is the Euclidean distance between GF values, a and b such that;

$$D_{a,b} = \sqrt{\sum_{i=1}^{j=3} \left(f_{a,j} - f_{b,j}\right)^2}$$

The single value D is used in analysis to compare GF values at the initial passage (p0), final passage (p5 or last passage of positive infection) and to compare the change in GF. We estimate the pairwise Euclidean distance between observations using the vegdist function from VEGAN package version 2.6.4 (Oksanen et al. 2022). We compare changes in the centroid (mean) and variance of the GF using a permutation-based approach, the permutational analysis of variance (PERMANOVA) (Anderson 2017). To test for differences in the variance (spread) we use the PERMDISP2 betadisper function. For pairwise comparisons with PERMANOVA and PERMDISP2, a Holm-Bonferroni correction for multiple comparisons was made (Holm 1979). R scripts of all analysis are available at Zenodo (10.5281/zenodo.10652647).

Low frequency variant detection in ancestral and evolved populations of CMV-i17F

Sequencing analysis was conducted in CLC genomics workbench v 22.0.1 (Qiagen). Reads were trimmed using Phred quality scores within 0.05 limits, with ambiguous nucleotides and

adapter trimming. Thereafter, ancestral CMV inoculum was mapped to the subgroup I reference isolate CMV-FNY (RNA1;RNA2;RNA3) (NC_002034; NC_002035; NC_001440) with standard settings. The consensus sequence was extracted with annotation and used as input for read mapping of evolved populations and low-frequency variant detection. For evolved populations, read trimming was as described and read mapping was done relative to the ancestral CMV inoculum RNA1-3 with standard settings. Variant detection was done using the low-frequency detection tool with low-frequency variant detection thresholds set as 15 reads, reference masking was set to ignore positions with coverage exceeding 100000, minimum read length of 20 nt, and to ignore broken pairs and non-specific matches. Coverage and count filters were set to a minimum coverage of 10, minimum count of 2 and minimum frequency of 1%. Quality filters were maintained as standard settings (minimum central quality: 20, min neighbourhood quality: 15), and the generated variant table as csv file per population was combined and used as input for further analysis.

Low-frequency variant data from ancestral and evolved CMV populations were analysed in R version 4.3.0.1 with a custom script available at Zenodo (10.5281/zenodo.10652647). We only detected small indels (< 5 bases), multiple nucleotide variation and single nucleotide variation. Therefore, the variant data were classified into four functional groups: (1) intergenic, (2) coding region: synonymous, (3) coding region: non-synonymous and (4) coding region: frameshift or stop codon. Next, we excluded any mutations detected in the ancestral population. We then considered three groups of mutations in the evolved populations, which were used to visualise the data and subsequent analyses of mutations. The main criterion for identifying these groups was mutation frequency (indicated in parenthesis): fixed mutations (frequency > 0.95), intermediate mutations (frequency > 0.05), and low-frequency mutations (frequency > 0.01). We used much more stringent criteria when filtering the ancestral population than for the evolved populations to exclude potential sequencing artefacts and genetic variation already present in the ancestral population. See Table S4 for a complete set of criteria.

Analysis of substitution rates

We analysed the rates at which mutations occurred to understand better what evolutionary forces were acting on sequence evolution. We considered the rate of non-synonymous substitutions (dN) and intergenic substitutions (dI) occurring for single nucleotide variants, both normalised by the rate of synonymous substitutions (dS). Both dN/dS and dI/dS indexes were calculated for the full genome with the approach and R code described in (Zwart et al. 2019). These analyses were performed separately for fixed, intermediate and low-frequency mutations.

Repeatability of mutations in evolved populations

To estimate the repeatability of mutations for evolved populations, we used the *H*-index, an approach and R code described by (Schenk et al. 2022) and are available at Zenodo (10.5281/zenodo.10652647). This measure quantifies the mutual fraction of shared mutations of two genotypes, i.e. the positional overlap between mutational events along the genome, and can accommodate mutations of all sizes (i.e., from point mutations to structural variation).

These analyses were performed separately for fixed, intermediate and low-frequency mutations.

Logistic regression on combined results

We performed logistic regression to identify variables which could explain low levels of virus titre, for the final timepoint analysed for each evolved virus population. We chose logistic regression because the outcome variable (*titre*) showed a bimodal distribution and two unambiguous groups with high or low titre, included as an indicator variable with values 0 and 1, respectively. In the model we included the GF distance (*D*) between the final time point and to the mean GF for first passage (p0) (*D*) as coordinate predictor variable, whether any non-synonymous mutations had occurred as binary predictive variable *(mutation*), and an interaction term, resulting in the model equation *titre ~ mutation * D*. The analysis was run using the glm() function in R, assuming a binomial error structure and logit link function. Given that we were specifically testing whether both predictive variables and the interaction term were associated with low titre, we performed one-sided Chi-square tests to determine whether the estimated coefficients were positive. R scripts are available at Zenodo (10.5281/zenodo.10652647).

Results and Discussion

Genome formulae of CMV isolates from subgroups I and II

The GF of multipartite viruses is highly variable, showing both stochastic variation and dependence on the host plant species. However, whether the GF is also dependent on the virus genotype is unknown. If the GF depends on virus genotype, during evolution the GF could change due to genetic changes in the virus. While these changes could occur due to selection for a different GF, they could also result from mutations that have been fixed due to genetic drift or due to pleiotropy, that is as a side effect of substitutions selected because they affect other viral traits. To consider these questions experimentally, we first explored whether the genome formula depends on virus genotype. We estimated the genome formula from high-throughput sequencing data for a number of CMV subgroup I and II isolates in *N. tabacum* (Figure 1).

When comparing CMV GFs for genetically distinct isolates, they appear to occupy different regions in GF space (Figure 2a). Some of these isolates contained a satellite virus, short RNAs that act as selfish replicators dependent on another virus and known to affect CMV accumulation during infection (Betancourt, Fraile, and Garcia-Arenal 2011), and which may affect the GF (Feng et al. 2012). When observing the GF of subgroup I and II isolates, there is an indication that there may be subgroup-specific differences (Table S3). For both subgroups, RNA3 has the highest level of accumulation. For subgroup 1, RNA 1 accumulates to a higher level than RNA2, whereas RNA2 accumulation is higher than RNA1 for subgroup 2 (Table S3). We know that GF variation in single infections of *N. tabacum* is high, but the measurement we report here for subgroup I isolate I17F is similar to previous ones (Figure

2b). Under the experimental conditions we used, most subgroup I isolates were in the same GF space, with the exception of isolate S4. By contrast, the majority of the subgroup II isolates are in a different GF space with higher levels of RNA2, with isolate K8 being the only exception. Despite the lack of experimental replication, our data are congruent with the previous measurements available and suggest that variation in virus genotype underlies variation in the genome formula, as there are differences between subgroups I and II.

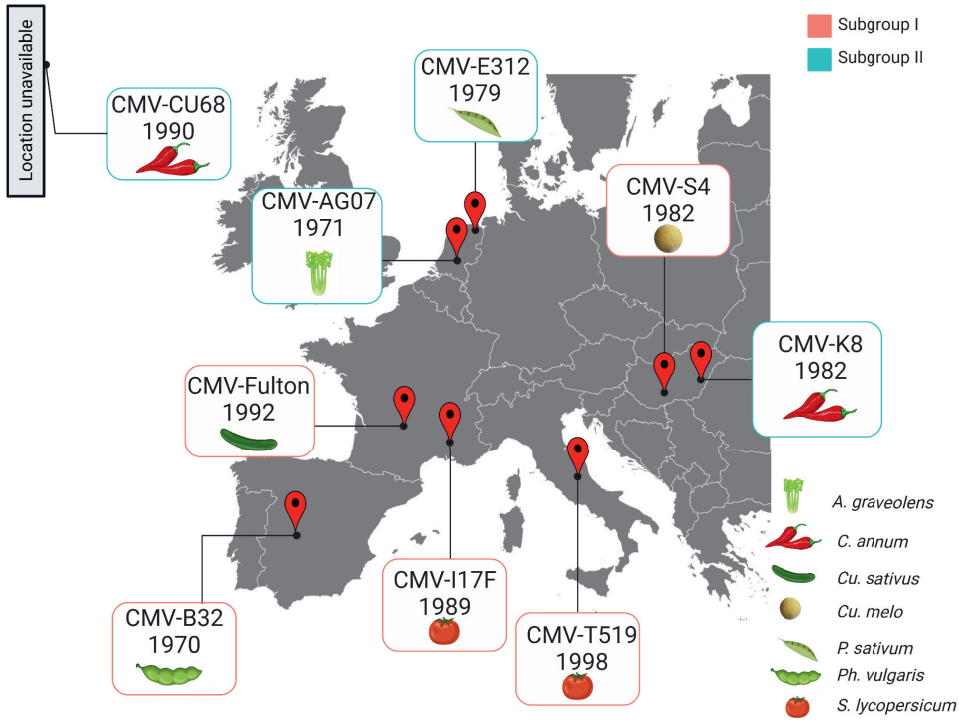


Figure 1. Cucumber mosaic virus isolates from phylogenetic subgroup I and II used in this study*. The name of the isolate and year of collection are given for each isolate, and symbols represent the host species according to the legend on the bottom right.*

Host affects the CMV-i17F genome formula

Previous research with FBNSV (Sicard et al. 2013), AMV (Wu et al. 2017) and CMV (Boezen, Johnson, et al. 2023) showed that the GF depends on host species. Before we study the evolutionary stability of the GF, we first need to test whether any of the host species induce rapid changes in the GF. These rapid changes within a single passage probably are not due to mutation, but simply result from virus-host interactions and may lead to increased viral accumulation. Single virus isolate CMV-i17F (Jacquemond and Lot 1981) was selected for all subsequent work in three host plant species: *A. thaliana*, *N. benthamiana* and *N. tabacum* (Figure 3).

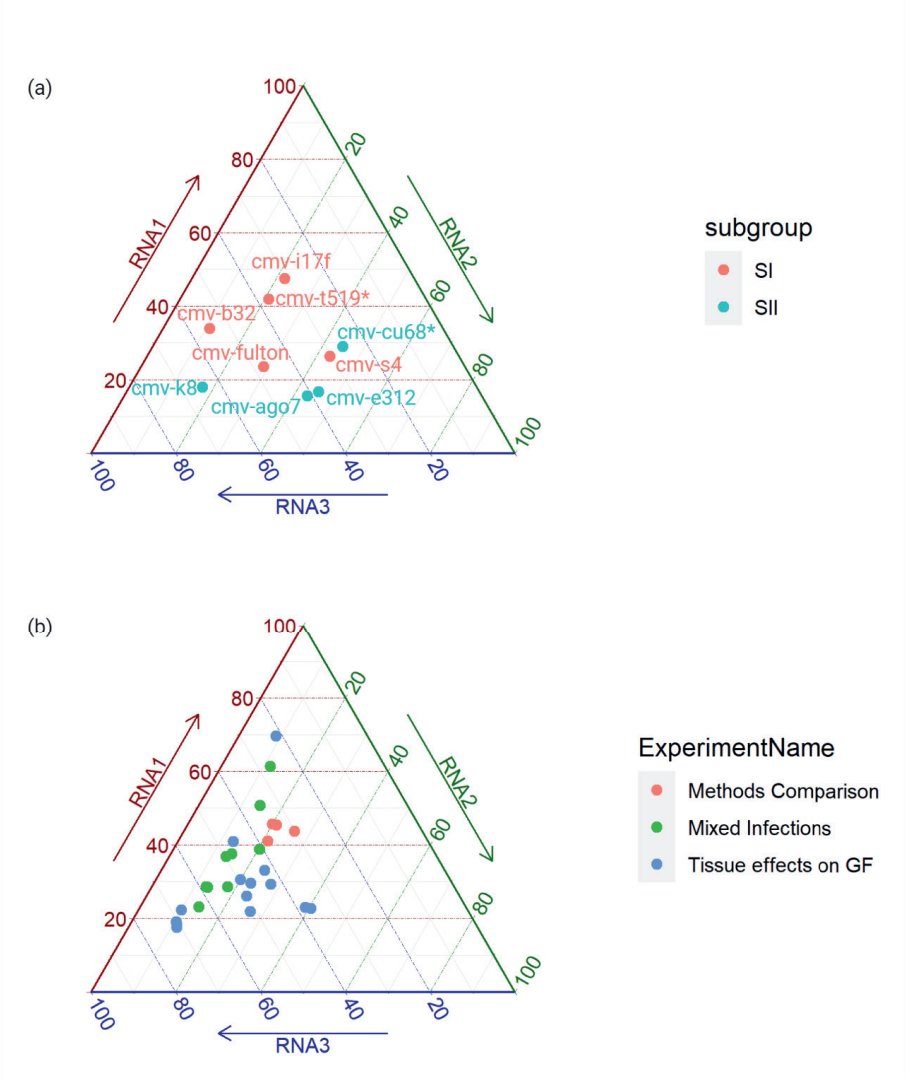


Figure 2. (a) GF of CMV isolates from subgroup I (red) and II (blue) after a single infection cycle*. The GF may be read on ternary diagrams as the intersection of three axes representing the three viral RNAS; x = relative frequency RNA3 (0 -100), y = relative frequency RNA1 (0 -100) z = relative frequency RNA2 (0 -100). GFs of subgroup I (n=5) (indicated in red) and subgroup II (n=4) (indicated in blue) isolates from a single infected N. tabacum at 14 days post infection (dpi). * Contain satellite virus.* ***(b) GF variation of CMV in N. tabacum.*** *The dataset is derived from 3 different experiments, of CMV-i17F infections in N.tabacum. GF measurements from 14 days post infection (dpi) infections from whole upper leaves (red dots), n=4 (Boezen, Johnson, et al. 2023), GFs for 14dpi N. tabacum infections from whole upper leaves (green dots), n=9 (Boezen, Vermeulen, et al. 2023) and n=14 observations from infections of apical leaf tissue from N. tabacum at 14dpi (blue circles) in (from Johnson, Grum-Grzhimaylo et al. Unpublished data).*

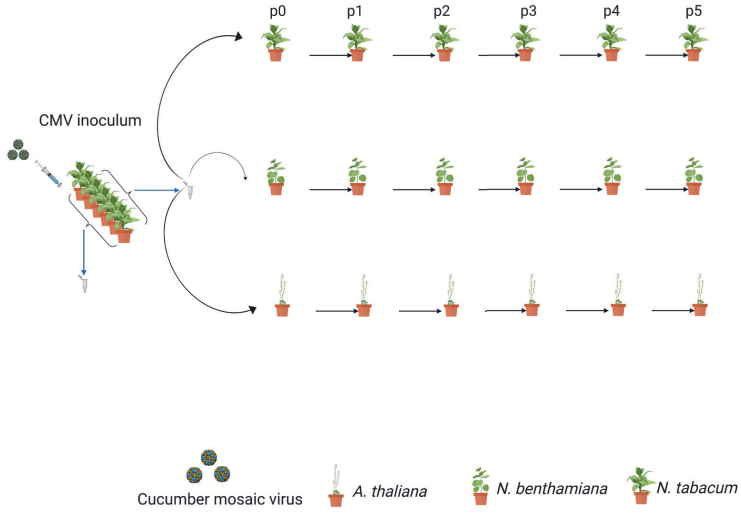


Figure 3. Serial passaging of genotype CMV-i17F in three host species. *Initial inoculation was done with a single dose of CMV-i17F obtained from N. tabacum at passage 0 (p0). Six replicate populations per host species were serially passaged, by randomly selecting a single plant per round. The GF was measured at the initial (p0) and final passage (p5 or last positive sample).*

We first performed a single round of passaging (resulting in passage p0). The GF of the ancestral population (i.e., the starting inoculum consisting of infected *N. tabacum* tissue) and of six individual p0 experimental populations consisting of one randomly selected infected plant each was determined by qPCR (Figure 4). A PERMANOVA analysis showed that there was a significant effect of host species on the GF ($F_{2,15}$ = 4.4023, P = 0.029). We did not find significant differences in the spread of the GF (PERMDISP2 test: $F_{2,15}$ = 0.4616, P = 0.6578). As PERMANOVA can give a significant result due to differences in centroid or spread, this result confirms that the centroid is host-species dependent. The inoculum GF is characterised by higher RNA3 copies than RNA1 and 2, respectively (Table 2). In all hosts, RNA2 had the lowest frequency. RNA1 had the highest frequency in *A. thaliana* and in *N. benthamiana* RNA3 had the highest frequency (Table 2). To confirm which hosts differed in their GF, pairwise comparisons of hosts were made with PERMANOVA, using a Holm-Bonferroni correction for multiple comparisons (see Methods and Materials). We did not find significant differences between *N. benthamiana* and *N. tabacum* ($F_{1,10}$ = *3.7411*, P = 0.1449), and between *N. benthamiana* and *A. thaliana* ($F_{1,10}$ = *1.7189*, P = 0.2173). There as a marginal difference between *N. tabacum* and *A. thaliana* ($F_{1,10}$ = *7.8803*, P = 0.05129). We, therefore, find that the GF of CMV-I17F is host dependent, with *A. thaliana* having the most distinct GF.

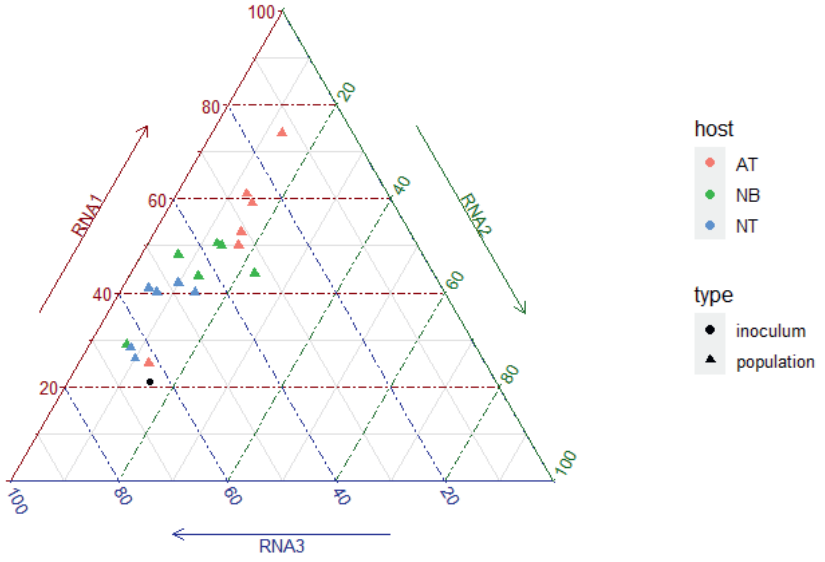


Figure 4. CMV-i17F has a host specific GF in A. thaliana, N. benthamiana and N. tabacum at p0. *GF of the p0 CMV-I17F inoculum (black circle), CMV infected A. thaliana (n=6, salmon triangle), N. benthamiana (n=6, green triangle) and N. tabacum (n=6, blue triangle) samples at 14 days post infection (dpi) at the initial passage stage (p0).*

Table 2. GF for CMV-i17F at initial passage (P0) in hosts species *A. thaliana*, *N. benthamiana* and *N. tabacum*. Numbers are mean ± standard deviation samples at 14 days post infection (dpi), for all hosts *n* = 6 per species. The GF is calculated as the mean value of the RNA segment relative to the sum of all RNA segments, RNA1 + RNA2 + RNA3 = 1.

Host	RNA1	RNA2	RNA3
Inoculum *(N. tabacum)*	0.21	0.15	0.63
A.thaliana	0.53 ± 0.16	0.14 ± 0.02	0.31 ± 0.16
N. benthamiana	0.44 ± 0.08	0.12 ± 0.06	0.43 ± 0.11
N. tabacum	0.36 ± 0.07	0.09 ± 0.03	0.54 ± 0.07

Whereas host-dependent differences in the GF have been reported for other plant viruses including FBNSV (Sicard et al. 2013) and AMV (Wu et al. 2017), Boezen, Johnson et al. (Boezen, Johnson, et al. 2023) reported no differences in the GF of CMV-i17F for the hosts *N. tabacum*, *N. benthamiana* and *Chenopodium quinoa*. In this study, CMV-I17F GF values in *N. tabacum* and *N. benthamiana* are similar to those previously described (Boezen, Johnson, et al. 2023). We did find a host-specific GF for CMV infection of *A. thaliana,* and our results therefore show that CMV's GF can vary over host species, as is the case for other multipartite

viruses. Nevertheless, in three host species CMV's GF is unchanged, and even in *A. thaliana*, the changes in CMV's GF are not as strong as those seen in AMV. For example, in *Capsicum annuum* and *Medicago sativa* RNA3 predominated, making up more than 80% of the AMV population (Wu et al. 2017).

Although more data over a broad range of hosts will be required as confirmation, we speculate that CMV may have a relatively stable GF over different host species, contributing to its extraordinary host range. Whereas the GF affords flexibility that may contribute to wide host ranges for multipartite viruses (Chapter 2), a degree of GF robustness may prevent extreme GF values with deleterious properties. Deleterious GF values could occur by short-sighted selection acting at one level of selection. For example, within-host selection on the GF could maximize rapid within-cell replication or between-cell spread, which are relevant fitness components as plant viruses typically display super-infection exclusion at the cellular level (Gutiérrez et al. 2015; Folimonova 2012). However, at higher levels of selection such as the between-organ and between-host level, these GF values may result in lower fitness because they have deleterious effects on fitness components that matter at this level of selection, such as virus titre and infectivity, the capacity of virus particles to cause infection. This trade-off between the effect of the GF at different levels of selection is plausible because any imbalance in the GF will always result in lower infectivity, provided all components are needed for infection and all virus particles have the same probability of entering a host cell (Sánchez-Navarro, Zwart, and Elena 2013) and Chapter 2 of this thesis. Therefore, GF-stabilizing mechanisms that prevent short-sighted, extreme GF values from occurring may be beneficial for enabling the virus population to expand within a novel host environment.

Experimental evolution of CMV in three hosts: extinctions and low virus titres

We investigated GF dynamics during serial passage in the three host species: *A. thaliana*, *N. benthamiana* and *N. tabacum.* We had six independent, replicate populations serially passaged five times in single infected plants. Per population, three to six plants were inoculated for each round of passaging to ensure having at least one infected plant. If for one population multiple plants tested positive for CMV infection by immunochromatography at the end of each passage, one infected plant was randomly chosen as the inoculum for the next round. One striking observation was the variability in the number of infected plants observed within and between species (Figure 5). Initial infectivity, the proportion of infected plants per passage, at passage 0 was variable across all three hosts. Furthermore, several populations went extinct before the final passage, namely 3 populations of *A. thaliana*, 2 populations of *N. benthamiana* and 1 population of *N. tabacum*. We considered populations to be extinctions after repeated 1 - 3 unsuccessful inoculations of the virus onto successive minimum of 6 host plants. *A. thaliana* was reinoculated without success for populations 1, 3 and 5 at passages 3, 5, and 4 onto 18, 6 and 23 plants, respectively. *N. benthamiana* populations 1 and 6 at passages 2 and 4 were reinoculated onto 19 and 12 plants, and *N. tabacum* population 6 at passage 2 was inoculated onto 18 plants without success. Two populations, population 4 of *A. thaliana* and population 5 of *N. tabacum*, proved difficult to passage because of low infectivity and were only successful at intermediate passages after several rounds of

inoculations of different plant cohorts. At the time of ending the experiment, these populations had not reached passage 5 and these populations are therefore considered incomplete but not extinct. Comparing the infectivity throughout the experiment, it appears that populations in *A. thaliana* and *N. tabacum* decline in their infectivity over the course of the experiment (Figure 6), whilst for *N. benthamiana* a slight increase can be observed. One key question that emerges from the passaging is why these extinctions occurred. Others have reported similar results: observed extinctions during serial passaging over 10 rounds with several CMV genotypes over 10 passages. Furthermore, they observed an increase in virus fitness in the restrictive plant host *Phaseolus vulgaris*, but not in the permissive hosts *C. sativus* and *S. lycopersicum* (Sacristán et al. 2005).

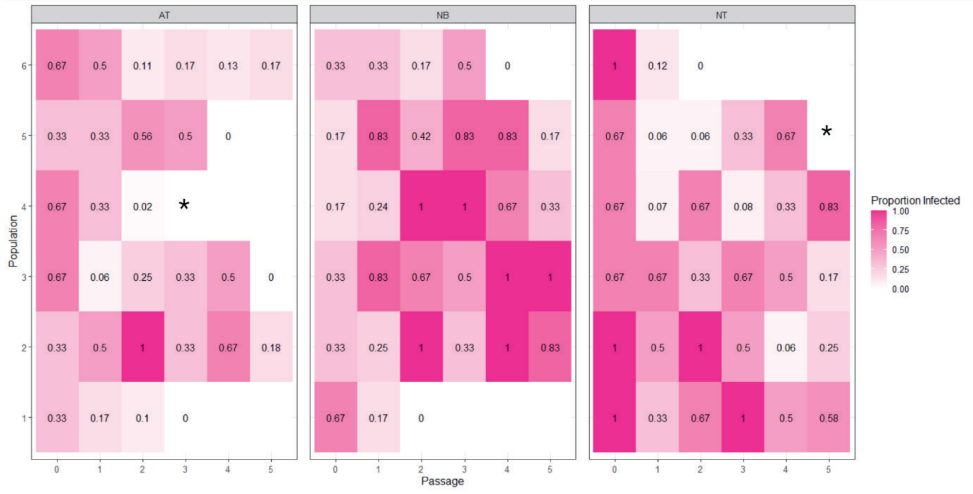


Figure 5. Proportion of positive CMV-i17F infected plants of A. thaliana, N. tabacum and N. benthamiana at each passage. *Populations which went extinct are marked with 0 to indicate at which passage round no positive infections were recorded. Three A. thaliana, two N. benthamiana and a single N. tabacum population went extinct during the course of the experiment. An asterisk indicates incomplete populations, which do not meet the criteria for extinction and did not reach p5.*

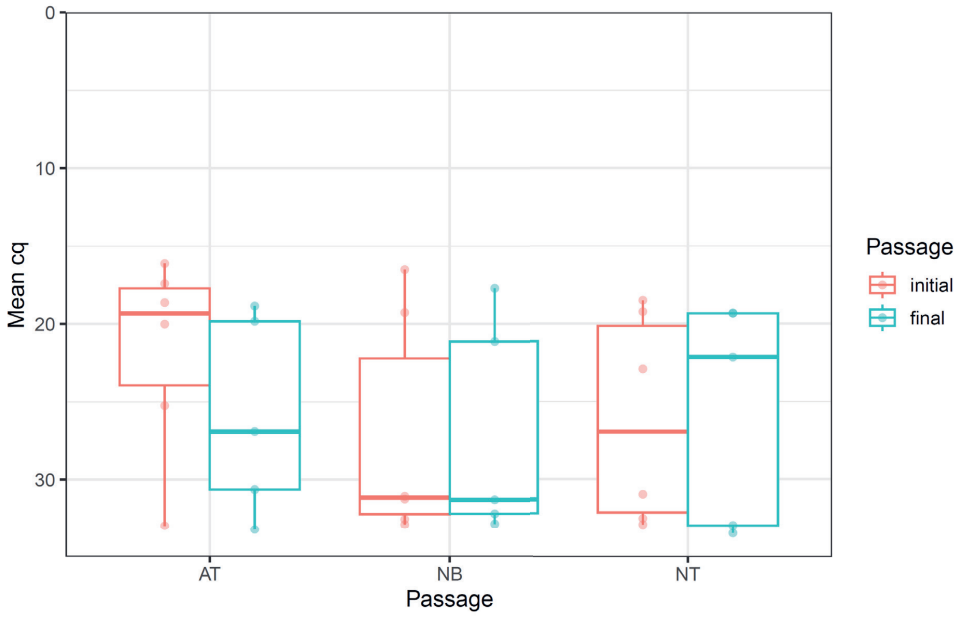


Figure 6. Virus titre remains stable from initial to final passage. *Virus titre is lower for populations which went extinct. For initial passage (p0) n=18, for final passage (p5 incl. extinctions) (n=16). At the final passage, n=6 populations were extinct. There were n=2 incompletely passaged populations, excluded from this analysis. On the y-axis is the mean RT-qPCR cycle quantification (Cq) for RNA1-3 per host at the initial passage and final passage per host, AT = A. thaliana, NB = N. benthamiana and NT = N. tabacum. We use Cq values as a proxy for virus titre; lower Cq values indicate higher virus titre.*

Next, we compared the virus titre for the initial and final timepoint for each population. Here, we consider the initial time point as passage 0. For the completed populations, we consider passage 5 as the final timepoint, whereas for the extinct populations, we consider the final passage in which CMV infection could be verified. Incomplete populations, i.e. those that we cannot consider extinct because insufficient attempts at infection were made, were excluded from the analysis. Overall, we find considerable variation in virus titre for both the initial and final timepoint, with some populations showing very low titre at either time point (Figure S1). We then considered titre at the final timepoint, separating the complete, incomplete and extinct populations. We see clear differences between the complete and extinct populations, with the extinct populations having significantly lower titre than the completed ones (Mann Whitney U Test: *W=46, P=0.02942*). Overall, extinction of populations was therefore associated with low virus titre (see also Table 4). Some of the completed populations do have low titres, but the collected tissue from passage 5 may also have very low infectivity as it was directly used for the analysis of segment accumulation and sequencing. Therefore, to fairly compare populations and to include information from as many populations as possible, we tried to account for the level of titre measured in the final passage rather than extinctions. Contrary to our expectations, there was no evidence of adaptive changes during the experiment beyond an initial host-dependent shift in the GF. Our experimental setup, therefore, does not permit us to explore the role of the GF in adaptive evolution. By contrast, we do want to understand what has caused low titre and the loss of infectivity, and in particular, whether changes in the GF could have been maladaptive. We have four hypotheses for why low titre occurred.

H0: Stochastic variation in virus titre. Infected plants show variation between plants in the levels of virus titre. The variation seen is not associated with changes in the properties of the virus we measure: the viral genome or the GF. We assume this stochastic variation arises from differences between plants, for example due to development of the immune system and plant size. Due to these chance effects, some populations with low titre will have a low probability of successful infection in the next round of passaging. Under this hypothesis, we expect to see only low virus titre and no further changes in the virus population.

H1: Deleterious GF changes in the absence of GF-affecting mutations. Here we speculate that the GF that maximises virus fitness is disrupted over time, in the absence of mutations that affect the GF. We see two mechanisms that could alter the GF in the absence of mutations, First, as discussed in the previous section, shorted-sighted selection on the GF at the within-host level could result in a GF that lowers fitness at the between-host level (i.e., infectivity). If selection on the GF plays a role, then there will be a degree of repeatability in the GF changes seen. Second, random changes in the GF could occur as a result of population bottlenecks, and these changes may lower the fitness of the population. We call these random changes GF drift. If GF drift cause deleterious GF changes, then we expect the repeatability of the GF changes over multiple virus populations to be less repeatable than in the case of the first selection-based hypothesis. Therefore, the degree of repeatability of the observed GF changes could suggest which of these two mechanisms is applicable.

H2: Deleterious mutations which do not influence the GF. Mutations that lower infectivity may increase in frequency in the population. As with the hypothesised GF changes, these mutations could become predominant due to genetic drift, mutation bias or as antagonistic pleiotropic side-effects of positive selection for traits that are only advantageous within the host. If such mutations are observed, it might be possible to identify the evolutionary force driving the occurrence of these mutations by considering their repeatability over populations, and the predicted type of mutation (i.e., loss-of-function or gain-of-function mutations). Under this hypothesis we will see mutations associated with low titre, but no appreciable differences in the GF.

H3: Mutations that predicate a deleterious GF shift. Under this hypothesis, *de novo* mutations in a population result in deleterious changes in the GF. As under *H2*, these mutations may fix for a variety of reasons, but the result is a deleterious GF shift. We expect to see high-frequency or fixed mutations in these populations, as well as a change in the GF.

Given the GF and mutation observations, we do not expect to find direct evidence for *H0*. Rather, given this set of four hypotheses, a lack of evidence for *H1, H2* and *H3* is considered evidence for *H0*. We acknowledge that the different hypotheses are not strictly mutually exclusive. For example, *H2* may result in a shift in the GF, but additional GF drift (i.e., *H1*) may be necessary to change the GF such that low titres occur. However, these hypotheses help us to consider the different data systematically, and thereby identify possible causes. Different explanations could also apply to different populations passaged in the same host, as all four hypotheses are based on stochastic processes. We therefore measured the GF for the evolved populations and sequenced the evolved genomes, to explore which of these four hypotheses has the most support.

Evolutionary changes in the GF

We first considered the GF of the evolved populations, at the final time point. PERMANOVA comparing the GF in the three host species showed a marginally significant effect of host on the GF (PERMANOVA: $F_{2,14}$ = 3.1707, P = 0.0359). The spread in the GF was not significantly different for different host species (PERMDISP2 test: $F_{2,14}$ =0.6439, P = 0.5367), confirming that the differences detected by the PERMANOVA concern the centroid of the GF. There are still significant differences in GF between species at the final timepoint. For all three species, GFs for the initial (p0) and final (p5) timepoints were roughly similar, both in terms of centroid and variance (Table S5; Figure 7). In *A. thaliana* and *N. benthamiana*, the GF does not appear to shift systematically (Figure 8). In contrast, the virus populations in *N. tabacum* appear to show some evidence of a directional shift in the GF, as the frequency of RNA2 increased in all populations (Figure 8).

To summarise the GF data and get an indication of whether GF changes contributed to low titre and extinction, we determined the Euclidean distance (*D*) between the mean GF value at p0 for a given host, and the GF for individual populations at p5 (Table 4). This value indicates whether the final GF deviates from the mean GF value for that species and is useful for identifying populations with a distinctive final GF value. The mean ±SEM of the GF distance *D* over all populations was 0.270 ±0.032, and the highest value in an individual high titre population was *D* = 0.38 *(N tabacum* population 4). Two low-titre populations had a *D* value higher than those seen in the high titre populations (*A. thaliana* population 1, *D* = 0.40; *N. benthamiana* population 2, *D* = 0.61; see Table 4), suggesting that the GF may have played a role in the extinction of these populations. Changes in the GF (*H1*) may therefore explain the low titres observed, but only for 2 out of 8 populations.

Mutations in the CMV genome

Genomics analysis of all the evolved CMV populations identified 53 *de novo* mutations. All mutations were single, or multi-nucleotide polymorphisms, and no large-scale changes were observed. We considered fixed (frequency > 95%), intermediate (> 5%) and low-frequency (> 1%) mutations (see Table S5) for data visualisation and subsequent analysis (Figure 9). For the fixed mutations, we identified common mutations present in two hosts. SNVs at position 2832 T>C were found in *A. thaliana* and *N. tabacum*, whilst a SNV located at position 2794 was only detected in population 3 of *N. benthamiana* (Figure 9). A non-synonymous mutation was found in the coding sequence of RNA2 in gene *2a*: 2017A>T and amino acid change T673S (Figure 9). This mutation was present in *A. thaliana* populations 3 and 5, and *N. tabacum* population 1, all of which went extinct.

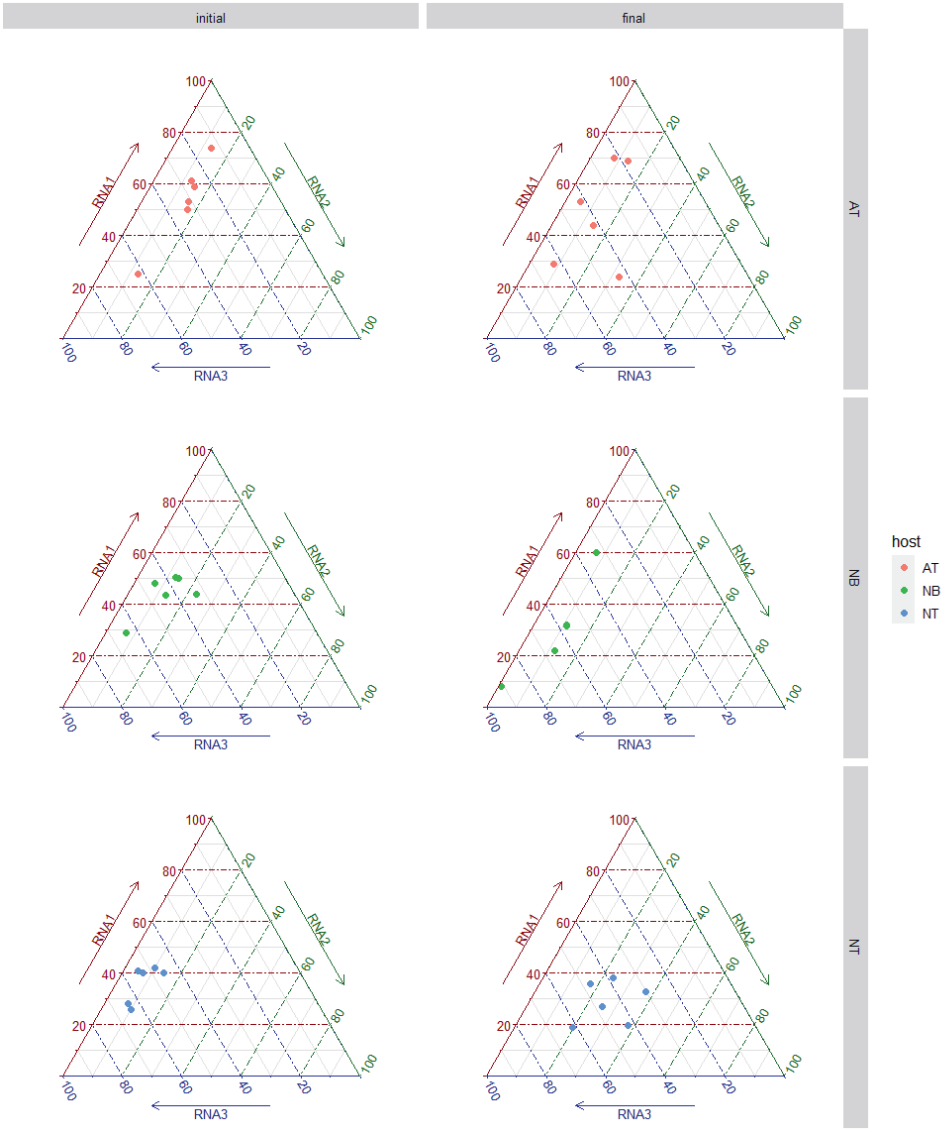


Figure 7. GF variation of CMV-i17F in host species A. thaliana, N. benthamiana and N. tabacum at 14 days post infection (dpi) before and after five successive serial passages.

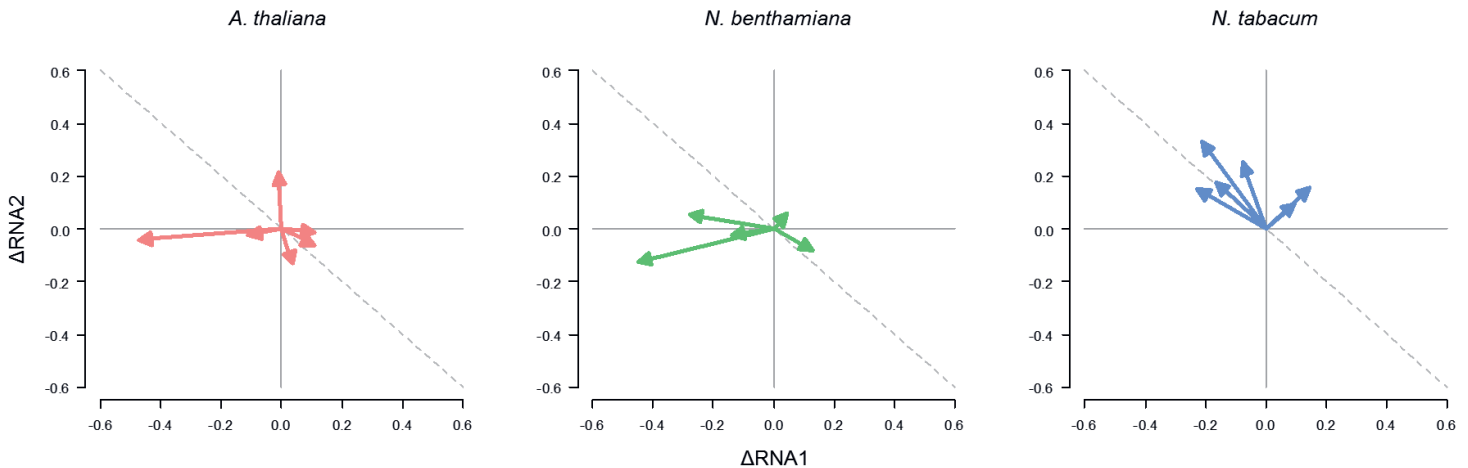


Figure 8. GF changes for each evolved population from initial to final passage. *For all three panels, the x-axis is the difference in the relative frequency of RNA1 between the first and final passage and the y-axis is the difference for RNA2. The arrow heads indicate the observed difference for each replicate, with a longer distance from the origin indicating a larger overall change. The distance between the arrow head and the dotted grey line (y = −x) indicates the difference in RNA3, and if the arrowhead is below the line RNA3 has decreased and vice versa. For A. thaliana and N. benthamiana, the magnitude of the change in GF is variable and its direction appears to be variable. For N. tabacum, the magnitude of GF change appears to be more constant and the direction of change is more similar, as RNA2 has increased for all populations.*

Overall, we see a considerable number of mutations that occur in two or more populations. We quantified this repeatability by estimating the *H*-index ($\underline{x} \pm SEM$) for the fixed (0.026 ±0.011), intermediate (0.053 ±0.022) and low-frequency mutations (0.034 ±0.016). These estimates confirm some parallel evolution occurs, but without similar estimates for other virus evolution studies, it is hard to draw conclusions from these measures. We also considered normalised substitution rates for non-synonymous (dN/dS) and intergenic (dI/dS) mutations. Synonymous mutations only occurred in the intermediate and low-frequency data, so we were able to calculate these indices only for these data (Intermediate mutations: dN/dS = 0.878, dI/dS = 14.756; low-frequency mutations: dN/dS = 0.512, dI/dS = 6.249). These results suggest neutral evolution in coding regions, as dN/dS ~ 1. In contrast, there was an over-representation of intergenic mutations, as dI/dS > 1. The mutation data are, therefore, characterised by modest levels of repeatability at the nucleotide level and at least a temporary excess of mutations in intergenic regions.

One striking pattern in the data is the repeated non-synonymous mutation T673S in coding region *2a*, the RNA-dependent RNA polymerase (RdRp), a mutation which occurs in three populations that went extinct. This was the only non-synonymous mutation that fixed (in *A. thaliana* populations 3 and 5) or was at intermediate frequency (in *N. benthamiana* population 1). The number of mutations detected per population was low (fixed: 0.28; intermediate frequency: 1.22; low frequency: 2.89). In the two populations that fixed T673S, the number of intermediate and low-frequency mutations was higher than in any other population (intermediate frequency, *A. thaliana* population 3 = 7, *A. thaliana* population 5 = 6; low frequency, *A. thaliana* population 3 = 12, *A. thaliana* population 5 = 25). The average dN/dS of all populations was ~ 1 for both intermediate and low-frequency mutations, and the estimates were dominated by mutations that occurred in these two populations. Therefore, we tentatively propose a trajectory in which there is fixation of a mutation in the RdRp, which may be followed by an excess of intermediate and low-frequency mutations that accumulated consistent with a neutral expectation. However, as only we have information on mutations at a single timepoint we cannot verify this sequence of events. These observations suggest genomes carrying T673S have become mutators. *N. benthamiana* population 1 also carries this mutation, but went extinct before this mutation went to fixation (mutation frequency = 74%), an explanation for the lack of other mutations detected in this population. The fact that T673S was present in three populations suggests that this mutation was under indirect positive selection, but also may have contributed to extinction of these populations, in support of *H2*. We speculate that the fitness effect of this mutation may be strongly context dependent, allowing it to be indirectly selected during within-host spread, while impeding between-host spread, perhaps by causing low titre due to an increased deleterious mutation load.

In summary, the genomics data suggest that mutations may play a role in the extinctions that occurred during passaging, as summarized in Table 4. The only non-synonymous mutation found at intermediate and high frequencies (i.e., 2017A>T in *2a*) was only found in extinct populations. By contrast, synonymous and intergenic mutations were found in six out of seven surviving, high titre populations, suggesting these mutations did not play a role in extinction. These data therefore provide support for *H2* in some populations, specifically for non-synonymous mutations. By contrast, there were no low titre populations with both appreciable GF changes and mutations, suggesting that there is no evidence for *H3.*

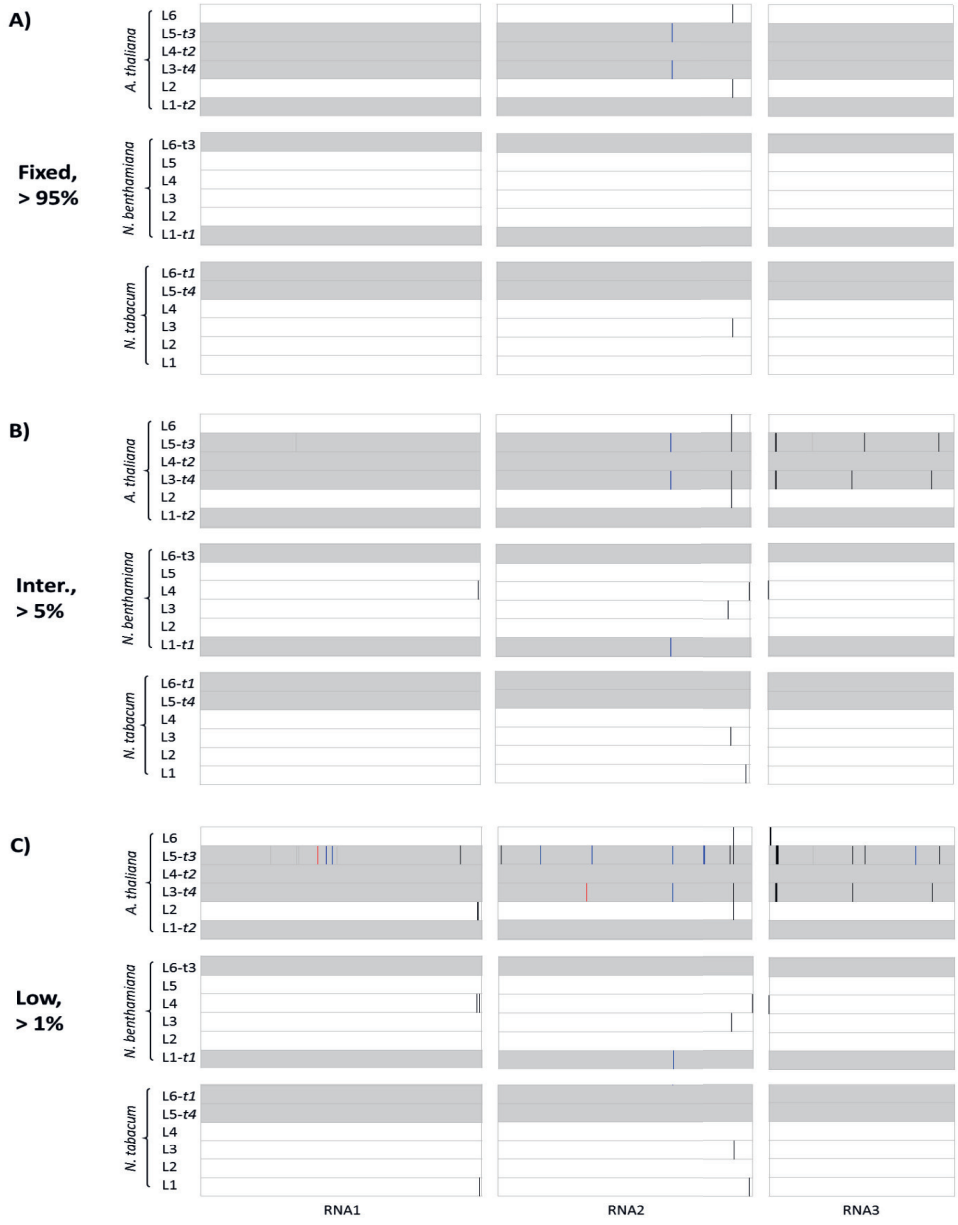


Figure 9. Mutations in CMV-i17F from evolved populations in A. thaliana, N. benthamiana and N. tabacum. *Mutations are shown that met criteria (see Table S5) for being fixed (panel A), intermediate frequency (panel B) or low frequency (panel C) are shown separately. Mutations are coded onto different CMV RNA segments as a single line along the genomic segment, with segment length oriented from starting position left to right. The colours of mutations indicate whether they are intergenic (black), synonymous (grey), non-synonymous (blue), or frameshifts or stop codons (red). Extinct populations are indicated in grey fill and the passage of extinction is indicated on the left.*

Effects of the genome formula and mutations on low titre

Finally, we considered both the GF and mutational data together in a quantitative analysis to determine how much support there is for the different hypotheses for extinction we have proposed *(H0* to *H3*). We performed a logistic regression to explain low titre and included the following explanatory factors: (1) the genome formula distance *D* to the mean GF at p0, as a test of *H1*, (2) the presence or absence of non-synonymous mutations as binary variable, as a test of *H2*, and (3) the interaction between *D* and the presence of non-synonymous mutations, as a preliminary indication of *H3*. We found marginally significant effects for both non-synonymous mutations ($\chi^2_{1,13}$ = 4.427, *P* = 0.018) and *D* ($\chi^2_{1,12}$ = 3,270, *P* = 0.035), whereas their interaction was not significant ($\chi^2_{1,11}$ < 0.001, *P* ~ 0.5). This analysis therefore provides preliminary support for *H1* and *H2*, whilst there is no support for *H3*. From a qualitative perspective, *H0* also has support because there are numerous extinct populations in which neither mutations nor GF changes occurred. As GF changes were implicated in extinction in only two lineages, we cannot ascertain whether these GF changes are repeatable, precluding firm conclusions on whether selection on the GF or GF drift cause these changes. However, the fact that the GF was stable over passages for most populations suggests that GF drift is the more likely explanation. In sum, we can conclude that stochastic events leading to low titres (*H0*), genome formula changes (*H1*) and mutation (*H2*) are likely to have played a role in the extinctions we have observed.

Concluding Remarks

Our results indicate that the CMV GF may depend on virus genotype, as there appear to be GF differences between CMV subgroups 1 and 2. In contrast to a previous report (Boezen, Johnson, et al. 2023), we show that the CMV GF does depend on host species. In particular, in *A. thaliana* we see a systematic shift in the GF after 2 weeks of infection, following infection with an inoculum from *N. tabacum*. We then evolved CMV in three host species, to explore the evolutionary dynamics of the GF and the interplay with mutations in the viral genome. We were surprised to find that serial passage of this virus proved difficult, and we consider 6 out of 18 populations to have gone extinct. These extinctions were marked by low virus titres, which we also saw in the final passage of 2 out of 9 populations for which five passages were completed. We therefore set out to understand why low virus titre and extinctions occurred during this experiment, by looking at changes in the GF and mutations in the viral genome. We hypothesised that low virus titres could occur due to stochastic demographic processes in individual plants (*H0*), deleterious GF shifts as a consequence of GF drift or within-host selection on the GF (*H1*), deleterious mutations affecting viral infectivity or yield that are unrelated to the GF (*H2*), and mutations that are deleterious because they cause a shift in the GF (*H3*).

We combined all the results obtained and considered what evidence there is for these three hypotheses, from a qualitative overview of the data (Table 4) and by logistic regression. The first conclusion we can draw is that there are clear differences between populations, as some populations provide support for *H0* (1 population), *H1* (2 populations) and *H2* (3 populations). We have not provided direct experimental evidence for fitness changes related to the GF or mutations, but rather found associations with low titre. Second, the same mutation (T673S in

genome segment *2a*) was found in three populations that went extinct. This mutation that was not detected in the ancestral population and appears to result in a mutator phenotype in the two populations in which it was fixed. This was the only non-synonymous mutation that occurred at intermediate and high frequencies, and the only mutation clearly associated with low titre. Some repeated mutations occurred in untranslated regions (3' UTR of RNA2), but they do not appear to affect titre as they are found in many high-titre populations. Third, changes in the genome formula that lead to high levels of CMV RNA3 (RNA3 > 0.6) are associated with low titre. Most completed populations with high titres have a more balanced final GF (RNA3 < 0.5), although one population did have an RNA3 level of 0.58. Therefore, the GF appears to play a role in some of these extinctions, although it appears to have remained stable in the majority of populations.

There are many aspects to this study that could be improved in future work. One major caveat is that we did not perform fitness assays, but simply measured titre during the serial passages. We chose this approach because a fitness assay would have required infecting new plants. Infecting these plants would have induced additional GF variation, precluding a direct link between the passaged population and fitness. Furthermore, fitness measurements are ideally performed by direct competitions with the ancestor (Zwart et al. 2014). In our case, the presence of the ancestral virus in a mixed population would further disrupt the GF and likely result in re-assorted subpopulations in the infected plants. Setting up fitness assays for experimental evolution with multipartite viruses remains challenging.

We set out to look for evidence that the GF may play a role in adaptive evolution of CMV in three plant species. However, our results suggest that shifts in the GF could be unrelated to adaptation or may even be deleterious and contribute to the extinction of virus populations. The cost to infectivity associated with a multipartite organisation has been described before (Fulton 1962; Iranzo and Manrubia 2012; Sánchez-Navarro, Zwart, and Elena 2013) and Chapter 2 of this thesis, but to our knowledge the deleterious consequences of GF shifts have not been described before. Whereas previous studies have highlighted the potential benefits of GF variation (Sicard et al. 2016; Zwart and Elena 2020), our results here highlight that GF may be costly and lead to the extinction of virus populations.

Table 4: Overview of results for complete and extinct populations (incomplete excl.). *N.benthamiana* population 5 has been excl. as the GF could not be measured for this population.

Fate[a]	Host	Population	Final passage	Relative log titre	GF distance[b]	Mutations		Hypothesis supported[f]			
						NS[c]	All[d]	H0	H1	H2	H3
Complete	*A.thaliana*	2	5	0.57	0.15	0	1				
		6	5	0.87	0.20	0	1				
	N. benthamiana	3	5	2.87	0.19	0	2				
		4	5	1.84	0.20	0	2				
	N. tabacum	1	5	1.21	0.29	0	1				
		3	5	2.06	0.20	0	1				
		4	5	2.07	0.38	0	0				
Low titre	*N. benthamiana*	2	5	-1.22	0.61	0	0		✓		
	N. tabacum	2	5	-2.19	0.32	0	0	✓			

Extinct	*A.thaliana*	1	2	-3.46	0.40	0	0		✓		
		3	4	-1.56	0.20	1	6			✓	
		5	3	-2.68	0.14	1	10			✓	
	N. benthamiana	1	1	-1.49	0.32	1	0			✓	
		6	3	-1.69	0.18	0	0	✓			
	N. tabacum	6	1	-2.05	0.26	0	0	✓			

[a] Indicates what happened to the population, with low titre indicating the final passage was completed but with a low measured titre. [b] The Euclidean distance between the mean GF at P0 for the respective host and GF in individual populations at P5 is given, to identify distinct GF values. [c] The number of non-synonymous mutations with a frequency > 5%. [d] All other mutations: synonymous and intergenic mutations with a frequency > 5%. [f] Check mark indicates support for a hypothesis on why extinction has occurred. We considered GF distance values greater than the highest value in the complete populations (0.38) as evidence for *H1*, the evidence of non-synonuymous mutattions as evidence for *H2*, and when both criteria are met there is evidence for *H3*. If there is no evidence for *H1-H3,* we consider this evidence for *H0.*

Supplementary 1

Table S1. Cucumber mosaic virus isolates from the virus collection at Wageningen Plant Research, (www.primediagnostics.com) Wageningen University and Research, Wageningen, The Netherlands

Subgroup	Isolate	Country	Host	Year	Reference
I	B32	Spain	*Phaseolus vulgaris*	1970	(Bos and Maat 1974)
	Fulton	France	*Cucumis sativus*	1992	NA
	i17F	France	*Solanum lycopersicum*	1975*	(Jacquemond and Lot 1981; Quiot et al. 1979; Nono-Womdim, Marchoux, and Gebre-Selassie 1991)
	S4	Hungary	*Cucumis melo*	1982	(Tóbiás, Maat, and Huttinga 1982)
	T519	Italy	*Solanum lycopersicum*	1998	NA
II	AGO7	The Netherlands	*Apium graveolens*	1971	(Bos 1973)
	CU68	Unknown	*Capsicum annum*	NA	NA
	E312	The Netherlands	*Pisum sativum*	NA	NA
	K8	Hungary	*Capsicum annum*	1982	(Tóbiás, Maat, and Huttinga 1982)

*Isolated from field surveys in the south of France and placed in INRAE collection (Quiot et al. 1979).

Table S2. Overview of RAW total number of reads per CMV isolate generated from RNA sequencing (Miseq)

Subgroup	CMV Isolate	Total Number of reads	Yield (mbp)	Average quality (Phred)	Mapped reads
Negative control		13,957,415	4126	36.24	n/a
SI	B32	10,357,597	3071	36.04	1,354,161
	Fulton	36,923,059	10879	35.97	8,648,711
	I17F	27,152,250	7941	36.08	10,679,111
	S4	14,224,810	4147	36.05	11,668,569
	T519	27,982,524	8252	36.06	11,834,974
SII	AGO7	11,471,307	3359	36.02	6,451,289
	CU68	12,135,304	3598	36.05	87,448
	E312	46,815,507	13891	35.91	25,086,989
	K8	13,683,817	4043	36.01	899,825

Table S3.GF of individual CMV isolates from subgroup I and II.

Subgroup	Isolate	RNA1	RNA2	RNA3
I	B32	0.34	0.11	0.55
	Fulton	0.24	0.29	0.47
	i17F	0.48	0.22	0.31
	S4	0.26	0.43	0.30
	T519	0.42	0.21	0.37
II	AGO7	0.16	0.43	0.41
	CU68	0.29	0.45	0.26
	E312	0.17	0.45	0.38
	K8	0.18	0.17	0.65
Mean GF SI*		0.35 ± 0.10	0.25 ± 0.12	0.40 ± 0.10
Mean GF SII*		0.21 ± 0.06	0.37 ± 0.14	0.42 ± 0.16

*Mean GF per subgroup ($\bar{x} \pm SD$), subgroup I, *n= 5* and subgroup II, *n=4* ($\bar{x} \pm SD$)

Table S4. Criteria for filtering mutations: Mutations that pass through the filtering in the ancestral population, are excluded from analyses on the evolved populations.

Group	Filtering of ancestral population			Filtering of evolved population		
	Coverage	Reads	Frequency	Coverage	Reads	Frequency
Fixed	>50	>5	>1.0%	>100	>10	>95.0%
Intermediate	>50	>5	>1.0%	>100	>10	>5.0%
Low	>50	>5	>0.2%	>100	>10	>1.0%

Coverage indicates the total coverage on the position of the mutation. Reads indicates the number of reads containing the mutation. Frequency indicates the frequency of the mutation, expressed as a percentage (i.e., A fixed mutation is 100%).

Table S5. GF of CMV-i17F at final passage in hosts species *A. thaliana n* = 6, *N. benthamiana n* = 5 and *N. tabacum n* = 6 at 14 days post infection (dpi). The GF is calculated as the mean value of the RNA segment relative to the sum of all RNA segments,
RNA1 + RNA2 + RNA3 = 1.

Host	RNA1	RNA2	RNA3
A. thaliana	0.48 ± 0.19	0.14 ± 0.10	0.36 ± 0.16
N. benthamiana	0.31 ± 0.19	0.08 ± 0.05	0.61 ± 0.21
N. tabacum	0.29 ± 0.08	0.27 ± 0.09	0.44 ± 0.10

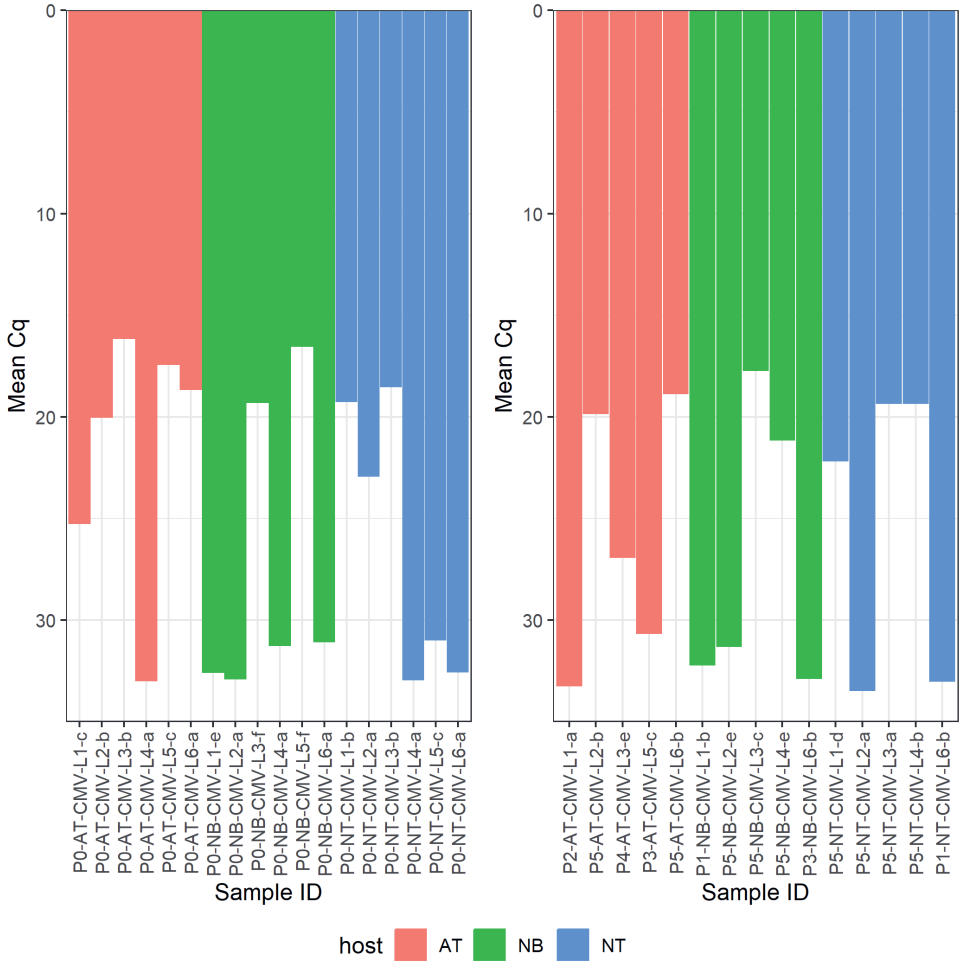


Figure S1. CMV titre variation per population at the initial timepoint (p0) and final timepoint (p5). *Incomplete lineages n=2 are excluded from this analysis. AT= A. thaliana, NB= N. bethamiana, NT= N. tabacum. On the x-axis are virus populations (numbered 1-6 per host) at the initial timepoint (passage 0, p0) and the final timepoint (passage 5, p5), or the last passage before extinction. On the y-axis is the mean RT-qPCR cycle quantification (Cq) for RNA1-3 per virus population at the initial passage and final passage per host. We use Cq values as a proxy for virus titre as they are inversely related; lower Cq values indicate higher virus titre.*

References

Anderson, Marti J. 2017. “Permutational Multivariate Analysis of Variance (PERMANOVA).” In *Wiley StatsRef: Statistics Reference Online*, 1–15. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat07841.

Bayer, Avraham, Greg Brennan, and Adam P. Geballe. 2018. “Adaptation by Copy Number Variation in Monopartite Viruses.” *Current Opinion in Virology*. Elsevier. https://doi.org/10.1016/j.coviro.2018.07.001.

Betancourt, Monica, Aurora Fraile, and Fernando Garcia-Arenal. 2011. “Cucumber Mosaic Virus Satellite RNAs That Induce Similar Symptoms in Melon Plants Show Large Differences in Fitness.” *The Journal of General Virology* 92 (8): 1930–38.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064.

Boezen, Dieke, Maritta Vermeulen, Marcelle Johnson, Rene Van Der Vlugt, Carolyn Malmstrom, and Mark Zwart. 2023. “Mixed Viral Infection Constrains the Genome Formula of Multipartite Cucumber Mosaic Virus.” *Frontiers in Virology* 3. https://doi.org/10.3389/fviro.2023.1225818.

Bos, L. 1973. “Identificatie van Enkele Virussen in Knolselderij.” *Gewasbescherming* 4: 15.

Bos, L., and D. Z. Maat. 1974. “A Strain of Cucumber Mosaic Virus, Seed-Transmitted in Beans.” *Netherlands Journal of Plant Pathology* 80 (4): 113–23.

Cone, Kelsey R., Zev N. Kronenberg, Mark Yandell, and Nels C. Elde. 2017. “Emergence of a Viral RNA Polymerase Variant during Gene Copy Number Amplification Promotes Rapid Evolution of Vaccinia Virus.” *Journal of Virology* 91 (e0142B-16). https://doi.org/10.1128/JVI.01428-16.

Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers, Jay Shendure, Adam P. Geballe, and Harmit S. Malik. 2012. “Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses.” *Cell* 150 (4): 831–41.

Feng, Junli, Leiyu Lai, Ruohong Lin, Chunzhi Jin, and Jishuang Chen. 2012. “Differential Effects of Cucumber Mosaic Virus Satellite RNAs in the Perturbation of MicroRNA-Regulated Gene Expression in Tomato.” *Molecular Biology Reports* 39: 775–84.

Folimonova, Svetlana Y. 2012. “Superinfection Exclusion Is an Active Virus-Controlled Function That Requires a Specific Viral Protein.” *Journal of Virology* 86 (10): 5554–61.

Fraile, Aurora, José Luis Alonso-prados, Miguel A. Aranda, Juan J. Bernal, José M. Malpica, and Fernando Garci. 1997. “Genetic Exchange by Recombination or Reassortment Is Infrequent in Natural Populations of a Tripartite RNA Plant Virus.” *Journal of Virology* 71 (2): 934–40.

Fulton, Robert W. 1962. “The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus.” *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.

Gutiérrez, Serafín, Elodie Pirolles, Michel Yvon, Volker Baecker, Yannis Michalakis, and Stéphane Blanc. 2015. “The Multiplicity of Cellular Infection Changes Depending on the Route of Cell Infection in a Plant Virus.” *Journal of Virology* 89 (18): 9665–75.

Hamilton, Nicholas E., and Michael Ferry. 2018. “Ggtern: Ternary Diagrams Using Ggplot2.” *Journal of Statistical Software* 87 (December): 1–17.

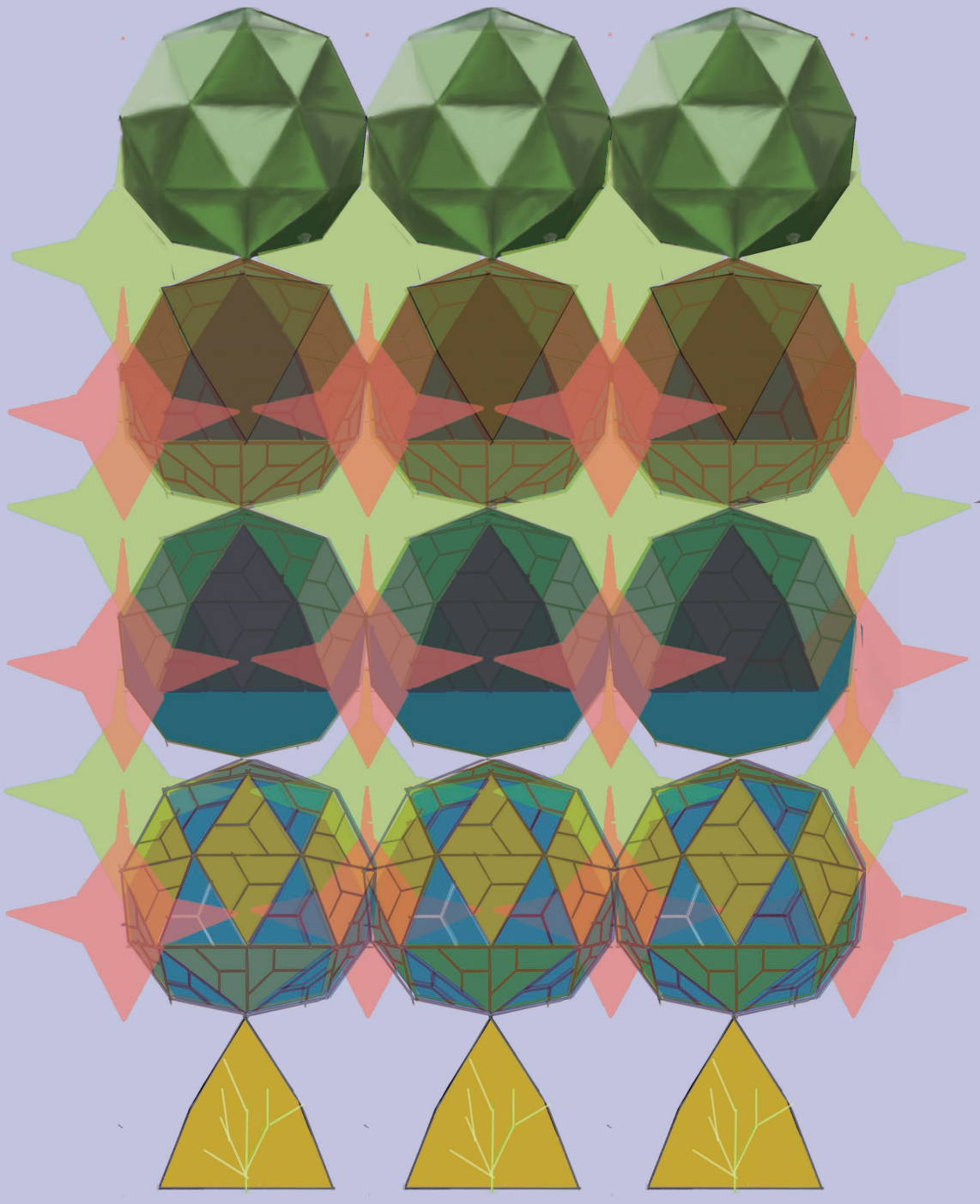
Holm, Sture. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics, Theory and Applications* 6 (2): 65–70.

Hu, Zhaoyang, Xiaolong Zhang, Wei Liu, Qian Zhou, Qing Zhang, Guohui Li, and Qin Yao. 2016. “Genome Segments Accumulate with Different Frequencies in Bombyx Mori Bidensovirus.” *Journal of Basic Microbiology* 56 (12): 1338–43.

Iranzo, Jaime, and Susanna C. Manrubia. 2012. “Evolutionary Dynamics of Genome Segmentation in Multipartite Viruses.” *Proceedings of the Royal Society B: Biological*

Sciences 279 (1743): 3812–19.
Jacquemond, Mireille, and Herve Lot. 1981. “L’ARN Satellite Du Virus de La Mosaïque Du Concombre I. - Comparaison de l’aptitude à Induire La Nécrose de La Tomate d’ARN Satellites Isolés de Plusieurs Souches Du Virus.” *Agronomie* 1 (10): 927–32.
Liefting, Lia W., David W. Waite, and Jeremy R. Thompson. 2021. “Application of Oxford Nanopore Technology to Plant Virus Detection.” *Viruses* 13 (8). https://doi.org/10.3390/v13081424.
Michalakis, Yannis, and Stéphane Blanc. 2020. “The Curious Strategy of Multipartite Viruses.” *Annual Review of Virology* 7 (1): 203–18.
Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. “The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner.” *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.
Mware, Benard Ouma. 2016. “Development of Banana Bunchy Top Virus Resistance in Bananas: RNAi Approach.” *PhD, Queensland University of Technology*. https://core.ac.uk/download/pdf/78100860.pdf.
Näsvall, Joakim, Lei Sun, John R. Roth, and Dan I. Andersson. 2012. “Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence.” *Science* 338 (6105): 384–87.
Nono-Womdim, R., G. Marchoux, and K. Gebre-Selassie. 1991. “Application Des Méthodes DAS et PAS ELISA Pour La Détection de Plusieurs Virus Infectant Le Piment.” *Phytopathologia Mediterranea* 30 (1): 14–22.
Nouri, Shahideh, Rafael Arevalo, Bryce W. Falk, and Russell L. Groves. 2014. “Genetic Structure and Molecular Variability of Cucumber Mosaic Virus Isolates in the United States.” *PloS One* 9 (5): 96582.
Ohshima, Kazusato, Kosuke Matsumoto, Ryosuke Yasaka, Mai Nishiyama, Kenta Soejima, Savas Korkmaz, Simon Y. W. Ho, Adrian J. Gibbs, and Minoru Takeshita. 2016. “Temporal Analysis of Reassortment and Molecular Evolution of Cucumber Mosaic Virus: Extra Clues from Its Segmented Genome.” *Virology* 487: 188–97.
Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O’hara, et al. 2022. *Vegan: Community Ecology Package* (version 2.6.4). Vegan: Community Ecology Package. https://cran.r-project.org/web/packages/vegan/index.html.
Ouedraogo, R. S., and M. J. Roossinck. 2019. “Chapter 18: Molecular Evolution.” In *Cucumber Mosaic Virus*, edited by Peter Palukaitis and Fernando García-Arenal, 207–15. Virology. The American Phytopathological Society.
Quiot, J. B., G. Marchoux, L. Douine, A. Vigouroux, and Others. 1979. “Ecology and Epidemiology of Cucumber Mosaic Virus in Southeast France. V. Role of Weeds in the Survival of the Virus.” In *Annales de Phytopathologie*, 11:325–48. Institut National de la Recherche Agronomique.
R Foundation for Statistical Computing. 2023. *R: A Language and Environment for Statistical Computing* (version 4.3.1). Vienna, Austria. https://www.r-project.org/.
Riehle, M. M., A. F. Bennett, and A. D. Long. 2001. “Genetic Architecture of Thermal Adaptation in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 525–30.
Roenhorst, Annelien. 2014. “Protocol for Mechanical Inoculation of Test Plants.” The Netherlands: National Plant Protection Organization.
Roossinck, M. J. 2001. “Cucumber Mosaic Virus, a Model for RNA Virus Evolution.” *Molecular Plant Pathology* 2 (2): 59–63.
Roossinck, Marilyn J. 2002. “Evolutionary History of Cucumber Mosaic Virus Deduced by Phylogenetic Analyses.” *Journal of Virology* 76 (7): 3382–87.
Sacristán, Soledad, Aurora Fraile, José M. Malpica, and Fernando García-Arenal. 2005. “An Analysis of Host Adaptation and Its Relationship with Virulence in Cucumber Mosaic Virus.” *Phytopathology* 95 (7): 827.
Sánchez-Navarro, Jesús A., Mark P. Zwart, and Santiago F. Elena. 2013. “Effects of the

Number of Genome Segments on Primary and Systemic Infections with a Multipartite Plant RNA Virus." *Journal of Virology* 87 (19): 10805–15.
Sandegren, Linus, and Dan I. Andersson. 2009. "Bacterial Gene Amplification: Implications for the Evolution of Antibiotic Resistance." *Nature Reviews. Microbiology* 7 (8): 578–88.
Schenk, Martijn F., Mark P. Zwart, Sungmin Hwang, Philip Ruelens, Edouard Severing, Joachim Krug, and J. Arjan G. M. de Visser. 2022. "Population Size Mediates the Contribution of High-Rate and Large-Benefit Mutations to Parallel Evolution." *Nature Ecology & Evolution* 6 (4): 439–47.
Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. "The Strange Lifestyle of Multipartite Viruses." Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819.
Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. "Gene Copy Number Is Differentially Regulated in a Multipartite Virus." *Nature Communications* 4: 2248.
Tóbiás, I., D. Z. Maat, and H. Huttinga. 1982. "Two Hungarian Isolates of Cucumber Mosaic Virus from Sweet Pepper (Capsicum Annuum) and Melon (Cucumis Melo): Identification and Antiserum Preparation." *Netherlands Journal of Plant Pathology* 88 (5): 171–83.
Tomanek, I., R. Grah, M. Lagator, A. M. C. Andersson, J. P. Bollback, G. Tkačik, and C. C. Guet. 2020. "Gene Amplification as a Form of Population-Level Gene Expression Regulation." *Nature Ecology & Evolution* 4 (4): 612–25.
Tomanek, Isabella, and Călin C. Guet. 2022. "Adaptation Dynamics between Copy-Number and Point Mutations." *ELife* 11 (December). https://doi.org/10.7554/eLife.82240.
Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. "Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus." *Scientific Reports* 7 (1): 1–15.
Yu, Nai-Tong, Hui-Min Xie, Yu-Liang Zhang, Jian-Hua Wang, Zhongguo Xiong, and Zhi-Xin Liu. 2019. "Independent Modulation of Individual Genomic Component Transcription and a Cis-Acting Element Related to High Transcriptional Activity in a Multipartite DNA Virus." *BMC Genomics* 20 (1): 573.
Zhao, Wan, Qianshuo Wang, Zhongtian Xu, Renyi Liu, and Feng Cui. 2019. "Distinct Replication and Gene Expression Strategies of the Rice Stripe Virus in Vector Insects and Host Plants." *The Journal of General Virology* 100 (5): 877–88.
Zimmer, Christoph T., William T. Garrood, Kumar Saurabh Singh, Emma Randall, Bettina Lueke, Oliver Gutbrod, Svend Matthiesen, et al. 2018. "Neofunctionalization of Duplicated P450 Genes Drives the Evolution of Insecticide Resistance in the Brown Planthopper." *Current Biology: CB* 28 (2): 268-274.e5.
Zwart, Mark P., Ghulam Ali, Elisabeth A. van Strien, Elio G. W. M. Schijlen, Manli Wang, Wopke van der Werf, and Just M. Vlak. 2019. "Identification of Loci Associated with Enhanced Virulence in Spodoptera Litura Nucleopolyhedrovirus Isolates Using Deep Sequencing." *Viruses* 11 (9). https://doi.org/10.3390/v11090872.
Zwart, Mark P., and Santiago F. Elena. 2020. "Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments." *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa022.
Zwart, Mark P., Anouk Willemsen, José Antonio Daròs, and Santiago F. Elena. 2014. "Experimental Evolution of Pseudogenization and Gene Loss in a Plant RNA Virus." *Molecular Biology and Evolution* 31 (1): 121–34.



Predicting the impact of virion architecture on the infectivity and evolution of segmented viruses

Marcelle L. Johnson[1,2], Erick Bermùdez-Méndez[3], Paul J. Wichgers Schreur[3], René A.A. van der Vlugt[2], J. Arjan G.M. de Visser[4], Mark P. Zwart[1]

[1] Netherlands Institute of Ecology (NIOO-KNAW), PO Box 50, 6700 AB, Wageningen, The Netherlands

[2] Laboratory of Virology, Wageningen University and Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

[3] Department of Virology and Molecular Biology, Wageningen Bioveterinary Research, Lelystad, The Netherlands

[4] Laboratory of Genetics, Wageningen University and Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

Submitted

Abstract

Viruses vary enormously in the large-scale organization of their complete and infectious extracellular particles, their virion architecture, which we here designate as types I-VI. Monopartite viruses (type I) encapsidate a single genome segment in a single particle. Multipartite viruses (type II) package their multiple genome segments individually, whilst selectively packaging segmented viruses (type III) package a complete set of their multiple genome segments into each virus particle. By contrast, non-selectively packaging segmented viruses show variation in the distribution of genome segments over virus particles. Some animal and plant viruses are thought to non-selectively package their genome segments, although to date the most convincing data have been obtained for animal viruses like Rift Valley fever virus (RVFV). For these viruses, the exact distribution of genome segments over virus particles is not known, and we, therefore, postulate architectures with variation in the identity of segments packaged (type IV), the number of segments packaged (type V), or both (type VI). In this study, we use mathematical models to explore the impact of virion architecture on virus infectivity and evolution, focusing on segmented viruses (types III-VI). First, we compare predicted infectivity of different genome architectures by considering the integral of dose-responses. For the non-selectively packaging viruses, types IV and VI had the lowest infectivity, suggesting that variation in the identity of the segments packaged has a higher cost than variation in the number of segments. Second, we obtain infectivity's when comparing predictions based on empirical data for RVFV. For this virus, the predicted cost to the infectivity of types IV and VI sometimes approached that of a multipartite virus (type II), whereas the type V had an appreciably lower cost. Finally, we simulate competitions between type III, IV, and V architectures, to explore whether benefits associated with changes in the frequency of genome segments can outweigh infectivity cost. Such benefits may arise from tuning gene expression to differing requirements imposed by the host. Types IV and V could outcompete type III under a broad range of conditions, suggesting that despite their lower infectivity non-selectively packaging viruses could be viable competitors.

Introduction

Viruses are obligate intracellular micro-parasites. To move between cells and hosts, most viruses produce virus particles: the viral hereditary material surrounded by a capsid, which in turn may be enclosed by a membrane in some viruses. The hereditary material of a virus can consist of RNA or DNA, and the number of viral genome segments varies between 1 and 12 (Gelderblom 1996) For a successful infection of a new host, most or all of the genome segments must be transmitted between hosts. The complete and infectious extracellular form of a virus is referred to as the virion. We refer to the gross organization of the virion, specifically the distribution of viral genome segments over virus particles, as the virion architecture. Viruses with multiple genome segments have different strategies to package these segments into virus particles. Classically, three virion architectures were recognized (Sicard et al. 2016). Monopartite viruses have a single genome segment, and therefore each virus particle with a (full-length) genome segment contains the complete hereditary information. Multipartite viruses have multiple genome segments and package them into separate virus particles, and therefore multiple virus particles are needed to transfer the complete hereditary information. Segmented viruses have multiple genome segments, but package a copy of each segment into every virus particle, and hence each virus particle contains the complete hereditary material. The distribution of virion architectures over host species is non-uniform, as shown in a recent overview (Michalakis and Blanc 2020). Monopartite viruses represent the largest group, infecting predominantly animals, plants, and bacteria. Segmented viruses infect mainly animal hosts (including insects), although there are also a considerable number of segmented plant and fungal viruses. Multipartite viruses are found nearly exclusively infecting plants and fungi (Michalakis and Blanc 2020). In this classical perspective, segmented viruses combine potential benefits of having multiple genome segments, such as the possibility for reassortment (Chao, Tran, and Tran 1997) and faster replication (Pressing and Reanney 1984), with the high infectivity of monopartite viruses. Faster replication would occur with smaller segment sizes when the availability of the replication complex is not a limiting factor (Sicard et al. 2016).

While these categories of virion architecture have been very useful for understanding viral diversity, in the past decade the lines between these categories have started to blur (Koonin et al. 2020). Multipartite viruses have been shown to have unbalanced genome-segment frequencies, and there are indications that this highly variable “genome formula” is an adaptive mechanism, maintaining viral mRNA levels in different host environments (Sicard et al. 2013; Gallet et al. 2022). The putative benefit of changes in the genome formula is the rapid regulation of gene expression in host environments demanding different levels of virus gene expression (Sicard et al. 2013). However, some segmented viruses can also have an unbalanced genome formula. Examples of these segmented viruses include the animal viruses Influenza A virus (IAV) (Brooke et al. 2013), bluetongue virus (BTV) (Moreau et al. 2020) and Rift Valley fever virus (RVFV) (Wichgers-Schreur et al. 2016), and the plant virus tomato spotted wilt virus (TSWV) (Kormelink et al. 1992). If some genome segments are more abundant than others, this may lead to the formation of incomplete virus particles. In some cases, including IAV and RVFV, virus populations with high frequencies of incomplete virus particles exist, and these incomplete particles contribute to transmission by complementing each other to introduce a complete genome (Bermúdez-Méndez et al. 2022). These segmented viruses do produce some complete virus particles, but the lower capacity for

transmission by incomplete particles makes them similar to multipartite viruses. Thus, these are expected to have lower infectivity than monopartite viruses but higher than that of multipartite viruses.

The exact distribution of genome segments over virus particles has not been described for most segmented viruses. Even in those rare cases when the identity of the segments present in virus particles has been carefully documented, the number of segment copies has not (Wichgers Schreur and Kortekaas 2016). Fluorescent *in situ* hybridization (FISH) may determine that only the *S* segment of RVFV is present in a virus particle, but not whether there are multiple copies of the segment present. Given the sparsity of information on segmented virus virion architectures and our intuition that these distributions will matter for infectivity, we set out to address this question with a modeling approach.

Virion architecture types

Before we explain the approach in detail, we first describe the six different types of virion architectures we will consider here, which we term types I-VI. An overview of the genome architectures we are considering is given in Figure 1.

Type I is the monopartite virion architecture: viruses with a single genome segment, each virus particle contains a full-length genome and is infectious. This is the most common virion architecture (Michalakis and Blanc 2020). Subtype IA is the multicopy-genome virion architecture. In some cases, viruses package multiple copies of their single genome segment into each virus particle (Rohrmann 2019). For example, Autographa californica multiple nucleopolyhedrovirus (AcMNPV) packages on average 4 nucleocapsids into each occlusion-derived virus particle (Zwart et al. 2008). In theory, it is possible to combine this feature of virion architecture with some of the other virion architectures considered here. However, as we are unaware of real-world examples and the predicted effect of the type IA architecture is very straightforward, we only consider this feature in conjunction with an unsegmented genome.

Type II are the multipartite viruses, viruses with segmented genomes that package each genome segment individually into a virus particle. Multipartite viruses are common among plant viruses (Michalakis and Blanc 2020), while only one bipartite virus has been identified in *Bombyx mori* bideonsovirus (BmBDV) *(Hu et al. 2016)* and putatively in culex mosquitoes (Ladner et al. 2016). It is generally recognized that multipartition has a cost for infectivity (Fulton 1962). For simplicity, we do not allow multipartite viruses to co-package any genome segments. While we are aware that such co-packaging may occur, in our framework such viruses should be considered segmented viruses.

Types III-VI are all considered segmented viruses: viruses that have multiple genome segments and can in principle package all of these segments in a single virus particle. Type III are the selective packagers. Although the genome consists of multiple segments, there are molecular mechanisms in place to ensure that one copy of each segment is packaged into every virus particle. This ensures that a complete genome is transmitted and each virus particle is fully infectious. A prototypical type III virus is the octapartite IAV. There are elaborate

mechanisms in place to ensure high packaging fidelity (Chou et al. 2012). These mechanisms include packaging signals in the untranslated regions of genomic RNA that result in specific RNA-RNA interactions between segments, and possibly interactions between genomic RNA and nucleoproteins (Li et al. 2021). Under some conditions considerable packaging errors occur (Farrell et al. 2023; Brooke et al. 2013; Diefenbacher, Sun, and Brooke 2018).

Types IV, V, and VI can collectively be termed the non-selective packagers. During packaging, both the identity and the number of each segment need to be selected to ensure the type III architecture (i.e., one copy of each segment). Here, we therefore consider that non-selective packagers can have variation in the identity of the segments packaged (type IV), the total number of segments packaged (type V), or both (type VI). There is clear evidence for the existence of non-selectively packaging segmented viruses, as most virus particles are found to be missing one or more genome segments. One well-studied example is the tri-segmented RVFV, in which only a small proportion of virus particles contain at least one copy of all three genome segments (Bermúdez-Méndez et al. 2022; Wichgers Schreur and Kortekaas 2016). In fact, approximately half of RVFV virus particles do not contain any viral genomic RNA (Bermúdez-Méndez et al. 2022; Wichgers Schreur and Kortekaas 2016). This suggests that this virus is not simply a selective packager with a considerable packaging error, but rather a true non-selective packager. Many members of the *Bunyavirales* are spread by arthropod vectors and can replicate in them (Boshra 2022). Their other hosts are plants or mammals, with occasional zoonosis in humans, and the plant-infecting members like TSWV are also likely to be non-selective packagers (Wichgers-Schreur et al. 2018). Although there are more segmented viruses of animals than plants ((Michalakis and Blanc 2020)), the models for non-selective packagers we describe therefore could be pertinent to both animal and plant viruses. To the best of our knowledge, the empirical distribution of segment numbers over virus particles has not been described, and in principle, this distribution could lead to a range of different virion architectures. To limit the possibilities, we explore and simultaneously contrast different architectures, here we have chosen three putative architectures we refer to as types IV to VI. We stress that these are purely hypothetical architectures that we use for exploration, in the absence of a quantitative description of read-world virion architectures.

For type IV, a fixed number of randomly sampled genome segments is packaged into each virus particle, with this number being equivalent to the total number of genome segment types. i.e., for a bi-segmented virus, each virus particle will contain two randomly sampled genome segments. This genome organization is plausible *a priori* if the virus particle has enough internal space to accommodate the full genome.

For type V, a variable number of genome segments is packaged into each virus particle, with the maximum number that can be packaged being equal to the number of genome segment types. However, the identity of the segments is regulated such that multiple copies of the same segment are never packaged into the same virus particle, i.e. for a bi-segmented virus, zero, one, or two genome segments can be packaged into each virus particle. If two segments are packaged then both segment types will be represented and the virus particle has the complete hereditary material.

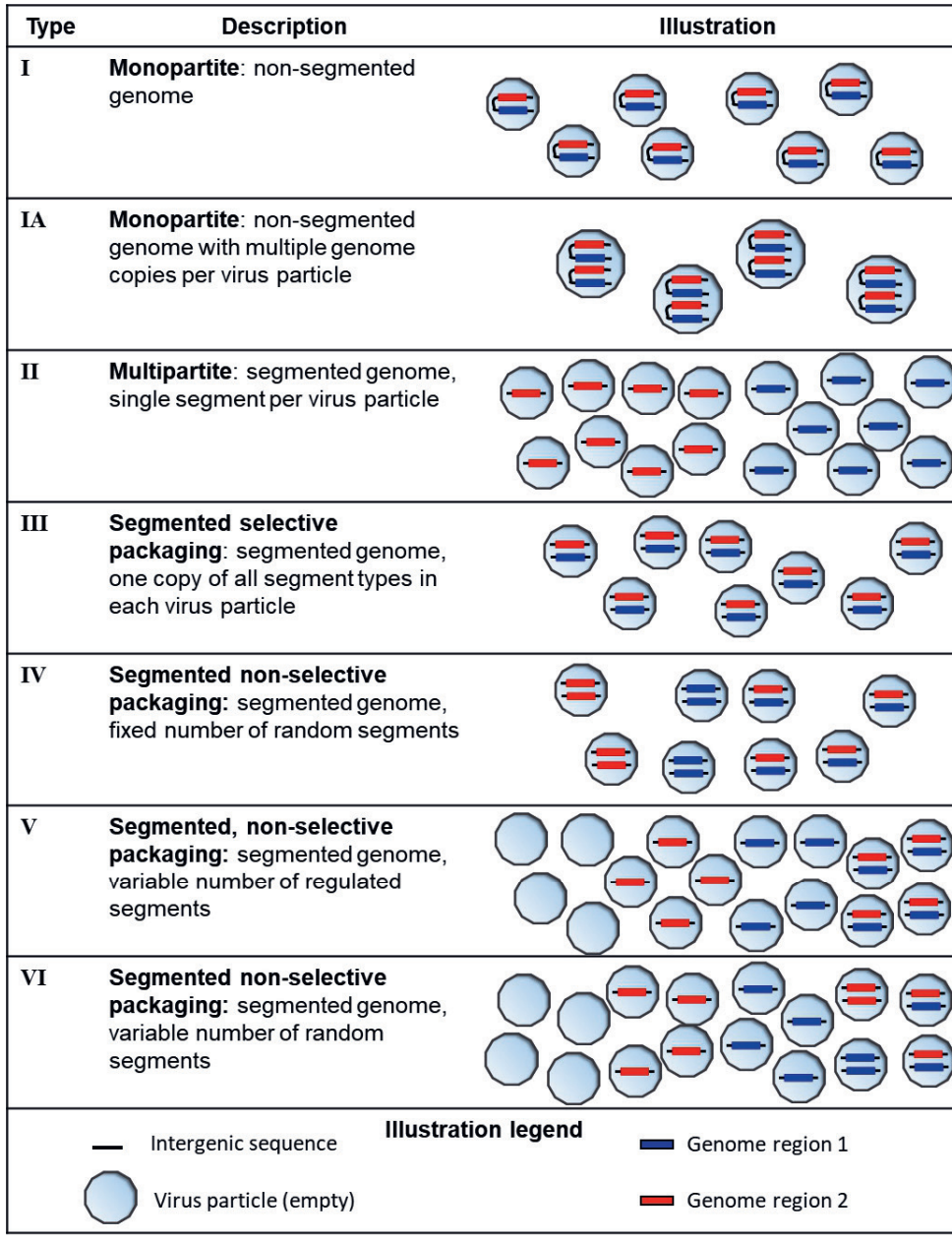


Figure 1. Overview of virion architectures. *Following our descriptions in the section Virion Architecture Types, we illustrate the different virion architectures considered here. The figure illustrates the identities of the segments packaged into virus particles, and also the absolute numbers produced under these different architectures, following the assumptions made in this study. Note that the two genomic regions are joined in a single segment for virion architectures* I *and* Ia.

For type VI, both the number of genome segments and the identity of the segments are randomly sampled. For a bi-segmeted virus, zero, one or two genome segments can be packaged into each virus particle. If two genome segments are packaged, they can either be of the same type or both types, and therefore the complete hereditary material may not always be present in a virus particle.

Scope and approach

Our goal is to predict the impact of virion architecture on viral infectivity and the evolution of the genome formula using computational approaches. First, we determine the cost to infectivity of the six different virion architectures. We make comparisons between extant virion architectures (type I - III) and hypothetical types (types IV-VI). Using empirical distributions of segment identity over virus particles for RVFV (Bermúdez-Méndez et al. 2022), we can consider the predicted cost to infectivity for this virus under the type IV-VI virion architectures. Finally, selective packagers will not be able to conserve any changes that occur in their genome formula, as the frequency of genome segments packaged is always perfectly balanced. Therefore, we simulate competition between selective and non-selective packagers to explore whether putative benefits associated with variation in the genome formula may outweigh the cost to infectivity.

Methods

We first provide an overview of the approach, model assumptions and the underlying rationale for these assumptions. We describe the analytical model used to make numerical predictions for dose-response (types I-III), and the simulation-based models used (types I-VI). We then describe methods used for determining infectivity cost for the empirical distribution of segment identities over virus particles for RVFV. Finally, we provide a description of the simulation models used to compete different segmented virus architectures against each other, and thereby the implications of changes in the genome formula for these competitions.

Overview of infectivity cost predictions

We assume the infection process consists of two steps: (i) initial invasion of the host by a virus particle (i.e. physical entry into an environment that supports replication), followed by (ii) productive infection: replication in the host if a complete genome is present, resulting in the formation of new infectious virus particles in the host. None of the approaches used are spatially explicit. We assume that mass action – i.e. the independent action hypothesis (IAH) – describes the kinetics of invasion for all virus particles (Zwart and Elena 2015). Given a dose of n virus particles and an invasion probability per particle ρ, the number of invading virus particles will be the product $\lambda = \rho n$. We will assume that all virus particles are equally invasive, regardless of their genome-segment content. However, to successfully replicate and have a productive infection, all κ genome segments of the virus need to be present in the host.

We assume virion architecture *per se* does not impose a cost on genome and virus particle production. If the genome is segmented, the number of full-length genomes produced is the same as for a monopartite virus. This situation corresponds to a situation in which there is a fixed pool of enzymes and nucleotides available to replicate the genome. Similarly, these genome segments can be distributed over any number of virus particles. This situation corresponds most closely to the production of rod-shaped virus particles, in which the number of capsid units is proportional to the length of the genome segment to be encapsidated (Solovyev and Makarov 2016). In practice, we realize that there may be a cost to distributing genome segments over a large number of virus particles, as this may require e.g. the production of more spike proteins in an enveloped virus. However, for simplicity, we do not consider such effects. Finally, we assume all genome segments are always present at equal frequencies when making cost calculations. We make this simplifying assumption (i) to restrict the scenarios we consider, and (ii) because a balanced GF will lead to the highest level of infection when we assume the same probability of invasion for all types of virus particles. Under the assumptions we have made (i.e., equal infectivity of all virus particle types), the highest level of infection will be achieved with a balanced genome formula. Only in the simulations of competition between different virion architectures do we relax this assumption and allow for an unbalanced genome formula.

Given these starting assumptions, no virion architecture will be more infectious than type I, monopartite viruses. Segmentation of the genome has no intrinsic benefit but it can lead to the loss of genome segments and the ensuing loss of the capacity for productive infection. We therefore compare all architectures to type I. We then normalize this difference by the difference between type I and a bi-segmented type II virus, as the multipartite virion architecture is generally considered to be a costly virion architecture (Sicard et al. 2013; Sánchez-Navarro, Zwart, and Elena 2013). To assess the effect of virion architecture on the capacity for infection, we consider the integral of the dose-response, considering dose on a logarithmic instead of a natural scale so that arbitrary model parameters (i.e., ρ) do not affect model predictions for the difference between integrals. This gives us the cost (C) of a virion architecture, for example for the type III architecture:

$$C_{III} = \frac{\int_{n=1}^{n_{max}} f_I\,(n)dn - \int_{n=1}^{n_{max}} f_{III}\,(n)dn}{\int_{n=1}^{n_{max}} f_I\,(n)dn - \int_{n=1}^{n_{max}} f_{II}\,(n)dn}$$

and likewise for architectures IV-VI.

Infectivity cost: numerical predictions from the analytical function

To make numerical predictions of dose-response from a function, we used a previously described approach (Sánchez-Navarro, Zwart, and Elena 2013). If κ types of virus particles, each carrying a unique genome segment, are needed for infection, and all virus particle types have an equal frequency and probability of infection, the dose-response can be given by:

$$I = \prod_{j=1}^{\kappa}(1 - e^{-\rho_j n_j}) = (1 - e^{-\rho n})^{\kappa}$$

where I is the proportion of infected hosts and *j* denotes the virus particle type. This approach works for the types I, IA, II, and III architectures. For types I, IA, and II, $\kappa = 1$. If the virion architecture affects n (we assume there is a limited pool of genome segments that can be packaged, but not restrictions on the number of virus particles produced), we refer to this correction as $\boldsymbol{\beta}$, resulting in an adjusted dose βn. Given type IA packages multiple full-length genome segments in each virus particle, this correction is $\beta = \frac{1}{\theta}$, where θ is the number of full-length genomes per virus particle. For the type II architecture, κ is equivalent to the number of genome segments (i.e., for a bi-segmented multipartite $\kappa = 2$ and for a tri-segmented multipartite $\kappa = 3$). For type III, we assume there are no packaging errors and hence $\kappa = 1$. In all cases, the integral is then calculated by numerical integration using the Gauss-Kronrod method from the Pracma package version 2.4.4 (Borchers 2023).

Simulation-based predictions for infectivity cost: types I, II and III

To make simulation-based predictions of dose response for all genome architectures we first need to determine the distribution of the total number of infecting virus particles. We consider all genome architectures so that we can compare numerical and simulation-based predictions for those virion architectures for which we can use both approaches (types I, IA, II and III). We let the number of infecting virus particles follow a Poisson distribution over hosts with a mean $\lambda = \rho n$. We use the rpois() function in R, rendering a realization φ. For types I, IA and III, when $\varphi > 0$ that host is infected. For type IA we need to adjust the dose, as done for the numerical prediction. For the type II architecture, we generate a realization of the number of infecting virus particles for each genome segment (i.e., virus particle type) φ_j, and hosts are only infected if for all κ realizations $\varphi_j > 0$. For this virion architecture, $\beta = \kappa$ as a complete segmented genome requires the same resources as an unsegmented genome.

Simulation-based predictions for infectivity cost: type IV

There are two approaches that can be used for modeling type IV architecture, which assumes non-selective packaging of a fixed number of genome segments. In the first approach (henceforth Model A); for simplicity we provide a description for a bi-segmented virus. We first draw a realization of the number of infecting virus particles φ. Next, we draw the number of copies of the first genome segment present (χ_1) from a binomial distribution with the rbinom() function. For this binomial distribution, the probability of success is 0.5 as both segments are equally abundant. We are describing a scenario in which κ segments are packaged randomly (with respect to segment identity) into each virus particle, so for a bi-segmented virus this results in 2φ trials for the binomial distribution. The number of copies of the second genome segment is $\chi_2 = 2\varphi - \chi_1$. A host is infected if for both realizations $\chi_j > 0$. The same approach can be extended to a tri-segmented virus. For the type IV architecture the total number of

virus particles produced is equivalent to type III, because the number of nucleocapsids per virus particle is fixed and hence more virus particles cannot be produced.

The second approach (henceforth Model B) used to model the type IV architecture is described elsewhere, where it is employed to model what we here term the type V virion architecture (Bermúdez-Méndez et al. 2022). We specify the frequency of virus particle types with different genome-segment contents, noting only whether each segment type is present or absent. For a bi-segmented virus with encapsidation up to 2 genome segments per virus particle, there are four possible virus particle types {segment 1, segment 2}: {0,0}, {1,0}, {0,1}, and {1,1}, with 0 indicating a segment is absent and 1 indicating it is present (with one or more copies). We represent the relative frequency r of each of these four types as a set $\{r_1, r_2, r_3, r_4\}$. We assume an even frequency of genome segments, and for the bi-segmented virus $r_2 = r_3$ and therefore we only present $\{r_1, r_2, r_4\}$. For type IV architecture, all virus particles contain 2 randomly selected segments, and therefore for a bi-segmented virus $\{r_1, r_2, r_4\} = \left\{0, \frac{1}{4}, \frac{1}{2}\right\}$. Likewise, for a tri-segmented virus there are eight types of virus particle possible: $\{r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8\}$ having a segment content {0,0,0}, {1,0,0}, {0,1,0}, {0,0,1}, {1,1,0}, {1,0,1}, {0,1,1} and {1,1,1}, respectively. As under our assumptions $r_2 = r_3 = r_4$ and $r_5 = r_6 = r_7$, we only present $\{r_1, r_2, r_5, r_8\}$, and for a tri-segmented type III virus $\{r_1, r_2, r_5, r_8\} = \left\{0, \frac{1}{27}, \frac{6}{27}, \frac{6}{27}\right\}$. To implement this model, we randomly assign each infecting virus particle to one of these types using the sample() function in R, weighting the probability of each outcome with the relative frequency of that virus-particle type. The host is only infected if all genome segments are present.

Simulation-based predictions for infectivity cost: type V

The type V virion architecture assumes a variable number of genome segments are packaged, but that only unique segment identities are packaged into a virus particle. This virion architecture was modeled with modified Models A and B. For Model A, after drawing the number of infecting virus particles (φ) as before (i.e., for type IV), we determine the genome-segment content of the virus particle by randomly drawing the number of genome segments present using a binomial distribution with a probability of success $1/\kappa$ and κ trials. i.e. On average, each virus particle contains one genome segment and the maximum number is the number of unique genome segments. For the bi-segmented virus, all two-segment virus particles will contain the complete genome under the assumptions made. For the tri-segmented virus, the genome-segment content of two-segment virus particles is determined using the sample() function without replacement of genome segment identity (i.e., each segment sampled is unique), whereas all three-segment virus particles contain the complete genome. As before, infection only proceeds in hosts in which all three segment types are represented.

For this architecture we need to account for the lower mean number of nucleocapsids packaged in virus particles, resulting in a larger pool of virus particles. Intuitively, on average one genome segment is packaged per virus particle, so bi-segmented and tri-segmented type V viruses produce twice and thrice as many virus particles as the type IV, respectively. Formally and more generally, $\beta = \kappa / \sum_{k=1}^{\omega} \Phi_j \psi_j$, where β is the correction to the total number

of virus particles produced, Φ is the frequency of a virus particle type, ψ is the total number of genome segments it contains (irrespective of segment identity), and ω is the total number of virus particle types. $\kappa \neq \omega$, as there are more types of virus particles than genome segments possible for non-selectively packaging segment viruses. In this case, given the model assumptions, we can obtain Φ values by determining the binomial probability mass function (PMF) for the distribution of genome segments over virus particles and dividing by the number of unique virus particles (in terms of segment content) with that number of segments. |In other words, for the bi-segmented type V virus, all particles with a total of two segments have both segment types, so there is only unique virus particle type whereas for a tri-segmented virus there are three combinations of genome segments that can be present in the particles with a total of two segments.

For Model B, it follows from elementary probability that for the bi-segmented virus all virus particle types are equally common and hence $\{r_1, r_2, r_4\} = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right\}$. For the tripartite virus, we use Φ_j as described above for the correction to the number of virus particles (i.e., β) resulting in $\{r_1, r_2, r_5, r_8\} \approx \{0.296, 0.148, 0.074, 0.037\}$. The number of virus particles can be corrected as for Model A (i.e., using β).

Simulation-based predictions for infectivity cost: type VI

The type VI virion architecture combines non-selective packaging of genome segments with variation in the number of segments packaged. β is identical to type V, as only the distribution of genome segment identity (but not the total number of genome segments) is different over virus particles. For Model A, as for type V, we randomly draw the number of infecting virus particles from a Poisson distribution and the number of genome segments per virus particle from a Binomial distribution, assuming a probability of success $1/\kappa$ and κ trials for the latter. We then need to set segment identity for those virus particles with more than 2 genome segments. For example, for a bi-segmented virus, we determine the total number of infecting virus particles with 2 genome segments, and then draw the number of successes χ_1 over the total number of genome segments present in this sub-population of virus particles.

For approach B, we use simple probabilistic arguments to determine the distribution of virus particle types. For example, for a bi-segmented virus following type VI architecture, virus particles with two genome segments have a probability of ¼ of containing only segment one, and likewise for segment 2. Hence, we can correct the expected frequency of virus particle identities such that $\{r_1, r_2, r_4\} = \left\{\frac{1}{4}, \left(\frac{1}{4} + \left[\frac{1}{4} \cdot \frac{1}{4}\right]\right), \left(\frac{1}{4} - \left[\frac{1}{2} \cdot \frac{1}{4}\right]\right)\right\} = \{0.25, 0.3125, 0.125\}$. Likewise, for the tri-segmented virus $\{r_1, r_2, r_5, r_8\} \approx \{0.296, 0.174, 0.057, 0.008\}$.

Infectivity cost: predictions for RVFV

To consider the implications of virion architecture for infectivity in a real-life example, we re-used previously published data on RVFV (Bermúdez-Méndez et al. 2022). For these data, the

distribution of genome segment identities over virus particles is known (i.e., which genome segment type is present or absent in a virus particle), whereas the distribution of the number of segments per identity has not been quantified (i.e., the number of copies of each genome segment type in a virus particle). We were only able to perform this calculation with Model B. The frequency of virus particle identities (e.g., $\{r_1, r_2, r_5, r_8\}$) is identical for all three architectures here, and only the correction for the number of virus particles (β) will be different across virion architectures. For the theoretical type IV and V architectures, we can calculate this correction directly given the frequency or virus particle identities. (E.g., Model IV assumes there are always three genome segments, so if only one segment type is present in a virus particle all three segments must have the same identity. For Model V, per definition there is only a single copy of each segment present in a virus particle.)

By contrast, the assumptions for Model VI will result in a specific distribution of the number of copies of a genome segment over virus particles, so we needed to verify if the empirical distribution was compatible with model predictions. We used grid searches to test whether varying the probability of success for the binomial distribution of genome segments per virus particle (previously set to $1/\kappa$ for theoretical comparisons) could reconcile model predictions with the data. We also considered whether any distribution of the virus particle types (i.e., not limited to having a binomially distributed number of genome segments per virus particle) could account for the data, again by performing a grid search over all possible relative frequencies of virus particle types. For both grid searches, we selected parameter values that minimized the sum of squares for predicted and observed relative frequencies of virus particle identities. As only the unrestrained distributions of virus particle types could account for the observed patterns under the type VI architecture, we used these predicted distributions of the number of segments per virus particle for making the correction to the number of virus particles.

Simulations of competition between selective and non-selective packagers

To explore the potential effects of the genome formula on the evolution of non-selectively packaging segmented viruses, we adapted a model of multipartite virus genome-formula evolution (Zwart and Elena 2020). This model was intended to explore the competition between cognate monopartite and bi-segmented multipartite viruses and identify conditions favorable for multipartite viruses. The model incorporates a function linking the intracellular GF and virus particle yield per cell. This function links the log of the ratio between the two segments, $r = log_{10}\left(\frac{f_1}{f_2}\right)$, to a level of virus particle accumulation using the normal probability density function. The optimal value of *r* for virus particle production is then the mean of the normal distribution (μ), while its variance (σ^2) determines how sensitive virus particle production is to any deviations from the optimum. By varying parameters μ and σ^2, we can test how the outcome of competition between monopartite and multipartite viruses will be affected as the intracellular genome formula has different effects on virus particle production. The model incorporates a fixed multiplicity of cellular infection and virus evolution in a regulated number of effectively infected cells. There is stochastic variation in the number of virus particles infecting cells and a function that links the intra-cellular genome formula and

virus particle yield produced. In the original study this model was used to study competition between a monopartite and multipartite viruses, but here we adapted this model to consider competition between a bi-segmented type III selective packager and a bi-segmented type IV or V non-selective packager. The type VI was excluded because it requires greater model complexity, while showing similar infection kinetics to the type IV. We tried to keep conditions identical to those previously reported, using the same fixed values or ranges of parameters as previously used (see Table 1 for an overview of model parameter values). In the below description, we highlight the changes made to the original model to represent these different virus variants.

We first explain competitions between types III and IV. The type III selective packager, whilst having a segmented genome, behaves exactly the same as the monopartite virus, by virtue of its perfect, selective packaging. The type IV non-selective packager produces virus particles with either two copies of the same segment (1,1 or 2,2), or one copy of both segments (1,2). We must therefore introduce a new class of virus particles (1,2) to the model, and account for the double genome segment content of the single-segment-type virus particles when calculating the genome formula. We first set initial frequencies of the virus particles of the selective packager (v_s) and non-selective packager virus variants (v_j) to $v_s = \frac{1}{2}$, $v_{1,2} = \frac{1}{4}$ and $v_1 = v_2 = \frac{1}{8}$. This starting population represents equal frequencies of the genome segments (f) of the type III and IV viruses ($f_{s,1} = f_{s,2} = f_{ns,1} = f_{ns,2} = \frac{1}{4}$), where ns indicates the non-selective packagers. Next, we must estimate the number of cells to be used to hold steady the number of effectively infected cells. Here we use the same approach as before (Zwart and Elena 2020, equation 1), considering the MOI, the fractions of cells invaded by the single segment type particles (v_1, v_2) and the fraction of cells invaded by virus particles with a complete genome ($v_s + v_{1,2}$) in place of the fraction of cells infected by the monopartite. We then alter the model to divide the Poisson-distributed total number of invading virus particles over the four virus-particle types based on their frequencies. For all simulations, we used the model variant with coinfection exclusion between the two viruses. We choose this variant of the model *a priori*, because in previous work on multipartite viruses, coinfection exclusion prevents the cognate monopartite virus from exploiting the multipartite virus's genome-formula-derived benefits. Coinfection exclusion is modeled as a stochastic process mediated by the second genome segment, therefore favoring the more abundant virus variant within a cell (Zwart and Elena 2020). Next, we can determine the within-cell genome formula (r) for each cell, which we consider on a $\log_{10}$ scale. As we assume there is no within-cell selection for segments, the frequency of genome segments that invaded the cell will be represented in the virus particles generated. For cells infected by the the type III virus, given there is coinfection exclusion and therefore a balanced genome-segment content in virus particles, per definition $r = log_{10}\left(\frac{f_{s,1}}{f_{s,2}}\right) = 0$. For cells infected by the type IV virus, $r = log_{10}\left(\frac{f_{s,1}}{f_{s,2}}\right) = log_{10}\left(\frac{2\varphi_1+\varphi_{1,2}}{2\varphi_2+\varphi_{1,2}}\right)$, recalling that φ is a realization of the number of infecting virus particles. The total virus particle yield can then be determined as before, using the normal probability density function with a mean μ and variance σ^2 (Zwart and Elena 2020). To partition this yield over virus particle types, we first use the binomial PMF to determine the frequencies of v_1 and v_2. For this PMF, the number of trials is two (i.e., the fixed number of genome segments per particle), the number of successes is two (i.e., we want to know the fraction of virus particles

containing only this segment type), and the probability of success is the relative frequency of segment $n,1$ or $n,2$ in that cell. Note that as the genome formula becomes more unbalanced, there will be fewer $v_{1,2}$ virus particles and the infectivity of the type IV sub-population will decrease. Finally, we modified the performance metric used in the original simulations on multipartite viruses to determine the winner of the competition if both viruses were maintained. We adjusted this metric to reflect the relative frequency of all segments associated with a variant.

For the competition between the type III and the type V viruses, for conciseness only we highlight the differences in the competition between types III and IV. Type V has the same number of classes of virus particles as type IV, but the number of genome segments encapsidated in these particles is different (i.e., these particles then contain only a single copy of the segment). Hence, to maintain ($f_{s,1} = f_{s,2} = f_{ns,1} = f_{ns,2} = \frac{1}{4}$), the initial frequencies of the virus particles needs to be $v_s = \frac{2}{5}$ and $v_1 = v_2 = v_{1,2} = \frac{1}{5}$. The infection process is identical to the type IV virus, but to determine the genome formula in infected cells we need to account for the differences in genome segment content: $r = log_{10}\left(\frac{f_{s,1}}{f_{s,2}}\right) = log_{10}\left(\frac{\varphi_1+\varphi_{1,2}}{\varphi_2+\varphi_{1,2}}\right)$. To model the packaging of genome segments into virus particles with an unbalanced genome formula is more complex for the type V than for the type IV. Recall there is a variable number of genome segments over virus particles, but if two segments are packaged they will represent the two segment types. If there is an unbalanced genome content, we reason that the bi-segmented type V virus will package virus particles according to its idealized scheme (I.e., $\{r_1, r_2, r_4\} = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right\}$) until the rarer variant is depleted, at which point only single-copy virus particles with the other segment will be generated. Prior to running the simulations of competition between viruses, we ran multiple simulations of the encapsidation process for different genome formula values. During the competition simulations, we randomly selected an iteration corresponding to the genome formula in that cell, allowing us to capture stochastic variation in the encapsidation process without having to run a simulation for each individual cell.

Table 1: Model parameter values

Section	Model parameter	Value or range	Comments
Dose response	ρ	0.001	Probability of cell entry per virus particle
	n	1 - 10^5	Dose range
	θ	2	Number of complete genome copies per virus particle for type IA virus
	Simulations	10^4	Number of individuals per dose
	Simulations Model B	10^3	Number of individuals per dose
Evolutionary Competitions	*C*	10^3	The number of effectively infected cells
	Max. passages	100	The maximum number of passages for which each individual simulation is allowed to proceed
	Simulations	10^3	The number of independent simulations for each set of conditions
	Probability of change in the environment	0.2	The probability that the environment changes in each round of passaging, i.e., that a new value of μ will be drawn.
	log10(MOI)	-2, -1.9, -1.8, (...) 2	Cellular multiplicity of infection (MOI) - the number of virus particles entering a cell
	ψ	0, 0.1, 0.2, (...) 2	Parameter that sets the range (0 ±ψ) from which values of μ can be drawn
	σ^2	0.01., 0.1, 1, 10	Parameter that determines sensitivity of virus accumulation to deviations from the optimal genome formula μ.

Results and discussion

Virion architecture is predicted to impact infectivity

By comparing the integrals of the dose response for different virion architectures, we could determine the cost to infectivity for each virion architecture over a fixed range of viral doses. Considerable research has suggested that multipartite viruses have a high cost to infectivity, due to the loss of genome segments during transmission (Fulton 1962; Gutiérrez and Zwart 2018; Sánchez-Navarro, Zwart, and Elena 2013). We used the comparison between the dose-response of a bi-segmented, multipartite virus (Type II, $\kappa = 2$) and monopartite virus (Type I), to normalize the predicted infectivity cost for all virion architectures as compared to the monopartite virus. The cost to infectivity for the monopartite is therefore per definition 0, whilst for the bi-segmented, multipartite virus it is 1.

We found that many virion architectures had a high predicted cost to infectivity (Table 2, see also Supplementary S1: Table S1). First, packaging multiple copies of a single-segment genome into each particle is very costly within our framework, leading to an infectivity that is considerably lower than that of the bi-segmented multipartite virus. Although this genome architecture is not common, it has been adopted by the multiple nucleopolyhedroviruses (in the genus *Alphabaculovirus*, double-stranded DNA insect viruses). For these viruses, each occlusion derived virion contains multiple nucleocapsids, each of which contains a copy of its monopartite genome (Slack and Arif 2007). This result immediately highlights an important caveat of our framework: we consider the implications of virion architecture by making simplifying assumptions on virus particle production and only consider genome completeness as a criterion for infectivity. In the real world, other biological factors will come into play. For multiple nucleopolyhedroviruses, packaging multiple genome copies (i.e., nucleocapsids) into each virus particle may enhance infectivity by allowing faster passage through the larval midgut, a key barrier to infection (Slack and Arif 2007). Our framework therefore cannot in itself predict whether a virion architecture is likely to occur, we can only predict the associated infectivity cost *ceteris paribus*.

Within this framework, genome segmentation is not necessarily associated with a cost to infectivity. Type III viruses have segmented genomes but selectively package them into virus particles, making their virus particle production and the infectivity of each virus particle equivalent to that of the monopartite virus (type I). Given this intuitive result, for a long time, it has been assumed that all segmented viruses would be selective packagers. Research on IAV has shown that a large pool of the viral population is composed of incomplete viral particles, missing virus genome segments, that have reduced infectivity compared to complete genomes but contribute to increasing overall virus infectivity (Brooke et al. 2013; Diefenbacher, Sun, and Brooke 2018; Farrell et al. 2023). We therefore chose to include an estimate of the cost to infectivity of the type III virion architecture with errors in packaging, under low (τ = 0.025) and high (τ = 0.25) error rates (Supplementary S2). By errors we understand the packaging of the wrong segment type, as the correct total number of segments is always packaged in the scenario we consider. When the error rate is low, the cost to infectivity is much lower than that incurred by other virion architectures (types IV, V or VI).

When the error rate is high, the cost to infectivity for the bi-segmented virus is similar to that of the type V non-selective packager, which packages a variable number of segments (Table S1). For very high error rates, a type III virion architecture with errors or a type IV virion architecture will become indistinguishable in terms of the realized distribution of segments over virus particles..

Our results confirm that all the other virion architectures for the segmented viruses have an infectivity cost compared to monopartite viruses and their selectively packaging relatives (Table 2). Surprisingly, the infectivity cost for two of these architectures - types IV and VI - approach the predictions for the multipartite viruses (type II). By contrast, the type V architecture has a considerably lower cost than the other two non-selective packagers and the multipartite viruses. We do not know why non-selective packaging has evolved, and indeed whether it may be associated with certain benefits that outweigh its costs. Putative evolutionary benefits of non-selective packaging and mechanisms of selective packaging of genome segments, may allow which virion architectures actually evolve. However, we predict that the type V virion architecture will have a much lower infectivity cost *ceteris paribus*, making it the most plausible virion architecture for the non-selective packagers within our framework.

We recognize that the virion architectures that we have postulated for the non-selective packaging segmented viruses (types IV, V and VI) are all hypothetical: to the best of our knowledge the distribution of genome segments over virus particles has not been quantified. We have therefore chosen three possible architectures to explore their effect on infectivity, while making some assumptions to constrain the variations considered. For example, we assume the mean number of genome segments per virus particle is 1 for types V and VI, to ensure a contrast with the type IV architecture. Other values for the mean number of genome segments per virus particle can be explored, and indeed we consider this possibility later for RVFV.

Why does the infectivity cost of the types IV and VI non-selective packagers approach that of the multipartite viruses? The cost to infectivity for multipartite viruses (type II) has a single cause under our assumptions. Given that nucleotide and enzyme pools constrain the number of virus particles produced and that multipartite viruses package only a single genome segment per virus particle, a bi-segmented multipartite virus may generate twice as many virus particles as the cognate monopartite. Therefore, cost arises because some hosts are only invaded by virus particles carrying one segment type, which cannot support replication. Abortive infections due to incomplete genomes can occur for type II, IV, V and VI architectures. Only the type IV and VI architectures result in the formation of multiple copies of the same segment in a single virus particle. Within our framework, these virus particles are wasteful because the additional copies of the same segment have no added benefit for infectivity, but do result in the formation of less particles. To illustrate these differences, consider the differences in the dose-response for type II and IV architectures (Figure 2). Type IV non-selective packagers have an infectivity advantage relative to type II multipartite viruses at low doses due to the occurrence of complete virus particles, which are completely absent in the multipartite virus population. At higher doses where the probability of infection by two virus particles becomes appreciable, the multipartite virus has higher infectivity relative to the type IV non-selective packager because its packaging is less wasteful.

Table 2: Theoretical cost to infectivity predictions. We give an overview of the cost of different virion architectures, as a function of the number of genome segments ($\kappa \in [1,2,3]$). All costs have been normalized by the difference in the integrals of the type I and II viruses, such that the cost of the type I is 0 and the cost of the type II is 1. Fields that correspond to combinations of virion architecture and segment number that are logically excluded are colored gray. Where possible we provide the numerical prediction of infectivity cost, if numerical predictions cannot be made we provide the Model A simulation-based prediction. Full results for all approaches are given in Table S1 (Supplementary S1), and illustrate that there was good agreement between the different modeling approaches used.

Category	Type	Cost, 1 segment	Cost, 2 segments	Cost, 3 segments
Monopartite: single genome copy per virus particle	I	0[b]		
Monopartite: multiple genome copies per virus particle, θ = 2	Ia	1.001[a]		
θ = 3		1.586[a,]		
Multipartite	II		1.000[b]	1.416[b]
Segmented, selective packaging	III		0[b]	0[b]
Segmented, variable segment identity	IV		0.832[c]	1.360[c]
Segmented, variable segment number	V		0.582[c]	1.135[c]
Segmented, variable segment identity & number	VI		0.971[c]	1.417[c]

[a] θ = 2 and 3, for two or three genome copies per virus particle, [b] numerical prediction, [c] Model A simulation-based prediction.

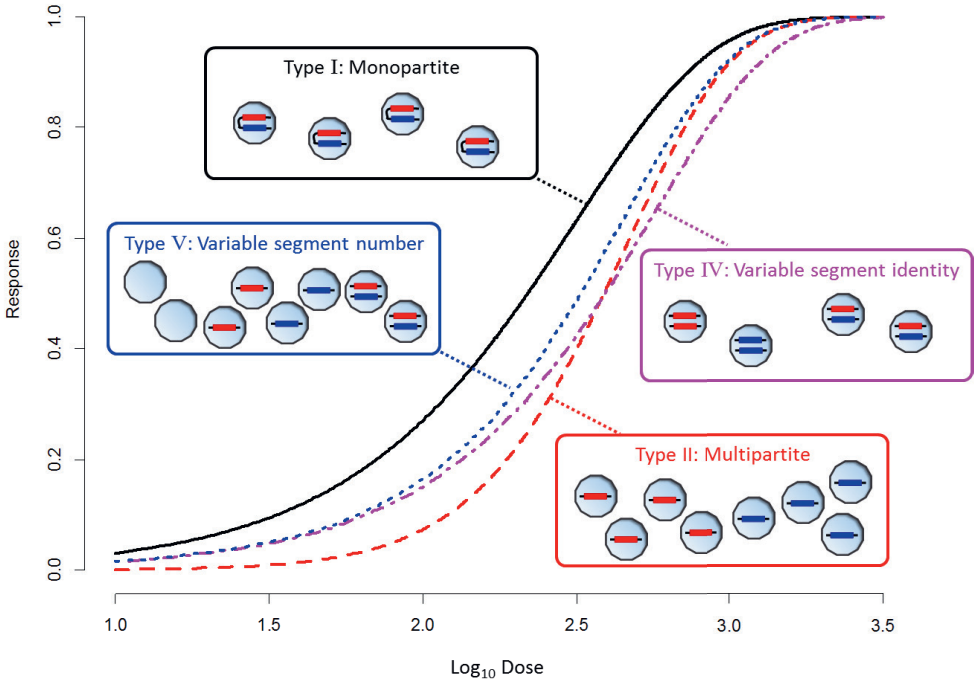


Figure 2. Dose response for different virion architectures. *To illustrate the different infectivity costs associated with virion architecture, we show dose-response curves for four virion architectures. The x-axis represents the log_{10}-transformed virus-particle dose, and the y-axis is the response (number of infected hosts). The type I (black, solid line) has the highest infectivity, followed by the type V (blue, dotted line). At low doses the type IV outperforms the type II, because the type IV contains some virus particles with a complete genome and the type II does not. By contrast, at high doses the type II outperforms the type IV, as it produces more virus particles and the probability of multiple virus hits has increased.*

Predictions for RVFV highlight the relevance of virion architecture

Our theoretical models predicted large differences in the infection cost for different non-selectively packaging segmented viruses. Although we think these virion architectures are plausible *prima facie*, they are hypothetical. To consider our model predictions for existing virion architectures of non-selectively packaging segmented viruses, we considered previous RVFV results. For virus particles originating from mammalian and insect cell lines, the

distribution of genome segment types over virus particles is known, although the number of copies of each segment has not been quantified (Bermúdez-Méndez et al. 2022). We could therefore use a simulation-based prediction for type IV - VI based on Model B (see Supplementary Text 1 for details) to determine infectivity: all virion architecture models have the same frequency of virus particle types, whereas the total number of virus particles produced depends on the virion architecture. For the type VI architecture, we infer the distribution of the total number of genome segments over virus particles using different approaches (Supplementary Text 1 and Figure S1).

Infectivity for RVFV was estimated for the virion architecture types IV, V and VI (Table 3), and we observe that the differences between virion architectures is more pronounced than for our theoretical examples (Table 2). For virus populations derived from mammalian cells, the frequency of virus particles with a complete genome is very low, and hence the cost to infectivity is greater under all virion architectures compared with virus particles derived from insect cells Overall, the cost to infectivity for the types IV and VI is consistently higher than that of a bi-segmented type II virus, and in the case of a type IV virus it is higher than that of a tri-segmented type II virus. In the previous study of Bermudez-Mendez et al (2022), the effect of non-selective packaging on virus infection was considered. However, the possibility that different virion architectures were compatible with the data was not considered, and the type V virion architecture was simply assumed when modeling infection (Bermúdez-Méndez et al. 2022). Our results here (Table 3) shows that the type V virion architecture is most consistent of that for RVFV and reinforces the relevance of virion architectures for virus infectivity. By combining cryo-EM, RNA sequencing and strand-specific qRT-PCR of the tri-segmented TSW, Yvon et al. (2023) show that virus particles contain a mix of genomic viral RNA co-packaged with complementary strands, and that within-host TSW accumulation and virus particle content produce comparable unequal genome formula ratios (Yvon et al. 2023). Demonstrating that at least for TSW and possibly for other viruses of the Bunyavirales, such as RVFV, that the virus particle content may not only be heterogenous for genomic virus segments but also for complementary strands. further highlighting the need for complete information. Highlighting the relevance of having complete, quantitative information on the distribution of genome segments over virus particles, a task which will be technically challenging.

Table 3: Infectivity cost predictions for RVFV. We give an overview of the infectivity cost of different segmented, non-selectively packaging virion architectures using the Model B simulation-based approach, based on a limited characterization of RVFV virus particles. Full results are given in Table S1.

Category	Type	Cost, RVFV Mammalian	Cost, RVFV Insectile
Variable segment identity	IV	1.975	1.311
Variable segment number	V	1.028	0.734
Variable segment identity & number	VI	1.386	1.061

Competitive fitness of non-selectively packaging viruses

For all multipartite viruses tested, the genome formula (i.e., the relative frequency of genome segments) is unbalanced (Sicard et al. 2013; Wu et al. 2017; Yu et al. 2019; Hu et al. 2016). Moreover, the genome formula of FBNSV and AMV was shown to be host dependent (Sicard et al. 2013; Wu et al. 2017) and values close to the equilibrium value are associated with higher accumulation (Sicard et al. 2013; Wu et al. 2017), suggesting that it can play a role in virus adaptation as tested in Chapter 4 and 5 of this thesis. For an idealized selectively packaging segmented virus (type III), the frequency of genome segments will be fixed during transmission: it will be reset to equilibrium in the population of virus particles produced by an infected cell. Therefore, even if beneficial genome formula changes occurred within a cell or host due to selection, they would not be transmitted. By contrast, the non-selectively packaging segmented viruses (types IV, V and VI) can acquire genome formula changes, if the frequency of the different particle types produced in a cell depends on the intra-cellular genome formula. We therefore adapted a model of competition between monopartite and multipartite viruses, to study competition between a selective packaging virus (type III) and two non-selective packagers (types IV and V). We chose to focus on the latter two architectures for two reasons: (i) they represent the two different mechanisms by which non-selective packaging occurs (loss of segment identity control or loss of control over the total number of segments during packaging) and, (ii) both these architectures have three classes of virus particles containing segments instead of five classes for the type VI architecture, resulting in lower model complexity. For the type V architecture, note that we had to establish the distribution of segments when the genome formula is unbalanced. Briefly, we let the ratio of single-segment to double-segment (i.e., with both segment types present) virus particles be 2:1 until one of the segments is depleted. This means that virus particles with both segments present will be made at the same rate as for a balanced genome formula until the rarer segment type is depleted, at which point the virus only produces particles containing single segments of the most frequent segment type.

For the competitions between types III vs IV and III vs. V, we found patterns reminiscent of the results for monopartite and multipartite viruses (Zwart and Elena 2020). The non-

selectively packaging viruses could displace the selectively packaging virus when virus particle yield was sensitive to the genome formula (i.e., when $\sigma^2 < 1$), when there was considerable variation in the environment in the optimum genome formula (high ψ values), and when MOI was in an intermediate range (Figures 3 and 4). However, the types IV and V could also displace the type III when virus particle yield was not sensitive to the genome formula (i.e., when $\sigma^2 \geq 1$), at high MOI values. When $\sigma^2 = 10$, this outcome occurs irrespective of ψ (range of values for μ) for both genome architectures). When $\sigma^2 = 1$, the type V displaces the the type III at low ψ values (Figure 4). Similar outcomes were not seen for competitions between monopartite and multipartite viruses, where the monopartite viruses always dominated in this parameter space (Zwart and Elena 2020). What could account for this unexpected outcome?

Based on previous results (Zwart and Elena 2020), we included coinfection exclusion in the model to avoid coinfections, so that the type III could not exploit the genome-formula associated benefits of the types IV and V. For low σ^2 values and high MOI values, the genome formula of the type IV and V viruses can vary without losses in infectivity or virus-particle production. Therefore, when genome formula drift (Gutiérrez and Zwart 2018) leads to an increase in the frequency of the second genome segment - which encodes for gene products facilitating coinfection exclusion - the types IV and V displace the type III through interference competition (Figure 4). To illustrate this effect, consider the genome formula dynamics for a number of conditions, in which a systematic upregulation of the second genome segment is seen for low σ^2 values (Figure 5). This emerging property of the genome formula was noted earlier for the simulated competitions between monopartite and multipartite viruses, but it was only manifest in small parameter space and had only a marginal effect on the success of the multipartite virus (Zwart and Elena 2020). By contrast, the types IV and V display this property over a large parameter space, presumably because their infectivity cost is lower than for the type II viruses. We do not know if segmented viruses exploit changes in the genome formula during adaptation in the real world. Clearly genome-formula variation exists for some segmented viruses (Bermúdez-Méndez et al. 2022; Moreau et al. 2020), and for IAV - which is considered to be a selective packager - segment frequency can be downregulated under some conditions (Sun and Brooke 2018; Brooke et al. 2014). Our results highlight the potential versatility of the genome formula, as this mechanism is used to adapt to conditions in ways we had not anticipated when developing these simulations.

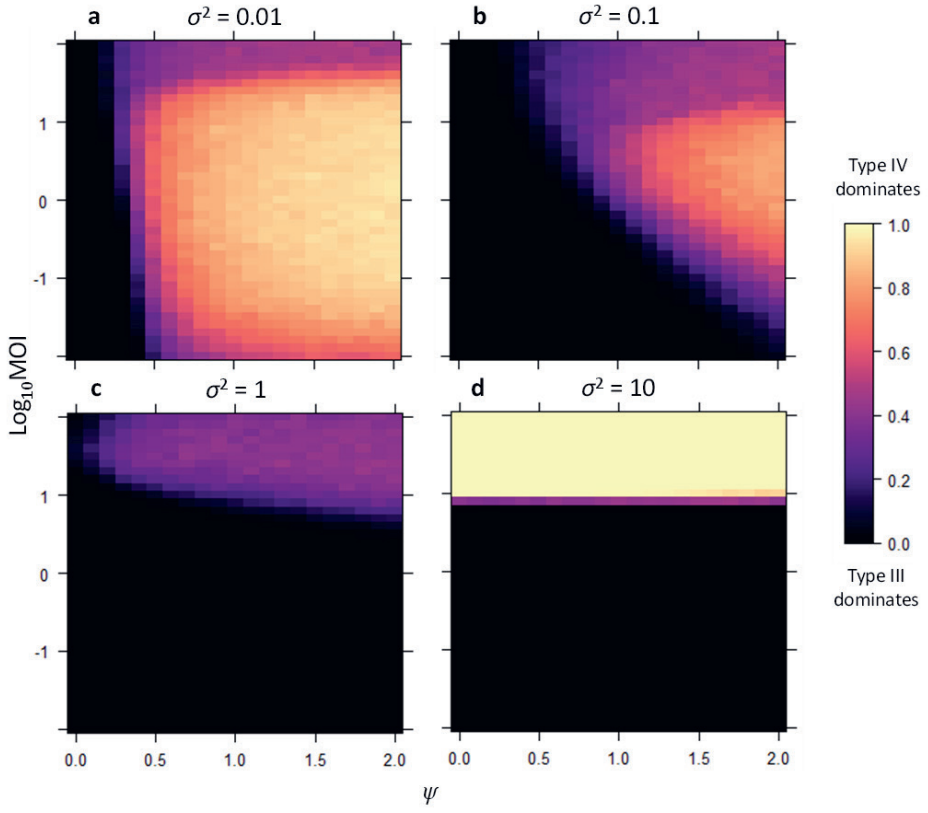


Figure 3. Competition between type III and IV virion architectures. *Competitions between viruses were run over a range of conditions including the magnitude of environmental heterogeneity in optimal virus gene expression (ψ), the cellular multiplicity of infection (MOI), and the sensitivity of virus yield to the genome formula (σ^2). The lower the value of σ^2, the greater the sensitivity to the genome formula. The heat indicates which virus predominated in the competitions.*

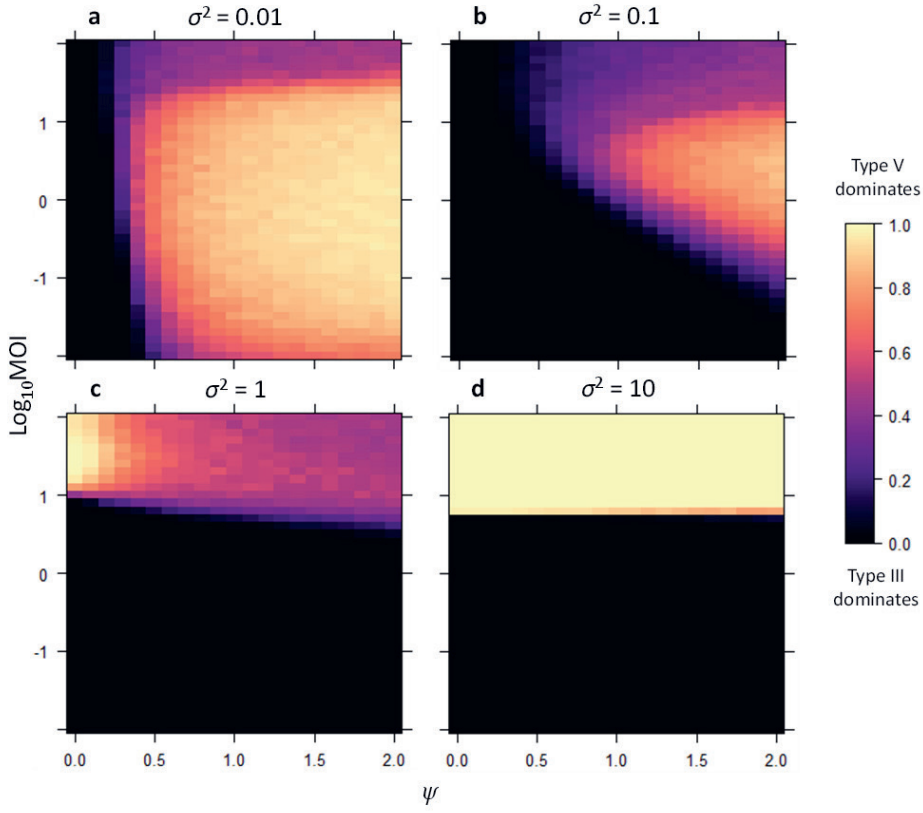


Figure 4. Competition between type III and V virion architectures. *Competitions between viruses were run over a range of conditions including the magnitude of environmental heterogeneity in optimal virus gene expression (ψ), the cellular multiplicity of infection (MOI), and the sensitivity of virus yield to the genome formula (σ^2). The lower the value of σ^2, the greater the sensitivity to the genome formula. The heat indicates which virus predominated in the competitions.*

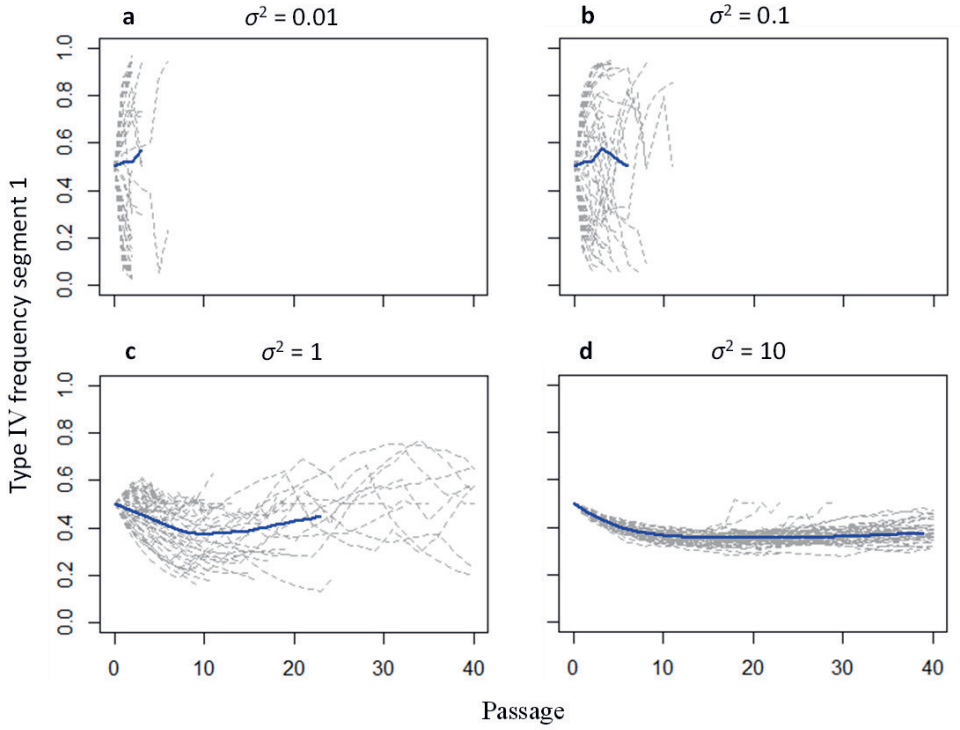


Figure 5. Genome formula dynamics of the type IV virus. *We illustrate the genome formula dynamics of a type IV virus during competitions with a type III virus, as we vary the sensitivity of virus yield to the genome formula (σ^2). For all panels, the x-axis is passage number and the y-axis is the frequency of genome segment 1 of the type IV virus. The solid blue line is the mean of 10^4 simulations, and the dotted grey lines are 50 individual simulations. The mean is only shown for those passages where there is a representative number of competitions (> 2000) in which neither virus (type III or IV) has fixed. For all simulations, the MOI is 10 and there is considerable environmental heterogeneity in optimal virus gene expression ($\psi = 2$). In panels A and B, there is high sensitivity of virus yield to the genome formula (low values of σ^2), and the environment dictates the genome formula, resulting in high variation between individual simulations. As the sensitivity to the genome formula decreases (increasing values of σ^2) in Panels C and D, the variation in the genome formula between simulations decreases and there is a lower frequency of segment 1, corresponding to an increase in segment 2 which codes for a coinfection exclusion function. By increasing the strength of cellular coinfection exclusion, the type IV displaces the type III virus.*

Concluding remarks

We have predicted the cost to infectivity for a broad range of existing (types I-III) and putative (types IV-VI) virion architectures. For those viruses with segmented genomes and non-selective packaging (types IV-VI), we find the lowest infectivity cost for the type V architecture. This virion architecture has stochastic variation in the total number of segments packaged,

whilst each package is unique. Although this virus cannot infect as efficiently as the monopartite (type I) and segmented, selective packager (type III), it is more efficient than the multipartite (type II) and alternative non-selective packaging types, which introduce stochastic variation in the identity of segments packaged (types IV and VI). Similar results were obtained for purely hypothetical virus populations, and when informing these models with observations of RVFV virus particles. When we simulated competition between selective and non-selective packagers, we found that both the types IV and V could outcompete the type III selective packaging virus under some conditions. These virion architectures therefore empower the adaptive benefits of the genome formula, possibly explaining their existence. Here we also found additional benefits for both virion architectures: type IV and V populations wielded genome-formula change in unexpected ways, enabled by the low cost to infectivity of this architecture. Overall, our results suggest that of the three segmented, non-selectively packaging virion architectures we have explored, the type V is the most likely to exist.

We do advise caution when interpreting the results of these models which have a number of simplifying assumptions. Firstly, a fixed amount of (ribo)nucleic acids for assembling genome segments as the only constraint for virus particle production in these models. Accordingly this assumption does not account for the cost for the total quantity of virus particles produced. We expect that this best reflects the situation of helical flexible filamentous viruses (and likely also the rigid rod-shaped viruses e.g. Tobamovirus) particles, where the number of capsid proteins needed to produce the particle is similar to the segment length (Solovyev and Makarov 2016). For viruses which have a more constrained morphology and rigid structure such as the icosahedral viruses (Hespenheide, Jacobs, and Thorpe 2004), and enveloped viruses this assumption might not hold. However, RVFV is an enveloped icosahedral virus that is known to produce empty virus particles (Bermúdez-Méndez et al. 2022; Wichgers Schreur and Kortekaas 2016), suggesting that also in this system the number of virus particles is not the limiting factor but rather the production of RNPs. Secondly, the model assumes that all viral doses are equally relevant when making comparisons across virion architecture types. We do this for simplicity and recognize that this is a broad generalization that does not account for the differences in real-world virus doses typically encountered by hosts. If a particular dose-response range is more relevant, this could have a substantial effect on model predictions. In Figure 2, there can be trade-offs at different doses between the virion architectures as seen for type II and type IV. The host-vector interaction will shape the infective doses in natural infections of the Bunyavirales, and provide insights on virion architectures for non-selective packaging viruses. Thirdly, there will be stochastic GF variation at the intracellular level from differences in virus replication. This will influence how the GF is inherited and reduce the proposed GF benefit of regulaitng segment copies for viral gene expression. Future work can include the intracellular GF variability to compare competitions between type III virion architecture with packaging errors and type IV and V non-selective packagers. We assume that infections can be initiated with a single copy of each genome segment (a balanced GF) however depending on the host, cell, or tissue environment it may be beneficial to have multiple copies of segments to accommodate differences in MOIs and will affect infection kinetics. This would favor an unbalanced GF at the within-host level and potentially at the between-host level as well. Lastly we assume that selectively packaging viruses have perfect packaging of segment type and number, however there are likely to be differences from stochastic process and differences in packaging efficiency.

We recognize that the virion architectures that we have postulated for the non-selective packaging segmented viruses (types IV, V, and VI) are all hypothetical: to the best of our knowledge, the distribution of genome segments over virus particles has not been quantified. We have, therefore, chosen three possible architectures to explore their effect on infectivity while making some assumptions to constrain the variations considered. For example, we assume the mean number of genome segments per virus particle is 1 for types V and VI, to ensure a contrast with the type IV architecture. Other values for the mean number of genome segments per virus particle can be explored, and indeed we had to consider this possibility for RVFV. We eagerly anticipate quantitative data on the distributions of virus segments over virus particles for segmented viruses, which allows testing of these predictions. If other non-selective virion architectures exist in the real world, we predict that these architectures must (i) directly or indirectly (e.g., by trade offs) enable other benefits, and/or (ii) occur because of constraints of packaging. E.g., non-selective packaging may always affect the identity of segments being packaged, in which case this putative virion architecture would not be possible.

Supplementary S1

Table S1. The costs for different virion architectures are presented for the different number of genome segments ($\boldsymbol{\kappa} \in [\mathbf{1}, \mathbf{2}, \mathbf{3}]$). We present the cost estimates for different approaches; numerical, simulations approach Model A and Model B. We show that there is agreement between calculations using the different approaches. The cost presented has been normalized by the difference in the integrals of the type I and II viruses, such that the cost of the type I is 0 and the cost of the type II is 1. Fields that correspond to combinations of virion architecture and segment number that are logically excluded are colored gray. For the type IA virus multiple copies of complete genome copies can be within a virus particle (θ), we estimate the cost for two or three genome copies per virus particle. The type III segmented selective packager can display high-fidelity packaging with a complete set of genome segments or there may be differences in packaging fidelity described by an error rate. We estimate the cost for type III packaging when the error rate (τ) is equal to 0, 0.025 and 0.25. The cost for the type III architecture reported in the main paper (E.g., Table 1) are for τ = 0.

Category	Type	Cost, 1 segment		Cost, 2 segments			Cost, 3 segments		
		Numerical	Model A	Numerical	Model A	Model B	Numerical	Model A	Model B
Monopartite: single genome copy per virus particle	I	0.000	0.000						
Monopartite: multiple genome copies per virus particle, θ = 2	Ia	1.001	1.003						
θ = 3	Ia	1.586	1.588						
Multipartite	II			1.000	1.000		1.416	1.417	
Segmented, selective packaging, τ = 0	III			0.000	0.000	0.000	0.000	0.001	0.000
τ = 0.025						0.066			0.107
τ = 0.25						0.544			0.791
Segmented, variable segment identity	IV				0.832	0.830		1.360	1.359
Segmented, variable segment number	V				0.582	0.585		1.135	1.134
Segmented, variable segment identity & number	VI				0.971	0.975		1.417	1.418

Predicted infectivity of RVFV for type VI virion architecture

β, a correction to the number of virus particles produced based on packaging, can be directly determined for the type IV and V architectures. However, the type VI architecture is more constrained: both the number of segments and the stochastic nature of their distribution over virus particles are specified under the model. We found that this architecture did not fit the data when we assumed a binomial distribution of the number of genome segments over virus particles with a mean of 1, or when we relaxed this assumption and fitted the mean value of genome segments per virus particle (Figure S1a,b). We therefore estimated the fractions of virus particles with 1, 2 or 3 types of genome segments that best fit the empirical distribution, without assuming any statistical distribution *a priori* (Figure S1c,d). This approach better reconciles the type VI virion architecture and the data, although there are still some discrepancies for the virus particles from insect cells. However, having a reasonable approximation of the distribution of the number of genome segments over virus particles, we can now estimate β for all three architectures and determine the cost to infectivity.

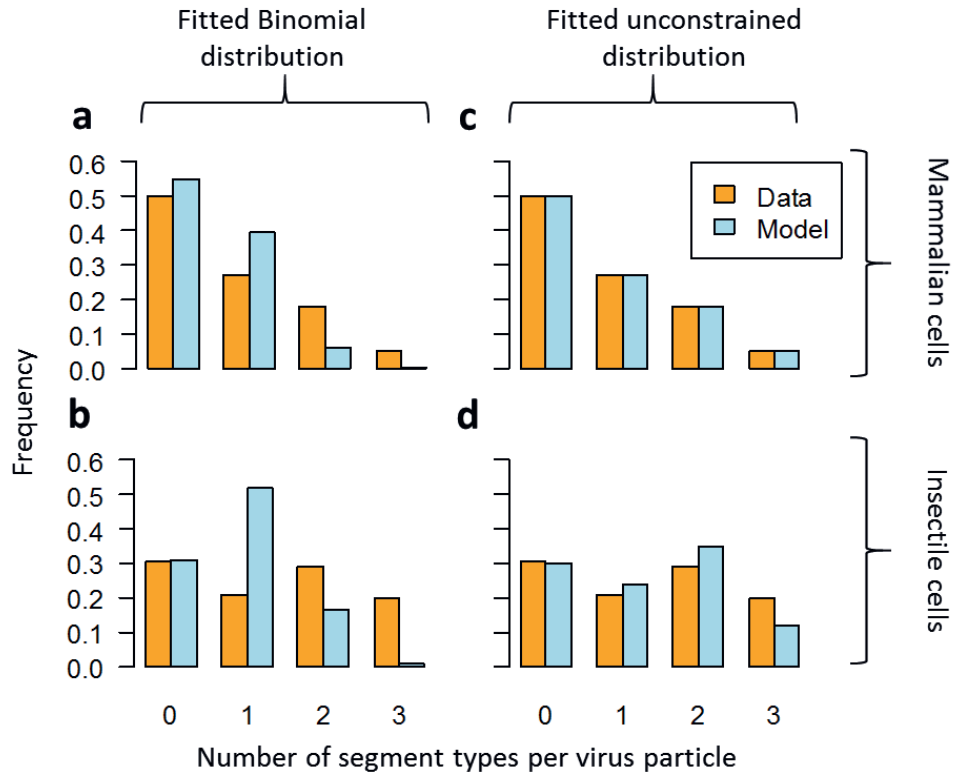


Figure S1. Observed and inferred distributions of RVFV genome segments over virus particles for type VI virion architecture. *For the tri-segmented RVFV, the distribution of segment types over virus particles is known, but the number of segment copies present has not been quantified. For the type IV and V virion architectures, the model assumptions automatically lead to the distribution of genome segments over virus particles. However, for the type VI, model parameter choices will affect how well the model matches the data. We must infer the distribution of the total number of genome segments over virus particles (as opposed to the distribution of genome segment types over particles) to know the relative number of virus particles that can be generated under the type VI virion architecture. Here we compare the observed (orange) and predicted (blue) distribution of the number of genome segment types over virus particles The comparison is between the observed distribution and type VI model predictions of the number of segment types per virus particle is made for mammalian (top panels) and insect cells (bottom panels). The two models considered for this distribution, are Model comparisons for panel (a) and (b) are for type IV virion architecture: binomial distribution of the total number of genome segments per virus particle (left panels) and for panel (c) and (d) type VI virion architecture: an unconstrained distribution (see Methods section) derived from fitting a model empirical data (right panels). For the Binomial model, there are clearly discrepancies between the model and both datasets, whereas the unconstrained distribution provides a better fit. For the type VI virion architecture*

Supplementary S2

The type III virion architecture describes segmented viruses which co-package their genome segments in virus particles. The process of segment packaging can be a highly orchestrated process ensuring that a single copy of each segment is included within the virus particle (Chou et al. 2012). This is reliant on high fidelity packaging, with no error, however experimental evidence has shown that there can be uneven packaging of genome segments due to difference in packaging fidelity (Brooke et al. 2013). The consequence of differences in packaging fidelity are virus particles which contain incomplete sets of genome segments and thereby introducing differences in the infectivity of virus particles based on their genome segment composition (Diefenbacher, Sun, and Brooke 2018; Farrell et al. 2023). Not only do differences in error rates introduce biases in virus particle's segment composition, these differences may also be advantageous for virus infection in different tissues and to escape host immune responses (Farrell et al. 2023; Vahey and Fletcher 2020). We therefore devised simulation-based approach to determine the cost to infectivity for the type III virion architecture when there is high, medium and low packaging fidelity.

Simulation-based predictions for infectivity cost: type III with packaging error

To determine the cost to infectivity of a type III virion architecture, we use a modified version of Model B (see: Methods) by providing the corresponding frequency of virus particle types (i.e., the set $\{r_1, r_2, [\dots]\}$) for a bi-segmented and and tri-segmented virus. For the type III architecture with packaging errors, the distribution of virus particle types needs to be predicted from the error rate. We assume (i) the error rate for packaging is τ (per segment packaged into a virus particle), and (ii) that the error rate for each segment packaging event is independent. For a bi-segmented virus, the expected frequencies of each virus particle type are given in Figure S1. There are four possible combinations of virus particle types $\{r_1, r_2, r_3, r_4\}$ describing the frequency of correct packaging of segments 1 and 2, and the frequency when either one is packaged with error. Note that if both segments 1 and 2 are packaged incorrectly, they are both of the wrong type and therefore a complete genome is present. Hence, all of the double-error fraction must be added to the correctly package fraction to determine the fraction of virus particle with complete genomes, $(1 - e^{-\tau})^2 e^{-2\tau}$. For a tri-segmented virus, having multiple packaging errors does not automatically lead to the presence of all genome segments, and only a subset of the multiple-error fractions restore the full genome. By considering all the possibilities, we can again work out the expected frequencies of all virus particle types (Figure S2). We determine predictions for low and high error rates. When τ = 0.025, for a bi-segmented virus $\{r_1, r_2, r_4\} = \{0, 0.024, 0.951\}$ and for a tri-segmented virus $\{r_1, r_2, r_5, r_8\}$ = $\{0, 1.49 \times 10^{-4}, 0.024, 0.928\}$. When τ = 0.25, for a bi-segmented virus $\{r_1, r_2, r_4\}$ = {0,0.172,0.655} and for a tri-segmented virus $\{r_1, r_2, r_5, r_8\}$ ={0, 0.010,0.156,0.504}.

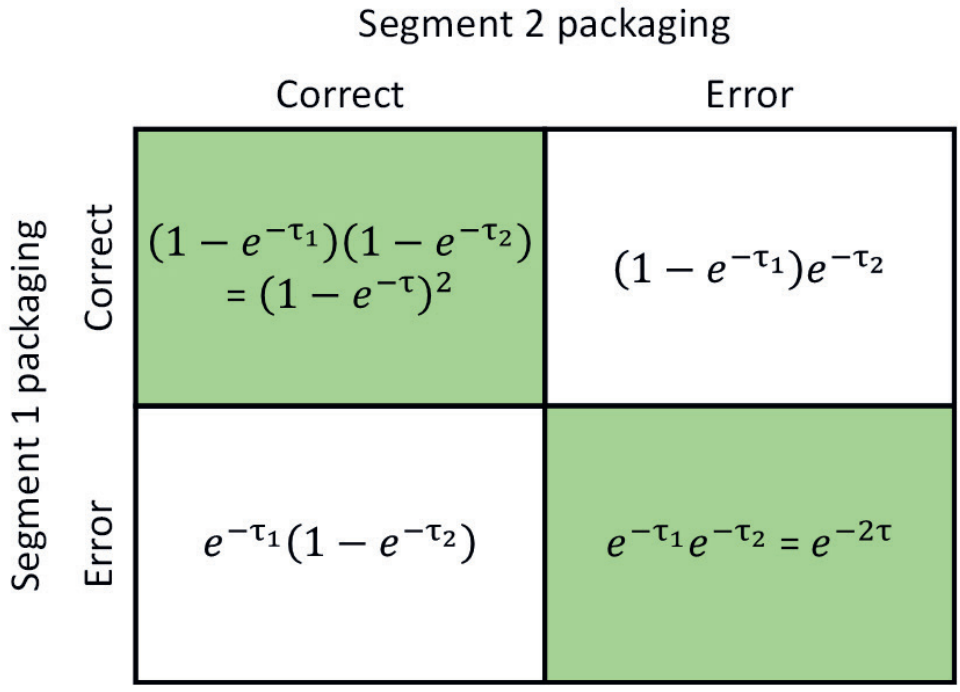


***Figure S2**. **Consequences of packaging errors for the distribution of genome segments over virus particles for a bi-segmented virus**. τ is the rate at which an incorrect segment is packaged, and the subscript indicates the rate for a specific segment for clarity (but $\tau_1 = \tau_2 = \tau$). In the green regions, both segment types are packaged into a virus particle. Note that for the bi-segmented virus, if both segments are packaged incorrectly the complete genome is present.*

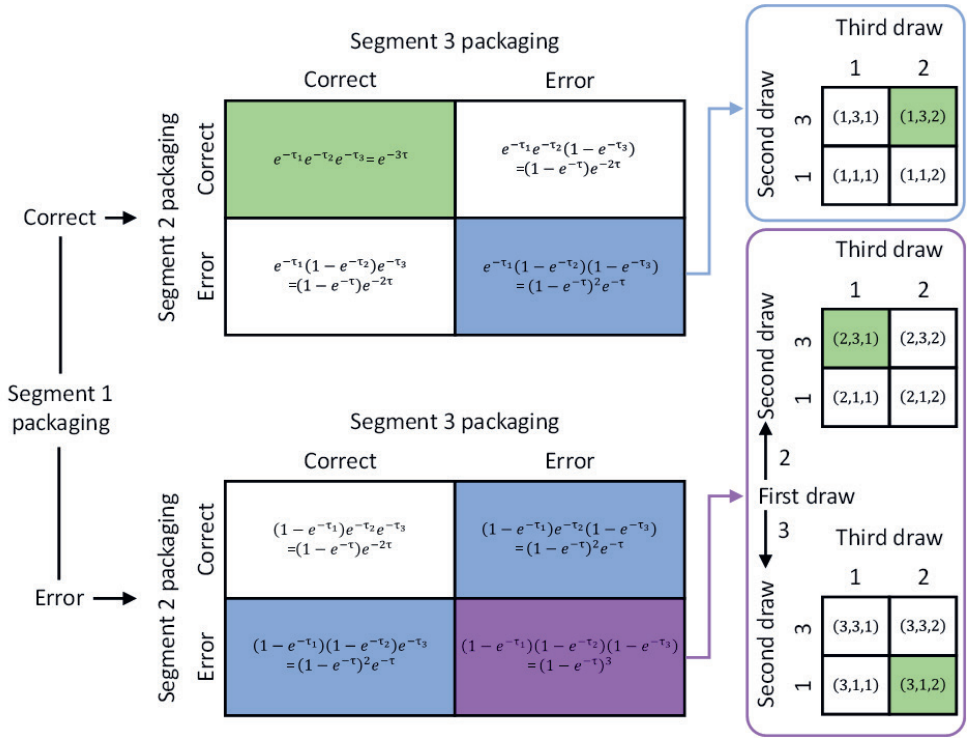


Figure S3. Consequences of packaging errors for the distribution of genome segments over virus particles for a tri-segmented virus. *τ is the rate at which an incorrect segment is packaged, and the subscript indicates the rate for a specific segment for clarity (but $\tau_1 = \tau_2 = \tau_3 = \tau$). The green color indicates the region in which all three segment types are packaged into a virus particle, where the blue and magenta regions indicate regions in which all segments are sometimes packaged into virus particles. In all three blue regions, two packaging errors are made and if the right combination of errors is made, all segments will still be present. This occurs in $\frac{1}{4}$ of instances in each of these three regions, as illustrated in the box with the blue outline in the top right for one scenario (i.e., the case in which the correct segment is packaged only for the first segment.) In the magenta region, three packaging errors occur, and in $\frac{2}{8} = \frac{1}{4}$ of instances a complete genome is present, as illustrated in the box with the magenta outline in the bottom right. Note that in the figure we focus on which fractions with errors contain a complete genome, but even when they do not lead to the completion of the genome, each fraction with errors must be added to the virus type it corresponds to.*

References

Bermúdez-Méndez, Erick, Kirsten F. Bronsvoort, Mark P. Zwart, Sandra van de Water, Ingrid Cárdenas-Rey, Rianka P. M. Vloet, Constantianus J. M. Koenraadt, Gorben P. Pijlman, Jeroen Kortekaas, and Paul J. Wichgers Schreur. 2022. "Incomplete Bunyavirus Particles Can Cooperatively Support Virus Infection and Spread." *PLoS Biology* 20 (11): e3001870.

Borchers, Hans W. 2023. *Pracma: Practical Numerical Math Functions* (version 2.4.4). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=pracma.

Boshra, Hani. 2022. "An Overview of the Infectious Cycle of Bunyaviruses." *Viruses* 14 (10). https://doi.org/10.3390/v14102139.

Brooke, Christopher B., William L. Ince, Jens Wrammert, Rafi Ahmed, Patrick C. Wilson, Jack R. Bennink, and Jonathan W. Yewdell. 2013. "Most Influenza a Virions Fail to Express at Least One Essential Viral Protein." *Journal of Virology* 87 (6): 3155–62.

Brooke, Christopher B., William L. Ince, Jiajie Wei, Jack R. Bennink, and Jonathan W. Yewdell. 2014. "Influenza A Virus Nucleoprotein Selectively Decreases Neuraminidase Gene-Segment Packaging While Enhancing Viral Fitness and Transmissibility." *Proceedings of the National Academy of Sciences of the United States of America* 111 (47): 16854–59.

Chao, Lin, Thu T. Tran, and Thutrang T. Tran. 1997. "The Advantage of Sex in the RNA Virus φ6." *Genetics* 147 (3): 953–59.

Chou, Yi-Ying, Reza Vafabakhsh, Sultan Doğanay, Qinshan Gao, Taekjip Ha, and Peter Palese. 2012. "One Influenza Virus Particle Packages Eight Unique Viral RNAs as Shown by FISH Analysis." *Proceedings of the National Academy of Sciences of the United States of America* 109 (23): 9101–6.

Diefenbacher, Meghan, Jiayi Sun, and Christopher B. Brooke. 2018. "The Parts Are Greater than the Whole: The Role of Semi-Infectious Particles in Influenza A Virus Biology." *Current Opinion in Virology* 33 (December): 42–46.

Farrell, Alex, Tin Phan, Christopher B. Brooke, Katia Koelle, and Ruian Ke. 2023. "Semi-Infectious Particles Contribute Substantially to Influenza Virus within-Host Dynamics When Infection Is Dominated by Spatial Structure." *Virus Evolution* 9 (1): vead020.

Fulton, Robert W. 1962. "The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus." *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.

Gallet, Romain, Jérémy Di Mattia, Sébastien Ravel, Jean-Louis Zeddam, Renaud Vitalis, Yannis Michalakis, and Stéphane Blanc. 2022. "Gene Copy Number Variations at the within-Host Population Level Modulate Gene Expression in a Multipartite Virus." *Virus Evolution* 8 (2): veac058.

Gelderblom, Hans R. 1996. *Structure and Classification of Viruses*. University of Texas Medical Branch at Galveston.

Gutiérrez, Serafín, and Mark P. Zwart. 2018. "Population Bottlenecks in Multicomponent Viruses: First Forays into the Uncharted Territory of Genome-Formula Drift." *Current Opinion in Virology* 33 (December): 184–90.

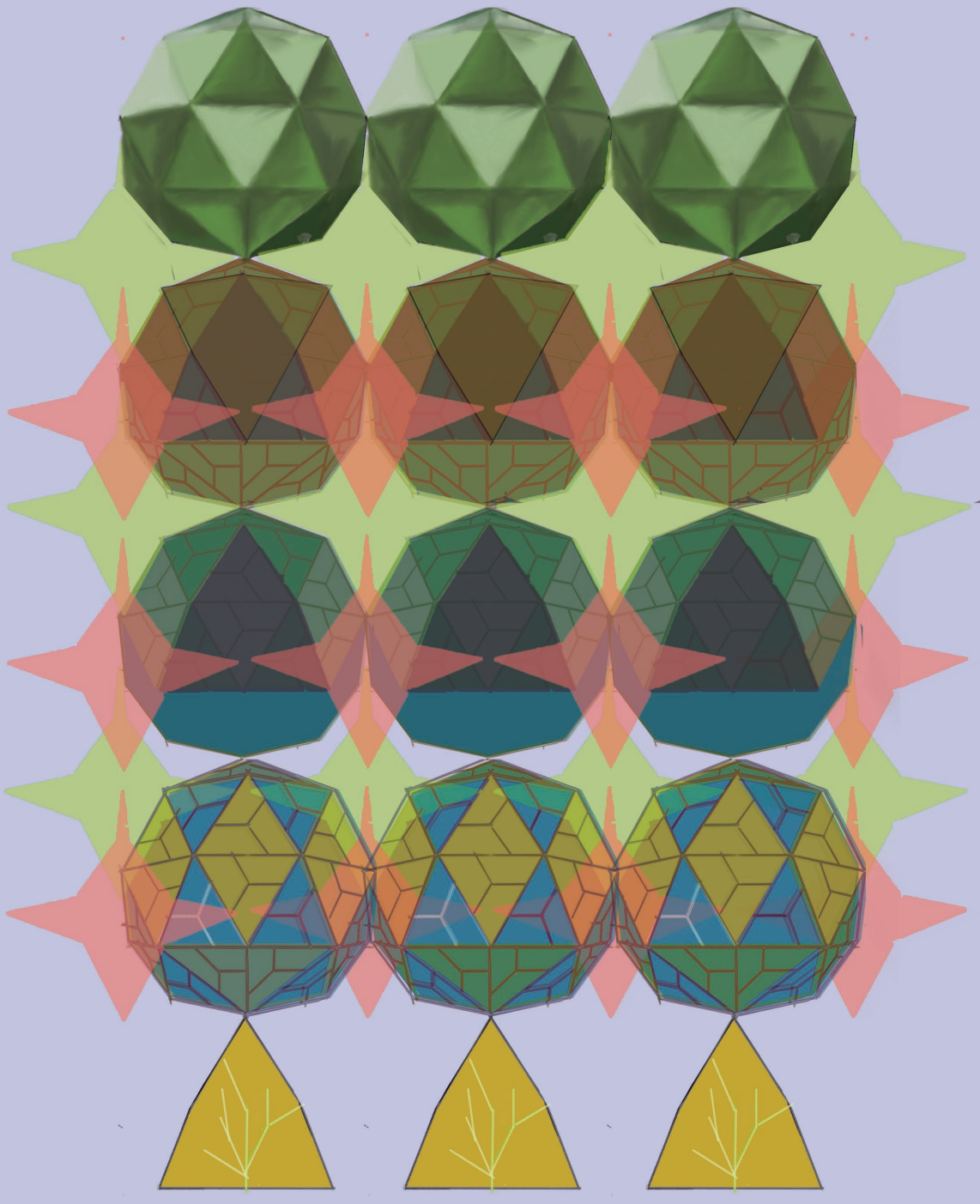
Hespenheide, B. M., D. J. Jacobs, and M. F. Thorpe. 2004. "Structural Rigidity in the Capsid Assembly of Cowpea Chlorotic Mottle Virus." *Journal of Physics. Condensed Matter: An Institute of Physics Journal* 16 (44): S5055.

Hu, Zhaoyang, Xiaolong Zhang, Wei Liu, Qian Zhou, Qing Zhang, Guohui Li, and Qin Yao. 2016. "Genome Segments Accumulate with Different Frequencies in Bombyx Mori Bidensovirus." *Journal of Basic Microbiology* 56 (12): 1338–43.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, Arvind Varsani, Yuri I. Wolf, Natalya Yutin, F. Murilo Zerbini, and Jens H. Kuhn. 2020. "Global Organization and Proposed Megataxonomy of the Virus World." *Microbiology and Molecular Biology Reviews:*

MMBR 84 (2). https://doi.org/10.1128/MMBR.00061-19.
Kormelink, R., P. de Haan, D. Peters, and R. Goldbach. 1992. “Viral RNA Synthesis in Tomato Spotted Wilt Virus-Infected Nicotiana Rustica Plants.” *The Journal of General Virology* 73 (Pt 3) (March): 687–93.
Ladner, Jason T., Michael R. Wiley, Brett Beitzel, Albert J. Auguste, Alan P. Dupuis, Michael E. Lindquist, Samuel D. Sibley, et al. 2016. “A Multicomponent Animal Virus Isolated from Mosquitoes.” *Cell Host & Microbe* 20 (3): 357–67.
Li, Xiuli, Min Gu, Qinmei Zheng, Ruyi Gao, and Xiufan Liu. 2021. “Packaging Signal of Influenza A Virus.” *Virology Journal* 18 (1): 36.
Michalakis, Yannis, and Stéphane Blanc. 2020. “The Curious Strategy of Multipartite Viruses.” *Annual Review of Virology* 7 (1): 203–18.
Moreau, Yannis, Patricia Gil, Antoni Exbrayat, Ignace Rakotoarivony, Emmanuel Bréard, Corinne Sailleau, Cyril Viarouge, et al. 2020. “The Genome Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner.” *Journal of Virology* 95 (1). https://doi.org/10.1128/JVI.01834-20.
Pressing, J., and D. C. Reanney. 1984. “Divided Genomes and Intrinsic Noise.” *Journal of Molecular Evolutionm* 20: 135–46.
Rohrmann, George F. 2019. *Baculovirus Molecular Biology*. National Center for Biotechnology Information (US).
Sánchez-Navarro, Jesús A., Mark P. Zwart, and Santiago F. Elena. 2013. “Effects of the Number of Genome Segments on Primary and Systemic Infections with a Multipartite Plant RNA Virus.” *Journal of Virology* 87 (19): 10805–15.
Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. “The Strange Lifestyle of Multipartite Viruses.” Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819.
Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. “Gene Copy Number Is Differentially Regulated in a Multipartite Virus.” *Nature Communications* 4: 2248.
Slack, Jeffery, and Basil M. Arif. 2007. “The Baculoviruses Occlusion-Derived Virus: Virion Structure and Function.” *Advances in Virus Research* 69: 99–165.
Solovyev, A. G., and V. V. Makarov. 2016. “Helical Capsids of Plant Viruses: Architecture with Structural Lability.” *The Journal of General Virology* 97 (8): 1739–54.
Sun, Jiayi, and Christopher B. Brooke. 2018. “Influenza A Virus Superinfection Potential Is Regulated by Viral Genomic Heterogeneity.” *MBio* 9 (5). https://doi.org/10.1128/mBio.01761-18.
Vahey, Michael D., and Daniel A. Fletcher. 2020. “Low-Fidelity Assembly of Influenza A Virus Promotes Escape from Host Cells.” *Cell* 180 (1): 205.
Wichgers Schreur, Paul J., and Jeroen Kortekaas. 2016. “Single-Molecule FISH Reveals Non-Selective Packaging of Rift Valley Fever Virus Genome Segments.” *PLoS Pathogens* 12 (8): e1005800.
Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. “Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus.” *Scientific Reports* 7 (1): 1–15.
Yu, Nai-Tong, Hui-Min Xie, Yu-Liang Zhang, Jian-Hua Wang, Zhongguo Xiong, and Zhi-Xin Liu. 2019. “Independent Modulation of Individual Genomic Component Transcription and a Cis-Acting Element Related to High Transcriptional Activity in a Multipartite DNA Virus.” *BMC Genomics* 20 (1): 573.
Yvon, Michel, Thomas L. German, Diane E. Ullman, Ranjit Dasgupta, Maxwell H. Parker, Sulley Ben-Mahmoud, Eric Verdin, et al. 2023. “The Genome of a Bunyavirus Cannot Be Defined at the Level of the Viral Particle but Only at the Scale of the Viral Population.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (48): e2309412120.
Zwart, Mark P., and Santiago F. Elena. 2015. “Testing the Independent Action Hypothesis of Plant Pathogen Mode of Action: A Simple and Powerful New Approach.”

Phytopathology® 105 (1): 18–25.
———. 2020. “Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments.” *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa022.
Zwart, Mark P., Monique M. van Oers, Jenny S. Cory, Jan W. M. van Lent, Wopke van der Werf, and Just M. Vlak. 2008. “Development of a Quantitative Real-Time PCR for Determination of Genotype Frequencies for Studies in Baculovirus Population Biology.” *Journal of Virological Methods* 148 (1-2): 146–54.



General Discussion

Exploring the costs and benefits of multipartition

Viruses have genomes that can be divided into one or several molecules called segments, and differ in the way these segments are packaged into particles. Monopartite viruses package their single nucleic acid molecule into a single particle, whilst in segmented viruses, segments are co-packaged into a single virus particle and transmitted together, ensuring that infections can be initiated with a single virus particle. Multipartite viruses have a genome divided into several segments, each of them packaged independently into virus particles. All segments are required for infection, thus presenting a conundrum: how can successful infections be initiated when there is a physical disconnect at each between-host transmission event? Monopartite viruses have single-hit infection kinetics, whereby a single infectious particle is required to initiate infection (Druett 1952; Bald 1937). Unlike monopartite viruses, multipartite viruses require complementation between segments to initiate infection and display multi-hit infection kinetics (Fulton 1962; Lauffer and Price 1945). This can be observed in the dose-response relationship with steeper gradients for segmented and multipartite viruses and a shift in the position of the curve (Fulton 1962; Lauffer and Price 1945). This poses a cost to transmission, as many segments and a higher dose are required to initiate infection (Iranzo and Manrubia 2012). It is a well-described dilemma for multipartite virus infection, and the cost of multipartition has not been determined empirically (Gutiérrez and Zwart 2018). We do not have an accurate estimate for the potential cost of transmission for virus genomes which differ in the number of genome segments and their genome organization. Simultaneously, we are aware that segmented viruses may have incomplete packaging of viral genome segments leading to virus populations in which the complete genome set is not co-transmitted (Nakatsu et al. 2018, 2016; Diefenbacher, Sun, and Brooke 2018). This would have an increase in the cost of transmission when compared to segmented viruses, which co-package a complete genome set.

Possible benefits of a multipartite virus genome organization have been an open question since their discovery, and proposed benefits to the segmentation of the genome include: (1) faster replication of shorter segments when polymerase is not a limiting factor (Nee 1987; Sicard et al. 2016), (2) increased genetic diversity by recombination and segment reassortment (Sicard et al. 2016), (3) increased virus particle stability (Ojosnegros et al. 2011) and (4) adaptive gene expression change via segment frequency (Sicard et al. 2013). The benefits of multipartition have recently been reviewed (Sicard et al. 2016; Michalakis and Blanc 2020; Lucía-Sanz and Manrubia 2017) and discussed in **Chapter 1** of this thesis. Benefits (1) and (2) are shared with segmented viruses (Sicard et al. 2016), whilst the proposed benefits of viral gene expression regulation by altering genome segment copies, “the genome formula” (GF), has been described for the octapartite faba bean necrotic stunt virus (FBNSV) (Sicard et al. 2013).

In the following sections, I provide an overview of the experimental results from this thesis and detail how these findings contribute to understanding: (1) the cost to transmission in multipartite and segmented viruses, (2) the role of GF drift in deleterious GF change, and (3) how GF change in CMV does not contribute to viral adaptation. I end this final thesis chapter with a discussion on what factors might contribute to the existence of multipartite viruses.

Quantifying the cost of infectivity in multipartite and segmented viruses

In **chapter 2** of this thesis, I quantify the cost of transmission for multipartite viruses by combining the change in the gradient of the dose-response relationship as well as the shift in the position of the curve. This was done by re-analyzing experimental data for infections of the tripartite alfalfa mosaic virus (AMV) in different *Nicotiana tabacum* hosts which constitutively express one or two AMV genes (Sánchez-Navarro, Zwart, and Elena 2013; Taschner et al. 1991). By infecting plants with AMV inoculations in which one or two segments are provided in excess, it is possible to quantify how changes in segment number alter the dose-response relationship. Our results, re-analyzing the dose-response relationship of AMV, show that the cost of transmission for multipartition is higher than predicted, because we considered both the gradient and position of the dose response curve.

Segmented viruses may be divided into two groups based on the packaging strategy: selective and non-selective packagers (Figure 1 in Chapter 6). Selective packagers (e.g., Type III according to the classification used in Chapter 6), such as the influenza A virus (IAV), are thought to package a full complement of genome segments into each virus particle (Chou et al. 2012; Nakatsu et al. 2018). The choreographed packaging of segments is mediated by packaging signals and RNA-RNA interactions (Li et al. 2021; Hutchinson et al. 2010; Goto et al. 2013). The efficiency of segmented virus packaging may vary and a considerable portion of the virus population may contain incomplete sets of the viral genome (Nakatsu et al. 2016; Diefenbacher, Sun, and Brooke 2018; Brooke et al. 2013). Non-selectively packaging segmented viruses differ in the manner in which segments are distributed over virus particles; the identity, the number of genome segments or both may vary across virus particles. This results in three hypothetical classifications that we considered: Type IV: segmented non-selective packager, a fixed number of genome segments of variable identity; Type V: segmented non-selective packager, variable number of genome segments and fixed segment identity, or both (Type VI: segmented non-selective packager, variable number of genome segments and variable segment identity). For the non-selective packaging viruses, a well-studied example is the Rift Valley fever virus (RVFV), which has been shown to produce a large fraction of virus particles that are devoid of any segments or contain an incomplete genome set (Bermúdez-Méndez et al. 2022; Wichgers Schreur and Kortekaas 2016). In **chapter 6,** I develop a modelling and simulation approach to quantify the cost to infectivity of these putative segmented virus genome organizations, which differ in their segment co-packaging strategies, and link them to what is known about the empirical distribution of genome segments over virus particles in RVFV**.**

Under the model we propose, selectively packaging segmented viruses (i.e., Type III) do not have a higher cost to transmission than monopartite viruses (Type I). When selective packagers have frequent errors in packaging fidelity, the cost to transmission approaches that of the non-selectively packaging virus (E.g., Type V). Furthermore, we show that some non-selectively packaging viruses (I.e., Type IV and VI) have a cost to transmission which is nearly equivalent to that of the multipartite viruses, and that at high doses the multipartite virus has a lower cost to infectivity than the non-selective packaging virus (I.e., Type IV). Combining the results from chapter 2 and chapter 6, we show that the cost to infectivity for multipartite viruses

is high but that in some postulated cases multipartite viruses may have a lower cost than segmented viruses. This suggests that the divisions between genome organizations are not as clear as initially described, and I propose that there may be a gradient in terms of the cost of transmission for the different genome organizations ranging from the least costly, monopartite viruses to the most costly, multipartite viruses (Figure 1).

The proposed gradient may be derived from differences in packaging strategy and the resulting distribution of genome segments over virus particles. Firstly, type III selective packagers have a cost of transmission which is equivalent to that of a monopartite virus as all segments are packaged and transmitted together. There may be differences in the fidelity of packaging, which may result in a decline in the virus particle content and an increase in the cost of transmission (pink arrow in Figure 1). For Type III segmented viruses there may be an increase in the cost of transmission as a function of the packaging error rate (represented as the increase in the light-blue shaded area occupied by Type III viruses in Figure 1). Furthermore, as the error rate increases packaging will eventually resemble that of the non-selective packagers (Type IV and V), whereby control is lost over the identity or number of segments packaged. Similarly, mechanisms which promote the packaging of near-complete virus particles, such as adhesions of particles to one another (Andreu-Moreno and Sanjuán 2018), may decrease the cost of transmission for multicomponent viruses. It becomes clear that, whilst the categories of genome organization are useful for understanding different packaging strategies for segmented and multipartite viruses, there may be instances where these strict boundaries may become less clear. .

Both of the approaches in chapter 2 and chapter 6 do not take into account ecological factors such as heterogeneity in host susceptibility, variation in exposure to viruses under real-world conditions, and virus transmission or other potential mechanisms which might mitigate the cost of transmission. Thus, whilst providing a useful framework for quantifying the cost of transmission in experimental settings, it does not provide an estimate of the cost to transmission in natural infections. I will now describe how some of the above factors may change the cost of transmission.

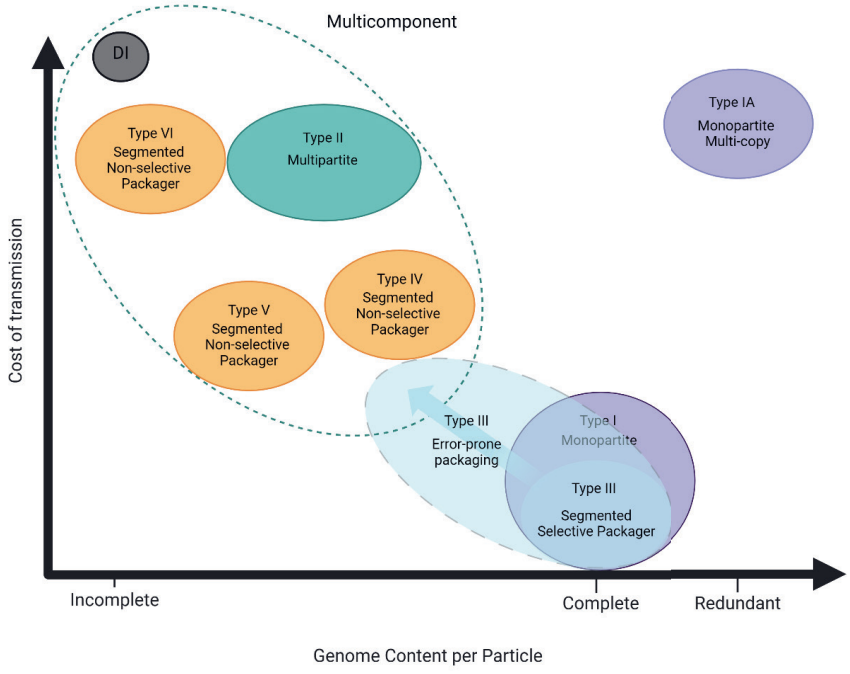


***Figure 1**. **A conceptual relationship between virus genome packaging strategy and the cost of transmission.** Monopartite viruses are indicated in purple: Type I and Type IA (monopartite viruses which package multiple copies of the virus genome in an occlusion body), segmented viruses (Type III selective packer in light blue and Non-selective packagers: Type IV – VI in orange) and in green the multipartite viruses (Type II). The light-blue arrow represents the Type III selective packager with errors. As the extent of error-prone packaging increases, the cost of transmission shifts towards the non-selective packaging segmented viruses. Defective interfering particles (DI) are represented in a dark brown circle, which may be associated with monopartite, segmented and multipartite viruses. The size of the ellipse is an approximation of the number of virus genomes which may employ this packaging strategy. Created with BioRender.com.*

The majority of multipartite viruses are found infecting plants (Michalakis and Blanc 2020; Lucía-Sanz and Manrubia 2017) and only a single insect virus (Bombyx mori bidensovirus) has been confirmed to be multipartite (Hu et al. 2013). Why would plants be a more suitable environment for multipartite viruses? Multipartite virus infection has relied on the tenet of complementation, that all virus genome segments be present within the same cell for an infection to be initiated. However, plant cells may present a unique environment for multipartite virus infection, as cells are connected to one another via plasmodesmata; junctions which physically connect cells one to one another (Faulkner 2018). Plant cells are continuously connected by the cytoplasm and microtubules, a feature known as the symplastic pathway (Faulkner 2018). Plasmodesmata are a pathway for the transport of water, photoassimilates,

essential nutrients and other macromolecules (Miras et al. 2022). This pathway is exploited by plant viruses to facilitate local cell-cell and long-distance movement (Miras et al. 2022; Heinlein 2015; Su et al. 2010). As part of within-host movement during infection, virus segments, virus ribonucleoprotein complexes and gene products may be transported via the symplastic pathway (Navarro, Sanchez-Navarro, and Pallas 2019; Lazarowitz and Beachy 1999; Kozieł, Julian Bujarski, and Otulak Kozieł 2023). This would potentially reduce the cost of transmission at the within-host level. Recently, it was shown that the octapartite FBNSV gene products may be found in cells where the respective genome segment is absent and that virus replication can occur across cells by complementation of cells containing at least one genome segment (Sicard et al. 2019). This represents the first observation that multipartite viruses do not require the full genome set within a cell for infection and replication. We developed a model to explore how gene product sharing may reduce the cost of transmission in **chapter 2**, finding that gene product sharing (ρ) is beneficial at moderate levels of sharing ($\rho < 0.5$), when MOI is low ($3 – 10^{0.5}$) and when virus replication is not sensitive to the GF ($\sigma^2 =10$). In contrast, earlier theoretical work suggests that multipartite viruses can outcompete monopartite viruses only when replication is sensitive to the GF (Zwart and Elena 2020). Combining these two results , therefore, suggests the two explanations (rapid adaptation in gene expression by changes in the GF and gene product sharing) are mutually exclusive. To date the GF has been measured at the whole plant level in systemic infection, it may be reasonable to assume that gene product sharing may be especially beneficial at the early infection stage when a small number of cells are infected and contain at least one virus genome segment. It remains to be seen to what extent gene product sharing may minimize the cost to within-host spread and how general gene product sharing is in other multipartite virus species.

The cost of transmission for multipartite viruses may be reduced by factors which influence vector transmission. Most plant viruses have arthropod vectors (Lucía-Sanz and Manrubia 2017). The total number of segments which are acquired during feeding will affect the rate of transmission, lowering the probability of missing segments (Betancourt et al. 2008; Ali et al. 2006). There will also be differences in the segment identities, as not all segments may be equally acquired during feeding. Combined, the total number of segments and the segment identity may affect the virus's genetic diversity in a starting population. For FBNSV the number of genome segments transmitted during aphid transmission differs per segment type and ranges between 3 - 7 copies (Gallet et al. 2018). Furthermore, the duration of aphid feeding and the number of aphids may influence how many segments are transmitted to a new host (Gallet et al. 2018). During longer acquisition times a larger pool of segments may be sampled, both in the total number and in segment type, thereby increasing the likelihood that a complete set of genome segments is transmitted. The number of aphids during acquisition and transmission increases the number of virus particles which may be transmitted, as several aphids may feed and re-inoculate the same plants. Combined, these factors will likely influence the exact conditions under which a multipartite virus is transmitted and therefore determine the magnitude of the fitness cost to infectivity. If virus doses are high, the between-host cost of transmission for multipartition may be limited, as suggested by other work (Valdano et al. 2019). However, real-world estimates of viral doses are unknown and this number is likely to be highly variable.

GF variation may be deleterious for CMV-i17F

A seminal study with FBNSV infections in *Vicia faba* and *Medicago truncatula* showed that genome segments accumulated to a host-specific ratio, the GF (Sicard et al. 2013). Furthermore, there was frequency-dependent selection towards an equilibrium the “setpoint genome formula”, which may be associated with higher virus titre (Sicard et al. 2013). The potential adaptive benefit of the GF linked to increased virus titre has also been described for the tripartite alfalfa mosaic virus (AMV) (Wu et al. 2017). These observations provide the first tentative link between the GF and a viral fitness component, virus titre. However, the exact nature of the link between the GF and virus titre remains unclear.

In this thesis the GF was measured in four different host species; *C. quinoa* (chapter 4), *A. thaliana, N. benthamiana and N. tabacum* (chapter 5). Although variation in the GF has only been reported for FBNSV (Sicard et al. 2013) and AMV (Wu et al. 2017), measurements of the GF variability are also limited to a small set of host plant species. The extent of GF variation is poorly understood and limits of the viable GF space (i.e. the full set of possible GF values for a virus that supports some level of virus replication) are unknown. Moreover, how the viable GF space varies within a single virus isolate and across different host species are also not known. Can we start to map the GF landscape for CMV, determining the GF space that can support infection and the relationship between the GF and virus titre? In other words, is it possible to obtain empirical GF fitness landscapes?

In **chapter 3**, I developed a method for quantifying GF variation; the genome formula distance (*D*). The Euclidean distance between two GF values. For this metric I make predictions of GF variation for two scenarios: (i) stochastic GF variation for different segments and (ii) a maximum GF drift, when there is a single population bottleneck. I use *D* to estimate GF variation under several different conditions; when comparing GF values from different experimental groups, and comparing GF variation between the inoculum and the infected tissues. I show by re-analyzing data from AMV (Wu et al. 2017), that the inoculum GF influences that found in infected leaves. To my knowledge, this is the first report that the GF may be transmitted from one host to another. and indicates that the GF may be heritable.

In **chapter 4** of this thesis, I investigate the relationship between the GF and virus titre in local lesion infections in *Chenopodium quinoa*. *C. quinoa* plants are infected with an isolate of the tripartite single-stranded RNA virus, cucumber mosaic virus-i17F (CMV-i17f). Plants were inoculated with CMV derived from infected *Nicotiana tabacum* and *N. benthamiana* in three separate experiments. I measured the GF in individual local lesions from each experiment by RT-qPCR and estimated virus titre from PCR cycle quantification (Cq) values, which are inversely related to virus titre. Using the GF variation metric (*D*), developed in **chapter 3** of this thesis, I was able to show that there is a high degree of GF variation in *C. quinoa* local lesions across experiments. As the three experiments differ in the inoculum source it is not possible to compare the experiments directly to one another, however I analyzed the distance between the inoculum and local-lesion GFs of inocula relative to that of the local lesions and found that there is a strong effect of inoculum on the GFs of local lesions. This is similar to results from the re-analysis of AMV (Wu et al. 2017) in chapter 3. Together these results show that the GF may be transmissible for AMV and CMV. Furthermore, it is in contrast to results

from FBNSV, where inoculum GF and inoculation procedure did not influence the observed GF in *V. faba* and *M. truncatula* plant species (Sicard et al. 2013). In the *C. quinoa* experiments mechanical transmission is used for inoculating CMV-i17F and this may play a role in the transmissibility of the GF.

In nature CMV is transmitted horizontally by aphids, most commonly by *Myzus persicae* and *Aphis gossypii.* This virus may also be vertically transmitted via infected seeds of host species (Jacquemond 2012). The experiments in chapter 4 used mechanical inoculation, a procedure which uses a high dose of the virus, often in excess to ensure successful infection. Thus, we can expect that that the bottleneck size is broad. However, analysis of GF variation in mechanical inoculation experiments 1 and 3 showed that for both experiments *D* was similar to predicted values corresponding to the maximum GF variability caused by a single bottleneck event. This result indicates that even at high virus doses, there is a narrow bottleneck for infections. This is likely due to the individual bottleneck size per local lesion, which is narrow. In nature CMV infections via vector transmission, have a narrow population bottleneck, as measured by considering genetic drift for a single genome segment (Ali et al. 2006; Betancourt et al. 2008). For CMV, a vector transmission bottleneck has been estimated to be on average ~ 3 individuals for two vectors, *A. gossypii* and *M. persicae* (Ali et al. 2006), and in another study was estimated as 1 – 2 individuals after transmission by the aphid *A. gossypii* (Betancourt et al. 2008). These bottleneck estimates appear to be in agreement with what is measured in Chapter 4, suggesting that although mechanical inoculation is at a high dose, experimental measurements of bottleneck size are comparable to those seen for vector transmission. Seed transmission ensures that CMV infection persists in an environment, as the infected seed bank may be maintained in a landscape patch which has a higher density of suitable host plants. In seed transmission of CMV, infection occurs primarily in the seed embryo and seed coat (Ali and Kobayashi 2010). This provides an early start for infection, as a small virus population is maintained and can develop over a longer time period during the course of infection (weeks or months) (Cobos et al. 2019). For GF transmission via seed, I speculate that there is likely a GF which is specific for the initial plant development stages and the potential for maintenance of an equilibrium GF. (Vitti et al. 2022) show that seed transmission of CMV in *N. tabacum* is most likely for the first generation of offspring. In the case of the GF, this would, therefore, represent that it is unlikely to have intergenerational transmission of the GF over longer periods of time. In both vector and seed transmission the bottleneck size may be similar or larger to that in mechanical transmission.

Furthermore, in **chapter 4** I was able to show that in experiment 1, which used an inoculum with high levels of CMV RNA1, there is bimodal GF variation. The existence of GF clusters is statistically supported in experiment 1, where the majority cluster (cluster 1) is associated with high virus titre compared to the minority cluster (cluster 2) which has a low virus titre. The low titre GF space identified in experiments 1 has not been observed for other multipartite viruses. In AMV and FBNSV the GF stabilizes around an equilibrium level (Wu et al. 2017; Sicard et al. 2013), which may be linked to increased virus titre. I also show that the low titre local lesions seen in experiment 1 are associated with a GF which is closer to that of the inoculum, indicating that the low titre GF space can be accessed when it is transferred from the inoculum. The bimodal GFs identified in chapter 4 were not observed for CMV-i17F infection in *C. quinoa* (Boezen, Johnson, et al. 2023) (Figure 2). These observations are based on whole leaves and not local lesions, thus representing the mean GF over the leaf from several lesions. The GF

identified in Boezen et. al (2023) is more balanced and is comparable to the dominant GF in chapter 4 experiment 1, which is associated with higher virus titre.

Local lesions are the consequence of the host hypersensitive response (HR), which activates programmed cell death at sites of infection to limit the spread of the virus. This spatial separation, combined with observation of high GF variation in experiments 1 and 3 approaching prediction for maximum GF drift, suggests that the variation is a consequence of a single bottleneck event. By combining the empirical GFs from experiments 1 – 3 and Boezen et al. (2023), it is possible to identify regions where CMV GFs are likely to occur and given the data for virus accumulation, begin to map the empirical viable GF space for CMV in *C. quinoa* analogous to traditional genetic fitness landscapes. Firstly, we observe that a broad GF space can be occupied for CMV-i17f local lesion infections (Figure 2). Secondly, most GFs are not located in a region with higher RNA2 levels (> 0.5), whilst RNA1 levels are rarely found below 0.2 and RNA3 levels don't exceed a frequency of 0.9. The majority of GF measures occur within a balanced central region which is the identified cluster 1 in experiment 1. The second identified region is that of cluster 2 in experiment 1, between 0.8 – 1.00 RNA1: 0 – 0.5 RNA2 and 0 -0.2 RNA3. Whilst the separation for these clusters is supported in experiment1, it is not supported in experiment 3. Data on virus accumulation (Figure 3 here and Chapter 4 of this thesis), show that these clusters may be associated with differences in virus titre.

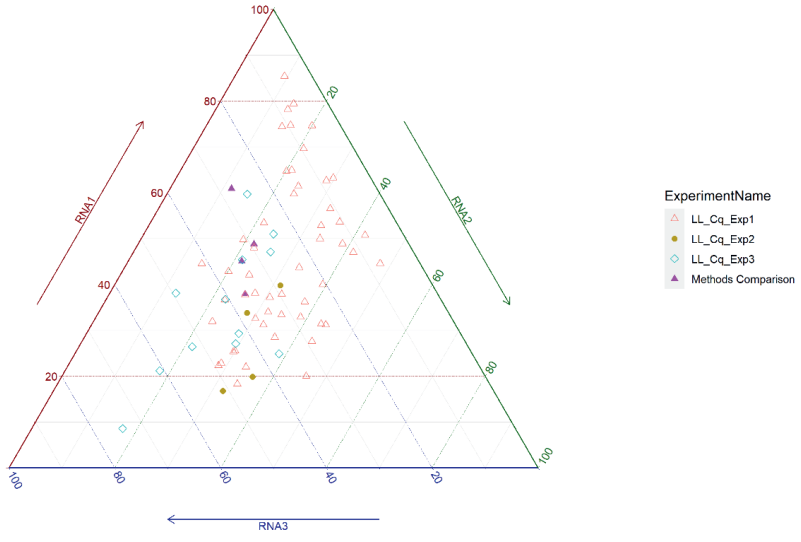


Figure 2. GF variation of CMV-i17F in C. quinoa local lesion and whole leaf infections. *Local lesion infections of CMV-i17f at 10 days post infection (dpi) from three experiments: Chapter 3, this thesis) and whole leaves of C. quinoa (Boezen, Johnson, et al. 2023) Local lesion infections from Chapter 3 of this thesis, experiment 1 (LL_Cq_Exp1: n = 52, red triangles), experiment 2 (LL_Cq_Exp2: n = 4, brown dots) and experiment 3 (LL_Cq_Exp3: n = 12, blue squares). CMV-i17F infections from whole leaves of C. quinoa (Boezen, Johnson, et al. 2023) (Methods Comparison: n = 4, 8 dpi, magenta triangles).*

The empirical fitness landscape of *C. quinoa* was investigated by using the GF distance metric (*D*), following the approach developed in chapter 3 of this thesis and reported earlier by (Wu et al. 2017). First, for *C. quinoa* experiments 1 – 3 (chapter 4 of this thesis), data was combined, and the mean GF was calculated. Thereafter, the distance to this mean GF (*D)* was calculated for each local lesion individually. The strength of the relationship between virus titre and GF was statistically tested by performing a Kendall rank correlation (Kendall 1938; Abdi 2007) with ties using the cor.test function in R 4.3.1 (R Foundation for Statistical Computing 2023). For the combined *C. quinoa* data, there is a weak, but significant positive relationship between *D* and the Cq value obtained by RT-qPCR, which is inversely related to virus titre (*Z = 2.473, p = 0.013, τ = 0.206*). Simplified, the greater *D* for individual local lesions, the lower the virus titre. We know from chapter 4 of this thesis, that the GF and virus titres in experiment 1 have a bimodal distribution and cluster into two groups. We can analyze the individual clusters to determine if GFs further from the centroid have lower virus titre. Combining the data from all three experiments, we contrast two groups: (1) the cluster 1 samples, as determined statistically for experiment 1 and including all samples from experiments 2 and 3, which fall in the same GF space, and (2) the cluster 2 samples from experiment 1 (Chapter 4, this thesis). This was done for the majority cluster 1, finding that there is no significant relationship between *D* and virus titre (Kendall rank correlation*: Z = 1.743, p = 0.081, τ = 0.179*). This indicates that cluster 1 has a broad GF space and a relatively stable high titre (Figure 3). When analyzing the local lesions grouped into cluster 2, we also find no statistically significant relationship between virus titre and *D* (*Z = -0.226, p = 0.821, τ = -0.034*), indicating that within the smaller cluster 2, the virus titre does not decrease with increasing GF distance. Combined, these results suggest that the GF space may be described as a mesa-like landscape: a broad central region where the GF can vary without appreciable consequences for fitness. There is an neighbouring narrow valley where no GF values have been measured, followed by a smaller mesa where considerable GF variation is allowed, but all GF values are associated with lower fitness. The fitness for all GF spaces other than the two mesas, including the narrow valley that separates them, is so low that we do not observe any populations there. I speculate that the lack of observations in this space may be due to selection constraining the GF to the two mesas. These results and the patterns inferred may be used as a starting point for future experiments to characterize the GF fitness landscape. Having observed the broad GF variability in *C. quinoa* local lesion infections of CMV, it is unknown whether a similar GF space may be found in other host species and whether there may be links to virus titres.

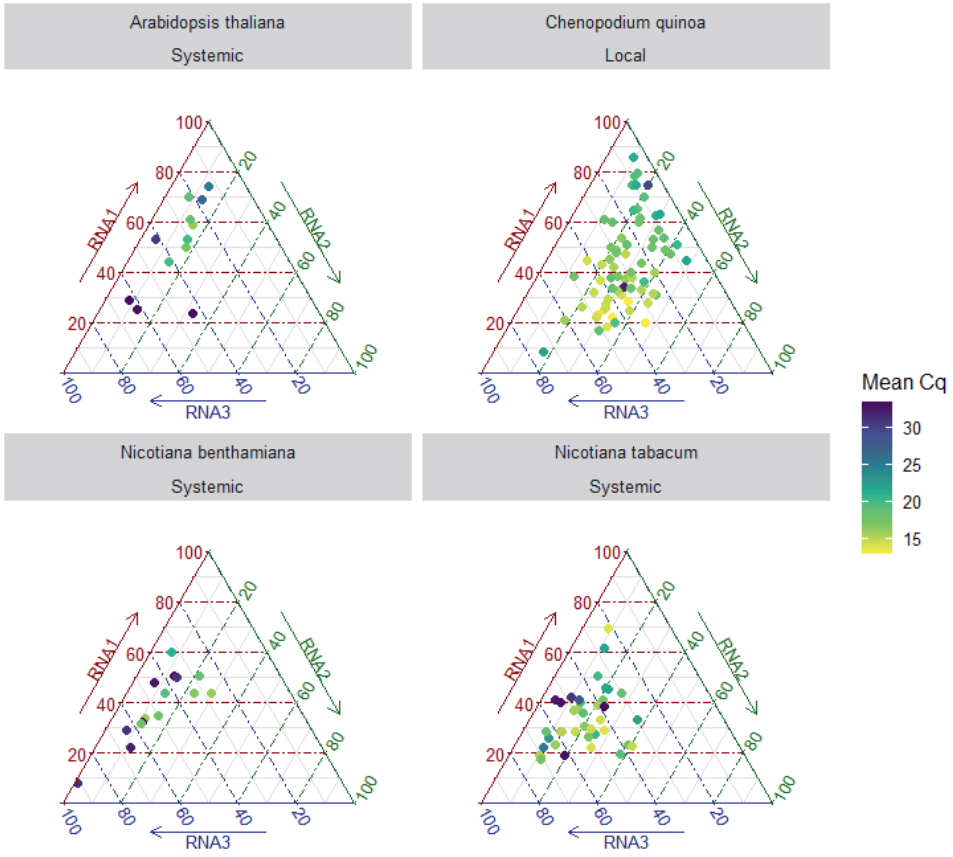


***Figure 3*. Empirical GF fitness landscape of CMV-i17f in systemic (14 days post infection) of *A. thaliana*, *N. benthamiana* and *N. tabacum* and local infections in *C. quinoa*.** *Virus titre is used as a proxy for viral fitness during the course of infection. Tire is defined as the cycle quantification (Cq) value from qPCR of CMV-i17f local lesion and systemic infections. Cq is inversely related to virus titre, such that lower values indicate higher virus titre whilst higher Cq values indicate lower virus titre. A. thaliana (n = 12, Chapter 5 of this thesis), local lesion infections in C. quinoa (10 dpi) (experiment 1: n = 52, experiment 2: n = 4 and experiment 3: n = 12, Chapter 4 of this thesis). CMV-i17F infections from whole leaves of C. quinoa (Boezen, Johnson, et al. 2023) (n = 4, 8 dpi). N. benthamiana (n = 12, from Chapter 5 of this thesis and n = 4 (Boezen, Johnson, et al. 2023). N. tabacum (n= 12, Chapter 5 of this thesis, n= 4 (Boezen, Johnson, et al. 2023), n = 9 from (Boezen, Vermeulen, et al. 2023) and n=14 from Johnson, Grum-Grzhimaylo et al. Unpublished data).*

In **chapter 5** of this thesis, I quantify the GF in three host species; *Arabidopsis thaliana*, *Nicotiana benthamiana* and *Nicotiana tabacum*; and the dynamics of GF change using an experimental evolution approach. I measure the GF in ancestral and evolved lines and attempt to link it to changes in virus titre, a virus fitness component. Results from this experiment are unexpected, as we expect to see adaptive GF changes whereby for different host species there is a host-specific GF and that it is linked to improved virus fitness (e.g. higher virus titre)

(Sicard et al. 2013; Wu et al. 2017). The results show that; (1) a host-specific GF shift is only observed for *A. thaliana* whilst *N. tabacum* and *N. benthamiana* have similar GFs to one another, (2) there is no clear adaptation of the GF in each host: that is, the initial host-specific GF in the first passage is unchanged at the final passage within a host species, (3) several virus populations go to extinction, and (4) there appears to be a decline in virus titre potentially associated with changes in the GF. I will begin by discussing the range of GF values measured for CMV-i17F in the three host species; *A. thaliana, N. benthamiana* and *N. tabacum* and compare and contrast this to observations in *C. quinoa*.

The CMV-i17F GF in *A. thaliana* appears to be restricted to an area with 0.3 -0.9 RNA1:0 – 0.2 RNA2:0.8 – 1.0 RNA3 (Figure 3). The majority of observations were found within this GF region, with a single observation found with RNA3 level at 0.65. Second, the axis of GF variation appears to be largely for RNA1. This suggests that in *A. thaliana* the GF may be more constrained compared to that found in *C. quinoa*. *N. benthamiana* has a GF space similar to that of *A. thaliana,* however GF variation along RNA1 is more restricted (GF values do not exceed a frequency of 0.6 for RNA1). The GF space occupied for *N. tabacum* is also found with variation along RNA1 axis, but exhibits more variation in RNA3 (0.3 – 0.95) and RNA2 (0 – 0.4). When comparing the GF space of *A. thaliana*, *N. benthamiana* and *N. tabacum* to that of *C. quinoa*, it is clear that in those host species the GF space appears more constrained. These differences may be formed partly by the differences in infection process at play. In *C. quinoa* local lesion infections restrict virus movement due to the antiviral hypersensitive response (HR) and cell death process (Lam, Kato, and Lawton 2001). In local lesions a small number of cells are infected (~10s to 100s) and the virus replication occurs within the confines of the necrosis zone. These infected cells will therefore undergo apoptosis and experience a reduction in cellular water, enzymatic degradation of proteins, and DNAse activity (Coll, Epple, and Dangl 2011).

During systemic virus infection, there has been virus replication and movement from the initially infected cells to other cells and tissues. Therefore, the GF space occupied in systemic infection is representative of several rounds of virus replication and potentially the selection of a GF that is optimized for a specific tissue type. By comparing these two scenarios, it becomes clear that the GF variation in local lesion infections is predicted to be higher due to the stochastic nature of the start of an infection and plant cell HR, whilst in systemic infection, GF variation may be due to bottlenecks during within-host movement and several rounds of virus replication. In FBNSV, diminishing GF variation is observed during the course of infection, as the GF converges on a host-specific equilibrium as the virus moves between leaves (Sicard et al. 2013). The results from chapter 5 (systemic infections in *N. tabacum, N. benthamiana* and *A. thaliana*) do not indicate that there may be convergence to an equilibrium GF occurring; however, these measurements are taken at a single time point and would not capture the dynamics of GF change within-host.

If the GF space in the three hosts (*N. tabacum, N. benthamiana* and *A. thaliana*) is similar to what is observed for *C. quinoa*, there may be regions which are characterized by higher or lower virus titre. In chapter 5, several virus populations went extinct and the GF may be associated with low virus titre. In *A. thaliana* and *N. benthamiana*, several virus populations have a lower virus titre; these are also GF values which are at the extreme or along the edge of the GF space. This may indicate that the changes in the GF have pushed these populations to an area of low accumulation. As described earlier for *C. quinoa*, the mean GF was

calculated for the three systemically infected hosts: *A. thaliana, N. benthamiana* and *N. tabacum.* The GF distance to the mean (*D)* was then determined for each host to test the hypothesis that populations with a larger GF distance will have a lower virus titre. In *A. thaliana*, there is a marginally significant positive relationship between *D* and virus titre (Kendall rank correlation: $Z = 2.268$, $p = 0.023$, $\tau = 0.504$), suggesting that there is a flat landscape for describing the GF fitness landscape. Viral fitness drops with increased distance from the mean (i.e. larger values of *D*), given for a few populations along the edge of the landscape. The majority of virus populations have similar virus titre and lower values for *D*.

There was no significant relationship between *D* and virus titre for *N. benthamiana* ($Z = 0$, $p = 1$, $\tau = 0$) and *N. tabacum* ($Z = 0.883$, $p = 0.377$, $\tau = 0.0988$). As with the earlier analogy of a fitness landscape, the shape of the GF space in *A. thaliana*, *N. benthamiana and N. tabacum* may be visualized more as occupying a narrower space. This is in contrast to what is observed for *C. quinoa* where there is a broad GF space with moderate levels of virus titre and a small area with low virus titre. The presence of low virus titre GFs within the GF space in *N. benthamiana* and *N. tabacum* suggests that it may be similar to holey adaptive landscapes observed in other systems (Gavrilets 1999).

GF stability in CMV-i17f

Sicard and colleagues (2013) put forward the hypothesis that the GF may allow viral copy number variation (CNV), positing that changes in viral genome segment copies have a direct effect on viral gene expression (Sicard et al. 2013). Thus, we can expect that changes in the GF correspond to changes to viral gene expression. Whilst the GF has been measured in other multipartite viruses, the accompanying changes in virus gene expression have not been measured (Wu et al. 2017; Sicard et al. 2013; Yu et al. 2019; Boezen, Johnson, et al. 2023).

Viral CNV has extensively been studied in the monopartite vaccinia virus (VACV) (Bayer, Brennan, and Geballe 2018). In VACV infection, amplifications of the *K3L* gene have been observed, with concomitant increases in expression of the cognate protein (Elde et al. 2012). CNV dynamics in VACV has been termed the “genomic accordion”, as there is *K3L* gene array amplification, followed by an increase in mutation supply and fixation of a beneficial H47R mutation, which increases VACV anti-phosphorylation activity and is followed by the collapse of the amplified region to single-copy (Elde et al. 2012). This process bears similarity to the innovation-amplification-divergence (IAD) model described in bacterial species (Näsvall et al. 2012). I hypothesize that the GF may follow genomic accordion and IAD dynamics during infection. The dynamics of GF change in time and over several infection and transmission cycles is unknown. Is there an adaptive role associated with changes of the GF? I propose that increased genome segment copies lead to increased gene dosage and mutation supply. That the short-term GF gene dosage benefit may be later replaced by a beneficial allele after the reduction in segment copies and the fixation of the beneficial allele in the viral population

.

In **chapter 5** I studied GF dynamics of the CMV in the three hosts; *N. tabacum, N. benthamiana* and *A. thaliana.* I find that virus infectivity is variable across hosts and extinctions occur in all hosts. I further examined how changes in virus titre may be linked to extinctions, showing that extinct populations had low virus titres just before extinction. I do not observe

adaptation of CMV to the host species (no significant changes in virus titre for any population), and a systematic host-specific GF shift is only observed *for A. thaliana*.

In chapter 5 I further investigate what factors may be contributing to virus extinctions and low titre. I explored four hypotheses *H0*: stochastic variation in virus titre between individual plants, *H1*: a deleterious GF shift in the absence of mutations which affect the GF, *H2*: deleterious mutations which do not affect the GF and lastly *H3*: a combination of the last two - that de novo mutations arise in the population via genetic drift which lead to a deleterious GF shift. Comparing the initial and final timepoint it is clear that the GFs are different between host species but that the GF does not systemically shift within a species, except in *N. tabacum*. I find evidence that in the case of 2 low-titre populations (one in *A. thaliana* and one in *N. benthamiana*), deleterious GF changes contributed to the extinction of these populations.

NGS analysis of the CMV populations identified 53 de novo mutations, either as single or multiple nucleotide polymorphisms. I observed a repeated non-synonymous mutation T673S in RNA2, encoding the RNA-dependent RNA polymerase (RdRp), in three extinct populations, comprising two populations in *A. thaliana* and one in *N. benthamiana*. I observe a higher number of low-frequency and intermediate mutations in two of these populations, than found for other populations, suggesting that the T673S may be a mutator mutation and may have contributed to extinctions. Combining low titre and mutation data in an analysis suggests that stochastic events leading to low tire (*H0*), deleterious GF change (*H1*) or deleterious mutations (*H2*) contributed to the observed extinctions in different populations.

In chapter 5, I do not observe adaptation of CMV via the GF and hypothesize that the limited host–specificity of the GF may be linked to the fact that CMV is a generalist pathogen. CMV is known to infect more than 1000 plant species (Roossinck 2001), and selection for host-specific GFs for such a broad number of species is unlikely. GF variation and host-specificity are likely to be influenced by within-host factors whilst remaining a trait under viral genetic control. Virus control of the GF may ensure the fine-tuning of virus replication, movement and packaging to host processes in a manner which promotes virus adaptation and improves virus fitness. For a generalist virus this may be a potential source of conflict, as host-specificity promotes viral within-host fitness and may reduce between-host fitness, as there is likely an upper limit on the number of host species for which unique host-specific GFs can be maintained.

I propose that for a generalist virus, such as CMV, there may be a trade-off between within-host and between-host GF specialization due to CMV adaptation to multiple hosts. At the within-host level I anticipate that selection for the GF will occur rapidly; that the initial infection may cause stochastic variation in the GF and after several rounds of virus replication there is a shift from the initial GF to an optimal GF, due to stabilizing selection, for a given host. This may improve overall within-host fitness; the ability to spread rapidly and infect many different tissues and cell types. At this level, virus replication within-host is prioritized and there is specialization.

At the between-host scale, the optimal GF may be one associated with higher virus titre and a balanced GF. A large number of virus particles and a representation of all segments would ensure transmission of all genome segments. The between-host transmission for CMV through aphids is non-specific as it occurs via nonpersistent transmission (Jacquemond 2012).

Results from Chapter 4 of this thesis suggest that there is a narrow bottleneck for CMV transmission of genome segments, in line with other studies that have observed considerable genetic drift acting on individual segments (Betancourt et al. 2008; Ali et al. 2006; Gallet et al. 2018).

The host-specificity of the GF is well described for FBNSV (Sicard et al. 2013), AMV (Wu et al. 2017) and BBTV (Yu et al. 2019). The GF has also been reported to change during the course of infection towards an equilibrium, the “setpoint GF” that appears to be associated with higher accumulation (Sicard et al. 2013). The equilibrium has been demonstrated to be linked to higher virus titre in AMV (Wu et al. 2017). Understanding GF dynamics at the within-host level will elucidate what viral and host processes drive GF changes and provide a more comprehensive map of the GF space, and its link to virus titre. In a preliminary study on the temporal and tissue development of the GF in *N. tabacum,* I found that the GF may differ for different tissues, an indication that different segment frequencies may be required for virus replication in these tissues (Figure 4). This is in keeping with recent work on beet necrotic yellow vein virus (BNYVV) that there are tissue-specific GFs and that these GFs differ in the virus titre (Dall’Ara et al. 2024). Strikingly, for the early timepoint (7 days post infection) the CMV-I17F GF is consistently found in a region of GF space that I have not observed in a considerable number of other experiments (e.g. Figure 2 and 3).

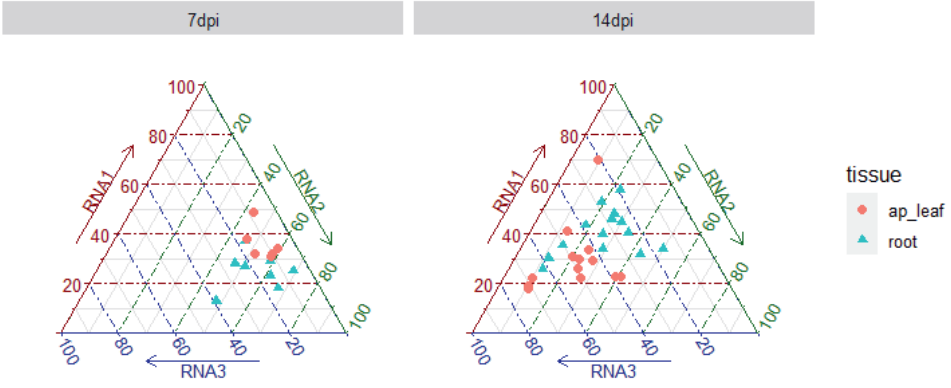


Figure 4. CMV-i17F GF space shifts spatiotemporally in N. tabacum*. CMV-i17F infection in N. tabacum at 7 days post infection (dpi) in apical leaf tissue (red circles, n = 6) and root tissues (blue triangles, n = 8) and at 14dpi in apical leaf (red circles, n = 14) and root tissues (blue triangles, n = 14) of (from Johnson, Grum-Grzhimaylo et al. Unpublished data).*

Why do multipartite viruses exist?

Multipartite viruses are a conundrum, as the cost of transmission is high and the benefits of genome segmentation and separate transmission are unclear. A considerable amount of research has been dedicated to understanding their emergence and the conditions which allow for a multipartite virus genome organization to develop. Several modelling approaches have been put forward to explain why multipartite viruses might arise, many of which rely on the occurrence of an initial fragmented form of a monopartite virus and the generation of a bi-segmented virus which requires complementation for replication (Iranzo and Manrubia 2012; Lucía-Sanz, Aguirre, and Manrubia 2018; Leeks et al. 2023; Park, Denha, and Higgs 2023). In order for a multipartite virus to arise, there would need to be an initial benefit for segmenting the genome. (Nee 1987) proposed that shorter genome segments have a faster replication than the full-length monopartite equivalent when polymerase activity is not a limiting factor. A second benefit is improved virus particle stability due to packaging of a shorter genome segment, as demonstrated for foot and mouth disease virus (FMDV) (Ojosnegros et al. 2011). Secondary benefits of segmentation include the ability to exchange genome segments by reassortment, combining beneficial mutations and removing deleterious mutations and the ability to regulate viral gene expression, by altering the ratio of genome segments in a host-specific manner (Sicard et al. 2013). I will first begin by discussing the different evolutionary models developed to explain the emergence of multipartite viruses, followed by the ecological factors and conditions which allow them to endure.

(Iranzo and Manrubia 2012) propose in a modelling approach that multipartite viruses are a product of the competition between two spontaneous viral genome segments which have different segment lengths. Segments are co-dependent and require the presence of the other for replication. The evolutionary advantage of genome segmentation is from the slower degradation of a shorter genome segment due to the increased stability of the virus particles compared to the longer monopartite ancestor (Iranzo and Manrubia 2012). They identify several scenarios for either monopartite or bipartite virus existence: (1) shorter segments will be at higher frequency when the degradation of a longer monopartite virus occurs more rapidly, (2) high MOI favours the bipartite virus, (3) both monopartite and bipartite forms of the virus can coexist when there is slower degradation of the monopartite virus and (4) for a highly multipartite virus to outcompete the monopartite virus, MOIs of >100 are required for genomes which consist of 5 segments or more. (Leeks et al. 2023) propose that multipartite viruses arose as a result of the presence of cheater genomes derived from a monopartite ancestor. In this scenario, cheater genomes do not produce a shared gene product with the monopartite virus but may take advantage of co-infections to use these products from the monopartite variant. The cheater genomes are segmented viruses which have deletions compared to the monopartite ancestor and can use the shared gene products to increase in frequency. Cheater genomes can increase in frequency when: (1) there is co-infection with a monopartite virus, as the cheater has faster replication due to shorter genome segments (Nee 1987), (2) duplicate cheater co-infection results in the production of a single gene product and (3) if two different cheater types co-infect a cell, there is complementation and virus replication. The co-infection of two different cheaters can occur more easily, facilitating multipartition to evolve.

Since we observe a large diversity of multipartite virus genomes for nucleic acid type (DNA/RNA) and strandedness (plus, minus or ambi-sense), the presence of phylogenetic groups comprised of both monopartite and bipartite viruses (e.g. the genus *Begomovirus*) and that many natural infections occur in the presence of satellite viruses and defective interfering particles, it is likely that multipartition arose on several occasions. The two evolutionary models presented for the existence of multipartite viruses (Iranzo and Manrubia 2012; Leeks et al. 2023) show it as a process driven by either the competition between genome segments of different lengths or from cheater genomes which parasitize a monopartite virus population. We know that here are virus species with genome segments of different lengths (e.g. CMV) and the presence of selfish replicators such as satellite viruses and defective interfering particles in virus populations (Nawaz-ul-Rehman et al. 2009; Mansourpour et al. 2021), thus there are plausible real-world scenarios which may represent processes similar to those presented by the models of multipartite virus evolution.

Most multipartite viruses infect plants and have arthropod vectors (Michalakis and Blanc 2020; Lucía-Sanz and Manrubia 2017). How may these ecological conditions contribute to their continued existence? (Valdano et al. 2019) test the effect of aphid feeding behavior on the transmission of multipartite viruses via a modelling approach and find that higher transmission of a multipartite virus is predicted between host species that are genetically similar and that with increasing host heterogeneity, there is a decline in transmission. Recently, (Di Mattia et al. 2022) showed for FBNSV that existing infections could be complemented by inoculations with missing segments, but only in the background of a replicating virus population. Complementation of virus segments in time between inoculations presents an opportunity as many viruses share common functional proteins. This raises the possibility that in mixed infections, missing functions may be complemented by another virus. This would reduce the cost of transmission as the specificity of virus segments will be reduced, and this is likely to occur for common functions such as virus movement or viral suppressors of RNA silencing.

Given that many multipartite viruses are vector-mediated plant-infecting viruses, there would be a cost associated with GF changes which are optimal in one host but not another. Vector transmission of multipartite viruses will play an important role in transmitting all segments, but also the frequency of segments. Begomoviruses have either monopartite or bipartite genomes, and the acquisition of a second genomic segment (DNA-B) has facilitated the host range expansion of the previously monopartite pepper yellow Mali virus (PepYVMLV) (Ouattara et al. 2022). The bipartite begomoviruses East African cassava mosaic Cameroon virus (EACMCV) and African cassava mosaic virus (ACMV) differ in GFs between source and donor plants compared to the whitefly vector, *Bemisai tabaci*, as well as each other (Kennedy et al. 2023). In a single infection, ACMV and EACMV maintained 0.6 (DNA-A: DNA:B) GF, with DNA-A as the least abundant. During co-infection, the GF of ACMV shifted, increasing by 1 order of magnitude, whilst EACMV shifted to a balanced GF. Thus, the between-host transmission cost is low for the whitefly vector (Kennedy et al. 2023).

Within a broader ecological context, it is known that mixed viral infections are common in plants and that opportunities for interaction are possible via common vector species. One such example is vector co-transmission by *A. craccivora* of the monopartite alfalfa leaf curl virus (ALCV) and FBNSV; where virus localization is similar in different aphid compartments, however, FBNSV accumulation was decreased (Di Mattia, Ryckebusch, et al. 2020). In the bipartite tomato mottle virus (ToMoV) virus transmission efficiency and host titer is reduced

when acquired sequentially or when co-inoculated with the monopartite tomato yellow leaf curl virus (TYLCV) during persistent transmission (McLaughlin et al. 2022). A similar approach using incomplete FBNSV infections (missing segments C, N or U4) which were complemented via sequential aphid transmission showed that the complete genome could be reconstituted within the vector midgut cells of *A. pisum* (Di Mattia et al. 2022). Additionally, we do not know if there are interactions between viruses when they are co-transmitted in a common vector (Tamborindeguy et al. 2023).

Concluding remarks and future outlook

In the current thesis, I investigated GF change at short and longer time scales to understand GF variation and evolution in different host species. I find that the GF of the generalist CMV is highly variable in the host *C. quinoa*, whilst showing less variation in *A. thaliana, N. benthamiana* and *N. tabacum*. Furthermore, CMV infection is characterized by a narrow bottleneck in which a small number of genome segments contribute to the development of the GF during infection. The GF of CMV is transmissible from one host to another, a first indication that the GF can have a between-host benefit. By investigating GF change in single and serial passage experiments I show that some GF changes may be deleterious and associated with lower virus titres whilst also showing that for experimental evolution of CMV the GF change is robust over several rounds of infection. I found preliminary evidence that genome segmentation in the form of segmented and multipartite viruses may be a general mechanism for increasing virus host range. I developed a framework for quantifying the cost of transmission for varying genome organization and show that in some cases segmented viruses may have a higher cost of transmission than multipartite viruses. Future research on the GF should investigate the genetic basis of GF control, as I observe that stochastic GF change and deleterious mutations may contribute to extinctions. Follow-up experiments can also further map the GF fitness landscape by including other measures of virus fitness, e.g. (1) the proportion of encapsidated versus non-encapsidated RNA to elucidate which components of the virus population contribute to GF transmission and (2) mapping virus infectivity to the GF space for different hosts – as you can expect that as the virus becomes specialized on a given host that there may be a trade-off in successful infection in other hosts.

References

Abdi, Hervé. 2007. “The Kendall Rank Correlation Coefficient.” In *Encyclopedia of Measurement and Statistics*, 508–10. Thousand Oaks, California: Sage. https://doi.org/10.4135/9781412952644.

Ali, Akhtar, and Michelle Kobayashi. 2010. “Seed Transmission of Cucumber Mosaic Virus in Pepper.” *Journal of Virological Methods* 163 (2): 234–37.

Ali, Akhtar, Hongye Li, William L. Schneider, Diana J. Sherman, Stewart Gray, Dawn Smith, and Marilyn J. Roossinck. 2006. “Analysis of Genetic Bottlenecks during Horizontal Transmission of Cucumber Mosaic Virus.” *Journal of Virology* 80 (17): 8345–50.

Andreu-Moreno, Iván, and Rafael Sanjuán. 2018. “Collective Infection of Cells by Viral Aggregates Promotes Early Viral Proliferation and Reveals a Cellular-Level Allee Effect.” *Current Biology: CB* 28 (20): 3212-3219.e4.

Bald, J. G. 1937. “The Use of Numbers of Infections For Comparing the Concentrations of Plant Virus Suspensions: Dilution Experiments with Purified Suspensions.” *The Annals of Applied Biology* 24 (1): 33–55.

Bayer, Avraham, Greg Brennan, and Adam P. Geballe. 2018. “Adaptation by Copy Number Variation in Monopartite Viruses.” *Current Opinion in Virology*. Elsevier. https://doi.org/10.1016/j.coviro.2018.07.001.

Bermúdez-Méndez, Erick, Kirsten F. Bronsvoort, Mark P. Zwart, Sandra van de Water, Ingrid Cárdenas-Rey, Rianka P. M. Vloet, Constantianus J. M. Koenraadt, Gorben P. Pijlman, Jeroen Kortekaas, and Paul J. Wichgers Schreur. 2022. “Incomplete Bunyavirus Particles Can Cooperatively Support Virus Infection and Spread.” *PLoS Biology* 20 (11): e3001870.

Betancourt, Mónica, Alberto Fereres, Aurora Fraile, and F. Garcia-Arenal. 2008. “Estimation of the Effective Number of Founders That Initiate an Infection after Aphid Transmission of a Multipartite Plant Virus.” *Journal of Virology* 82 (24): 12416–21.

Boezen, Dieke, Marcelle L. Johnson, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064.

Boezen, Dieke, Maritta Vermeulen, Marcelle Johnson, Rene Van Der Vlugt, Carolyn Malmstrom, and Mark Zwart. 2023. “Mixed Viral Infection Constrains the Genome Formula of Multipartite Cucumber Mosaic Virus.” *Frontiers in Virology* 3. https://doi.org/10.3389/fviro.2023.1225818.

Brooke, Christopher B., William L. Ince, Jens Wrammert, Rafi Ahmed, Patrick C. Wilson, Jack R. Bennink, and Jonathan W. Yewdell. 2013. “Most Influenza a Virions Fail to Express at Least One Essential Viral Protein.” *Journal of Virology* 87 (6): 3155–62.

Chou, Yi-Ying, Reza Vafabakhsh, Sultan Doğanay, Qinshan Gao, Taekjip Ha, and Peter Palese. 2012. “One Influenza Virus Particle Packages Eight Unique Viral RNAs as Shown by FISH Analysis.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (23): 9101–6.

Cobos, Alberto, Nuria Montes, Marisa López-Herranz, Miriam Gil-Valle, and Israel Pagán. 2019. “Within-Host Multiplication and Speed of Colonization as Infection Traits Associated with Plant Virus Vertical Transmission.” *Journal of Virology* 93 (23). https://doi.org/10.1128/JVI.01078-19.

Coll, N. S., P. Epple, and J. L. Dangl. 2011. “Programmed Cell Death in the Plant Immune System.” *Cell Death and Differentiation* 18 (8): 1247–56.

Dall’Ara, M., Y. Guo, D. Poli, D. Gilmer, and C. Ratti. 2024. “Analysis of the Relative Frequencies of the Multipartite BNYVV Genomic RNAs in Different Plants and Tissues.” *The Journal of General Virology* 105 (1). https://doi.org/10.1099/jgv.0.001950.

Di Mattia, Jérémy, Babil Torralba, Michel Yvon, Jean-Louis Zeddam, Stéphane Blanc, and Yannis Michalakis. 2022. “Nonconcomitant Host-to-Host Transmission of Multipartite

Virus Genome Segments May Lead to Complete Genome Reconstitution." *Proceedings of the National Academy of Sciences of the United States of America* 119 (32): e2201453119.
Diefenbacher, Meghan, Jiayi Sun, and Christopher B. Brooke. 2018. "The Parts Are Greater than the Whole: The Role of Semi-Infectious Particles in Influenza A Virus Biology." *Current Opinion in Virology* 33 (December): 42–46.
Druett, H. A. 1952. "Bacterial Invasion." *Nature* 170 (4320): 288–288.
Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers, Jay Shendure, Adam P. Geballe, and Harmit S. Malik. 2012. "Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses." *Cell* 150 (4): 831–41.
Faulkner, Christine. 2018. "Plasmodesmata and the Symplast." *Current Biology: CB* 28 (24): R1374–78.
Fulton, Robert W. 1962. "The Effect of Dilution on Necrotic Ringspot Virus Infectivity and the Enhancement of Infectivity by Noninfective Virus." *Virology*. https://doi.org/10.1016/0042-6822(62)90038-7.
Gallet, Romain, Frédéric Fabre, Gaël Thébaud, Mircea T. Sofonea, Anne Sicard, Stéphane Blanc, and Yannis Michalakis. 2018. "Small Bottleneck Size in a Highly Multipartite Virus during a Complete Infection Cycle." *Journal of Virology* 92 (14). https://doi.org/10.1128/JVI.00139-18.
Gavrilets, Sergey. 1999. "A Dynamical Theory of Speciation on Holey Adaptive Landscapes." *The American Naturalist* 154 (1): 1–22.
Goto, Hideo, Yukiko Muramoto, Takeshi Noda, and Yoshihiro Kawaoka. 2013. "The Genome-Packaging Signal of the Influenza A Virus Genome Comprises a Genome Incorporation Signal and a Genome-Bundling Signal." *Journal of Virology* 87 (21): 11316–22.
Gutiérrez, Serafín, Yannis Michalakis, Manuella Van Munster, and Stéphane Blanc. 2013. "Plant Feeding by Insect Vectors Can Affect Life Cycle, Population Genetics and Evolution of Plant Viruses." *Functional Ecology* 27 (3): 610–22.
Gutiérrez, Serafín, and Mark P. Zwart. 2018. "Population Bottlenecks in Multicomponent Viruses: First Forays into the Uncharted Territory of Genome-Formula Drift." *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2018.09.001.
Heinlein, Manfred. 2015. "Plasmodesmata: Channels for Viruses on the Move." In *Plasmodesmata: Methods and Protocols*, edited by Manfred Heinlein, 25–52. New York, NY: Springer New York.
Hu, Zhaoyang, Guohui Li, Guangtian Li, Qin Yao, and Keping Chen. 2013. "Bombyx Mori Bidensovirus: The Type Species of the New Genus Bidensovirus in the New Family Bidnaviridae." *Chinese Science Bulletin = Kexue Tongbao* 58 (36): 4528–32.
Hutchinson, Edward C., Johann C. von Kirchbach, Julia R. Gog, and Paul Digard. 2010. "Genome Packaging in Influenza A Virus." *The Journal of General Virology* 91 (Pt 2): 313–28.
Iranzo, Jaime, and Susanna C. Manrubia. 2012. "Evolutionary Dynamics of Genome Segmentation in Multipartite Viruses." *Proceedings of the Royal Society B: Biological Sciences* 279 (1743): 3812–19.
Jacquemond, Mireille. 2012. "Cucumber Mosaic Virus." Edited by Gad Loebenstein and Hervé Lecoq. *Advances in Virus Research* 84 (January): 439–504.
Kendall, M. G. 1938. "A New Measure of Rank Correlation." *Biometrika* 30 (1–2): 81–93.
Kennedy, George G., William Sharpee, Alana L. Jacobson, Mary Wambugu, Benard Mware, and Linda Hanley-Bowdoin. 2023. "Genome Segment Ratios Change during Whitefly Transmission of Two Bipartite Cassava Mosaic Begomoviruses." *Scientific Reports* 13 (1): 10059.
Kozieł, Edmund, Józef Julian Bujarski, and Katarzyna Otulak Kozieł. 2023. "Chapter 16 - Plant Cell Apoplast and Symplast Dynamic Association with Plant-RNA Virus Interactions as a Vital Effect of Host Response." In *Plant RNA Viruses*, edited by

Rajarshi Kumar Gaur, Basavaprabhu L. Patil, and Ramasamy Selvarajan, 311–28. Academic Press.
Lam, Eric, Naohiro Kato, and Michael Lawton. 2001. “Programmed Cell Death, Mitochondria and the Plant Hypersensitive Response.” *Nature* 411 (6839): 848–53.
Lauffer, M. A., and W. C. Price. 1945. “Infection by Viruses.” *Archives of Biochemistry* 8 (December): 449–68.
Lazarowitz, S. G., and R. N. Beachy. 1999. “Viral Movement Proteins as Probes for Intracellular and Intercellular Trafficking in Plants.” *The Plant Cell* 11 (4): 535–48.
Leeks, Asher, Penny Grace Young, Paul Eugene Turner, Geoff Wild, and Stuart Andrew West. 2023. “Cheating Leads to the Evolution of Multipartite Viruses.” *PLoS Biology* 21 (4): e3002092.
Li, Xiuli, Min Gu, Qinmei Zheng, Ruyi Gao, and Xiufan Liu. 2021. “Packaging Signal of Influenza A Virus.” *Virology Journal* 18 (1): 36.
Lucía-Sanz, Adriana, Jacobo Aguirre, and Susanna Manrubia. 2018. “Theoretical Approaches to Disclosing the Emergence and Adaptive Advantages of Multipartite Viruses.” *Current Opinion in Virology* 33 (December): 89–95.
Lucía-Sanz, Adriana, and Susanna Manrubia. 2017. “Multipartite Viruses: Adaptive Trick or Evolutionary Treat?” *Npj Systems Biology and Applications* 3 (1): 34.
Mansourpour, Mahsa, Romain Gallet, Alireza Abbasi, Stephane Blanc, Akbar Dizadji, and Jean-Louis Zeddam. 2021. “Effects of an Alphasatellite on Life Cycle of the Nanovirus Faba Bean Necrotic Yellows Virus.” *Journal of Virology*, November, JVI0138821.
Michalakis, Yannis, and Stéphane Blanc. 2020. “The Curious Strategy of Multipartite Viruses.” *Annual Review of Virology* 7 (1): 203–18.
Miras, Manuel, Mathieu Pottier, T. Moritz Schladt, J. Obinna Ejike, Laura Redzich, Wolf B. Frommer, and Ji-Yun Kim. 2022. “Plasmodesmata and Their Role in Assimilate Translocation.” *Journal of Plant Physiology* 270 (March): 153633.
Nakatsu, Sumiho, Shin Murakami, Keiko Shindo, Taisuke Horimoto, Hiroshi Sagara, Takeshi Noda, and Yoshihiro Kawaoka. 2018. “Influenza C and D Viruses Package Eight Organized Ribonucleoprotein Complexes.” *Journal of Virology* 92 (6). https://doi.org/10.1128/JVI.02084-17.
Nakatsu, Sumiho, Hiroshi Sagara, Yuko Sakai-Tagawa, Norio Sugaya, Takeshi Noda, and Yoshihiro Kawaoka. 2016. “Complete and Incomplete Genome Packaging of Influenza A and B Viruses.” *MBio* 7 (5). https://doi.org/10.1128/mBio.01248-16.
Näsvall, Joakim, Lei Sun, John R. Roth, and Dan I. Andersson. 2012. “Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence.” *Science* 338 (6105): 384–87.
Navarro, Jose A., Jesus A. Sanchez-Navarro, and Vicente Pallas. 2019. “Key Checkpoints in the Movement of Plant Viruses through the Host.” *Advances in Virus Research* 104 (July): 1–64.
Nawaz-ul-Rehman, Muhammad Shah, Shahid Mansoor, Rob W. Briddon, and Claude M. Fauquet. 2009. “Maintenance of an Old World Betasatellite by a New World Helper Begomovirus and Possible Rapid Adaptation of the Betasatellite.” *Journal of Virology* 83 (18): 9347–55.
Nee, Scan. 1987. “The Evolution of Multicompartmental Genomes in Viruses.” *Journal of Molecular Evolution* 25: 277–81.
Ojosnegros, S., J. García-Arriaza, C. Escarmís, S. C. Manrubia, and C. Perales. 2011. “Viral Genome Segmentation Can Result from a Trade-Off between Genetic Content and Particle Stability.” *PLoS Genetics* 7 (3): 1001344.
Park, Hyunjin, Saven Denha, and Paul G. Higgs. 2023. “Evolution of Bipartite and Segmented Viruses from Monopartite Viruses.” *Viruses* 15 (5). https://doi.org/10.3390/v15051135.
R Foundation for Statistical Computing. 2023. *R: A Language and Environment for Statistical Computing* (version 4.3.1). Vienna, Austria. https://www.r-project.org/.

Roossinck, M. J. 2001. “Cucumber Mosaic Virus, a Model for RNA Virus Evolution.” *Molecular Plant Pathology* 2 (2): 59–63.

Sánchez-Navarro, Jesús A., Mark P. Zwart, and Santiago F. Elena. 2013. “Effects of the Number of Genome Segments on Primary and Systemic Infections with a Multipartite Plant RNA Virus.” *Journal of Virology* 87 (19): 10805–15.

Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. “The Strange Lifestyle of Multipartite Viruses.” Edited by Tom C. Hobman. *PLoS Pathogens* 12 (11): e1005819.

Sicard, Anne, Elodie Pirolles, Romain Gallet, Marie-Stéphanie Vernerey, Michel Yvon, Michel Peterschmitt, Serafin Gutierrez, Yannis Michalakis, and Stéphane Blanc. 2019. “A Multicellular Way of Life for a Multipartite Virus.” *ELife* 8 (March): e43599.

Sicard, Anne, Michel Yvon, Tatiana Timchenko, Bruno Gronenborn, Yannis Michalakis, Serafin Gutierrez, and Stéphane Blanc. 2013. “Gene Copy Number Is Differentially Regulated in a Multipartite Virus.” *Nature Communications* 4: 2248.

Su, Shengzhong, Zhaohui Liu, Cheng Chen, Yan Zhang, Xu Wang, Lei Zhu, Long Miao, Xue-Chen Wang, and Ming Yuan. 2010. “Cucumber Mosaic Virus Movement Protein Severs Actin Filaments to Increase the Plasmodesmal Size Exclusion Limit in Tobacco.” *The Plant Cell* 22 (4): 1373–87.

Tamborindeguy, Cecilia, Fernando Teruhiko Hata, Rúbia de Oliveira Molina, and William Mário de Carvalho Nunes. 2023. “A New Perspective on the Co-Transmission of Plant Pathogens by Hemipterans.” *Microorganisms* 11 (1). https://doi.org/10.3390/microorganisms11010156.

Taschner, P. E., A. C. van der Kuyl, L. Neeleman, and J. F. Bol. 1991. “Replication of an Incomplete Alfalfa Mosaic Virus Genome in Plants Transformed with Viral Replicase Genes.” *Virology* 181 (2): 445–50.

Valdano, Eugenio, Susanna Manrubia, Sergio Gómez, and Alex Arenas. 2019. “Endemicity and Prevalence of Multipartite Viruses under Heterogeneous Between-Host Transmission.” *PLoS Computational Biology* 15 (3): e1006876.

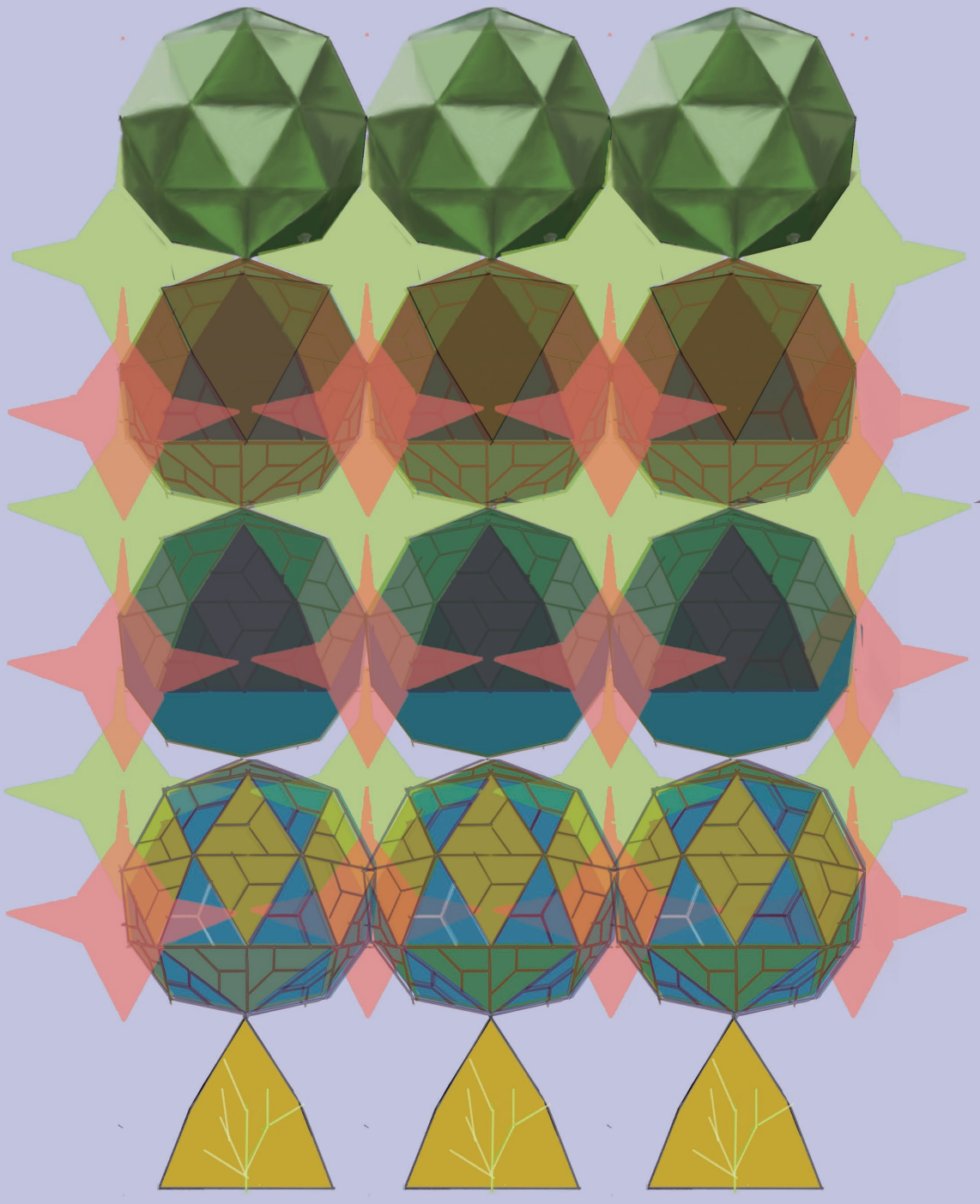
Vitti, Antonella, Israel Pagán, Brigida Bochicchio, Angelo De Stradis, Pasquale Piazzolla, Antonio Scopa, and Maria Nuzzaci. 2022. “Cucumber Mosaic Virus Is Unable to Self-Assemble in Tobacco Plants When Transmitted by Seed.” *Plants* 11 (23). https://doi.org/10.3390/plants11233217.

Wichgers Schreur, Paul J., and Jeroen Kortekaas. 2016. “Single-Molecule FISH Reveals Non-Selective Packaging of Rift Valley Fever Virus Genome Segments.” *PLoS Pathogens* 12 (8): e1005800.

Wu, Beilei, Mark P. Zwart, Jesús A. Sánchez-Navarro, and Santiago F. Elena. 2017. “Within-Host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus.” *Scientific Reports* 7 (1): 1–15.

Yu, Nai-Tong, Hui-Min Xie, Yu-Liang Zhang, Jian-Hua Wang, Zhongguo Xiong, and Zhi-Xin Liu. 2019. “Independent Modulation of Individual Genomic Component Transcription and a Cis-Acting Element Related to High Transcriptional Activity in a Multipartite DNA Virus.” *BMC Genomics* 20 (1): 573.

Zwart, Mark P., and Santiago F. Elena. 2020. “Modeling Multipartite Virus Evolution: The Genome Formula Facilitates Rapid Adaptation to Heterogeneous Environments.” *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa022.



Summary

Virus genomes consist of RNA or DNA in single or several molecules packaged together or separately into virus particles. They may be classified as monopartite viruses, comprised of a single genome segment which is individually packaged. Segmented viruses have genomes composed of several segments which are packaged together into a single virus particle. Lastly, there are the multipartite viruses, having several genome segments which are individually packaged into virus particles and transmitted. The differences in the number of genome segments and individual packaging of multipartite viruses influence the likelihood of infection, as viruses with more segments will require higher doses to initiate an infection. In this thesis, I investigate the costs and benefits of a multipartite virus strategy combining theoretical and experimental approaches.

In **Chapter 1**, I provide an overview of historical and contemporary research on multipartite viruses. I discuss the cost of multipartition for between-host transmission and the proposed benefits. There are several proposed benefits for multipartition; faster replication of shorter genome segments when polymerase is abundant, increased genetic diversity via reassortment, increased virus particle stability and gene expression regulation by segment copy number change – the genome formula (GF). I discuss the hypothesis of viral gene expression regulation by the GF, supporting evidence in the literature, and the potential role of the GF in local adaptation. I discuss potential factors which may affect the GF and present the model system for this thesis, cucumber mosaic virus (CMV).

In **Chapter 2**, I present a review of the state of the art in addressing the costs and benefits of the multipartite virus genome strategy. I develop a quantitative approach for estimating the cost to transmission of multipartition by analyzing the changes in shape and position of the dose-response curve. I test this by reanalyzing experimental infections of the tripartite alfalfa mosaic virus (AMV) in which there is the constitutive expression of one or two viral genome segments by host plants. I show that the cost of transmission for multipartite viruses is higher when including the change in position of the dose-response curve. Experiments with faba bean necrotic stunt virus (FBNSV) have shown that infection can occur by complementation across neighboring cells, where genome segments may be absent. I model the relationship between the GF and viral gene product sharing, showing that gene product sharing may minimize the cost to transmission of multipartite viruses. I also show that the benefits of gene product sharing and gene expression regulation bY the GF may be mutually exclusive. I next pose the question if genome segmentation may be a means for expanding virus host range. By analyzing virus-host databases I show that genome segmentation indeed may contribute to wider host ranges in segmented and multipartite viruses.

In **Chapter 3**, I develop an approach for quantitatively analyzing GF variation in multipartite and segmented viruses. I use the GF distance (*D*), the Euclidean distance between any two GF values to quantify GF variation. *D* allows for comparing GF variation between multipartite viruses which have different segment numbers, nucleic acid type and infection process. I estimate theoretical GF variation for two scenarios; firstly when there is random GF variation ($\overline{D}_{a,b}^{rand}$) and the maximum GF variation when there is GF drift after a single bottleneck ($\overline{D}_{a,b}^{drift}$). I calculate *D* for empirical GF measurements from three multipartite viruses; CMV, AMV and

FBNSV. Re-analysis of AMV data shows that the GF is transmissible, as the GF in the inoculated leaf is more similar to the inoculum than expected by chance.

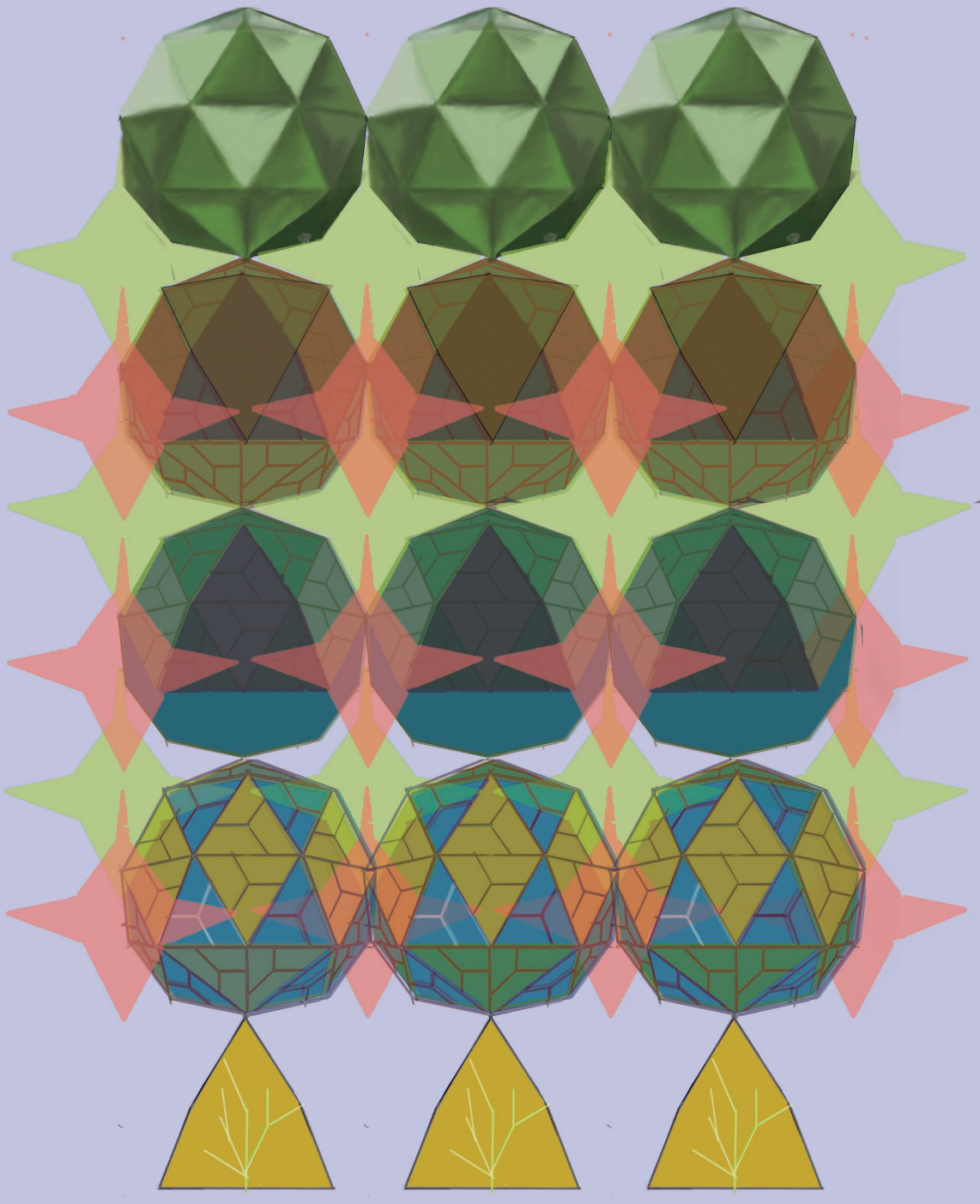
In **Chapter 4**, I experimentally determine GF variation in CMV local lesion infections, infections which are restricted to a small number of cells, in the host *Chenopodium quinoa*. To this end I inoculate *C. quinoa* plants in three experiments with the isolate CMV-i17F derived from *Nicotiana benthamiana* or *Nicotiana tabacum,* and determine the GF by RT-qPCR. Results show that GF variation is high across experiments and I show that GF variation (*D*) is high in local lesions, approaching that of maximum variation from a single population bottleneck ($\overline{D_{a,b}^{drift}}$) . I also show that there is an inoculum effect on the GF in local lesions, suggesting that the GF may be transmissible. When examining GF variation across experiments, I show that for experiment 1 the GF observations may be divided into two groups. The first group is a large majority cluster with a central balanced GF, and a second minority cluster occupies a narrow GF space with variation in the axis of RNA1 and RNA2 and with low RNA3. The two clusters showed significant differences in virus titre, where the majority cluster had a higher virus titre than the minority one. This is the first evidence of multiple clusters for the GF within a single host, and the association with differences in viral fitness.

In **Chapter 5**, I hypothesize that the GF segment copy number changes allow for viral gene expression regulation and local adaptation of a multipartite virus in different hosts. I infect *Arabidopsis thaliana, N. benthamiana* and *N. tabacum* with CMV-i17F and serially passage the virus for five rounds. I determined the GF by RT-qPCR and estimated virus fitness by using virus titre as a fitness measure. I use sequence analysis of the evolved populations to determine GF and mutation interactions during local adaptation. My results show that CMV-I17F has a variable GF, in which *Nicotiana* hosts are similar whilst the GF of *A. thaliana* is distinct. In all hosts, there is an extinction of virus populations associated with a decline in virus titre. Analysis of sequencing data identified point mutations in several populations, most commonly on RNA 2a, the viral RNA-dependent RNA polymerase, and untranslated regions. Extinctions appear linked to the presence of the mutation in RNA2a or shifts in the GF.

In **Chapter 6**, I investigate the cost of genome organization and packaging strategy on infectivity. Multipartite viruses package genome segments individually into virus particles, whilst segmented viruses package all genome segments together in a single virus particle. For segmented viruses, there may be error-prone packaging, wherein not all genome segments are packaged together in a virus particle. In addition, there may be non-selective packaging of genome segments, where there may be differences in the number of genome segments packaged and the identity of segments. I develop a mathematical modelling approach to quantify the cost to infectivity using the change in gradient of the dose-response relationship and the shift in position. Our results show that the cost to infectivity is highest for the non-selective packaging segmented viruses. I re-analyse empirical data for Rift valley fever virus (RVFV) and show that it may have a cost to infectivity similar to that of a multipartite virus.

In **Chapter 7** I synthesize my results from each chapter and place them within a broader context. I further develop the cost of transmission for multipartite and segmented viruses. I present a conceptual overview of the genome particle content and the cost of transmission, showing that error-prone segmented viruses may display transmission dynamics similar to

non-selective packaging segmented viruses. I synthesize available CMV-i17F GFs in *A. thaliana, C. quinoa, N. benthamiana* and *N.tabacum* hosts and the relationship to virus titre, discussing the empirical GF fitness landscape. I show that in *C. quinoa*, there may be two mesa-like features in the GF landscape: a larger mesa with a broad central GF space that is associated with high virus titre and a second narrower mesa associated with lower virus titre and with GFs having variation along the axis of RNA2, high RNA1 and low RNA3. *N. benthamiana* and *N. tabacum* do not display differences in viral fitness based on the GF space occupied by CMV-I17f, whilst infections in *A. thaliana* are characterized by a flat fitness landscape in which virus populations with a greater GF distance have lower virus titre. I end with a discussion on why multipartite viruses exist and suggest that future research should focus on characterizing the GF fitness landscape for encapsidated and unencapsidated RNA to identify which component of the GF space of virus population is transmitted.



List of Publications

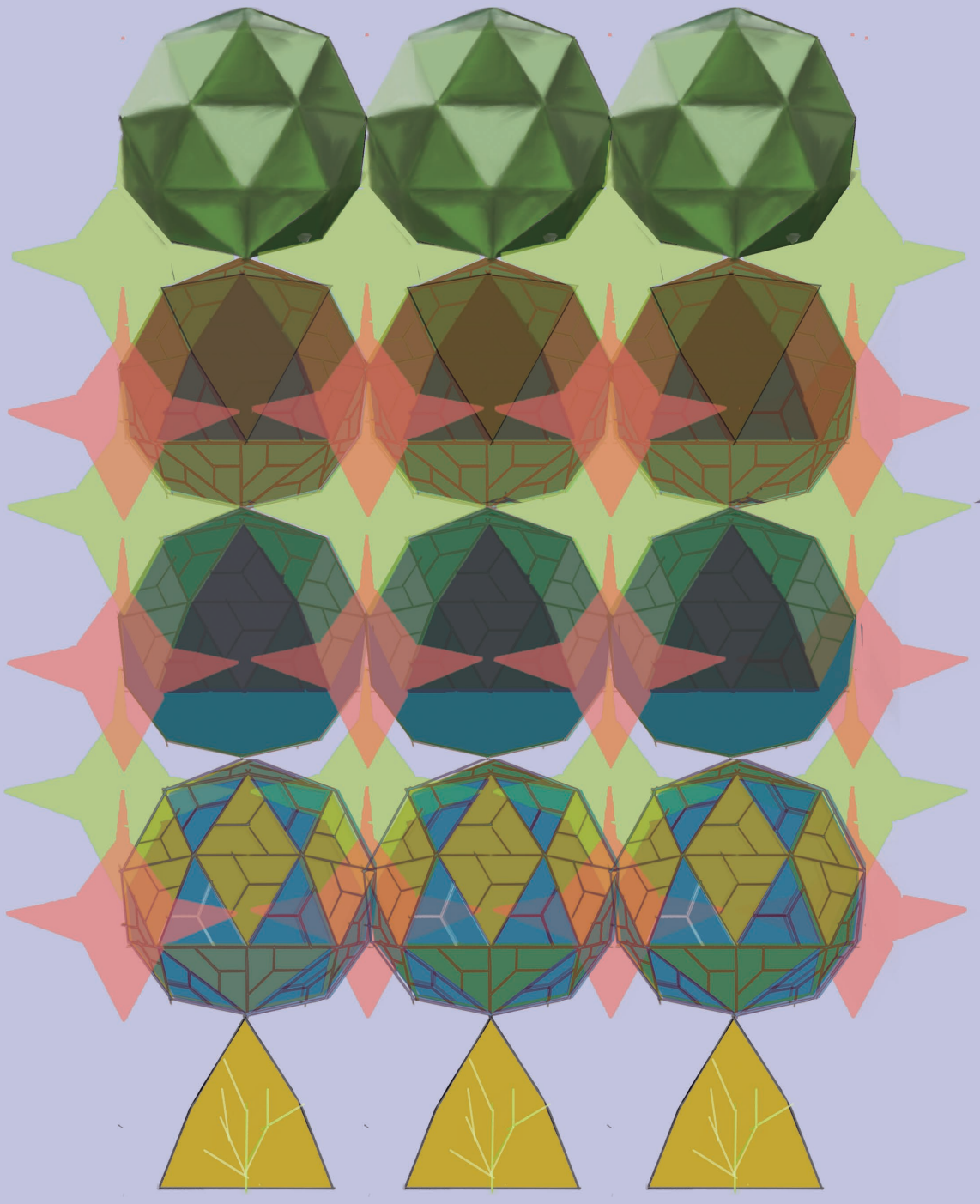
Johnson, Marcelle L., and Mark P. Zwart. 2024. “Robust Approaches to the Quantitative Analysis of Genome Formula Variation in Multipartite and Segmented Viruses.” *Viruses* 16 (2): 270. doi:10.3390/v16020270.(**Chapter 3 of this thesis**)

Boezen, Dieke, **Marcelle L. Johnson**, Alexey A. Grum-Grzhimaylo, René Aa van der Vlugt, and Mark P. Zwart. 2023. “Evaluation of Sequencing and PCR-Based Methods for the Quantification of the Viral Genome Formula.” *Virus Research*, February, 199064. doi:10.1016/j.virusres.2023.199064.

Boezen, Dieke, Maritta Vermeulen, **Marcelle Johnson**, Rene Van Der Vlugt, Carolyn Malmstrom, and Mark Zwart. 2023. “Mixed Viral Infection Constrains the Genome Formula of Multipartite Cucumber Mosaic Virus.” *Frontiers in Virology* 3. doi:10.3389/fviro.2023.1225818.

Wortel, Meike T., Deepa Agashe, Susan F. Bailey, Claudia Bank, Karen Bisschop, Thomas Blankers, Johannes Cairns, Enrico Sandro Colizzi, Davide Cusseddu, Michael M. Desai, Bram van Dijk, Martijn Egas, Jacintha Ellers, Astrid T. Groot, David G. Heckel, **Marcelle L. Johnson**, Ken Kraaijeveld, Joachim Krug, Liedewij Laan, Michael Lässig, Peter A. Lind, Jeroen Meijer, Luke M. Noble, Samir Okasha, Paul B. Rainey, Daniel E. Rozen, Shraddha Shitut, Sander J. Tans, Olivier Tenaillon, Henrique Teotónio, J. Arjan G. M. de Visser, Marcel E. Visser, Renske M. A. Vroomans, Gijsbert D. A. Werner, Bregje Wertheim, and Pleuni S. Pennings. 2023. “Towards Evolutionary Predictions: Current Promises and Challenges.” *Evolutionary Applications* 16 (1): 3–21. doi:10.1111/eva.13513.

Zwart, Mark P., Stéphane Blanc, **Marcelle Johnson**, Susanna Manrubia, Yannis Michalakis, and Mircea T. Sofonea. 2021. “Unresolved Advantages of Multipartitism in Spatially Structured Environments.” *Virus Evolution* 7 (1): veab004. doi:10.1093/ve/veab004.



Acknowledgements

Completing a PhD would not be possible without the guidance, collaboration and support of colleagues, friends and family. Thank you all for your contributions both large and small, they have certainly not gone unnoticed.

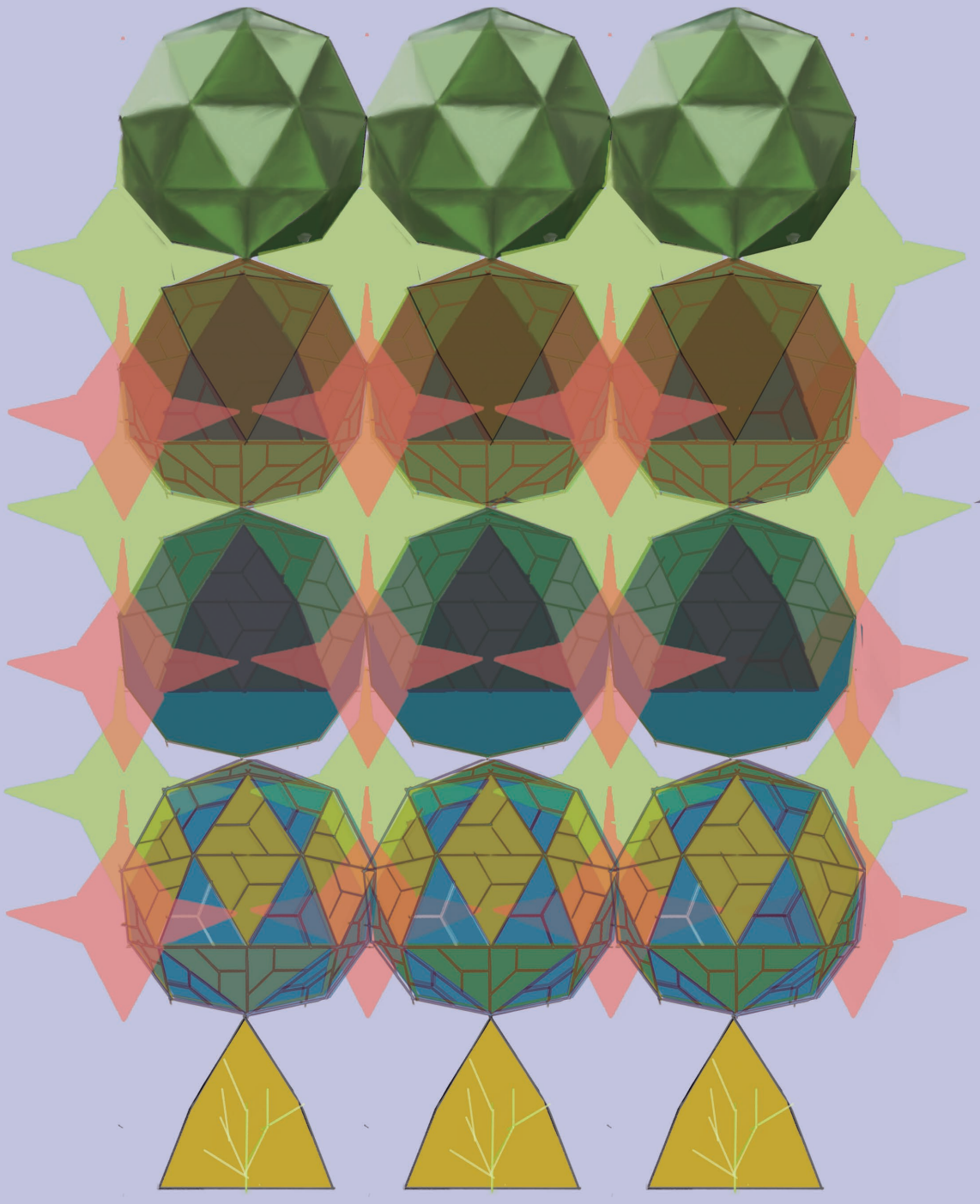
Firstly, I express my deepest gratitude to my supervisory team, Mark, René and Arjan, for their invaluable guidance and advice. Mark, I've cherished my time in the "new" Virus Ecology research group at the NIOO and the opportunity to be a part of the early success and contribute to the research on multipartite viruses. Your open communication has allowed me to express my ideas freely, even the offbeat ones, and the chance to test different ideas in the greenhouse has been a true privilege. Our many colourful whiteboard sessions over the years have been a source of inspiration. René, your practical advice and availability for a quick chat have been a lifeline. Arjan, your support and our long discussions on evolutionary theory have been instrumental in my PhD journey.

I wish to extend warm wishes to my paranymphs, Dieke Boezen and Azkia Nurfikari. Your professionalism and friendship have enriched my PhD experience. Dieke, I am grateful to have you as a friend and to have spent many days watering plants in the greenhouse, drinking coffee, discussing feminist literature and Dutch politics. I wish you success as a science program manager, and I do not doubt that you'll contribute to a more equitable and scientifically diverse research ecosystem. Azkia, you are one of the kindest people I know. Thank you for your open-heartedness, positive attitude, and endless hours in Reddit forums on niche topics. I'll never get these hours back, but they made me laugh and cry simultaneously. I look forward to seeing your transformation into a virologist extraordinaire. I would like to thank my PhD cohort in the Department of Microbial Ecology: German, Letusa, Cristina, Raul, Linda, Lena, Dimitris, Kang, Han, Stijn, Eline, Stalin, Muhammed, Ana and Dario. I will always remember turning the PhD corner into the Amazon, singing Disney jams with gusto on a Thursday morning, puzzling and many interesting and fun coffee chats. I warmly remember our PhD weekend trip, which included long walks, board games, music horses and all the many Friday borrels together. I'd like to extend my sincerest thanks to the many PostDocs: Mahedere, Marcio, Viviane, Ben, Ohana, Jie and Christina, and Technicians Agaat and Saskia, who were an important part of making the PhD enjoyable.

I thank the Virus Ecology group for the many meaningful discussions and support throughout the PhD. Thank you, Alex, for your advice and guidance in the lab and funny t-shirts. Roos, thanks for your friendliness and welcoming spirit. Dimitris, I look forward to your interesting PhD results and success. I thank Maria Hundscheid, Simone Weidner, Elisabeth van Strien, my internship and master thesis students, Sebastien Theil, Jos de Kleijn, Bo Hartman, and Elisa Miranda Vasquez.

I thank my colleagues from the Laboratory of Virology and Monique van Oers for their openness and thoughtful discussions of my research results. I especially thank my fellow PhDs Simone Gasque and Sharella Schop for their friendship.

I would like to express my heartfelt thanks to my partner, my family and my in-laws. To my in-laws, Leontien and Rob, thank you for your interest in my research and kindness over the years. Pim, I would not have gotten far without your love and support. Thank you for reminding me that life is for living; making time for cooking, gardening, and travelling. I've enjoyed many adventures with you and can't wait to start our Aruban escapades and build a life together. My parents, George and Lillian, thank you for your advice and for always supporting my academic career. You have given me the confidence to pursue my dreams and to stay true to myself. To my sister, Charelle, we are two peas in a pod, and I couldn't have done this without you. For my family, I am because you are.



About the Author

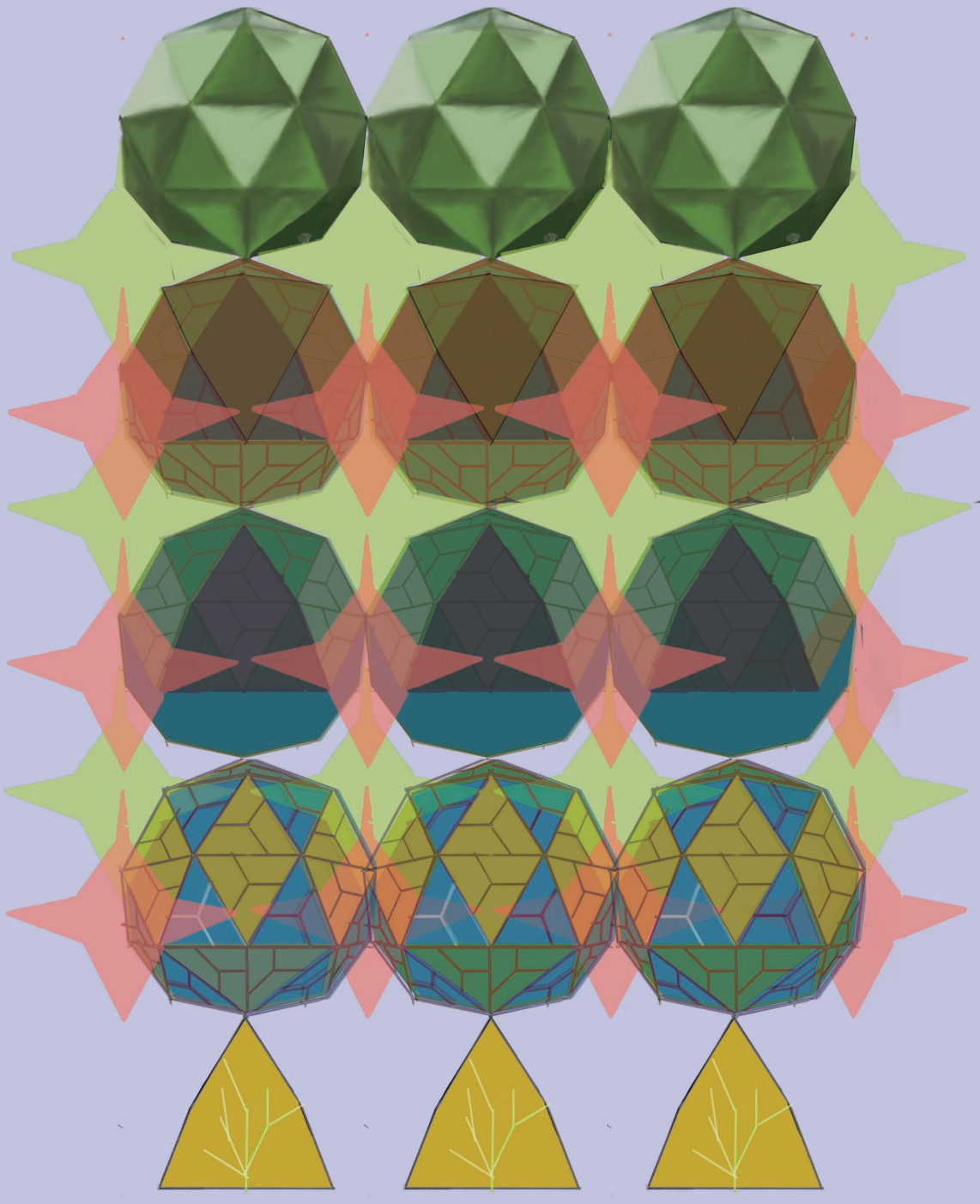
Marcelle Lauren Johnson was born on 23 April 1990 in Johannesburg, South Africa. She grew up in Johannesburg and spent much of her life surrounded by nature, often enjoying family trips to national parks and exploring the African savannah landscape. The nearby nature reserve and travel across South Africa formed her love of plants, animals, and the outdoors.

In 2009, she started a BSc at the University of Witwatersrand, majoring in geography, ecology, and environment and conservation. Thereafter, in 2014, she completed her honours thesis at the School of Animal, Plant and Environmental Sciences, combining plant ecology and evolutionary biology with a thesis titled “Cold Tolerance Strategies of two Helichrysum Species to Alpine Species in the Drakensberg Alpine Centre (DAC)” supervised by Prof. Glynis Goodman-Cron and Dr Kershree Padayachee. She graduated with distinction and used her passion for research, plant science and a commitment to sustainability to build her career. She completed an internship at the Gauteng Department of Agriculture and Rural Development, within the Department of Air Quality to model greenhouse gas emissions from regional landfills, waste processing facilities and metal industrial processes. Her keen interest in the green transition saw her contribute to legislation on air quality, greenhouse gas emissions, and accounting systems.

She returned to her passion for scientific research in 2015, completing an MSc in Plant Biology at the Swedish University of Agricultural Sciences (SLU) in Uppsala, Sweden. During this time, Marcelle was granted a Swedish Institute (SI) study scholarship and an SLU tuition fees scholarship. Here, she delved into the field of plant virology, completing a thesis, “Barley yellow dwarf-associated virus infection in winter wheat and oat in southern Sweden”, under the guidance of Prof. Anders Kvarnheden. Marcelle's expertise in plant virology and plant breeding was further honed during her research stays as part of the European Plant Breeding College 2015 - 2016 (now the Erasmus+ Master in Plant Breeding) at the University of Ghent, Belgium, and at the UniLaSalle Beauvais in France. Her collaboration in the development of a breeding programme on the African orphan crop Cleome gynandra is a testament to her practical skills and her ability to apply her knowledge in real-world scenarios.

In 2018, she joined the VIDI project of Dr Mark Zwart at the Netherlands Institute of Ecology (NIOO-KNAW) and as part of the Laboratory of Virology and Genetics at Wageningen University of Research in The Netherlands with co-promotors Prof. René van der Vlugt and Prof. Arjan de Visser. She researched the evolution of the genome formula of multipartite plant viruses, combining greenhouse experiments and laboratory and modelling approaches.

Marcelle's time in The Netherlands was not solely dedicated to her academic pursuits. She also took on service positions, demonstrating her deep-rooted commitment to the community. As the PhD representative within the Department of Microbial Ecology at the NIOO-KNAW, and a member of the Inclusion, Diversion, Equity and Accessibility committee, Marcelle actively contributed to the betterment of her academic environment. Her dedication to community service is a testament to her character and her desire to make a positive impact beyond her research.



PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review/Project Proposal (4.5 ECTS)

- Unravelling the evolutionary significance of the tripartite genome of Cucumber mosaic virus

Post-graduate courses (5.1 ECTS)

- Advanced statistics: design of eexperiments; WIAS en PE&RC (2018)
- The carpentries workshop: genomics data; NIOO-KNAW/UvA/WUR (2018)
- Data Carpentry with Python; Netherlands eScience Center (2021)
- Evolutionary biology in Guarda (2022)

Invited review of journal manuscripts (3 ECTS)

- PLoS Genetics: multipartite virus gene expression (2019)
- Nature Microbiology: SARS-CoV-2 within host variation (2021)
- Nature Communications: Influenza A within host evolution (2022)

Competence, skills and career-oriented activities (5 ECTS)

- Workshop: storytelling for academics; Wageningen in'to Languages/NIOO (2018)
- PhD & Postdoc Career development event: map your future; KNAW (2018)
- Competence assessment; WUR WGS (2018)
- Theatre skills for science communication; Artesc/NIOO (2019)
- Brain-friendly working & writing; WGS (2020)
- FameLab Wageningen; WUR (2020)
- The choice: unbox your PhD process & take charge of your performance; WGS (2020)
- Workshop on data visualization; YoungWUR/information is beautiful (2020)
- Career orientation; WUR WGS (2021)
- Successful grant writing; KNAW (2022)
- Let's go viral; Biodiversity XL (2022)

Scientific integrity/ethics in science activities (0.9 ECTS)

- Scientific integrity; NIOO-KNAW (2018)
- Scientific integrity; WGS (2020)

PE&RC Annual meetings, seminars and PE&RC weekend/retreat (0.9 ECTS)

- PE&RC Weekend for first years (2018)

Discussion groups/local seminars or scientific meetings (10 ECTS)

- Wageningen evolution & ecology seminars (2018-2023)
- NIOO Eco-evo meetings (2018-2023)
- NIOO Virus ecology meeting (2019-2023)
- ESCV Virtual meeting on Covid-19 (2020)
- Netherlands association of virus ecology meeting (2022)

International symposia, workshops and conferences (6.6 ECTS)

- Wageningen PhD symposium; poster presentation; Wageningen, the Netherlands (2019)
- Lorentz Center workshop predicting evolution; poster presentation; Leiden, the Netherlands (2019)
- 15th International symposium on plant virus epidemiology; poster presentation; Madrid, Spain (2022)
- Biodiversity XL event; NIOZ Texel, the Netherlands (2022)
- Netherlands annual ecology meeting; Lunteren, the Netherlands (2022)

Lecturing/supervision of practicals/tutorials (1.5 ECTS)

- MBO Internship student (2021-2022)

BSc/MSc thesis supervision (9 ECTS)

- Measuring the absolute quantity of the RNA particles in cucumber mosaic virus using RT-qPCR
- Determination of variation in genome formula of a multipartite RNA virus between different isolates and host species
- The genome formula of cucumber mosaic virus (CMV-i17f) in *Chenopodium quinoa*

The research described in this thesis was carried out at the Department of Microbial Ecology at the Netherlands Institute of Ecology (NIOO-KNAW) and was financially supported by a grant from the Dutch Research Council (NWO 016.VIDI.171.0161) To Mark Zwart. This is NIOO thesis number 217.

Cover design by Rebecca Arredondo.

Layout by Marcelle Johnson

Printed by ProefschriftMaken