Gene copy number variations at the within-host population level modulate gene expression in a multipartite virus

Romain Gallet[1,2], Jérémy Di Mattia[1], Sébastien Ravel[1], Jean-Louis Zeddam[1], Renaud Vitalis[2], Yannis Michalakis[3#] and Stéphane Blanc[1#*]

Affiliations

[1]: PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

[2]: CBGP, Univ Montpellier, CIRAD, INRAE, institut Agro, IRD, Montpellier, France.

[3]: MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France

[#] YM and SB equally contributed to this work

*Correspondence to Stéphane Blanc:

UMR PHIM, CIRAD TA-A54/K, Campus International de Baillarguet, 34398 Montpellier cedex 05, France

e.mail: stephane.blanc@inrae.fr tel: +33499624804

Running title: Copy number variation and gene expression in multipartite viruses

Competing interest: The authors declare no competing interest. This work was funded by French national institutions INRAE, CNRS, IRD, the French national agency for funding research ANR, and Montpellier University of excellence (MUSE)

Abstract

Multipartite viruses have a segmented genome, with each segment encapsidated separately. In all multipartite virus species for which the question has been addressed, the distinct segments reproducibly accumulate at a specific and host-dependent relative frequency, defined as the 'genome formula'. Here, we test the hypothesis that the multipartite genome organization facilitates the regulation of gene expression via changes of the genome formula, and thus via gene copy number variations. In a first experiment, the faba bean necrotic stunt virus (FBNSV), whose genome is composed of eight DNA segments each encoding a single gene, was inoculated into faba bean or alfalfa host plants, and the relative concentrations of the DNA segments and their corresponding mRNAs were monitored. In each of the two host species, our analysis consistently showed that the genome formula variations modulate gene expression, the concentration of each genome segment linearly and positively correlating to that of its cognate mRNA but not of the others. In a second experiment, twenty parallel FBNSV lines were transferred from faba bean to alfalfa plants. Upon host switching, the transcription rate of some genome segments changes but the genome formula is modified in a way that compensates for these changes and maintains a similar ratio between the various viral mRNAs. Interestingly, a deep-sequencing analysis of these twenty FBNSV lineages demonstrated that the host-related genome formula shift operates independently of DNA-segment sequence mutation. Together, our results indicate that nanoviruses are plastic genetic systems, able to transiently adjust gene expression at the population level in changing environment, by modulating the copy number but not the sequence of each of their genes.

Introduction

Multipartite viruses are intriguing genetic systems whose biology is difficult to explain within the current conceptual framework of virology [1–3] . Their genomes consist of two or more nucleic acid segments, the most striking feature being their individual encapsidation into distinct virus particles. Such a split genome packaging has an obvious cost, which is the increased risk of segment loss at each transmission event, and thus of failed inoculation. Even though the means by which the myriads of multipartite viral species efficiently manage this cost are poorly uncovered [4,5], it is unanimously acknowledged that they all face the same issue of maintenance of the genome integrity [1,3,6]. Counterpart benefits of multipartite genome architecture, in contrast, are highly debated and as yet no proposition reached a consensus [6–8]. Most hypotheses recognize the smaller size of the genome segments as conferring an advantage to the system, either through faster replication [9], mutation escape [10], genetic exchange via segment reassortments [11], or particle stability [12]. These proposals have several drawbacks: i) most do not explain the separate encapsidation of the distinct segments; ii) they are not specific to multipartite genome architecture and similarly apply to viruses encapsidating all segments together; last, iii) none constrains the relative frequency of the segments, which should thus evolve toward the situation of minimum cost where all segments accumulate at equal copy number. Available studies estimating the relative amount of distinct genome segments in hosts infected by multipartite viruses together indicate that this situation of minimum cost is never reached [13–17]. Although other explanations are imaginable, one possibility is that the actual benefits in these viral systems are related to the differential accumulation of the distinct segments. We and others accordingly proposed that multipartite viruses can tune gene expression in fluctuating environments by modifying gene (or segment) copy number [3,14] and that this capacity could be adaptive [18].

Gene copy number (GCN) and copy number variations (CNV), defined as variations of the number of copies of one or several genes across individuals, have a strong impact on gene expression and phenotypes in all organisms [19]. That CNV-induced changes in gene expression contribute to adaptation in fluctuating environments, particularly upon colonization of new niches and host switching in host-pathogen interactions, has been consistently demonstrated through experimental evolution of fungi [19], bacteria [20] and viruses [21]. Some general features related to CNV are highlighted by the corresponding literature. First, there is a rampant generation of copy number polymorphism, sometimes occurring at a rate higher than the mutation rate, that selection can act on within a population [22,23]. In all cases, CNV polymorphism is generated by recombination and DNA repair machineries, preferentially acting on specific features of sequences flanking the amplified regions. Accordingly, and depending on these flanking sequences, some genome regions are more

prone to amplification than others [20,21,23,24]. Second, depending on the regulatory network within a genome, even small-scale amplification can have large effects. The amplification of one region can either increase or decrease the expression of specific genes, located within or outside this region [22,25], with drastic non-linear changes and even bifurcation in the behavior of the network when amplification thresholds are reached [25]. Experimental observations of such non-linear effects of gene amplification have been reported and are discussed in [25]. Third, CNV repeatedly proved immediately adaptive due to a simple gene dosage effect [20–22]; i.e. following environmental changes, the expression of a gene may be deregulated (maladapted) and a simple adjustment of its copy number (gene dosing) alleviates the defect and the corresponding GCN variant is selected for [19,23,26]. Fourth, such a coarse mechanism of amplification-mediated gene expression tuning (AMGET) [27] is based on gene expression heterogeneity within the population, and can evolve rapidly at a pace where transcriptional regulation has no chance to emerge/adapt de novo [19,22,24,26–28]. Fifth, gene amplification is costly [20–22]. Consequently, whenever selection pressure is removed, when the organism is back in the benign non-restrictive environment or when a regulatory mutation occurs, the extra copies of the gene are rapidly deleted. This phenomenon of genome expansion and contraction, named “genomic accordion”, has been empirically observed in fungi, bacteria and viruses [19,23,26]. Sixth, and finally, the fleeting nature of genomic accordion often leaves no sequence signature, and therefore its importance in the adaptation and evolution of pathogenic fungi, bacteria and viruses has likely been underestimated [20,21,27].

Related studies on viruses have focused on monopartite large double stranded DNA viruses such as phage T4 [29], baculoviruses [30], herpesviruses [31], poxviruses [26] or even giant viruses [32], because they can accommodate relatively large genome size variation. It is commonly assumed that physical packaging constraints do not allow such genome size variations for other viruses, as for example RNA or ssDNA viruses, and thus preclude any possibility of genomic accordion-like adaptation processes [21]. In this context, it is astounding that multipartite viruses have not been envisaged as potentially specialized genetic systems for amplification-mediated gene expression tuning, and this is the hypothesis we experimentally addressed in this study. Because each segment is separately packaged in its own virus particle, the genome-length constraint on segment copy number is totally absent in a multipartite architecture, opening the way to GCN-regulated gene expression and GCN-driven adaptation for ssDNA/RNA viruses where it is usually deemed impossible. Similarly, because genome segments are by definition physically separated, each could be amplified independently with no requirement for recombination/nucleic acid repair machineries. Precisely because no sequence rearrangement appears necessary, the system may not require any sequence modification at all and therefore be extremely conductive to CNV. Would this be so, the importance of CNV in the way of life of multipartite viruses may have long been overlooked because of the total

absence of genomic (sequence) signatures, even the transient ones involved in CNV and genomic accordion in other organisms.

Using the faba bean necrotic stunt virus (FBNSV, family *Nanoviridae*), where each of the eight genome segments encodes a single gene and where each segment is encapsidated individually, we provide support for the use of segment amplification-mediated gene expression tuning as an everyday lifestyle. We earlier reported that the FBNSV segments each accumulate in specific amounts, reproducibly yielding a host-dependent frequency pattern designated as the genome formula [14]. We then speculated that the copy number of each of the segments could contribute to the control of gene expression, and others theoretically supported the idea that a genome formula producing a gene expression pattern better adapted to a given environment can be selected for extremely rapidly [18]. Here, we empirically show that the copy number of each DNA segment correlates positively and linearly to the concentration of its encoded mRNA, but rarely (if at all) to that of the other viral mRNAs, indicating that gene copy number variations drive gene expression in FBNSV. This effect of the genome formula on gene expression is consistently verified in two distinct host species. Further and unanticipatedly, our results reveal that the genome formula modifications observed upon host switching compensate for distinct rates of mRNA production and maintain a relatively constant stoichiometry in the viral transcriptome. Finally, high-throughput sequencing of twenty parallel viral lines demonstrates that the host-dependent FBNSV genome formula shift is not associated with positive/negative selection of sequence variants but rather illustrates a *bona fide* mutation-free copy number variation.

Materials and Methods

Viral strain and plant infection procedures

In all experiments, we used the FBNSV isolate JKI-2000 provided by the Gronenborn lab and described in [33]. Faba bean (*Vicia faba, cv "Seville"*) plants were agroinoculated with cultures of *Agrobacterium tumefaciens* COR308 strain, each carrying a pBin19 plasmid containing a tandem repeat of one of the 8 FBNSV segments. All 8 *A. tumefaciens* cultures were mixed together at equal proportions and inoculated into plants as described in [14]. For practical reasons, alfalfa (*Medicago truncatula*) plants were infected via aphid transmission as described earlier [14]. We have previously shown that the same genome formula is reached whether plants are infected through agroinoculation or aphid transmission and whatever the initial frequency of inoculated segments [14].

Experiment 1

Overview: Concomitant quantification of viral DNAs and mRNAs

The genome formula was characterized as the median relative frequency of each segment across several plant replicates. The idea of this experiment was to embrace the across-replicate variation in the relative frequency of each segment in order to see whether it is translated into across-replicate variation in the relative frequency of the corresponding mRNAs. We thus estimated the relative concentrations of both viral DNA segments and viral mRNAs in each plant sample analyzed. For plant viruses, the viral gene expression is stopped at some point of the infection in fully infected tissues [34]. Consequently, to ensure capturing the transient expression of mRNAs, we repeated this experiment at two different time points. The first replicate (Trial A) was performed on 16 faba bean and 28 alfalfa plants. Samples were collected at different dates, on the first day where the individual infected plants showed symptoms, *i.e.* 10 to 15 days post infection (dpi) for faba bean and 13 to 18 dpi for alfalfa. The second replicate (Trial B) was performed on 21 faba bean and 20 alfalfa plants, and samples were collected at one single later date for each plant species, once all plants of the species expressed symptoms, *i.e.* 21 days post infection (dpi) for faba bean and 20 dpi for alfalfa. In both trials, the infection of each of these plants with FBNSV was independent.

Extraction of single stranded DNA and mRNA from each plant sample

On each infected plant, an apical leaflet was sampled. Approximately 100 mg of leaf tissue was placed in a microtube containing two sterile glass beads, and frozen immediately in liquid nitrogen. Samples were homogenized mechanically using a mixer mill MM 301 (four cycles of 20 seconds at 30 Hz).

To extract nucleic acids, 900 μL of GHLC buffer (6.5 M guanidinium hydrochloride, 100 mM Tris-HCl pH 8.0, 100 mM sodium acetate pH 5.5, 0.1 M β-mercaptoethanol) were added to the homogenized samples. Tubes were vortexed and then centrifuged at 10,000 *g* for 10 minutes at 4°C in a 5415R Eppendorf (Hamburg, Germany) centrifuge. Nine hundred microliters of TRI Reagent (Sigma-Aldrich) warmed at 65°C were added to supernatants. Tubes were vortexed gently over three cycles of 30 seconds, and 200 μL of chloroform were added. After vortexing, samples were incubated for ten minutes at room temperature, and centrifuged at 12,000 *g* for 15 minutes at 4°C. For each sample, 1,200 μL of aqueous upper phase containing nucleic acids were retrieved and divided in two tubes each containing 600 μL. These 600 μL were mixed with 560 μL of cold (-20°C) isopropanol and centrifuged at 12,000 *g* for 20 minutes at 4°C. Supernatants were discarded and pellets washed with 70% ethanol at 4°C. Finally, nucleic acids were resuspended in 50 μL RNAse free water, and the two

tubes for each sample pooled back together to obtain 100 μL of nucleic acid solution. This nucleic acid extraction step allowed retrieving both single stranded DNA and mRNA from the plant tissues.

Estimation of the genome and transcriptome formulas by qPCR

Quantitative PCRs were directly performed on these nucleic acid samples in order to infer the FBNSV genome (DNA) formula. Inferring the viral transcriptome (RNA) formula by qPCR was more tedious as it first required complete removal of the viral DNA and then reverse transcription of the mRNAs into cDNA.

Total elimination of the viral DNA could be achieved by using two treatments, a DNAse I digestion followed by the purification of mRNAs. The DNAse digestion was conducted by mixing 16 μL of nucleic acid sample with 2 μL of 10X DNAse buffer (400 mM Tris-HCl pH 8.0, 100 mM $MgSO_4$ and 10 mM $CaCl_2$) and 2 μL of DNAse I (Promega). This mix was incubated at 37°C for 30 minutes. A subsequent 10 minutes incubation at 65°C inactivated the DNAse. After digestion, the Dynabeads mRNA purification kit (Ambion - ThermoFisher) was used on the nucleic acid samples following the manufacturer's recommendations. Control qPCRs were performed on these samples and confirmed the complete degradation/elimination of viral DNA (Figure S1 in Supporting Information).

For production of the cDNAs, 10 μL of mRNA samples were mixed with 1 μL dNTP [10 mM] incubated for 5 minutes at 65°C and later placed on ice. A mix composed of 4 μL of 5x buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl and 15 mM $MgCl_2$), 2 μL of DTT [100 mM] and 40 units of RNAsin ribonuclease inhibitor (Promega) was added to the mRNA sample and incubated for 2 minutes at 42°C. Two hundred units of SuperScript™ II Reverse Transcriptase (RT) (Invitrogen) were added to the mix, followed by an additional 50 minutes incubation at 42°C. The RT was inactivated with a final 15 minutes incubation at 70°C. The newly formed cDNAs were diluted 10 times so that the buffer does not affect the following qPCR reactions.

All qPCR reactions (40 cycles of 95°C for 10 s, 60°C for 10 s and 72°C for 10 s) were carried out using a LightCycler 480 thermocycler (Roche) and the LightCycler FastStart DNA Master Plus SYBR green I kit (Roche), following the manufacturer's instructions. The nucleic acid sample (1.2 μL of a 10-fold dilution of either total nucleic acid extracts or cDNA preparation) was added to the qPCR mix (5 μL of Roche 2x qPCR mastermix, 3.5 μL of H_2O, 0.3 μL of primer mix, 8.8 μL total) after distribution in 384-well microtiter plates. Primers [14] were used at a final concentration of 0.3 μM for amplifications of the C, M, S segments and 0.5 μM for amplifications of the N, R, U1, U2, U4 segments.

Serial dilutions of plitmus28 plasmids each carrying one of the eight FBNSV segment [33] were placed on each qPCR plate (8 serial dilutions per PCR plate in total, one for each FBNSV segment). These were used as an internal control in order to draw a standard curve for each segment

and for each qPCR plate, alleviating any potential bias related to between-qPCR plate variations. Fluorescence data were first analyzed with the LinRegPCR program [35] and later converted into ng of DNA by using the standard curves. Both DNA and RNA formulas could then be inferred by computing the relative proportions of each segment or mRNA as described in [14]. All qPCR reactions were duplicated (two wells on the same PCR plate).

Statistical analyses

To investigate the relationship between gene expression and the concentration of DNA we first calculated the Pearson correlations between the frequency of each segment and the frequency of its corresponding mRNA in each host plant and trial. Because of the large number of correlations and tests we applied the Benjamini-Hochberg False Discovery Rate (FDR) correction to the correlations across all segments for each host plant species and trial. These results are reported in Figure 1 and Supplementary Table S1. We then calculated the Pearson correlations between the frequency of each segment and that of the seven non-cognate mRNA, but in this case we did not apply FDR corrections, as further commented in the Results section. We used the R software version 3.1.3 [36] to calculate all these correlations.

For further characterizing the relationship between the frequency of each genome segment and that of its cognate mRNA, we compared linear and quadratic model fits to the data and applied model selection using the Akaike Information Criterion (AIC) as described in [37,38].

To study how the different factors (segment, host plant, DNA formula) interact, we modeled the concentration of mRNA of each segment as a function of the segment, the host plant and the concentration of DNA of the segment. Performing such analyses on relative frequencies of mRNA and DNA would provide intuitively interpretable results on intuitively normalized quantities: the frequencies. Unfortunately, such an analysis would be flawed by the fact that the frequencies of the DNA and mRNA of different segments within each replicate are not independent, since they sum to one; and because of this the regression coefficients linking them would also be correlated since their mean should also equal one. We thus opted for the following approach: (*i*) to investigate the interaction between the DNA formula and the host plant species we performed separate analyses on each segment, modeling the logit frequency of the segment's mRNA as a function of the host plant species, the logit of the frequency of the segment's DNA and their interaction; these analyses are reported in Supplementary Table S2; (*ii*) we run a full model on the concentrations, and not the frequencies, of the DNA and mRNA of each segment, because the concentrations are not parametrically constrained. To comply with analysis of variance assumptions these concentrations were first transformed using the Johnson Sb transformation. The transformed values were analyzed in a mixed linear model whose dependent variable was the concentration of mRNA, and the

explanatory variables were 'replicate', declared as a random factor, and 'segment', 'host plant' and 'concentration of DNA' declared as fixed variables (and all the multiple interactions among the latter three declared as fixed variables). This analysis is reported in Table S3 in Supporting Information. The analyses mentioned in this paragraph were performed using JMP 13.2.1 (SAS Institute 2016).

A distance between DNA formulas and between RNA formulas was calculated to compare the situation in faba bean and in alfalfa host plants. This distance between formulas was calculated as follows:

$$d = \Sigma \, |f_i^{faba} - f_i^{alfalfa}|$$

where f_i is the relative frequency of the *i*th segment (or mRNA) in the formula.

All distances between DNA formulas in faba bean and in alfalfa and between RNA formulas in faba bean and in alfalfa were calculated (16 x 28 = 448 distances between DNA formulas and 448 distances between RNA formulas in Trial A; 21 x 20 = 420 distances between DNA formulas and 420 distances between RNA formulas in Trial B). As the formula of each plant was used several times to calculate all possible distances (e.g., the genomic formula of the faba bean plant 1 was used 28 times to calculate all distances between this formula and all alfalfa formulas), not all distances in the dataset are independent. In order to take this pseudo-replication into account, we analyzed these distances with a mixed model with the factor "faba bean plant identity" and "alfalfa plant identity" as random factors and the "nucleic acid" (DNA vs. mRNA) and time (Trial A vs. Trial B) as fixed effect factors. This statistical analysis was performed with JMP 13.2.1 (SAS Institute 2016).

Experiment 2

Overview: Monitoring viral polymorphism in populations passing from faba bean to alfalfa

This experiment has been described in a previous technical paper estimating the various possible quantitative biases during amplification steps and ultra-deep sequencing of these viral populations [39]. In the present study, the same experiment and thus the same deep sequencing dataset are used to monitor polymorphism in 20 independent viral populations passing from faba bean to alfalfa host plants.

Briefly, 15 aphids were placed on each of 20 FBNSV-infected faba bean plants, three weeks post-infection. Three days later, 10 of these aphids were used from each plant to transmit the FBNSV to a set of 20 alfalfa plants, thus creating 20 independent viral populations. During this experiment, total DNA extraction was performed on systemically infected faba bean (21 days post infection, just before aphids where placed on the plants) and alfalfa plants (26 days post inoculation by the aphids). qPCR were first performed on all 40 DNA extracts in order to measure the FBNSV genome formula in the two host species. Then, a rolling circle amplification (RCA, amplifying single stranded circular

FBNSV DNA segments) was performed in order to enrich the samples with viral DNA sequences, and the 40 RCA products were sent for deep sequencing (for full details see [39]). The full sequence data set is available upon request.

Candidate mutations for genome formula variation

To be considered a mutation impacting the FBNSV genome formula when the virus is passed from faba bean to alfalfa, the mutation should (*i*) show a significant increase in frequency between faba bean and alfalfa samples, beyond that expected under drift alone; (*ii*) this change in frequency should be consistent across replicates; and (*iii*) this increase should correlate with the variation in the genome formula. We describe below how mutations under selection have been searched for. The other two requirements, repeated occurrence in parallel viral populations and correlation with genome formula changes, are reported in the Results section.

In order to identify mutations whose frequency changed between faba bean and alfalfa samples beyond what is expected under drift alone, we tested for the homogeneity of temporal differentiation across nucleotide sites for each viral population passed from faba bean to alfalfa, using a procedure inspired by Goldringer and Bataillon [40]. The rationale of this analysis is that if all sites are selectively neutral, they should provide identically distributed estimates of temporal differentiation. However, if some sites are targeted by selection (or if they are in linkage disequilibrium with selected variants), then some heterogeneity in site-specific measures of temporal differentiation should be observed. To identify those sites that show outstanding differentiation compared to neutral expectation, we simulated the dynamics of nucleotide frequency change between the faba bean and the alfalfa samples, conditionally on the initial nucleotide counts in the faba bean sample and on the strength of genetic drift during the experiment.

To that end, we first estimated the haploid effective size of the viral population (N_e) using approximate Bayesian computation (ABC) (see, e.g., [41]). Because each segment is transmitted independently and since the genome formula may reflect different rates of genetic drift during transmission [42], ABC analyses were performed (and therefore N_e estimates were computed) independently for each segment. The data consisted of the observed number of A, T, C and G counts obtained by deep-sequencing in all forty FBNSV populations (20 in faba bean, 20 in alfalfa). Yet, to lessen the impact of sequencing errors in deep-sequencing data, we discarded all the variants with an observed frequency of the most frequent allele (MAF, computed as the overall frequency across the faba bean and the alfalfa samples) falling above 0.97, thereby assuming a variant calling threshold of 0.03 (see, e.g. [43]). We ended up with 269 polymorphic sites (out of 7,907 sites x 20 replicates = 158,140 sites), corresponding to 173 unique sites. For each segment-specific analysis, all polymorphic sites (with MAF $\leq$ 0.97) were pooled.

We then simulated individual nucleotide frequency trajectories, as follows: suppose that we observe a vector $\mathbf{y} \equiv (y_A, y_C, y_G, y_T)$ of nucleotide counts, out of the total coverage $n_{fb} \equiv y_A + y_C + y_G + y_T$ in the faba bean sample. We assume that these observed counts correspond to a (biallelic) SNP with sequencing errors, and we denote by y_{fb} the counts for the major (most frequent) allele. We further assume (following [44]) that y_{fb} is drawn from a binomial distribution $B(n_{fb}, \pi_{fb})$ where π_{fb} is the (unknown) allele frequency of the major allele in the faba bean population. Assuming a (uniform) Beta(1,1) prior distribution for π_{fb}, and using the Bayes inversion formula, the posterior distribution of π_{fb} is distributed as Beta(y_{fb} + 1,n_{fb} - y_{fb} + 1). For each nucleotide site and for each ABC simulation, we therefore draw the initial allele frequencies in the faba bean sample $\tilde{\pi}_{fb}$, at random from a Beta(y_{fb} + 1,n_{fb} - y_{fb} + 1) distribution. We then draw "pseudo-observed" allele counts using a random binomial draw from $B(n_{fb}, \tilde{\pi}_{fb})$. This procedure allows accounting for the sampling variance in initial allele frequencies. Then, we simulate τ generations of drift, using successive binomial draws with parameters N_e (the segment-specific effective population size) and the nucleotide frequencies in the previous generation. In the last generation, a sample of nucleotide counts is drawn from a binomial distribution with parameters n_M (the total observed coverage in the alfalfa sample) and $\tilde{\pi}_M$ (the simulated nucleotide frequencies in the last generation). In what follows, we considered a single generation of drift (i.e., $\tau = 1$). Finally, sequencing errors were modeled (for both the faba bean and the alfalfa samples) by means of multinomial draws, with probabilities (1 - ε) not to mutate, and $\varepsilon/3$ to mutate to any other state. For each segment, a total of 1,000,000 ABC simulations were performed assuming a log-uniform prior for N_e in the [1; 1,000] range and a log-uniform prior for the error rate ε in the [0.001; 0.1] interval. To avoid any bias, all simulations with a major allele frequency larger than or equal to 0.97 were discarded. The summary statistics considered to compare observed and simulated data were the mean, variance, skewness and kurtosis of (*i*) single-locus estimates of F_{ST} [45] computed between the faba bean and the alfalfa samples at each SNP (with a major allele frequency ≤ 0.97) within a segment ; (*ii*) the allele frequency difference of the most frequent allele between the faba bean and the alfalfa samples at each SNP within a segment. Posterior distributions of N_e and ε were computed using the abc package for R [46] with the local linear regression model [47] and a tolerance threshold of 0.001.

In a second step, for each segment and for each variant, we tested the null hypothesis that the locus-specific differentiation measured at this focal marker was only due to genetic drift. For this purpose, we computed the expected distribution of F_{ST} at each site, conditional upon the estimated effective population size for the segment, the inferred error rate, and the allele frequencies at the focal site in the faba bean sample. To do so, we simulated individual nucleotide frequency trajectories following the same rationale as for the ABC simulations, drawing N_e and ε estimates from

their ABC posterior distributions. For each simulated trajectory, we computed site-specific estimates of temporal F_{ST} from the simulated nucleotide counts at the initial and last generation. The whole procedure was repeated at least 1,000,000 times for each of the 269 polymorphic sites. Finally, we assigned a *p*-value to each site, computed as the proportion of simulations giving a site-specific estimate of F_{ST} larger than or equal to the observed value at the focal nucleotide site. As above, all simulations with a major allele frequency larger than or equal to 0.97 were discarded.

All codes and R scripts, as well as the SNP counts data, specifically developed and used for these analyses are publicly accessible at (https://doi.org/10.57745/ILFCP4).

Results

Gene copy number drives gene expression in FBNSV

To investigate whether the FBNSV gene expression is affected by GCN, we assessed whether variation of the relative concentration of the viral mRNA produced by each segment across different individual plants of a given host species could be explained by variation of the genome formula across these same individual plants (Experiment 1 described in Materials and Methods). In each plant sample analyzed, we thus estimated the relative concentrations of both viral DNA segments and their cognate mRNAs, that we hereafter respectively designate genome formula and transcriptome formula. It has been shown in various biological systems that the viral gene expression is stopped at some point of the infection [34]. To maximize our chances to capture the transient expression of viral mRNAs, we thus repeated this experiment at two different time points: early in Trial A, as soon as infection symptoms were visible on each individual plant, and later in Trial B, at the same time post-infection for all individual plants once all had exhibited symptoms. Because an mRNA half-life can be short, we were aware that the two trials could differ in their capacity to potentially reveal a correlation between the genome and transcriptome formulas.

Figure 1 (and Table S1) shows that the relative frequency of each of the eight mRNAs of the FBNSV is positively correlated to that of its encoding segment in Trial A, both in faba bean and in alfalfa host plants (except for the S segment in faba bean for which the correlation is not significant). Trial B provided consistent observations, with six and four segments, respectively on faba bean and alfalfa, showing significant positive relationships (Figure S2 and Table S1). The segment-by-segment analyses identified statistically significant effects of either the DNA formula or its interaction with the host plant species for all segments in Trial A and for six segments in trial B (Table S2), further indicating that a change in a segment frequency and thus of the genome formula induces a change of the gene expression.

The slopes of the linear regressions between mRNA and DNA relative frequencies vary with both the segments and the host species (Figure 1). To assess the statistical significance of this slope variation across hosts we analyzed the plant species effect on the DNA/mRNA correlation for each segment separately. A statistically significant effect was observed for segments C, R and U1 in both trials and for segment M in Trial A (Table S2), indicating that these segments are differentially expressed in the two host plant species. The slope variation across segments is further supported by the statistically significant segment-by-plant interaction in the full model using mRNA and DNA concentrations (Table S3).

The relationship between gene copy number and gene expression is remarkably simple

Two observations indicate a simple relationship between genome formula and gene expression in this viral system. First, the relative abundance of any specific genome segment does not strikingly depart from a simple positive and linear relationship with that of its cognate mRNA. Second, most correlation tests between the frequency of any specific segment and that of each of the seven non-cognate viral mRNAs proved non-significant.

To substantiate the first observation, we verified whether incorporating quadratic terms in the regressions better explains the data than the regressions reported in Figure 1 and S2, which only contain terms linear in DNA concentration. Across all trials in faba bean and alfalfa, this proved very rarely true, *i.e.* for 5 regressions out of 32 (for full detail see Table S4 in Supporting Information). Adding a quadratic term explained the data better solely for C, N and R in faba bean Trial A (in the case of segments N and R, after removing the point with the highest DNA concentration -rightmost in Figure 1- this was no longer true), for N in alfalfa Trial A, for none of the segments in faba bean Trial B, and for U4 in alfalfa Trial B (here also, after removing the rightmost point the quadratic term is no longer statistically significant).

For the second observation, we calculated all possible Pearson's correlations between viral DNAs and mRNAs in faba bean and alfalfa and in trials A and B (256 correlation tests; see Table S5 in Supporting Information). Concentrations of genome segments near systematically correlated positively with those of their cognate mRNAs, as already presented in the previous section, but rarely with non-cognate mRNAs. More specifically, in faba bean Trial A, 87.5% (7/8) cognate correlations were statistically significant vs. 14% (8/56) non-cognate (one-tailed Fisher exact test $p < 0.0001$). In faba bean Trial B, the corresponding numbers were 75% (6/8) for cognate vs. 23% (13/56) for non-cognate (one-tailed Fisher exact test $p = 0.0055$), in alfalfa Trial A, 100% (8/8) for cognate vs. 16% (9/56) for non-cognate (one-tailed Fisher exact test $p < 0.0001$), and for alfalfa Trial B, 50% (4/8) for cognate vs 5% (3/56) for non-cognate (one-tailed Fisher exact test $p = 0.0033$). In order to make our

conclusions as conservative as possible, no corrections for multiple tests and related false discovery rate were performed in this analysis.

All together, these results suggest that changes in the frequency of a given FBNSV genome segments positively and linearly affects the expression of the corresponding gene, while poorly affecting the others.

Different genome formulas in faba bean and alfalfa produce similar transcriptome formulas

We plotted and compared genome and transcriptome formulas when estimated from faba bean and from alfalfa plants (Figure 2 for Trial A, Figure S3 for trial B). As already observed in a previous study [14], the FBNSV genome formulas in faba bean and alfalfa are clearly distinct. However, the transcriptome formulas observed in the two host species appear more similar. To confirm this observation, we compared the distance between genome formulas and between transcriptome formulas in these two hosts (see Materials and Methods). Our statistical analysis formally established that the distance between faba bean and alfalfa transcriptome formulas was significantly smaller than the distance between faba bean and alfalfa genome formulas in both trials (Table 1 for trial A and Table S6 for trial B). These results demonstrate that while the relative copy number of the genome segments changes drastically when FBNSV switches from faba bean to alfalfa, the relative proportions (or stoichiometry) of the eight mRNAs tend to be conserved. This interesting observation is further discussed later.

Looking for adaptive mutations in the FBNSV sequence

To investigate whether the change in genome formula when FBNSV switches hosts is due, or not, to selection of mutations in the sequence of one or several segments, we re-analyzed deep-sequencing data from 20 independent FBNSV populations passed from faba bean to alfalfa (Experiment 2 described in Materials and Methods). Figure S4 shows that, just like in Experiment 1 and in earlier reports [14,39], the FBNSV genome formula was clearly different in faba bean and alfalfa, confirming that the expected host-dependent genome formula shift has occurred.

The modification of a phenotype during viral infection could either have a genetically determined basis or be due to a plastic response. To distinguish between these two possibilities, we aimed at identifying mutations showing outstanding differentiation between faba bean and alfalfa samples (as compared to what is expected under genetic drift alone) that could be interpreted as evidence of selection and whose frequency variation across host species could explain the genome formula variation.

Over the 7907 nucleotide positions in the concatenated FBNSV genome and the 20 replicated viral populations monitored in faba bean and alfalfa samples, we detected 269 variants (*i.e.*, with a

major allele frequency ≤ 0.97), corresponding to 173 distinct sites. From the ABC analysis, we then inferred the effective population size for each segment, as well as the error rate (Table S7), and used these estimates to simulate allele frequency dynamics in order to test whether the extent of differentiation observed in our viral lines passing from faba bean to alfalfa could be explained by drift only. It is noticeable that these estimates of effective population size for each segment, though using a totally distinct approach, are very similar to those reported earlier [42]. Interestingly, we found only 8 sites at which the observed differentiation departed from the expected distribution under neutrality ($p \leq 0.01$). Among these eight sites, two were revealed in two out of the twenty parallel viral lines and six were revealed only once. The position of these sites on the FBNSV genome segments, whether they are in coding regions, synonymous or non, is indicated in Table S8. Figure 3A illustrates that the frequency of each of the corresponding mutations can follow very diverse trajectories in the 20 parallel FBNSV lines, either increasing, decreasing or not changing at all, pleading against a deterministic process.

We finally tested whether the frequency variations (observed between faba bean and alfalfa samples) of each of these 8 candidate mutations were correlated to changes of the genome formula. For this, we calculated the distance between the genome formula in faba bean and that in alfalfa for each of the 20 FBNSV lines, and plotted these distances against the variation of the mutation frequency in each corresponding line (Figure 3B). All regressions proved non-significant, further confirming that even the extremely rare sites identified as eventually showing higher differentiation than expected under drift alone cannot account for the genome formula shift of FBNSV. We thus conclude that this is a mutation independent process, and whether it is to be considered a plastic or genetically-driven phenomenon is not trivial and is further discussed in the next section.

Discussion

After the discovery of the genome formula of nanoviruses [14,17], additional studies performed on other multipartite viruses showed that their genomic segments also accumulate at different frequencies [15], in a host-dependent manner [16]. While this phenomenon appears general in multipartite [48] and perhaps even in segmented viruses [49,50], the mechanisms leading to the establishment of the genome formula as well as its actual function remain a mystery. We first hypothesized [7,14] that the multipartite genome architecture would allow the adjustment of gene expression through the modulation of the GCN, and this proposition was further discussed [16,48–50] and even theoretically supported [18] by others. The proposed process [7,14,48] is that infection sites could randomly differ in the proportions of the different segments, and that within-host selection would act on this variation to favor the replication/dissemination of those sites with a

genome formula producing the gene expression pattern better adapted to the specific host. According to this process, viral populations of a given virus genotype could rapidly converge to the genome formula that is best adapted to a given environment [18]. What is appealing with this hypothesis on the mechanism that can generate the set-point genome formula is that it provides an astonishingly versatile means to regulate gene expression that perfectly matches, or even magnifies, the general conclusion enounced from studies on CNV in other organisms: gene amplification is based on rampant generation of copy number polymorphism and allows rapid and graded response for populations in heterogeneous and changing environments, which can tune gene expression when promoters are not adequately regulated at a pace where regulatory sequences have no time to evolve [19–21,27]. Baseline experimental support for such a function of FBNSV genome formula lies in three key points (*i*) the demonstration of a correlation between GCN and gene expression -- *i.e.* a correlation between the relative segment frequencies and those of their respective mRNA --, (*ii*) the ability to adjust the GCN when the environment changes, and (*iii*) the demonstration that this is a mutation-free process, confirming that no *de novo* regulatory sequences have evolved. Thus, our results demonstrate the functional role of the genome formula and its variation, though its potential role in adaptation to host switches, or potential other physiological conditions of its hosts, remains to be empirically demonstrated.

The first point is consistently verified in both faba bean and alfalfa. In the 'early' Trial A, the only segment that did not show a statistically significant correlation was segment S in the faba bean background (*i.e.* one non-significant correlation out of 16 – Figure 1 and Table S1). This is probably due to the relative scarcity of this segment in the faba bean DNA formula, which leads to overall low relative frequencies of both S mRNA and DNA, and consequently low between-plant variation. In trial B, statistically significant correlations could be observed in 10 instances out of 16. As already commented in both the Materials and Methods and the Results, we anticipated a possible distinct turnover for DNA segments and for their cognate mRNAs, which might bias the assessment of their correlated accumulation at some stages of the infection. Despite this potential drawback apparently affecting more trial B, the DNA segment frequencies proved to significantly impact the cognate mRNA production in most cases. Since the set point genome formula has been reported [14] or modeled [18] to be reached early during the onset of the infection and then remain constant, we assume that there is more experimental noise in Trial B, related to shorter lived mRNA when compared to viral DNA.

The fact that different FBNSV segments had different levels of mRNA production (different slopes for different DNA/mRNA regressions Figure 1) may simply reflect a different efficiency of the segments' respective promoter, as earlier reported for related banana bunchy top virus (BBTV, genus *Babuvirus*, family *Nanoviridae*) [51], or different mRNA half-lives. In FBNSV and in nanoviruses in

general, though totally uncharacterized, each gene likely has its own promoter strength because sequences flanking the transcription start are not highly conserved across segments. We note, however, that whatever the regulatory sequences on distinct FBNSV segments, the effect of gene copy number variation reported here impacts gene expression pattern. More, interestingly, the mRNA production could also vary for a given segment between the faba bean and alfalfa backgrounds (Figure 1), indicating that its promoter may not be equally compatible with the host plant species' respective transcriptional machinery or that the stability of mRNA may vary across hosts. Considering that the FBNSV genome formula is different and modulates the expression of the viral mRNAs in the two host plant species, and that a given segment does not produce the same amount of mRNA in these two environments, one intuitively expects the relative proportion of the distinct viral mRNAs (transcriptome formula) to also greatly vary, at least reflecting differences observed at the DNA level and perhaps even more. Surprisingly, however, our results reveal that the transcriptome formula tends to be more conserved in the two hosts. We believe this observation supports the second key point listed in the first paragraph of the Discussion. The potential importance of the stability of gene expression pattern and how copy number variations can maintain a dosage balance between interacting genes has earlier been discussed [22]. Here, the interactions between FBNSV and the mRNA metabolism machineries in faba bean and in alfalfa should modify the viral mRNA frequency pattern (different slope of segments DNA/mRNA regressions in the two hosts). Our results suggest that the modification of the genome formula in the two hosts allows the maintenance of a dosage balance between FBNSV genes, resulting in a similar transcriptome formula.

The third point lies in the demonstration that the host-related genome formula shift is a mutation-free phenomenon, which both greatly advances our understanding of this genetic system and adds one additional enigma. The advance is the discovery that the genome formula changes of a given viral isolate upon host switching are plastic and not traceable on a sequence basis. In diverse organisms, it is a classical view that gene amplification is often transient and followed by gene contraction (genomic accordion) ultimately leaving no genomic signature [19,23,26,27]. Remarkably, for multipartite viruses (at least for FBNSV), not only gene expansion/contraction leaves no genomic signature, but it does not even require transient sequence modification. The other face of the coin is that this discovery adds one dimension to the puzzle: In monopartite viruses, bacteria and eucaryotes it is a genome that is modified with a portion amplified. The corresponding genotype then represents a genetic innovation that is selected for or against. In multipartite viruses, when the genome formula changes, what exactly is the genetic innovation? A group of interacting segments may represent the genetic innovation. Indeed, as discussed above, we earlier proposed that such

group of interacting segments could be the unit of selection [7,14]. This possibility has been theoretically formalized and supported [18,48] but experimental evidence is still missing.

One additional observation that we found particularly intriguing is the type of relationship between FBNSV GCN and gene expression. As already documented in the Introduction, the impact of gene duplication/amplification on gene expression has been empirically reported in eucaryotes, procaryotes, and monopartite large dsDNA viruses where, to the best of our knowledge, a correlated increase of the expression of the corresponding gene has been evidenced but not characterized in detail. In their seminal theoretical paper, Mileyko and colleagues [25] considered only two to three interacting genes, all located collinearly in the same amplified region, thus all similarly amplified. Despite these simple virtual gene networks, and as already explained in the introduction, a remarkable diversity of possible gene expression changes was revealed. In our experimental system, with all eight FBNSV genes physically separated on distinct genome segments, and all differentially amplified, considering the complete lack of data on the interaction network between these genes, we had no ground for educated guesses but we did not expect something as simple as what we observed. The copy number of each of the DNA segments has a positive and linear relationship with the production of its cognate mRNA with little impact on the expression of other viral genes. Again inspired by and in line with the same theoretical study [25], we propose that the FBNSV may have evolved away from gene-amplification thresholds leading to drastic changes/bifurcation in the behavior of the expression network. Such a simple behavior of the gene expression network might be a condition for this virus to operate amplification-mediated gene expression tuning as an everyday lifestyle without jeopardizing the system at every possible change of the genome formula.

All the above considerations highlight the conceptual problem of what exactly constitutes the genome of multipartite viruses: should it be just the set of the sequences of their segments or should a definition of their genome also include the number of copies of each segment? The genome formula shift accompanying a host switch could be viewed as an indication of viral plasticity (as assumed in the previous paragraph). The situation is, however, unclear: the genome formula shift is in essence a modification of the copy number of specific genes. In this respect, it is not conceptually different from the 'genomic accordion' process. In the case of the monopartite viruses, bacteria and eucaryotes it is obvious to everyone that the determinism of the phenotypic adaptation to the environmental challenge via "genomic accordions" is genetically based. The fact that the genes of multipartite viruses are carried by separate segments offers them great flexibility in the number of copies. This flexibility muddles our conceptual characterizations: because we have difficulty in defining what their genome really is, we have difficulty in deciding on the nature of their responses to environmental changes. Is it genetic, plastic, epigenetic? At this point we cannot but leave this

question open, only the unravelling of the mechanisms underlying genome formula changes, the identification of the unit of selection in these systems, will provide the answer.

Acknowledgements

We are grateful to Sophie Leblaye for plant production. This work was supported by the French Agence Nationale de la Recherche (ANR) grants “Nano” (ANR-14-CE02-0014) and “Nanovirus” (ANR-18-CE92-0028-01), and by Montpellier Université d’excellence, MUSE, project BLANC-MUSE2020-Multivir. RG, RV and SB acknowledge support from INRAE Plant health and environment division, SR from CIRAD, JLZ from IRD and YM from CNRS and IRD.

Competing interests

The authors declare no competing interest for this work.

References

1. Iranzo J, Manrubia SC. Evolutionary dynamics of genome segmentation in multipartite viruses. Proc Biol Sci. 2012;279: 3812–9.

2. Lucía-Sanz A, Manrubia S. Multipartite viruses: adaptive trick or evolutionary treat? NPJ Syst Biol Appl. 2017;3: 34. doi:10.1038/s41540-017-0035-y

3. Michalakis Y, Blanc S. The Curious Strategy of Multipartite Viruses. Annu Rev Virol. 2020;7: 203–218. doi:10.1146/annurev-virology-010220-063346

4. Gilmer D, Ratti C, Michel F. Long-distance movement of helical multipartite phytoviruses: keep connected or die? Curr Opin Virol. 2018;33: 120–128. doi:10.1016/j.coviro.2018.07.016

5. Sicard A, Pirolles E, Gallet R, Vernerey M-S, Yvon M, Urbino C, et al. A multicellular way of life for a multipartite virus. eLife. 2019;8. doi:10.7554/eLife.43599

6. Lucía-Sanz A, Aguirre J, Manrubia S. Theoretical approaches to disclosing the emergence and adaptive advantages of multipartite viruses. Curr Opin Virol. 2018;33: 89–95. doi:10.1016/j.coviro.2018.07.018

7. Sicard A, Michalakis Y, Gutiérrez S, Blanc S. The Strange Lifestyle of Multipartite Viruses. PLoS Pathog. 2016;12: e1005819. doi:10.1371/journal.ppat.1005819

8. Zwart MP, Blanc S, Johnson M, Manrubia S, Michalakis Y, Sofonea MT. Unresolved advantages of multipartitism in spatially structured environments. Virus Evol. 2021;7: veab004. doi:10.1093/ve/veab004

9. Nee S. The evolution of multicompartmental genomes in viruses. J Mol Evol. 1987;25:

277–81.

10. Pressing J, Reanney DC. Divided genomes and intrinsic noise. J Mol Evol. 1984;20: 135–46.

11. Chao L. Levels of selection, evolution of sex in RNA viruses, and the origin of life. J Theor Biol. 1991;153: 229–246.

12. Ojosnegros S, Garcia-Arriaza J, Escarmis C, Manrubia SC, Perales C, Arias A, et al. Viral genome segmentation can result from a trade-off between genetic content and particle stability. PLoS Genet. 2011;7: e1001344.

13. French R, Ahlquist P. Characterization and engineering of sequences controlling in vivo synthesis of brome mosaic virus subgenomic RNA. J Virol. 1988;62: 2411–2420. doi:10.1128/JVI.62.7.2411-2420.1988

14. Sicard A, Yvon M, Timchenko T, Gronenborn B, Michalakis Y, Gutierrez S, et al. Gene copy number is differentially regulated in a multipartite virus. Nat Commun. 2013;4: 2248.

15. Hu Z, Zhang X, Liu W, Zhou Q, Zhang Q, Li G, et al. Genome segments accumulate with different frequencies in Bombyx mori bidensovirus. J Basic Microbiol. 2016;56: 1338–1343. doi:10.1002/jobm.201600120

16. Wu B, Zwart MP, Sánchez-Navarro JA, Elena SF. Within-host Evolution of Segments Ratio for the Tripartite Genome of Alfalfa Mosaic Virus. Sci Rep. 2017;7: 5004. doi:10.1038/s41598-017-05335-8

17. Mansourpour M, Gallet R, Abbasi A, Blanc S, Dizadji A, Zeddam J-L. Effects of an alphasatellite on life cycle of the nanovirus Faba bean necrotic yellows virus. J Virol. 2021; JVI0138821. doi:10.1128/JVI.01388-21

18. Zwart MP, Elena SF. Modeling multipartite virus evolution: the genome formula facilitates rapid adaptation to heterogeneous environments†. Virus Evol. 2020;6: veaa022. doi:10.1093/ve/veaa022

19. Todd RT, Selmecki A. Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. eLife. 2020;9: e58349. doi:10.7554/eLife.58349

20. Elliott KT, Cuff LE, Neidle EL. Copy number change: evolving views on gene amplification. Future Microbiol. 2013;8: 887–899. doi:10.2217/fmb.13.53

21. Bayer A, Brennan G, Geballe AP. Adaptation by copy number variation in monopartite viruses. Curr Opin Virol. 2018;33: 7–12. doi:10.1016/j.coviro.2018.07.001

22. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci. 2012;279: 5048–5057. doi:10.1098/rspb.2012.1108

23. Belikova D, Jochim A, Power J, Holden MTG, Heilbronner S. “Gene accordions” cause genotypic and phenotypic heterogeneity in clonal populations of Staphylococcus aureus. Nat

Commun. 2020;11: 3526. doi:10.1038/s41467-020-17277-3

24. Sasani TA, Cone KR, Quinlan AR, Elde NC. Long read sequencing reveals poxvirus evolution through rapid homogenization of gene arrays. eLife. 2018;7: e35453. doi:10.7554/eLife.35453

25. Mileyko Y, Joh RI, Weitz JS. Small-scale copy number variation and large-scale changes in gene expression. Proc Natl Acad Sci U A. 2008;105: 16659–64.

26. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, et al. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell. 2012;150: 831–41.

27. Tomanek I, Grah R, Lagator M, Andersson AMC, Bollback JP, Tkačik G, et al. Gene amplification as a form of population-level gene expression regulation. Nat Ecol Evol. 2020;4: 612–625. doi:10.1038/s41559-020-1132-7

28. Brennan G, Kitzman JO, Rothenburg S, Shendure J, Geballe AP. Adaptive gene amplification as an intermediate step in the expansion of virus host range. PLoS Pathog. 2014;10: e1004002. doi:10.1371/journal.ppat.1004002

29. Wu CH, Black LW. Mutational analysis of the sequence-specific recombination box for amplification of gene 17 of bacteriophage T4. J Mol Biol. 1995;247: 604–617.

30. Ardisson-Araújo DMP, da Silva AMR, Melo FL, Dos Santos ER, Sosa-Gómez DR, Ribeiro BM. A Novel Betabaculovirus Isolated from the Monocot Pest Mocis latipes (Lepidoptera: Noctuidae) and the Evolution of Multiple-Copy Genes. Viruses. 2018;10: E134. doi:10.3390/v10030134

31. Arias C, Weisburd B, Stern-Ginossar N, Mercier A, Madrid AS, Bellare P, et al. KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. PLoS Pathog. 2014;10: e1003847. doi:10.1371/journal.ppat.1003847

32. Shukla A, Chatterjee A, Kondabagil K. The number of genes encoding repeat domain-containing proteins positively correlates with genome size in amoebal giant viruses. Virus Evol. 2018;4: vex039. doi:10.1093/ve/vex039

33. Grigoras I, Timchenko T, Katul L, Grande-Perez A, Vetten HJ, Gronenborn B. Reconstitution of authentic nanovirus from multiple cloned DNAs. J Virol. 2009;83: 10778–87.

34. Gutiérrez S, Pirolles E, Yvon M, Baecker V, Michalakis Y, Blanc S. The Multiplicity of Cellular Infection Changes Depending on the Route of Cell Infection in a Plant Virus. J Virol. 2015;89: 9665–9675. doi:10.1128/JVI.00537-15

35. Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, Van Den Hoff MJB, et al. Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res. 2009;37: e45. doi:10.1093/nar/gkp045

36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R

Foundation for Statistical Computing; 2021. Available: https://www.R-project.org/

37. Burnham KP, Anderson DR. Model selection and multimodel inference. A practical information-theoretic approach. Second. New York: Springer; 2002.

38. Bolker BM. Ecological models and data in R. Princeton: Princeton University Press; 2008.

39. Gallet R, Fabre F, Michalakis Y, Blanc S. The Number of Target Molecules of the Amplification Step Limits Accuracy and Sensitivity in Ultradeep-Sequencing Viral Population Studies. J Virol. 2017;91. doi:10.1128/JVI.00561-17

40. Goldringer I, Bataillon T. On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. Genetics. 2004;168: 563–568. doi:10.1534/genetics.103.025908

41. Beaumont MA. Approximate Bayesian Computation in Evolution and Ecology. Annu Rev Ecol Evol Syst. 2010;41: 379–406. doi:10.1146/annurev-ecolsys-102209-144621

42. Gallet R, Fabre F, Thébaud G, Sofonea MT, Sicard A, Blanc S, et al. Small Bottleneck Size in a Highly Multipartite Virus during a Complete Infection Cycle. J Virol. 2018;92. doi:10.1128/JVI.00139-18

43. Leonard AS, Weissman DB, Greenbaum B, Ghedin E, Koelle K. Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. J Virol. 2017 [cited 20 Jan 2022]. doi:10.1128/JVI.00171-17

44. Frachon L, Libourel C, Villoutreix R, Carrère S, Glorieux C, Huard-Chauveau C, et al. Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. Nat Ecol Evol. 2017;1: 1551–1561. doi:10.1038/s41559-017-0297-1

45. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38: 1358–1370.

46. Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012;3: 475–479. doi:10.1111/j.2041-210X.2011.00179.x

47. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002;162: 2025–2035. doi:10.1093/genetics/162.4.2025

48. Gutiérrez S, Zwart MP. Population bottlenecks in multicomponent viruses: first forays into the uncharted territory of genome-formula drift. Curr Opin Virol. 2018;33: 184–190. doi:10.1016/j.coviro.2018.09.001

49. Diefenbacher M, Sun J, Brooke CB. The parts are greater than the whole: the role of semi-infectious particles in influenza A virus biology. Curr Opin Virol. 2018;33: 42–46. doi:10.1016/j.coviro.2018.07.002

50. Moreau Y, Gil P, Exbrayat A, Rakotoarivony I, Bréard E, Sailleau C, et al. The Genome

Segments of Bluetongue Virus Differ in Copy Number in a Host-Specific Manner. J Virol. 2020;95: e01834-20. doi:10.1128/JVI.01834-20

51. Yu N-T, Xie H-M, Zhang Y-L, Wang J-H, Xiong Z, Liu Z-X. Independent modulation of individual genomic component transcription and a cis-acting element related to high transcriptional activity in a multipartite DNA virus. BMC Genomics. 2019;20: 573. doi:10.1186/s12864-019-5901-0

Tables

Table 1: Statistical analysis of the distance between genome and transcriptome formulas in faba bean and alfalfa (Experiment 1, Trial A). The nature of the nucleic acid (NatNA), DNA or mRNA, was a fixed factor. We used the individual plant, faba bean or alfalfa, and its interaction with the nucleic acid as random factors to account for pseudoreplication. See Materials and Methods for more explanations.

Model adjusted R^2= 0.8351

Fixed Effect Tests

Source	Nparm[a]	DF[b]	DFDen[c]	F Ratio	Prob > F	
NatNA	1	1	38.59	48.1650	<.0001	***

[a]: number of parameter, [b]: degrees of freedom, [c]: denominator degrees of freedom

REML Variance Component Estimates

Random effect	Variance component	Std Error	95% lower	95% upper	Wald p-value	% of total
faba	0.0017	0.0014	-0.0011	0.0044	0.2334	9.416
alfalfa	0.0004	0.0015	-0.0026	0.0034	0.7895	2.291
NatNA*faba	0.0033	0.0013	0.0008	0.0058	0.0095	18.657
NatNA*alfalfa	0.0072	0.0020	0.0032	0.0112	0.0004	40.424
Residual	0.0052	0.0006	0.0048	0.0057		29.211
Total	0.0177	0.0021	0.0143	0.0226		100.000

Figures legends

Figure 1: Correlations between the relative frequency of FBNSV genome segments and that of their respective mRNAs. The data used here are those from Trial A, Experiment 1 (see Materials and Methods). Each panel shows the correlation between the relative frequency of an FBNSV segment and the relative frequency of the corresponding mRNA. Data points, linear regressions, correlation coefficients and *p*-values are shown in blue and red for FBNSV infecting faba bean and alfalfa respectively. '***', '**', and '*' correspond to *p*-value ≤ 0.001, 0.01 and 0.05, respectively. The dotted line illustrates a slope of 1.

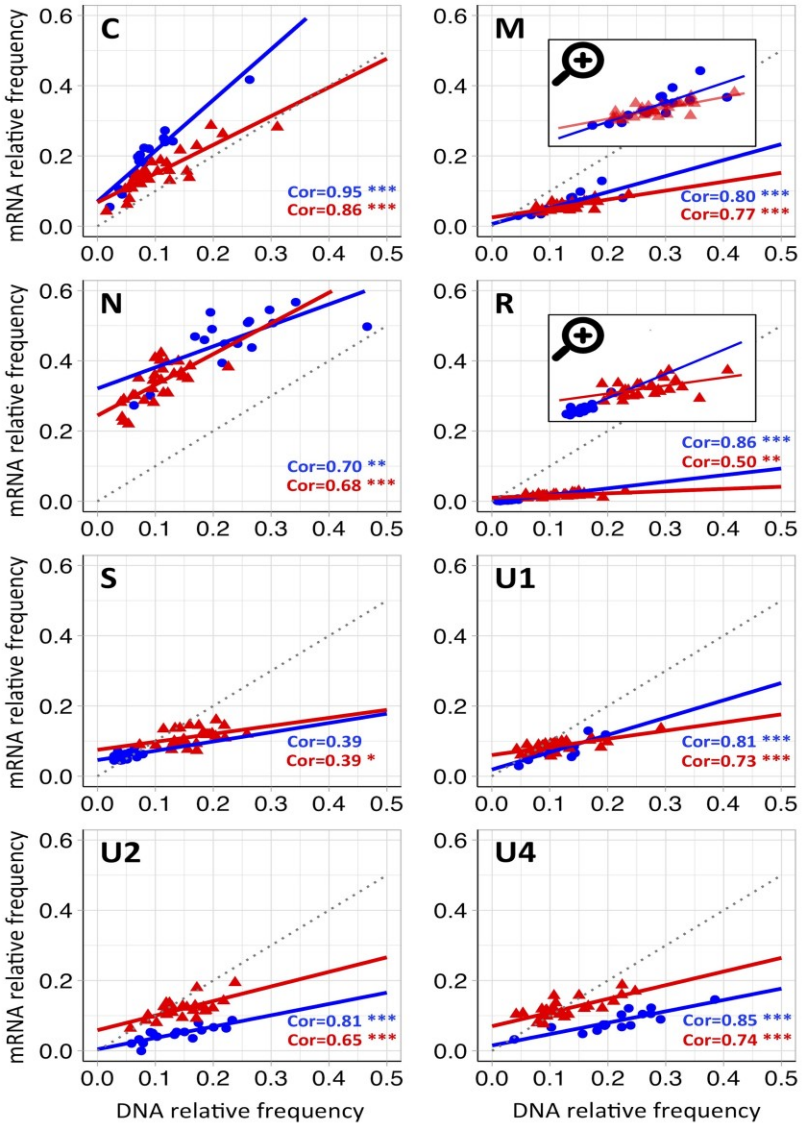

Figure 2: Radar plot of FBNSV genome and transcriptome formulas in trial A (Experiment 1). The median relative frequencies of each FBNSV segment (left) or of their corresponding transcripts (right) are represented on one of the eight axes composing the radar plot (formulas calculated from the 16 faba bean and 28 alfalfa plants in trial A). The FBNSV formulas observed in faba bean and alfalfa are represented in blue and red respectively. Standard deviations are represented by colored bars. The distances between the transcriptome formulas observed in faba bean and alfalfa are significantly smaller than those between the corresponding genome formulas (Table 1).

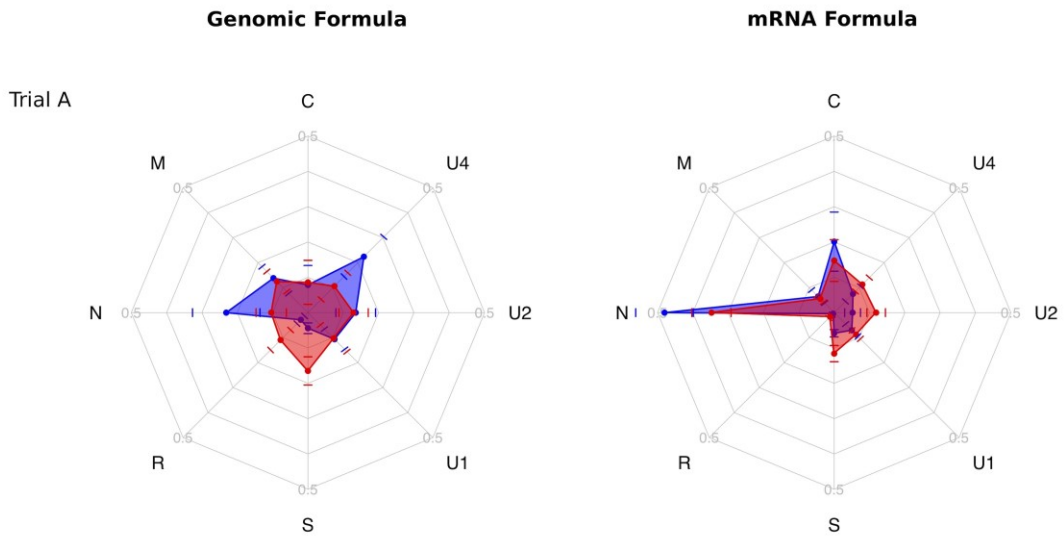

Figure 3: **No adaptive mutations can explain the host-dependent genome formula shift of FBNSV**
Panel A shows the changes in frequency of mutations at the eight sites detected as possibly under selection in at least one of the twenty parallel FBNSV lines passing from faba bean to alfalfa. Each viral population is represented with a specific color. All 20 populations are represented in all graphs; when less than 20 populations are visible, it is because several are superimposed. For all eight mutations, the frequency sometimes increases, decreases or do not change, depending on the population considered. Panel B shows the distances between the genome formula in faba bean and that in alfalfa (ΔGF) in the twenty parallel FBNSV lines plotted against the frequency of mutations (Δf) at the eight sites detected as possibly under selection. The *p*-values of the regressions are indicated in each case.