Predicting the stability of homologous gene duplications in a plant RNA virus

Anouk Willemsen[1], Mark P. Zwart[1,2], Pablo Higueras, Josep Sardanyés[3,4], Santiago F. Elena[1,5]

[1]Instituto de Biología Molecular y Celular de Plantas (IBMCP), Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Campus UPV CPI 8E, Ingeniero Fausto Elio s/n, 46022 València, Spain.

[2]Institute of Theoretical Physics, University of Cologne, Zülpicher Straße 77, 50937 Cologne, Germany.

[3]ICREA Complex Systems Laboratory, Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Spain.

[4]Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas-Universitat Pompeu Fabra), Passeig Maritim de la Barceloneta 37, 08003 Barcelona, Spain.

[5]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

*Authors for correspondence: A.W. (anwil1@ibmcp.upv.es) or S.F.E. (sfelena@ibmcp.upv.es)

Abstract

One of the striking features of many eukaryotes is the apparent amount of redundancy in coding and non-coding elements of their genomes. Despite the possible evolutionary advantages, there are fewer examples of redundant sequences in viral genomes, particularly those with RNA genomes. The low prevalence of gene duplication in RNA viruses most likely reflects the strong selective constraints against increasing genome size. Here we investigated the stability of genetically redundant sequences and how adaptive evolution proceeds to remove them. We generated plant RNA viruses with potentially beneficial gene duplications, measured their fitness and performed experimental evolution, hereby exploring their genomic stability and evolutionary potential. We found that all gene duplication events resulted in a loss of viability or significant reductions in fitness. Moreover, upon evolving the viable viruses and analyzing their genomes, we always observed the deletion of the duplicated gene copy and maintenance of the ancestral copy. Interestingly, there were clear differences in the deletion dynamics of the duplicated gene associated with the passage duration, the size of the gene and the position for duplication. Based on the experimental data, we developed a mathematical model to characterize the stability of genetically redundant sequences, and showed that the fitness of viruses with duplications is not enough information to predict genomic stability as a recombination rate dependent on the genetic context – the duplicated gene and its position – is also required. Our results therefore demonstrate experimentally the deleterious nature of gene duplications in RNA viruses, and we identify factors that constrain the maintenance of duplicated genes.

Key words: Gene duplication, Genome stability, Experimental evolution, Virus evolution

Introduction

Gene duplication results in genetic redundancy; in other words, the existence of genetic elements that encode for the same function. It is a powerful process that can regulate gene expression, increase the genetic and environmental robustness of organisms, and act as a stepping stone to the evolution of new biological functions. Therefore, it is not surprising that gene duplication is a frequent phenomenon in many organisms (Zhang 2003; Andersson & Hughes 2009).

There are few examples of genetic redundancy in viral genomes. In general, viral genomes tend to be highly streamlined, with limited intergenic sequences and in many cases overlapping open reading frames (ORFs), suggesting genome size is under strong selection (Lynch 2006). RNA viruses typically have smaller genomes than DNA viruses, and consequently there is an extreme low prevalence of gene duplication in RNA viruses (Belshaw et al. 2007; Belshaw et al. 2008; Simon-Loriere & Holmes 2013). For the reverse-transcribing viruses, three different gene duplication events have been reported within the *Retroviridae* family (LaPierre et al. 1999; Kambol et al. 2003; Tristem et al. 1990). This low prevalence of gene duplication in retroviruses is surprising, since repeated sequence elements of endogenous retroviruses are thought to mediate genomic rearrangements, including gene duplication (Hughes & Coffin 2001). For the ss(-)RNA viruses, two different tandem gene duplications have been reported (Walker et al. 1992; Blasdell et al. 2012; Gubala et al. 2010; Simon-Loriere & Holmes 2013) within the *Rhabdoviridae* (infecting vertebrates, invertebrates and plants). For the ss(+)RNA viruses, single duplication events have been reported for three different domains: (*i*) a tandem duplication of the coat protein gene (*CP*) within the *Closteroviridae* (infecting plants) (Boyko et al. 1992; Fazeli & Rezaian 2000; Tzanetakis et al. 2005; Tzanetakis & Martin 2007; Kreuze et al. 2002; Simon-Loriere &

Holmes 2013); (*ii*) a tandem duplication of the genome-linked protein gene (*VPg*) in *Foot-and-mouth disease virus* from the *Picornaviridae* (infecting vertebrates) (Forss & Schaller 1982); and (*iii*) a duplication of the third segment, generating an additional one, in *Beet necrotic yellow vein viru*s from the *Benyviridae* (Simon-Loriere & Holmes 2013). To date, no cases of gene duplication in dsRNA viruses have been reported.

The variation in genome size and structure indicates that gene duplication must have played a role in the early diversification of virus genomes. However, the rapid evolution of RNA viruses and the potential fitness costs associated with harboring additional genetic material probably makes it unlikely to detect viruses with duplications, or even the signatures of recent duplication events. Strong selective constraints against increasing genome sizes are thought to play a role in the lack of gene duplications that we nowadays observe in RNA viruses (Holmes 2003; Belshaw et al. 2007; Belshaw et al. 2008). One of these constraints is the high mutation rates of RNA viruses, which is approximately one mutation per genome and per replication event (Sanjuán et al. 2010). This limits the probability of copying without errors a genome above the length limit imposed by Eigen's error threshold (Eigen 1971): the inverse of the per site mutation rate. Another constraint is the need for fast replication due to strong within-cell and within-host competition (Turner & Chao 1998). An increase in genome size is therefore likely to increase the number of deleterious mutations that occur per genome during each round of replication, and to slow down the replication process. On the other hand, the small and streamlined RNA virus genomes also limit sequence space for the evolution of novel functions, and in turn adaptation to environmental changes.

Here, we therefore consider experimentally the evolutionary fate of gene duplications in viral genomes, in terms of their effects on fitness, the stability of the duplicated gene and the evolvability of these viruses. We experimentally explore four cases of homologous duplication of genes within the *Tobacco etch virus* (TEV; genus *Potyvirus,* family *Potyviridae*) genome (Revers & García 2015): (*i*) the multi-functional protein (HC-Pro) involved in aphid transmission, polyprotein cleavage, genome amplification, and suppression of RNA silencing, (*ii*) the main viral protease (NIa-Pro), (*iii*) the viral RNA-dependent RNA polymerase (NIb), and (*iv*) the coat protein (CP). Potyviruses are a particularly interesting system for studying the evolution of gene duplications, as they encode a single polyprotein that is auto catalytically processed into the mature gene products. For each complete (+)RNA, as well as frame-shifted transcripts for which translation terminates at P3-PIPO, there will be isostoichiometric expression of all genes. Assuming there are no unknown mechanisms that regulate gene expression, the scope for the regulation of gene expression in potyviruses could therefore be very limited. Gene duplication may represent a way to bypass these constraints and achieve higher expression of specific genes.

We speculated that the duplication of these four proteins might have widely different impacts on TEV fitness. As HC-Pro is a multifunctional protein, two copies of HC-Pro could lead to specific improvement of one or more of its functions. This potential improvement could possibly be caused by two mechanisms. Firstly, by simply producing more protein there could be an immediate benefit for one of HC-Pro's functions. Secondly, there could be improvement of protein function when the duplicated virus is evolved, because the two gene copies can diverge and specialize on different functions. Higher levels of NIa-Pro may result in a more efficient processing of the polyprotein, making more mature viral proteins available faster for the replication process. As potyviruses have only a limited number of post-translational mechanisms for regulating gene expression levels, we predicted that the overproduction of NIa-Pro will alter the equilibrium concentrations of all the different mature peptides and thus have a major impact in TEV fitness. Higher levels of NIb may result in higher levels of transcription and faster replication of the virus and this could lead to higher levels of genome accumulation and potentially the within-host spread of infection by a greater number of virions. The cellular multiplicity of infection (MOI), which has been estimated to be as low as 1.14 virions per infected cell for TEV (Tromas et al. 2014a), might even increase. Higher levels of CP expression could allow for the encapsidation of more genomic RNA molecules without affecting the accumulation of all other mature peptides. However, in all these cases completion of the infectious cycle would still depend on the cytoplasmic amount of other limiting viral (*e.g.,* P1, P3, CI, and VPg) or host proteins.

The duplication events that we explore here could therefore conceivably have beneficial effects on TEV replication, perhaps offsetting the costs inherent to a larger genome and thereby increasing overall fitness. Moreover, especially in the case of HC-Pro, they could perhaps lead to the evolution of greatly improved or novel functions. However, given the scarcity of gene duplications in RNA viruses, we expected that the fitness costs of duplication are likely to be high, and that one of the two gene copies would be rapidly lost. If further mutations could potentially help accommodate the duplicated gene, then this could lead to interesting evolutionary dynamics: will the duplicated gene be lost or will beneficial mutations that lead to stable maintenance of the gene occur first? (Zwart et al. 2014) Moreover, as they could potentially disrupt correct processing of the polyprotein, the possibility that some of the duplications would not be viable in the first place could also not be discounted (Majer et al. 2014). To address these issues we have constructed four viruses with gene duplications and tested their viability. We subsequently evolved these viruses and determined the stability of the duplicated gene, as well as looking for signals of accommodation of the duplicated gene. Finally, we built a mathematical model to estimate key parameters from the experimental data, such as the recombination rates responsible for the deletion of duplicated genes, and to explore the evolutionary dynamics and stability conditions of the system.

Materials and Methods

Viral constructs, virus stocks and plant infections

The TEV genome used to generate the virus constructs, was originally isolated from *Nicotiana tabacum* plants (Carrington et al. 1993). In this study five different variants of TEV were used containing single gene duplications. Two of these virus variants, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ were generated in a previous study (Willemsen et al. 2016). The other three variants were generated in this study: TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV-CP$_{10}$-CP$_{11}$.

TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV-CP$_{10}$-CP$_{11}$ were generated from cDNA clones constructed using plasmid pMTEVa, which consists of a TEV infectious cDNA (accession: DQ986288, including two silent mutations, G273A and A1119G) flanked by SP6 phage RNA promoter derived from pTEV7DA (GenBank: DQ986288). pMTEVa contains a minimal transcription cassette to ensure a high plasmid stability (Bedoya & Daròs 2010). The clones were constructed using standard molecular biology techniques, including PCR amplification of cDNAs with the high-fidelity Phusion DNA polymerase (Thermo Scientific), DNA digestion with *Eco*31I (Thermo Scientific) for assembly of DNA fragments (Engler et al. 2009), DNA ligation with T4 DNA ligase (Thermo Scientific) and transformation of *E. coli* DH5α by electroporation. Sanger sequencing confirmed the sequences of the resulting plasmids.

The plasmids of TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV-CP$_{10}$-CP$_{11}$ were linearized by digestion with *Bgl*II prior to *in vitro* RNA synthesis using the mMESSAGE mMACHINE® SP6 Transciption Kit (Ambion), as described in Carrasco et al. (2007). The third true leaf of 4-week-old *N. tabacum* L. cv Xanthi *NN* plants was mechanically inoculated with varying amounts (5 µg - 30 µg) of transcribed RNA. All symptomatic tissue was collected 7 dpi (days post inoculation) and stored at −80 ºC as stock tissue.

Serial passages

For the serial passage experiments, 500 mg homogenized stock tissue was ground into fine powder using liquid nitrogen and a mortar, and resuspended in 500 µl phosphate buffer (50 mM KH_2PO_4, pH 7.0, 3% polyethylene glycol 6000). From this mixture, 20 µl were then mechanically inoculated on the third true leaf of 4-week old *N. tabacum* plants. At least five independent replicates were performed for each virus variant. At the end of the designated passage duration (3 or 9 weeks) all leaves above the inoculated leaf were collected and stored at −80 ºC. For subsequent passages the frozen tissue was homogenized and a sample of the homogenized tissue was ground and resuspended with an equal amount of phosphate buffer (Zwart et al. 2014). Then, new *N. tabacum* plants were mechanically inoculated as described above. The plants were kept in a BSL-2 greenhouse at 24 ºC with 16 h light.

Reverse transcription polymerase chain reaction (RT-PCR)

The wild-type TEV produces characteristic symptoms in the host plant. However, some of the altered genotypes show few or no symptoms and virus infection had to be confirmed by RT-PCR. To confirm infection and to determine the stability of the duplicated genes, RNA was extracted from 100 mg homogenized infected tissue using the InviTrap Spin Plant RNA

Mini Kit (Stratec Molecular). Reverse transcription (RT) was performed using M-MuLV reverse transcriptase (Thermo Scientific) and the reverse primer 5'-CGCACTACATAGGAGAATTAG-3' located in the 3'UTR of the TEV genome. PCR was then performed with *Taq* DNA polymerase (Roche) and primers flanking the region containing the duplicated gene copy (supplementary Table S1, Supplementary Material online). To test whether the ancestral gene copy was intact this region was also amplified for TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses, where the duplicated genes are not located in tandem (supplementary Table S1, Supplementary Material online). PCR products were resolved by electrophoresis on 1% agarose gels. For those virus populations in which we detected deletions during the evolution experiment, we estimated the genome size based on the amplicon size and the genome size of the ancestral viruses.

Fitness assays

The genome equivalents per 100 mg of tissue of the ancestral virus stocks and all evolved lineages were determined for subsequent fitness assays. The InviTrap Spin Plant RNA Mini Kit (Stratec Molecular) was used to isolate total RNA from 100 mg homogenized infected tissue. Real-time quantitative RT-PCR (RT-qPCR) was performed using the One Step SYBR PrimeScript RT-PCR Kit II (Takara), in accordance with manufacturer instructions, in a StepOnePlus Real-Time PCR System (Applied Biosystems). Specific primers for the *CP* gene were used; forward 5'-TTGGTCTTGATGGCAACGTG-3' and reverse 5'-TGTGCCGTTCAGTGTCTTCCT-3'. The StepOne Software v.2.2.2 (Applied Biosystems) was used to analyze the data. The concentration of genome equivalents per 100 mg of tissue was then normalized to that of the sample with the lowest concentration, using phosphate buffer.

For the accumulation assays, 4-week-old *N. tabacum* plants were inoculated with 50 µl of the normalized dilutions of ground tissue. For each ancestral and evolved lineage, at least three independent plant replicates were used. Leaf tissue was harvested 7 dpi. Total RNA was extracted from 100 mg of homogenized tissue. Virus accumulation was then determined by means of RT-qPCR for the *CP* gene of the ancestral and the evolved lineages. For each of the harvested plants, at least three technical replicates were used in the RT-qPCR.

To measure within-host competitive fitness, we used TEV carrying an enhanced green fluorescent protein (TEV-eGFP) (Bedoya & Daròs 2010) as a common competitor. TEV-eGFP has proven to be stable up to six weeks (using 1- and 3-week serial passages) in *N. tabacum* (Zwart et al. 2014), and is therefore not subjected to appreciable eGFP loss during our 1-week long competition experiments. All ancestral and evolved viral lineages were again normalized to the sample with the lowest concentration, and 1:1 mixtures of viral genome equivalents were made with TEV-eGFP. The mixture was mechanically inoculated on the same species of host plant on which it had been evolved, using three independent plant replicates per viral lineage. The plant leaves were collected at 7 dpi, and stored at −80 ºC. Total RNA was extracted from 100 mg homogenized tissue. RT-qPCR for the *CP* gene was used to determine total viral accumulation, and independent RT-qPCR reactions were also performed for the *eGFP* sequence using primers forward 5’-CGACAACCACTACCTGAGCA-3’ and reverse 5’-GAACTCCAGCAGGACCATGT-3’. The ratio (R) of the evolved and ancestral lineages to TEV-eGFP is then $R = (n_{CP} - n_{eGFP})/n_{eGFP}$, where n_{CP} and n_{eGFP} are the RT-qPCR measured copy numbers of *CP* and *eGFP*, respectively. Within-host competitive fitness can then be estimated as $W = \sqrt[t]{R_t/R_0}$, where R_0 is the ratio at the start of the experiment and R_t the ratio after t days of competition (Carrasco et al. 2007). Note that the method for determining R only works well when the frequency of the common is below ~0.75. This limitation was not problematic though, since in these experiments the fitness of the evolved virus populations remained the same or increased. The statistical analyses comparing the fitness between lineages were performed using R v.3.2.2 (R Core Team 2014) and IBM SPSS Statistics version 23.

Sanger sequencing

For those evolved virus populations in which deletions were detected by RT-PCR, the exact positions of these deletions were determined. The genomes were partly sequenced by the Sanger method. RT was performed using AccuScript Hi-Fi (Agilent Technologies) reverse transcriptase and a reverse primer outside the region to be PCR-amplified for sequencing (supplementary Table S2, Supplementary Material online). PCR was then performed with Phusion DNA polymerase (Thermo Scientific) and primers flanking the deletions (supplementary Table S1, Supplementary Material online). Sanger sequencing was performed at GenoScreen (Lille, France: www.genoscreen.com; last accessed April 10, 2016) with an ABI3730XL DNA analyzer. For TEV-HCPro$_2$-HCPro$_3$, six sequencing reactions were done per lineage using the same outer reverse primer as used for PCR amplification plus five inner primers (supplementary Table S2, Supplementary Material online). For TEV-NIaPro$_2$-NIaPro$_8$, three sequencing reactions were done per lineage using three inner primers (supplementary Table S2, Supplementary Material online). For TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$, six sequencing reactions were done per lineage using the same two outer primers as used for PCR amplification plus four inner primers (supplementary Table S2, Supplementary Material online). For TEV-CP$_{10}$-CP$_{11}$, two sequencing reactions were done per lineage using the same two outer primers as used for PCR amplification (supplementary Table S2, Supplementary Material online). Sequences were assembled using Geneious v.8.0.3 (www.geneious.com; last accessed April 10, 2016) and the start and end positions of the deletions were determined. Based on the ancestral reference sequences, new reference sequences were constructed containing the majority deletion variant for each of the evolved lineages.

Illumina sequencing, variants, and SNP calling

For Illumina next-generation sequencing (NGS) of the evolved and ancestral lineages, the viral genomes were amplified by RT-PCR using AccuScript Hi-Fi (Agilent Technologies) reverse transcriptase and Phusion DNA polymerase (Thermo Scientific), with six independent replicates that were pooled. Each virus was amplified using three primer sets generating three amplicons of similar size (set 1: 5’-GCAATCAAGCATTCTACTTCTATTGCAGC-3’ and 5’-TATGGAAGTCCTGTGGATTTTCCAGATCC-3’; set 2: 5’-TTGACGCTGAGCGGAGTGATGG-3’ and 5’-AATGCTTCCAGAATATGCC-3’; set 3: 5’-TCATTACAAACAAGCACTTG-3’ and 5’-CGCACTACATAGGAGAATTAG-3’). Equimolar mixtures of the three PCR products were made. Sequencing was performed at GenoScreen. Illumina HiSeq2500 2×100bp paired-end libraries with dual-index adaptors were prepared along with an internal PhiX control. Libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina Inc.). Sequencing quality control was performed by GenoScreen, based on PhiX error rate and Q30 values.

Read artifact filtering and quality trimming (3’ minimum Q28 and minimum read length of 50 bp) was done using FASTX-Toolkit v.0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/index.html, last accessed April 10, 2016). De-replication of the reads and 5’ quality trimming requiring a minimum of Q28 was done using PRINSEQ-lite v.0.20.4 (Schmieder & Edwards 2011). Reads containing undefined nucleotides (N) were discarded. As an initial mapping step, the evolved sequences were mapped using Bowtie v.2.2.6 (Langmead & Salzberg 2012) against their corresponding ancestral sequence: TEV (GenBank accession number KX137149), TEV-HCPro$_2$-HCPro$_3$ ancestral (GenBank accession number KX137150), TEV-NIaPro$_2$-NIaPro$_8$ ancestral (GenBank accession number KX137151), TEV-NIb$_1$-NIb$_9$ ancestral (GenBank accession number KT203712), TEV-NIb$_2$-NIb$_9$ ancestral (GenBank accession number KT203713), and against the evolved lineages including the corresponding deletions in the lineages where they are present. Subsequently, mutations were detected using SAMtools’ mpileup (Li et al. 2009) in the evolved lineages as compared to their ancestral lineage. At this point, we were only interested in mutations at a frequency > 10%. Therefore, we present frequencies as reported by SAMtools, which has a low sensitivity for detecting low-frequency variants (Spencer et al. 2014).

After the initial pre-mapping step, error correction was done using Polisher v2.0.8 (available for academic use from the Joint Genome Institute) and consensus sequences were defined for every lineage. Subsequently, the cleaned reads were remapped using Bowtie v.2.2.6 against the corresponding consensus sequence for every lineage. For each new consensus, Single nucleotide polymorphisms (SNPs) within each virus population were identified using SAMtools’ mpileup and VarScan v.2.3.9 (Koboldt et al. 2012). For SNP calling maximum coverage was set to 40000 and SNPs with a frequency < 1% were discarded.

Modeling the stability of gene insertions

We developed a mathematical model to fit with the experimental data for the 3-week and 9-week passages. We were particularly interested in better understanding the frequency of virus populations in which viruses with deletions in the duplicated gene were present or had been fixed. The model consists of two coupled ordinary differential equations:

$$\frac{dA}{dt} = aA\left(1 - \frac{A+\beta B}{\kappa}\right) - \delta A \qquad (1)$$

$$\frac{dB}{dt} = bB\left(1 - \frac{\alpha A+B}{\kappa}\right) + \delta A \qquad (2)$$

where A is the number of virions containing a gene duplication, B is the number of virions with a reversion to a single copy, a is the initial growth rate of A, b is the initial growth rate of B, $\beta = b/a$ is a constant for determining the effect of the presence of B on replication of A (Solé et al. 1998), $\alpha = 1/\beta$ is a constant for determining the opposing effect of A on B, κ is the time-dependent carrying capacity of the host plant, and δ is the recombination rate per genome and replication at which the extra copy of the gene is removed from A to produce B. We assume that κ increases linearly over time, being proportional to the estimated weight of collected plant tissue (2 g for the whole plant at inoculation, 200 g for the collected leaves at 9 weeks). At the start of each round of infection, there is a fixed bottleneck size of λ. The number of infecting virions of A is determined by a random draw from a Binomial distribution with a probability of success $\lambda A/(A + B)$ and a size λ, and the number of infecting virions of B is then λ minus this realization from the Binomial distribution.

Estimates of most model parameters could be obtained from previous studies (Table 1). An estimate of b has been made (Zwart et al. 2012), whilst a can be determined knowing the competitive fitness of the virus with duplication relative to the wild-type virus. The value of b used is 1.344, and a values are 1.175, 1.234, 1.185 and 1.134 for TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$, respectively. The only model parameter that needed to be estimated from the data is δ. To obtain an estimate of this parameter, we implemented the model as described in equations 1 and 2 in R 3.1.0. For each dataset to which we wanted to fit the model, we first simply ran the model for a wide range of recombination rates: considering all values of $\log(\delta)$ between −20 and −0.1, with intervals of 0.1. One-thousand simulations were run for each parameter value.

To fit the model to the data, we considered model predictions of the frequency of three kinds of virus populations over time: (*i*) those populations containing only the full-length ancestral virus with a gene duplication (X_1), (*ii*) those populations containing only variants with a genomic deletion removing the artificially introduced second gene copy (X_2), and (*iii*) those populations containing a mixture of both variants (X_3). Due to the modeling of recombination and selection as deterministic processes, the model predicts recombinants will be ubiquitous. However, in order to be reach appreciable frequencies and eventually be fixed, recombinants must reach a high enough frequency that they will be sampled during the genetic bottleneck at the start of infection. Moreover, we used a PCR-based method with inherent limits to its sensitivity to characterize experimental populations. For these two reasons, we assumed that the predicted frequency of A must be greater than 0.1 and less than 0.9 to be considered a mixture. We then compared model predictions for the frequency of the three different kinds of virus populations with the data by means of multinomial likelihoods. The likelihood of the number of occurrences of these three stochastic variables denoting observations of a particular kind of population (X_1, X_2, X_3) follows a Multinomial distribution with probabilities p_1, p_2 and p_3 ($\sum_{i=1}^{3} p_i = 1$). The multinomial probability of a particular realization (x_1, x_2, x_3) is given by:

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{(\sum_{i=1}^{3} x_i)!}{\prod_{i=1}^{3} x_i!} \prod_{i=1}^{3} p_i^{x_i}.$$

The estimate of δ is then simply that value that corresponds to the lowest negative log-likelihood (NLL), for the entire range of δ values tested. We first fitted the model with a single value of δ to all the data (Model 1; 1 parameter). Next, we fitted the model with a virus-dependent value of δ, but one which is independent of passage duration (Model 2; 4 parameters). We then fitted the model with δ value dependent on passage duration, but the same for each virus (Model 3; 2 parameters). Finally, we fit the model to each experimental treatment separately (Model 4; 8 parameters). For all these different model fittings, 95% fiducial estimates of δ were obtained by fitting the model to 1000 bootstrapped datasets.

The numerical solutions of the differential equations 1 and 2 used to characterize the dynamical properties, build the phase portraits and to obtain the transient times (supplementary Text S1, Supplementary Material online) have been obtained using a fourth-order Runge-Kutta method with a time step size 0.1.

Results

Genetic redundant constructs and the viability of the resulting viruses

To simulate the occurrence of duplication events within the TEV genome (Figure 1A), different TEV genotypes were constructed using four genes of interest (Figure 1). Each of these genotypes therefore represents a single gene duplication event. Where necessary, the termini of the duplicated gene copies were adjusted, such that the proteins can be properly translated and processed. Cleavage sites are provided, similar to the original proteolytic cleavage sites at the corresponding positions. A description of every duplication event will be given in the same order as these genes occur within the TEV genome.

First, for duplication of the multifunctional *HC-Pro* gene, a second copy of *HC-Pro* was inserted in the second position within the TEV genome, between the *P1* serine protease gene and the original *HC-Pro* copy, generating a tandem duplication (Figure 1B). This position is a common site for the cloning of heterologous genes (Zwart et al. 2011). Second, a copy of the *NIa-Pro* main viral protease gene was introduced between *P1* and *HC-Pro* (Figure 1C). Third, two genotypes containing a duplication of the *NIb* replicase gene were generated (Willemsen et al. 2016), where a copy of the *NIb* gene was inserted at the first position (before *P1*) and the second position in the TEV genome (Figure 1D). Fourth, for duplication of the *CP* we introduced a second copy at the tenth position between *NIb* and *CP,* generating a tandem duplication (Figure 1E). Henceforth we refer to these five genetic redundant viruses as TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, TEV-NIb$_2$-NIb$_9$, and TEV-CP$_{10}$-CP$_{11}$, respectively, with subscripts denoting the intergenic positions of the duplicated gene in question.

The viability of these viruses was tested in *N. tabacum* plants, by inoculating plants with approximately 5 µg *in vitro* generated transcripts. TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$, were found to infect *N. tabacum* plants, as determined by RT-PCR on total RNA extracted from these plants. After performing multiple viability tests, TEV-CP$_{10}$-CP$_{11}$ demonstrated to have a very low infectivity and high amounts (> 20 µg) of RNA are needed for infection to occur. Performing RT-PCR of the region containing two *CP* copies, we detected either (*i*) a band corresponding to the wild-type virus, indicating that upon infection with RNA the second *CP* copy is deleted immediately, or (*ii*) a band that indicates the two *CP* copies are present. Taking the latter as a starting population for experimental evolution, within the first passage, we detect a band corresponding to the wild-type virus, in six out of eight lineages, and we did not detect any infection in the remaining two lineages. When sequencing the region containing the deletions in the different lineages, using Sanger technology, exact deletions of the second *CP* copy were observed. We discontinued further experiments on this virus due to the extreme instability of the second *CP* copy.

Evolution of genetically redundant viruses

After reconstitution of TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$ from infectious clones, these viruses containing gene duplications were evolved in *N. tabacum* plants. All viruses were evolved for a total of 27 weeks, using nine 3-week passages and three 9-week passages with at least five independent lineages for each passage duration. In the starting population of TEV-HCPro$_2$-HCPro$_3$ we observed mild symptoms. However, in lineages from the first 3- and 9-week passages, the plants rapidly became as symptomatic as those infected by the wild-type virus. At the start of the evolution experiment, TEV-NIaPro$_2$-NIaPro$_8$ also displayed only mild symptoms and infection appeared to expand slower than for the wild-type TEV. However, in the first 9-week passage symptoms became stronger, similar to the wild-type virus, as the virus expanded through the plant. These stronger symptoms were also observed in the subsequent 9-week passages. Increases in symptom severity were also observed for the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses (Willemsen et al. 2016).

Partial and complete deletions of the duplicated gene copy were detected by RT-PCR (Figure 2), but never in the ancestral gene. Deletion of the duplicated gene copy of the TEV-HCPro$_2$-HCPro$_3$ variant occurred rapidly after infection of the plants; after one passage the gene duplication could no longer be detected by RT-PCR (Figure 2A). No deletions were detected in the TEV-NIaPro$_2$-NIaPro$_8$ lineages using the shorter 3-week passages (Figure 2B). Deletions were not detected in the first 9-week passage either, but in the second passage partial or complete deletions did occur. Mixed populations that contain virions with a deletion together with virions that have maintained the intact duplicated copy, are mainly present in the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages (Figures 2C and 2D). However, TEV-NIb$_1$-NIb$_9$ loses the duplicated copy faster (Figure 2C; second 3-week passage, and first 9-week passage) than TEV-NIb$_2$-NIb$_9$ (Figure 2D; third 3-week passage, and second 9-week passage). Based on the majority deletion variants observed by RT-PCR, genome size was estimated for every passage (Figure 3). Comparing the genome size of the different viral genotypes in Figure 3, there are clear differences in the time until the duplicated gene copy is deleted. The duplicated *HC-Pro* copy appears the least stable, while the duplicated *NIa-Pro* copy appears to be the most stable. For TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$, there are lineages that contain deletions that lead to a genome size smaller than that of the wild-type TEV.

Viruses with a gene duplication have reduced fitness which cannot always be fully restored after deletion

For both the ancestral and evolved virus populations, we measured within-host competitive fitness (Figure 4) and viral accumulation (Figure 5). Comparing the ancestral viruses containing a gene duplication to the ancestral wild-type virus (solid circles in Figures 4 and 5), we observed statistically significant decreases in competitive fitness (Figure 4A; TEV-HCPro$_2$-HCPro$_3$: $t_4 = 8.398$, $P = 0.001$. Figure 4B; TEV-NIaPro$_2$-NIaPro$_8$: $t_4 = 12.776$, $P < 0.001$. Figure 4C; TEV-NIb$_1$-NIb$_9$: $t_4 = 6.379$, $P = 0.003$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.348$, $P = 0.001$). Statistically significant decreases in accumulation levels were also observed for

TEV-HCPro$_2$-HCPro$_3$, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ (Figure 5A; TEV-HCPro$_2$-HCPro$_3$: $t_4 = 3.491$, $P = 0.0251$. Figure 5C; TEV-NIb$_1$-NIb$_9$: $t_4 = 45.097$, $P < 0.001$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.650$, $P < 0.001$). However, there was no difference in accumulation for the virus with a duplication of *NIa-Pro* (Figure 5B; TEV-NIaPro$_2$-NIaPro$_8$: $t_4 = 2.099$, $P = 0.104$). All the virus with duplications therefore have a reduced within-host competitive fitness, and three out of four viruses also have reduced accumulation. None the possible benefits of these gene duplications therefore can compensate for their costs.

After evolving the viruses with gene duplications using three 9-week passages, within-host competitive fitness of all four viruses increased (open circles in Figure 4), compared to their respective ancestral viruses (Figure 4; asterisks indicate a significant increase, *t*-test with Holm-Bonferroni correction). Within-host fitness was similar to the evolved wild-type TEV for evolved lineages of both TEV-HCPro$_2$-HCPro$_3$ (Figure 4A; Mann-Whitney $U = 23$, $P = 0.432$) and TEV-NIaPro$_2$-NIaPro$_8$ (Figure 4B; Mann-Whitney $U = 20$, $P = 0.151$). On the other hand, the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages did not reach wild-type virus within-host fitness levels (Figure 4C; TEV-NIb$_1$-NIb$_9$: Mann-Whitney $U = 0$, $P = 0.008$; TEV-NIb$_2$-NIb$_9$: Mann-Whitney $U = 0$, $P = 0.008$). The within-host fitness of evolved lineages was also compared by means of a nested ANOVA (Table 2), allowing the independent lineages (at least 5) to be nested within the genotype and the independent plant replicates (3) to be nested within the independent lineages within the genotype. The nested ANOVA confirms that there is indeed an effect of the genotype for the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses (Table 2 and Willemsen et al. 2016), while for the TEV-HCPro$_2$-HCPro$_3$ and TEV-NIaPro$_2$-NIaPro$_8$ no effect was found. In summary, the fitness of TEV-HCPro$_2$-HCPro$_3$ and TEV-NIaPro$_2$-NIaPro$_8$ clearly increases to levels similar to the wild-type, whilst fitness did not increase for TEV-NIb$_2$-NIb$_9$ and TEV-NIb$_1$-NIb$_9$.

Together with within-host fitness, virus accumulation also increased significantly for the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ virus lineages (Figure 5C; asterisks), when compared to their respective ancestral viruses. However, accumulation levels did not increase significantly for most of the evolved lineages of the TEV-HCPro$_2$-HCPro$_3$ and TEV-NIaPro$_2$-NIaPro$_8$ genotypes. Nevertheless, these two genotypes have much higher initial accumulation levels than the genotypes with a duplication of the *NIb* gene. When comparing the accumulation levels of the evolved lineages to those of the wild-type (again, using lineage as the replication unit), TEV-HCPro$_2$-HCPro$_3$ (Figure 5A; Mann-Whitney $U = 20$, $P = 0.755$), TEV-NIaPro$_2$-NIaPro$_8$ (Figure 5B; Mann-Whitney $U = 11$, $P = 0.841$), and TEV-NIb$_2$-NIb$_9$ (Figure 5C; Mann-Whitney $U = 3$, $P = 0.056$) do reach wild-type accumulation levels, whilst TEV-NIb$_1$-NIb$_9$ does not (Figure 5C; Mann-Whitney $U = 0$, $P = 0.008$). Comparing the accumulation levels of the evolved lineages by means of a nested ANOVA (Table 2) confirms that there is an effect of the genotype for the TEV-NIb$_1$-NIb$_9$ virus (Table 2 and Willemsen et al. 2016), while no effect for the TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV-NIb$_2$-NIb$_9$ viruses was found.

When comparing the within-host competitive fitness of the evolved TEV-NIaPro$_2$-NIaPro$_8$ 9-week lineages to the 3-week lineages, we found that there is a linear relationship between genome size and within-host competitive fitness (Figure 6; Spearman's rank correlation $\rho = -0.795$, 10 d.f., $P = 0.006$). The evolved 9-week lineages, that contain genomic deletions, have a significant higher within-host competitive fitness (Mann-Whitney $U = 4.5$, $P < 0.001$) than the evolved 3-week lineages without deletions.

Genome sequences of the evolved lineages

All evolved and ancestral lineages have been fully sequenced using the Illumina technology. The sequences of the ancestral lineages were used as an initial reference for the evolved lineages. Furthermore, for the lineages where deletions were detected by RT-PCR (Figure 3), parts of the genome were sequenced by Sanger to determine the exact deletion sites. The majority deletions variants were used to construct new reference sequences for each of the evolved TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$ lineages that contain deletions. After an initial mapping step, mutations were detected in the evolved lineages as compared to their corresponding ancestor (Materials and Methods).

Beside the large genomic deletions, different patterns of adaptive evolution were observed for each viral genotype (Figure 7 and Table 3). For the evolved TEV-HCPro$_2$-HCPro$_3$ virus a convergent nonsynonymous mutation was found in 3/7 9-week lineages in the *P1* gene (A304G), however, this mutation was also present in 1/5 9-week lineages of TEV. Another convergent nonsynonymous mutation was found in 3/7 9-week lineages in the *P3* gene (U4444C), known to be implicated in virus amplification and host adaptation (Revers & García 2015). For the evolved TEV-NIaPro$_2$-NIaPro$_8$ virus, fixed convergent nonsynonymous mutations were found in the duplicated *NIa-Pro* (C1466U) copy in 4/5 3-week lineages, and in *6K1* (A4357G) in 3/5 3-week lineages. The latter mutation was also fixed in 1/5 3-week TEV lineages. For the evolved TEV-NIb$_1$-NIb$_9$ virus a fixed convergent nonsynonymous mutation was found in the pseudogenized *NIb* copy (A1643U) in 2/5 3-week lineages. For the evolved TEV-NIb$_2$-NIb$_9$ virus one fixed synonymous mutation was found in the multifunctional CI protein (C6531U) in 2/5 9-week lineages. Other convergent mutations in all virus genotypes were found in *VPg*, *NIa-Pro* and *NIb* genes, however these mutations were also found in 2 or more lineages of the wild-type virus (Figure 7). Therefore, we do not consider these genotype-specific mutations as adaptive.

After remapping the cleaned reads against a new defined consensus sequence for each lineage, we looked at the variation within each lineage. Single nucleotide polymorphisms (SNPs) were detected at a frequency as low as 1%. In the evolved TEV-HCPro$_2$-HCPro$_3$ lineages, a total of 633 SNPs were detected, with a median of 45 (range 27 - 247) per lineage. There is no clear difference in the number of SNPs comparing the evolved 3-week lineages, median 48.50 (35 - 247), and the evolved 9-week lineages, median 44 (27 - 56). However, there is one 3-week lineage (3WL6) that accumulated a much higher number of SNPs (247) compared to the other TEV-HCPro$_2$-HCPro$_3$ lineages. In the evolved TEV-NIaPro$_2$-NIaPro$_8$ lineages, a total of 421 SNPs were detected, with a median of 43.50 (11 - 103) per lineage. The evolved 3-week lineages that have not lost the duplicated *NIa-Pro* copy accumulated more SNPs per lineage, median 53 (40 - 59), compared to the evolved 9-week lineages, median 33 (11 - 103). However, one lineage of the 9-week lineages (9WL2) also accumulated a much higher number of SNPs (103) compared to the other TEV-NIaPro$_2$-NIaPro$_8$ lineages. In the evolved wild-type TEV lineages, a total number of 402 SNPs were detected, with a median of 35 (17 - 63) SNPs per lineage. Like for TEV-HCPro$_2$-HCPro$_3$, for TEV there is no clear difference in the number of SNPs comparing the evolved 3-week lineages, median 34 (32 - 50), and the evolved 9-week lineages, median 36 (17 - 63). The data for the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages can be found in Willemsen et al. (2016), where a total 301 and 220 SNPs were detected with medians of 36 (27 - 45) and 23.5 (4 - 44) per lineage, respectively. In all four virus genotypes as well as in the wild-type virus, most of the SNPs were present at low frequency (SNPs < 0.1: TEV-HCPro$_2$-HCPro$_3$ = 89.8%; TEV-NIaPro$_2$-NIaPro$_8$ = 84.8%; TEV-NIb$_1$-NIb$_9$ = 83.3%; TEV-NIb$_2$-NIb$_9$ = 81.2%; TEV = 85.2%), with a higher prevalence of synonymous (TEV-HCPro$_2$-HCPro$_3$: 54.7%, TEV-NIaPro$_2$-NIaPro$_8$: 57.7%, TEV-NIb$_1$-NIb$_9$: 66.4%, TEV-NIb$_2$-NIb$_9$: 64.5%, TEV:

59.7%) versus nonsynonymous changes (supplementary Figure S1, Supplementary Material online). Moreover, a percentage of the nonsynonymous changes for TEV-HCPro$_2$-HCPro$_3$ (16.7%) and TEV-NIaPro$_2$-NIaPro$_8$ (7.3%) as well as the wild-type TEV (14.8%), are actually leading to stop codons and therefore unviable virus variants. For both TEV-HCPro$_2$-HCPro$_3$ and TEV-NIaPro$_2$-NIaPro$_8$ the difference in the distribution of synonymous versus nonsynonymous SNP frequency is significant (Kolmogorov-Smirnov test; TEV-HCPro$_2$-HCPro$_3$: D = 0.219, P < 0.001; TEV-NIaPro$_2$-NIaPro$_8$: D = 0.151, P = 0.009), whilst for TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ (Willemsen et al. 2016) and the wild-type virus this is not significant (Kolmogorov-Smirnov test; TEV: D = 0.084, P = 0.555). For more details on the frequency of the SNPs within every lineage, see supplementary Tables S3, S4 and S5 (Supplementary Material online).

Genomic stability of TEV with duplications of homologous genes

To better understand the evolutionary dynamics of viruses with gene duplications, we developed a simple mathematical model of virus competition and evolution. Based on amplicon sizes, the genome size for all evolved lineages was estimated for every passage (Figure 3). Our model attempts to account for these data, and specifically how long the duplicate gene copy is maintained in the virus population. The model we developed describes how a population composed initially of only virus variants with a gene duplication (variant A) through recombination, selection and genetic drift acquires and eventually fixes a new variant that only retains the original copy of the duplicated gene (variant B). The model includes a genetic bottleneck at the start of each round of passaging (*i.e.,* the initiation of infection in the inoculated leaf), with a fixed total number of founders and binomially distributed number of founders for variants containing the gene duplication. Following this genetic bottleneck, there is deterministic growth of both variants as well as deterministic recombination of *A* into *B*. The main question we addressed is whether knowing the fitness of duplicated viruses (*i.e., a*) is sufficient information to predict the stability of the inserted gene. Or do the data support a context-dependent recombination rate, with the context being (*i*) identity and position of the duplication, (*ii*) passage length, or (*iii*) both? We considered four different situations that are represented in the following models:

Model 1: one recombination rate for all conditions (1 parameter);

Model 2: virus-genotype-dependent recombination rate (4 parameters);

Model 3: passage-duration-dependent recombination rates (2 parameters);

Model 4: virus-genotype- and passage-duration-dependent recombination rates (full model, 8 parameters).

The model estimates of δ are given in Table 4. Note that the parameter is often a minimum (when the virus is very unstable) or a maximum value (when the virus is very stable). If the optimum is represented by more than one parameter value, the mean of these values is given. When comparing these models, we found that Model 2 is the best-supported model (Table 5). Thus, only a genotype-dependent recombination rate is required to account for the data. The results strongly suggest that knowing the fitness of a virus with a gene duplication is not sufficient information for predicting genomic stability. Rather, as the recombination rate is dependent on the genetic context, the supply of recombinants which have lost the duplicated gene will vary greatly from one genotype to another. High recombinogenic sites will remove the second copy fast, while low recombinogenic sites will preserve the copy for longer periods of time, after which it will be unavoidably removed, as confirmed by the numerical analysis of equations 1 and 2 (supplementary Text S1, Supplementary Material online).

On the other hand, passage duration has a strong effect on the observed stability of gene duplications. However, model selection shows that this phenomenon can be sufficiently explained by considering the combined effects of selection and genetic drift, and without invoking passage-duration-dependent recombination rates. Given that recombination and selection are deterministic in the model, deletion variants will always arise during infection. However, depending on the rates of recombination and selection, these deletion variants may not reach a high enough frequency to ensure they are sampled during the genetic bottleneck at the start of each round of infection. This effect will be much stronger in the 3-week passages – since there will be fewer recombinants and less time for selection to increase their frequency – explaining why for some viruses there is such a marked difference in the observed genomic stability for different passage lengths.

The deterministic dynamics of the evolutionary model of stability of genomes containing gene duplications have also been investigated (supplementary Text S1, Supplementary Material online). These analyses were performed on the model as described in equations 1 and 2, albeit with a time-independent carrying capacity. The stability of three fixed points was analyzed: (*i*) the extinction of both *A* and *B*, (*ii*) the domination of *B* over the population, and (iii) the coexistence of *A* and *B* in the population. The fixed points analysis indicates that when $a < b$ and $\delta > a - b$, the *B* virus subpopulation will outcompete the *A* subpopulation. This parametric combination is the one obtained using the biologically meaningful parameter values shown in Table 1. This means that, under the model assumptions, the population of virions containing a gene duplication is unstable, and the population will be asymptotically dominated by virions containing a single gene copy. Notice that, the rate of recombination (δ) is also involved in this process of outcompetition (see the bifurcation values calculated in supplementary Text S1, Supplementary Material online). Specifically, if $\delta > a - b$, coexistence of the two virion types is not possible. Interestingly, the model also reveals the existence of a transcritical bifurcation separating the scenario between coexistence of the two types of virions and dominance of *B*. Such a bifurcation can be achieved by tuning δ as well as by unbalancing the fitness of *A* and *B* virus types. This bifurcation gives place to a smooth transition between these two possible evolutionary asymptotic states (*i.e.*, unstable *A* population and coexistence of *A* and *B* populations). The bifurcation will take place when $a = b + \delta$, or when δ is above a critical threshold, $\delta_c = a - b$, that can be calculated from the mathematical model (supplementary Text S1, Supplementary Material online).

Finally, we characterized the time to extinction of the *A* subpopulation. We characterized these extinction times as a function of viral fitness and of δ. Such a time is found to increase super-exponentially as the fitness of the *A* subpopulation, approaches the fitness of the *B* subpopulation. This effect is found for low values of recombination rates, similar to those shown in Table 4. As expected, increases in δ produce a drastic decrease in the time needed for the single-copy population to outcompete the population of viruses with the duplicated genes (supplementary Text S1, Supplementary Material online).

Discussion

Genetic redundancy is thought to be evolutionary unstable in viruses due to the costs associated with maintaining multiple gene copies. Here we tested the stability of duplicated sequences that might contribute to the enhancement of a virus function or even exploration of new functions. Overall, none of the duplication events explored appeared to be beneficial for TEV, both in terms of their immediate effects and second-order effects on evolvability. Gene duplication resulted in either an inviable virus or a significant reduction in viral fitness. In all cases the duplicated gene copy, rather than the ancestral gene copy, was deleted during long-duration passages. The earlier detection and more rapid fixation of deletion variants during longer-duration passages is congruent with results from a previous study (Zwart et al. 2014), where deletions of *eGFP* marker inserted in the TEV genome were usually observed after a single 9-week passage, but were rarely spotted even after nine 3-week passages. On the other hand, the highly diverging results for genome stability obtained here for the different viruses, suggests that passage duration is not the main factor determining whether gene duplication will be stable. Therefore, the size of the duplicated gene, the nature of the gene, and/or the position for duplication could play a role in the stability of genomes with a duplication.

The observation that gene duplication events result in decreases in fitness is not unexpected. The high mutation rate of RNA viruses is likely to constrain genome size (Holmes 2003), given that most mutations are deleterious (Sanjuán et al. 2004; Elena et al. 2006). This imposes the evolution of genome compression, where overlapping reading frames play a major role (Belshaw et al. 2007; Belshaw et al. 2008). In our model virus we speculate that the fitness cost can be related to three processes: (*i*) the increase in genome size, (*ii*) the extra cost of more proteins being expressed in the context of using more cellular resources, and (*iii*) a disturbance in correct polyprotein processing. Although all the duplications considered here could conceivably have advantages for viral replication or encapsidation, our results suggest that the any such advantages are far outweighed by the costs associated with a larger genome, increased protein expression or the effects on polyprotein processing. However, the duplication of *NIa-Pro* does not affect the viral accumulation rate. This could be explained by the fact that the *NIa-Pro* gene is much smaller than the other duplicated genes. Consequently, in conditions where selection has the least time to act between bottleneck events associated with infection of a new host (3-week passages), no deletions were observed. In the long-passage experiment selection has more time to act and increase the frequency of beneficial *de novo* variants, allowing them to be sampled during the bottleneck at the start of the next round of infection. In addition, the size of the gene duplication also seems to play a role. But what about the position and the nature of the duplicated gene? When duplicating the same gene, *NIb*, to either the first or second position in the TEV genome, we see clear differences in the deletion dynamics and fitness measurements (Figures 3, 4, 5 and Willemsen et al. 2016). Comparing the duplication and subsequent deletion of *HC-Pro*, *NIa-Pro* and *NIb* at the same second position, we observe that both accumulation and within-host competitive fitness cannot be completely restored by the virus that originally had two copies of the *NIb* replicase gene, whilst viruses that originally had two copies of the *HC-Pro* gene or the main viral protease *NIa-Pro* gene do restore their fitness after deletion. However, since our evolutionary experiments were limited to approximately half a year, we cannot rule out complete restoration of fitness over longer time periods.

At the sequence level, there were some convergent single-nucleotide mutations, although in most cases these occur only in a small fraction of lineages. The transient presence of the duplicated *NIa-Pro* copy in the 3-week lineages does seem to be linked to an adaptive mutation. However, our fitness measurements suggest that the cost of gene duplication cannot be overcome by this single nucleotide mutation. The main change in the evolved lineages is the deletion of the duplicated gene copy. However, some deletions extend beyond the duplicated gene copy, including the N-terminal region of the HC-Pro cysteine protease, similar to results obtained by previous studies (Dolja et al. 1993; Zwart et al. 2014; Willemsen et al. 2016). The N-terminal region of HC-Pro is implicated in transmission by aphids (Thornbury et al. 1990; Atreya et al. 1992) and is not essential for viral replication and movement (Dolja et al. 1993; Cronin et al. 1995). Our experimental setup does not involve transmission by aphids, however, we do not observe this deletion when evolving the wild-type virus. Moreover, we only observe this deletion when the position of gene duplication or insertion (Zwart et al. 2014) is before HC-Pro, suggesting that gene insertion and subsequent deletion at this position facilitates recombination to an even smaller genome size.

By fitting a mathematical model of virus evolution to the data, we find that knowing only fitness is not enough information to predict the stability of the duplicated genes: as we have shown, there is a context-dependent recombination rate, and specifically, the identity and position of the duplication also play a role. Given that the supply rate of variants with large deletions will be driven largely by homologous recombination, we had expected stability to depend on the genetic context. The estimates of the recombination rate per genome and generation in this study are far lower than previously reported for TEV, which was estimated to be 3.427×10^{-5} per nucleotide and generation (Tromas et al. 2014b), which translates into 0.3269 per genome and generation. The estimates of this study (TEV-HCPro$_2$-HCPro$_3$: 1.000×10^{-3}; TEV-NIaPro$_2$-NIaPro$_8$: 2.239×10^{-10}; TEV-NIb$_1$-NIb$_9$: 1.259×10^{-3}; TEV-NIb$_2$-NIb$_9$: 3.548×10^{-5}) are much closer to the per nucleotide estimate. This large discrepancy could be related to two factors. First, Tromas et al. (2014b) considered recombination between two highly similar genotypes, which requires consideration of many details of the experimental system, including the rate at which cells will be coinfected by these genotypes. On the other hand, considering these details should lead to a general estimate of the recombination rate (as opposed to the rate at which two different genotypes will recombine in a mixed infection) and hence this explanation is not very satisfactory. Second, only a small fraction all recombination events will render viruses with a conserved reading frame, and a suitable deletion size: large enough to have appreciable fitness gains and be selected, but small enough not to disrupt the surrounding cistrons or polyprotein processing. Therefore, the parameter δ described here is in fact the rate at which this particular subclass of recombinants occurs. This subclass is likely to be only a small fraction of all possible recombinants, and hence it is quite reasonable that these two estimates of the recombination rate vary by several orders of magnitude.

In addition to gene duplications, the model developed in this study can be applied to predict the stability of other types of sequence insertions, such as those brought about by horizontal gene transfer. Understanding the stability of gene insertions in genomes is highly relevant to the understanding genome-architecture evolution, but it also has important implications for biotechnological applications, such as heterologous expression systems. Our results here suggest that the fitness costs of extraneous sequences may not be a good predictor of genomic stability, in general. Therefore, in practical terms, it could be advisable to empirically test the stability of *e.g.* a viral construct, rather than make assumptions on stability based on parameters such as replication or accumulation.

Acknowledgements

We thank Francisca de la Iglesia and Paula Agudo for excellent technical assistance.

References

Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. Annu. Rev. Genet. 43:167–195. doi: 10.1146/annurev-genet-102108-134805.

Atreya CD, Atreya PL, Thornbury DW, Pirone TP. 1992. Site-directed mutations in the potyvirus HC-PRO gene affect helper component activity, virus accumulation, and symptom expression in infected tobacco plants. Virology. 191:106–111. doi: 10.1016/0042-6822(92)90171-K.

Bedoya LC, Daròs J-A. 2010. Stability of *Tobacco etch virus* infectious clones in plasmid vectors. Virus Res. 149:234–240. doi: 10.1016/j.virusres.2010.02.004.

Belshaw R, Pybus OG, Rambaut A. 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 17:1496-1504. doi: 10.1101/gr.6305707.

Belshaw R, Gardner A, Rambaut A, Pybus OG. 2008. Pacing a small cage: mutation and RNA viruses. Trends Ecol. Evol. 23:188–193. doi: 10.1016/j.tree.2007.11.010.

Blasdell KR et al. 2012. Kotonkan and Obodhiang viruses: African ephemeroviruses with large and complex genomes. Virology. 425:143–153. doi: 10.1016/j.virol.2012.01.004.

Boyko VP, Karasev AV, Agranovsky AA, Koonin EV, Dolja VV. 1992. Coat protein gene duplication in a filamentous RNA virus of plants. Proc. Natl. Acad. Sci. USA. 89:9156–9160. doi: 10.1073/pnas.89.19.9156.

Carrasco P, Daròs J-A, Agudelo-Romero P, Elena SF. 2007. A real-time RT-PCR assay for quantifying the fitness of *Tobacco etch virus* in competition experiments. J. Virol. Methods. 139:181–188. doi: 10.1016/j.jviromet.2006.09.020.

Carrington JC, Haldeman R, Dolja VV, Restrepo-Hartwig MA. 1993. Internal cleavage and trans-proteolytic activities of the VPg-proteinase (NIa) of *Tobacco etch potyvirus in vivo*. J. Virol. 67:6995–7000.

Cronin S, Verchot J, Haldeman-Cahill R, Schaad MC, Carrington JC. 1995. Long-distance movement factor: a transport function of the potyvirus helper component proteinase. Plant Cell. 7:549–559. doi: 10.1105/tpc.7.5.549.

Dolja VV, Herndon KL, Pirone TP, Carrington JC. 1993. Spontaneous mutagenesis of a plant potyvirus genome after insertion of a foreign gene. J. Virol. 67:5968–5975.

Eigen M. 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften. 58:465–523.

Elena SF, Carrasco P, Daròs J-A, Sanjuán R. 2006. Mechanisms of genetic robustness in RNA viruses. EMBO Rep. 7:168–173. doi: 10.1038/sj.embor.7400636.

Engler C, Gruetzner R, Kandzia R, Marillonnet S. 2009. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One. 4:e5553. doi: 10.1371/journal.pone.0005553.

Fazeli CF, Rezaian MA. 2000. Nucleotide sequence and organization of ten open reading frames in the genome of grapevine leafroll-associated virus 1 and identification of three subgenomic RNAs. J. Gen. Virol. 81:605–615. doi: 10.1099/0022-1317-81-3-605.

Forss S, Schaller H. 1982. A tandem repeat gene in a picornavirus. Nucleic Acids Res. 10:6441–6450. doi: 10.1093/nar/10.20.6441.

Gubala A et al. 2010. Ngaingan virus, a macropod-associated rhabdovirus, contains a second glycoprotein gene and seven novel open reading frames. Virology. 399:98–108. doi: 10.1016/j.virol.2009.12.013.

Holmes EC. 2003. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. 11:543–546. doi: 10.1016/j.tim.2003.10.006.

Hughes JF, Coffin JM. 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat. Genet. 29:487–489. doi: 10.1038/ng775.

Kambol R, Kabat P, Tristem M. 2003. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, Xenopus laevis. Virology. 311:1–6. doi: 10.1016/S0042-6822(03)00263-0.

Koboldt DC et al. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568–576. doi: 10.1101/gr.129684.111.

Kreuze JF, Savenkov EI, Valkonen JPT. 2002. Complete genome sequence and analyses of the subgenomic RNAs of *Sweet potato chlorotic stunt virus* reveal several new features for the genus Crinivirus. J. Virol. 76:9260–9270. doi: 10.1128/JVI.76.18.9260-9270.2002.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359. doi: 10.1038/nmeth.1923.

LaPierre LA, Holzschu DL, Bowser PR, Casey JW. 1999. Sequence and transcriptional analyses of the fish retroviruses walleye epidermal hyperplasia virus types 1 and 2: evidence for a gene duplication. J. Virol. 73:9393–403.

Li H et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079. doi: 10.1093/bioinformatics/btp352.

Lynch M. 2006. Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. 60:327–349. doi: 10.1146/annurev.micro.60.080805.142300.

Majer E et al. 2014. Relocation of the NIb gene in the *Tobacco etch potyvirus* genome. J. Virol. 88:4586–4590. doi: 10.1128/JVI.03336-13.

Majer E, Daròs J-A, Zwart M. 2013. Stability and fitness impact of the visually discernible Rosea1 marker in the *Tobacco etch virus* genome. Viruses. 5:2153–2168. doi:

10.3390/v5092153.

Martínez F, Sardanyés J, Elena SF, Daròs J-A. 2011. Dynamics of a plant RNA virus intracellular accumulation: stamping machine vs. geometric replication. Genetics. 188:637–646. doi: 10.1534/genetics.111.129114.

R Core Team. 2014. R: A language and environment for statistical computing. http://www.r-project.org/ (last accessed April 10, 2016).

Revers F, García JA. 2015. Molecular Biology of Potyviruses. Adv. Virus Res. 92:101–199. doi: 10.1016/bs.aivir.2014.11.006.

Sanjuan R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. USA. 101:8396–8401. doi: 10.1073/pnas.0400146101.

Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. J. Virol. 84:9733–9748. doi: 10.1128/JVI.00694-10.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27:863–864. doi: 10.1093/bioinformatics/btr026.

Simon-Loriere E, Holmes EC. 2013. Gene duplication is infrequent in the recent evolutionary history of RNA viruses. Mol. Biol. Evol. 30:1263–1269. doi: 10.1093/molbev/mst044.

Solé RV, Ferrer R, González-García I, Quer J, Domingo E. 1998. Red Queen dynamics, competition and critical points in a model of RNA virus quasispecies. J. Theor. Biol. 198:47-59.

Spencer DH et al. 2014. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J. Mol. Diagn. 16:75–88. doi: 10.1016/j.jmoldx.2013.09.003.

Thornbury DW, Patterson CA, Dessens JT, Pirone TP. 1990. Comparative sequence of the helper component (HC) region of *Potato virus Y* and a HC-defective strain, *Potato virus C*. Virology. 178:573–578. doi: 10.1016/0042-6822(90)90356-V.

Tristem M, Marshall C, Karpas A, Petrik J, Hill F. 1990. Origin of vpx in lentiviruses. Nature. 347:341–342. doi: 10.1038/347341b0.

Tromas N, Zwart MP, Lafforgue G, Elena SF. 2014a. Within-host spatiotemporal dynamics of plant virus infection at the cellular level. PLoS Genet. 10:e1004186. doi: 10.1371/journal.pgen.1004186.

Tromas N, Zwart MP, Poulain M, Elena SF. 2014b. Estimation of the *in vivo* recombination rate for a plant RNA virus. J. Gen. Virol. 95:724-732. doi: 10.1099/vir.0.060822-0.

Turner PE, Chao L. 1998. Sex and the evolution of intrahost competition in RNA virus phi6. Genetics. 150:523–532.

Tzanetakis IE, Martin RR. 2007. Strawberry chlorotic fleck: Identification and characterization of a novel Closterovirus associated with the disease. Virus Res. 124:88–94. doi: 10.1016/j.virusres.2006.10.005.

Tzanetakis IE, Postman JD, Martin RR. 2005. Characterization of a Novel Member of the Family *Closteroviridae* from Mentha spp. Phytopathology. 95:1043–1048. doi: 10.1094/PHYTO-95-1043.

Walker PJ et al. 1992. The genome of *Bovine ephemeral fever rhabdovirus* contains two related glycoprotein genes. Virology. doi: 10.1016/0042-6822(92)90165-L.

Willemsen A, Zwart MP, Tromas N, Majer E. 2016. Multiple barriers to the evolution of alternative gene orders in a positive-strand RNA virus. Genetics. 202:1503–1521. doi: 10.1534/genetics.115.185017.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298. doi: 10.1016/S0169-5347(03)00033-8.

Zwart MP, Daròs J-A, Elena SF. 2011. One is enough: in vivo effective population size is dose-dependent for a plant RNA virus. PLoS Pathog. 7:e1002122. doi: 10.1371/journal.ppat.1002122.

Zwart MP, Daròs J-A, Elena SF. 2012. Effects of potyvirus effective population size in inoculated leaves on viral accumulation and the onset of symptoms. J. Virol. 86:9737–9747. doi: 10.1128/JVI.00909-12.

Zwart MP, Willemsen A, Daròs JA, Elena SF. 2014. Experimental evolution of pseudogenization and gene loss in a plant RNA virus. Mol. Biol. Evol. 31:121–134. doi: 10.1093/molbev/mst175.

Figure Legends

Fig. 1. Schematic representation of the different TEV genotypes containing gene duplications. The wild-type TEV (A) codes for 11 mature peptides, including P3N-PIPO embedded within the P3 protein at a +2 frameshift. Five different viral genotypes containing a single gene duplication were constructed. Second copies of *HC-Pro* (B), *NIa-Pro* (C) and *NIb (D)* were introduced between *P1* and *HC-Pro*. A second copy of *NIb* was also introduced before *P1* (D). And a second copy of *CP* was introduced between *NIb* and *CP* (E). For simplification P3N-PIPO is only drawn at the wild-type TEV.

Fig. 2. Deletion detection along the evolution experiments. RT-PCR was performed on the region containing a duplication in the viral genotypes (A-D). Either an intact duplicated copy (white boxes), a deletion together with an intact duplicated copy (light-grey boxes), or a partial or complete loss of the duplicated copy (dark-grey boxes) were detected.

Fig. 3. The reduction in genome size over time. The different panels display how the genome size of the different viral genotypes with gene duplications (A-D) changes along the evolution experiments. The dotted grey lines indicate the genome sizes of the wild-type virus (below) and the ancestral viruses (above). The genome sizes of the 3-week lineages are drawn with dashed black lines and open symbols, and those of the 9-week lineages are drawn with continuous blue lines and filled symbols.

Fig. 4. Within-host competitive fitness of the evolved and ancestral lineages. Fitness (*W*), as determined by competition experiments and RT-qPCR of the different viral genotypes with respect to a common competitor; TEV-eGFP. The ancestral lineages are indicated by filled circles and the evolved lineages by open circles. The different viral genotypes are color coded, where the wild-type virus is drawn in green. The asterisks indicate statistical significant differences of the evolved lineages as compared to their corresponding ancestral lineages (*t*-test with Holm-Bonferroni correction).

Fig. 5. Virus accumulation of the evolved and ancestral lineages. Virus accumulation, as determined by accumulation experiments and RT-qPCR at 7 dpi of the different viral genotypes. The ancestral lineages are indicated by filled circles and the evolved lineages by

open circles. The different viral genotypes are color coded, where the wild-type virus is drawn in green. The asterisks indicate statistical significant differences of the evolved lineages as compared to their corresponding ancestral lineages (*t*-test with Holm-Bonferroni correction).

Fig. 6. The relationship between genome size and within-host competitive fitness. The pink filled circle represents the within-host competitive fitness of the ancestral TEV-NIaPro$_2$-NIaPro$_8$ and the green filled circle that of the ancestral wild-type TEV. The black open circles represent the evolved 3-week (right) and 9-week (left) TEV-NIaPro$_2$-NIaPro$_8$ lineages. The evolved 9-week lineages, that contain genomic deletions, have a significant higher within-host competitive fitness (Mann-Whitney $U = 4.5$, $P < 0.001$) than the evolved 3-week lineages without deletions. A linear regression has been drawn to emphasize the trend in the data.

Fig. 7. Genomes of the ancestral and evolved lineages. Mutations were detected using NGS data of the evolved virus lineages as compared to their ancestral lineages. The square symbols represent mutations that are fixed (> 50%) and the circle symbols represent mutations that are not fixed (< 50%). Filled symbols represent nonsynonymous substitutions and open symbols represent synonymous substitutions. Black substitutions occur only in one lineage, whereas color-coded substitutions are repeated in two or more lineages, or in a lineage from another virus genotype. Note that the mutations are present at different frequencies as reported by SAMtools. Grey boxes indicate genomic deletions in the majority variant.

Supplementary files

Supplementary file 1.

This file contains supplementary Table S1 with primers flanking duplicated and ancestral gene regions from 5' to 3' and supplementary Table S2 with primers for Sanger sequencing from 5' to 3'.

Supplementary file 2.

This file contains supplementary Figure S1 with the distribution of SNP frequencies in the evolved TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV lineages.

Supplementary file 3.

This file contains supplementary Tables S3, S4 and S5 with the within population sequence variation of the evolved and ancestral TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$ and TEV lineages, respectively.

Supplementary file 4.

This file contains supplementary Text S1 where the numerical analysis of equations 1 and 2 is fully presented. The file also contains the figures referred in the supplementary text.

Data Deposition

The sequences of the ancestral viral stocks were submitted to GenBank with accessions TEV-NIb_1-NIb_9: KT203712; TEV-NIb_2-NIb_9: KT203713; TEV-$HCPro_2$-$HCPro_3$: KX137150; TEV-$NIaPro_2$-$NIaPro_8$: KX137151; TEV: KX137149).

Funding

This work was supported by the John Templeton Foundation [grant number 22371 to S.F.E.]; the European Commission 7th Framework Program EvoEvo Project [grant number ICT-610427 to S.F.E.]; the Spanish Ministerio de Economía y Competitividad (MINECO) [grant numbers BFU2012-30805 and BFU2015-65037-P to S.F.E.]; the Botín Foundation from Banco Santander through its Santander Universities Global Division [J.S.]; and the European Molecular Biology Organization [grant number ASTF 625-2015 to A.W]. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Table 1. Model parameters

Parameter	Value	Explanation
λ	500	Number of founders of infection (Zwart et al. 2014).
$\kappa_{t\,=\,9\,weeks}$	4×10^9	Final value time-varying carrying capacity (9 weeks post-infection), weight of leaves multiplied by carrying capacity as estimated (Zwart et al. 2012).
b	1.344	Initial growth rate (per generation) for virus with single gene copy (Zwart et al. 2012).
a	φb	Initial growth rate for virus with a duplicated gene, where φ is the relative fitness of the virus with duplications compared to the virus with a single gene copy (see results)
g	2.91	Generations per day (Martínez et al. 2011).
β	b/a	The effect of *A* the replication of *B*
α	a/b	The effect of *B* the replication of *A*

References

Martínez F, Sardanyés J, Elena SF, Daròs J-A. 2011. Dynamics of a plant RNA virus intracellular accumulation: stamping machine vs. geometric replication. Genetics. 188:637–646. doi: 10.1534/genetics.111.129114.

Zwart MP, Daròs J-A, Elena SF. 2012. Effects of potyvirus effective population size in inoculated leaves on viral accumulation and the onset of symptoms. J. Virol. 86:9737–47. doi: 10.1128/JVI.00909-12.

Zwart MP, Willemsen A, Daròs JA, Elena SF. 2014. Experimental evolution of pseudogenization and gene loss in a plant RNA virus. Mol. Biol. Evol. 31:121–134. doi: 10.1093/molbev/mst175.

Table 2. Nested ANOVA s on within-host competitive fitness and viral accumulation

Genotype	Trait	Source of Variation	SS	df	MS	*F*	*P*
TEV-HCPro$_2$-HCPro$_3$	Competitive fitness	Genotype	0.049	1	0.049	0.786	0.396
		Lineage within Genotype	0.624	10	0.062	9.947	< 0.001
		Plant within Lineage within Genotype	0.150	24	0.006	347.244	< 0.001
		Error	0.001	72	1.81×10^{-5}		
	Accumulation	Genotype	0.008	1	0.008	0.080	0.783
		Lineage within Genotype	1.012	10	0.101	1.005	0.467
		Plant within Lineage within Genotype	2.417	24	0.101	51.187	< 0.001
		Error	0.142	72	0.002		
TEV-NIaPro$_2$-NIaPro$_8$	Competitive fitness	Genotype	0.155	1	0.155	4.534	0.066
		Lineage within Genotype	0.274	8	0.034	4.285	0.004
		Plant within Lineage within Genotype	0.160	20	0.008	503.714	< 0.001
		Error	0.001	60	1.59×10^{-5}		
	Accumulation	Genotype	0.246	1	0.246	1.199	0.305
		Lineage within Genotype	1.643	8	0.205	2.224	0.070
		Plant within Lineage within Genotype	1.847	20	0.092	323.631	< 0.001
		Error	0.017	60	0.001		
TEV-NIb$_1$-NIb$_9$	Competitive fitness	Genotype	1.308	1	1.308	49.734	< 0.001
		Lineage within Genotype	0.212	8	0.026	3.145	0.018
		Plant within Lineage within Genotype	0.169	20	0.008	175.319	< 0.001
		Error	0.003	58	4.81×10^{-5}		
	Accumulation	Genotype	5.374	1	5.374	36.006	< 0.001
		Lineage within Genotype	1.194	8	0.149	0.939	0.507
		Plant within Lineage within Genotype	3.178	20	0.159	263.098	< 0.001
		Error	0.036	60	0.001		
TEV-NIb$_2$-NIb$_9$	Competitive fitness	Genotype	1.796	1	1.796	36.175	< 0.001
		Lineage within Genotype	0.397	8	0.050	4.207	0.004
		Plant within Lineage within Genotype	0.236	20	0.012	519.611	< 0.001
		Error	0.001	60	2.27×10^{-5}		
	Accumulation	Genotype	0.824	1	0.824	3.728	0.090
		Lineage within Genotype	1.771	8	0.221	3.439	0.012
		Plant within Lineage within Genotype	1.290	20	0.065	110.478	< 0.001
		Error	0.034	59	0.001		

Table 3. Adaptive convergent mutations within each virus genotype

Virus genotype	nt change at ancestral position	aa change	gene	nt position in gene
TEV-HCPro$_2$-HCPro$_3$	A304G	I→V	P1	160
	U4444C	S→P	P3	625
TEV-NIaPro$_2$-NIaPro$_8$	C1466U	S→L	NIa-Pro$_2$	410
	A4357G	I→V	6K1	148
TEV-NIb$_1$-NIb$_9$	A1643U	Y→F	NIb$_1$	1499
TEV-NIb$_2$-NIb$_9$	C6351U	Y→Y	CI	1173

nt: nucleotide; aa: amino acid

Table 4. Model parameter estimates for deterministic recombination rate

Model	Estimates of Log$_{10}$[δ] (Lower 95% fiducial limit, upper 95% fiducial limit)			
1	−6.2 (*)			
2	2HCPro ≥ −3.0 (*)	2NIaPro = −9.65 (−10.1, −9.0)	2NIb1 ≥ −2.9 (*)	2NIb2 = −4.45 (−5.2, −3.7)
3	3W = −6.2 (*)	9W = −9.65 (−10.1, −7.5)		
4	2HCPro 3W ≥ −3.0 (*)	2NIaPro 3W ≤ −9.0 (*)	2NIb1 3W ≥ −2.9 (*)	2NIb2 3W = −4.45 (−5.5, −3.7)
	2HCPro 9W ≥ −10.5 (*)	2NIaPro 9W = −9.6 (−10.1, −7.5)	2NIb1 9W ≥ −10.1	2NIb2 9W ≥ −11.5 (*)

* indicates the fiducial limit is identical to the parameter estimate, also when the parameter estimate is a range, 2HCPro: TEV-HCPro$_2$-HCPro$_3$; 2NIaPro: TEV-NIaPro$_2$-NIaPro$_8$; 2NIb1: TEV-NIb$_1$-NIb$_9$; 2NIb2: TEV-NIb$_2$-NIb$_9$; 3W: 3-week passages; 9W: 9-week passages

Table 5. Model selection for models with deterministic recombination

Model	Parameters	NLL	AIC	ΔAIC	Akaike Weight
1	1	254.284	510.569	443.470	0.000
2	4	29.549	67.099	−0.000	0.982
3	2	226.669	457.339	390.240	0.000
4	8	29.549	75.099	8.000	0.018

Figure 1

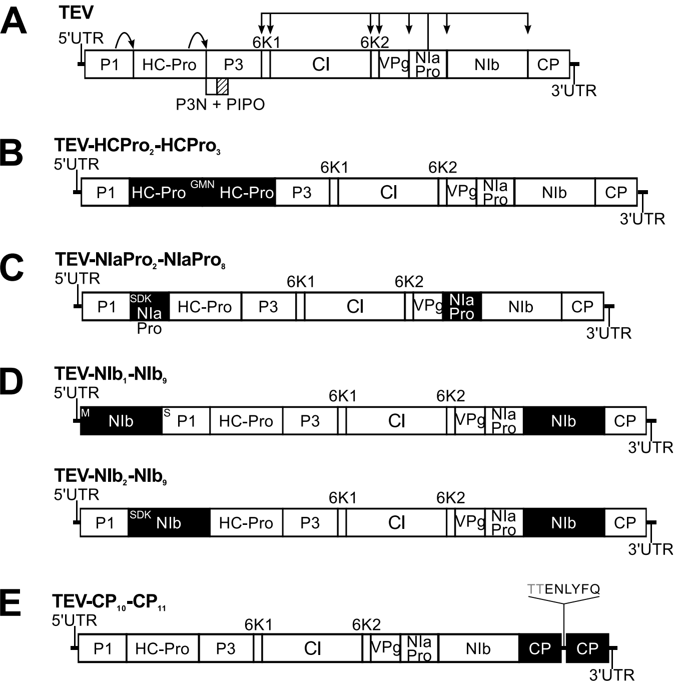
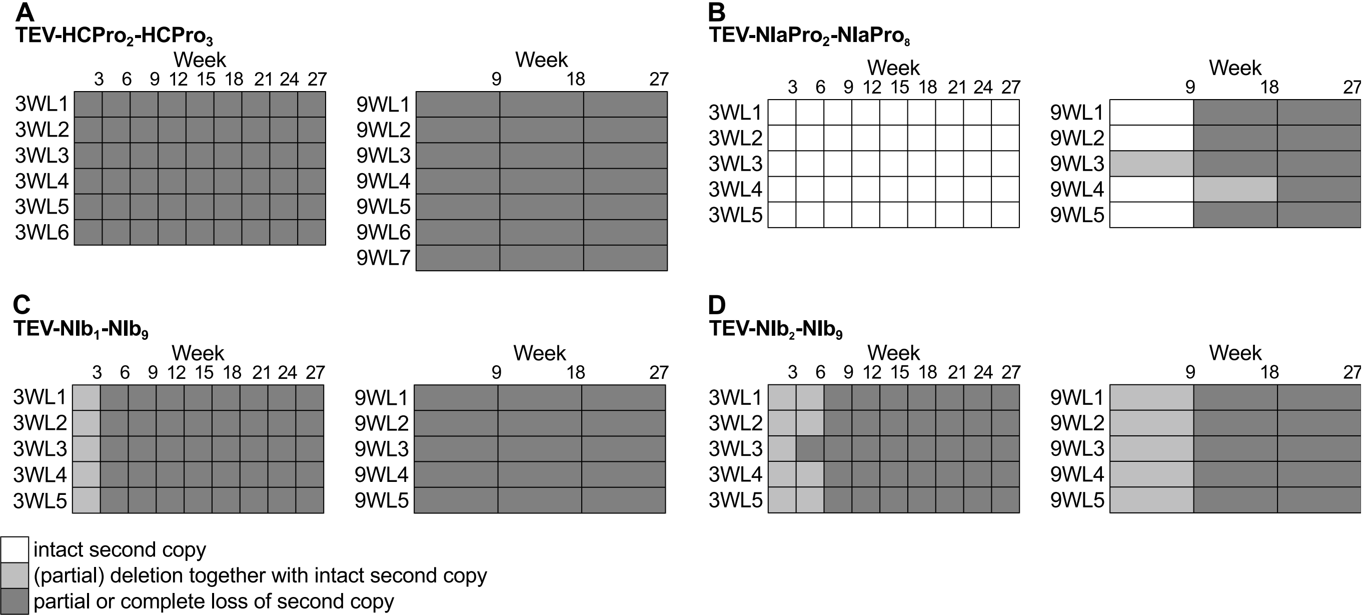

A
TEV
5'UTR
P1
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
3'UTR
P3N + PIPO
B
TEV-HCPro$_2$-HCPro$_3$
5'UTR
P1
HC-Pro
GMN
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
3'UTR
C
TEV-NIaPro$_2$-NIaPro$_8$
5'UTR
P1
SDK
NIa Pro
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
3'UTR
D
TEV-NIb$_1$-NIb$_9$
5'UTR
M
NIb
S
P1
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
3'UTR
TEV-NIb$_2$-NIb$_9$
5'UTR
P1
SDK
NIb
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
3'UTR
E
TEV-CP$_{10}$-CP$_{11}$
TTENLYFQ
5'UTR
P1
HC-Pro
P3
6K1
CI
6K2
VPg
NIa Pro
NIb
CP
CP
3'UTR

Figure 2



A

TEV-HCPro$_2$-HCPro$_3$

	Week								
	3	6	9	12	15	18	21	24	27
3WL1	■	■	■	■	■	■	■	■	■
3WL2	■	■	■	■	■	■	■	■	■
3WL3	■	■	■	■	■	■	■	■	■
3WL4	■	■	■	■	■	■	■	■	■
3WL5	■	■	■	■	■	■	■	■	■
3WL6	■	■	■	■	■	■	■	■	■

	Week		
	9	18	27
9WL1	■	■	■
9WL2	■	■	■
9WL3	■	■	■
9WL4	■	■	■
9WL5	■	■	■
9WL6	■	■	■
9WL7	■	■	■

B

TEV-NIaPro$_2$-NIaPro$_8$

	Week								
	3	6	9	12	15	18	21	24	27
3WL1	□	□	□	□	□	□	□	□	□
3WL2	□	□	□	□	□	□	□	□	□
3WL3	□	□	□	□	□	□	□	□	□
3WL4	□	□	□	□	□	□	□	□	□
3WL5	□	□	□	□	□	□	□	□	□

	Week		
	9	18	27
9WL1	□	■	■
9WL2	□	■	■
9WL3	░	■	■
9WL4	□	░	■
9WL5	□	■	■

C

TEV-NIb$_1$-NIb$_9$

	Week								
	3	6	9	12	15	18	21	24	27
3WL1	░	■	■	■	■	■	■	■	■
3WL2	░	■	■	■	■	■	■	■	■
3WL3	░	■	■	■	■	■	■	■	■
3WL4	░	■	■	■	■	■	■	■	■
3WL5	░	■	■	■	■	■	■	■	■

	Week		
	9	18	27
9WL1	■	■	■
9WL2	■	■	■
9WL3	■	■	■
9WL4	■	■	■
9WL5	■	■	■

D

TEV-NIb$_2$-NIb$_9$

	Week								
	3	6	9	12	15	18	21	24	27
3WL1	░	░	■	■	■	■	■	■	■
3WL2	░	░	■	■	■	■	■	■	■
3WL3	░	■	■	■	■	■	■	■	■
3WL4	░	░	■	■	■	■	■	■	■
3WL5	░	░	■	■	■	■	■	■	■

	Week		
	9	18	27
9WL1	░	■	■
9WL2	░	■	■
9WL3	░	■	■
9WL4	░	■	■
9WL5	░	■	■

□ intact second copy

░ (partial) deletion together with intact second copy

■ partial or complete loss of second copy

Figure 3

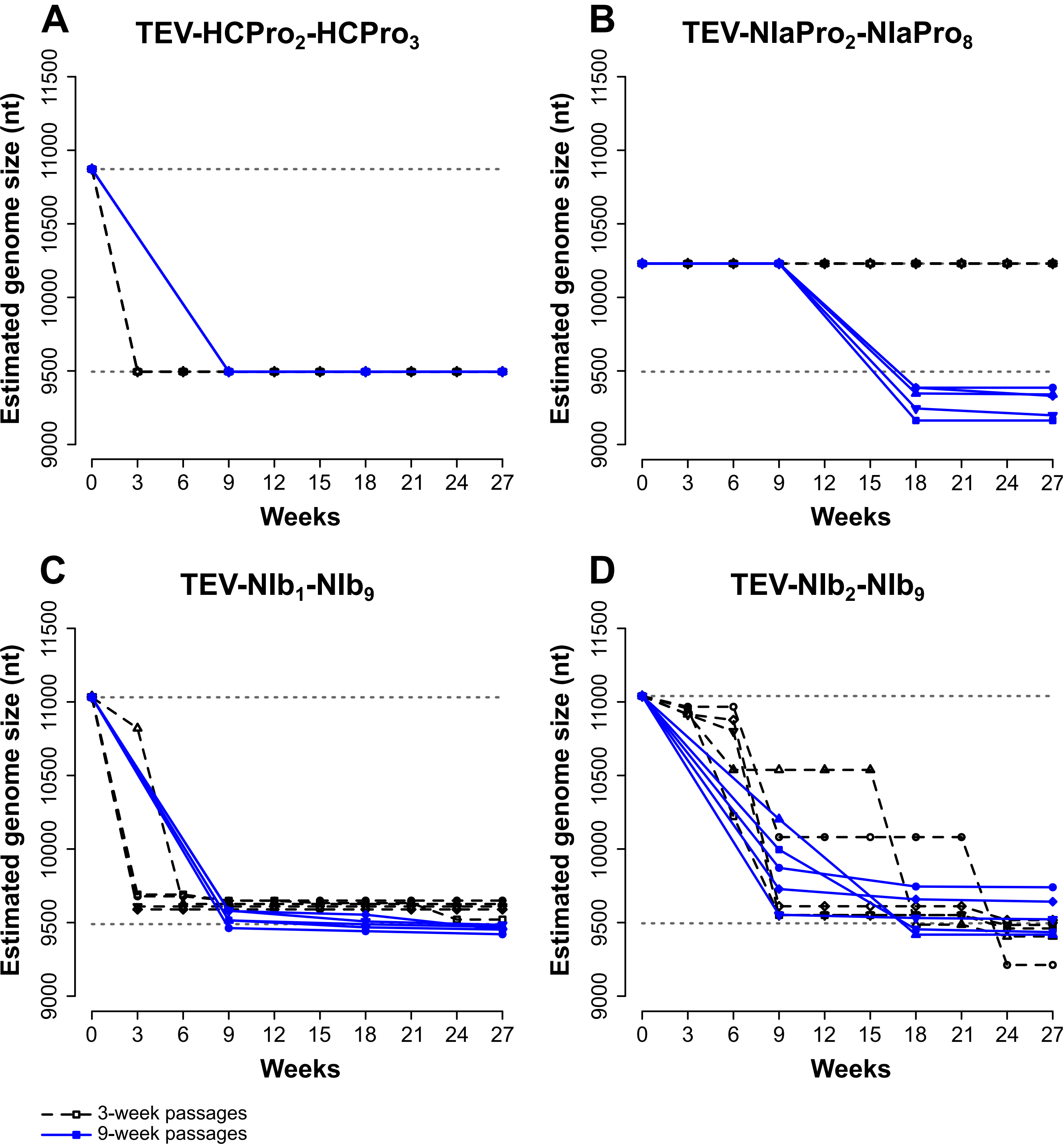

A
TEV-HCPro$_2$-HCPro$_3$
B
TEV-NIaPro$_2$-NIaPro$_8$
C
TEV-NIb$_1$-NIb$_9$
D
TEV-NIb$_2$-NIb$_9$
Estimated genome size (nt)
9000
9500
10000
10500
11000
11500
Weeks
0
3
6
9
12
15
18
21
24
27
3-week passages
9-week passages

Figure 4

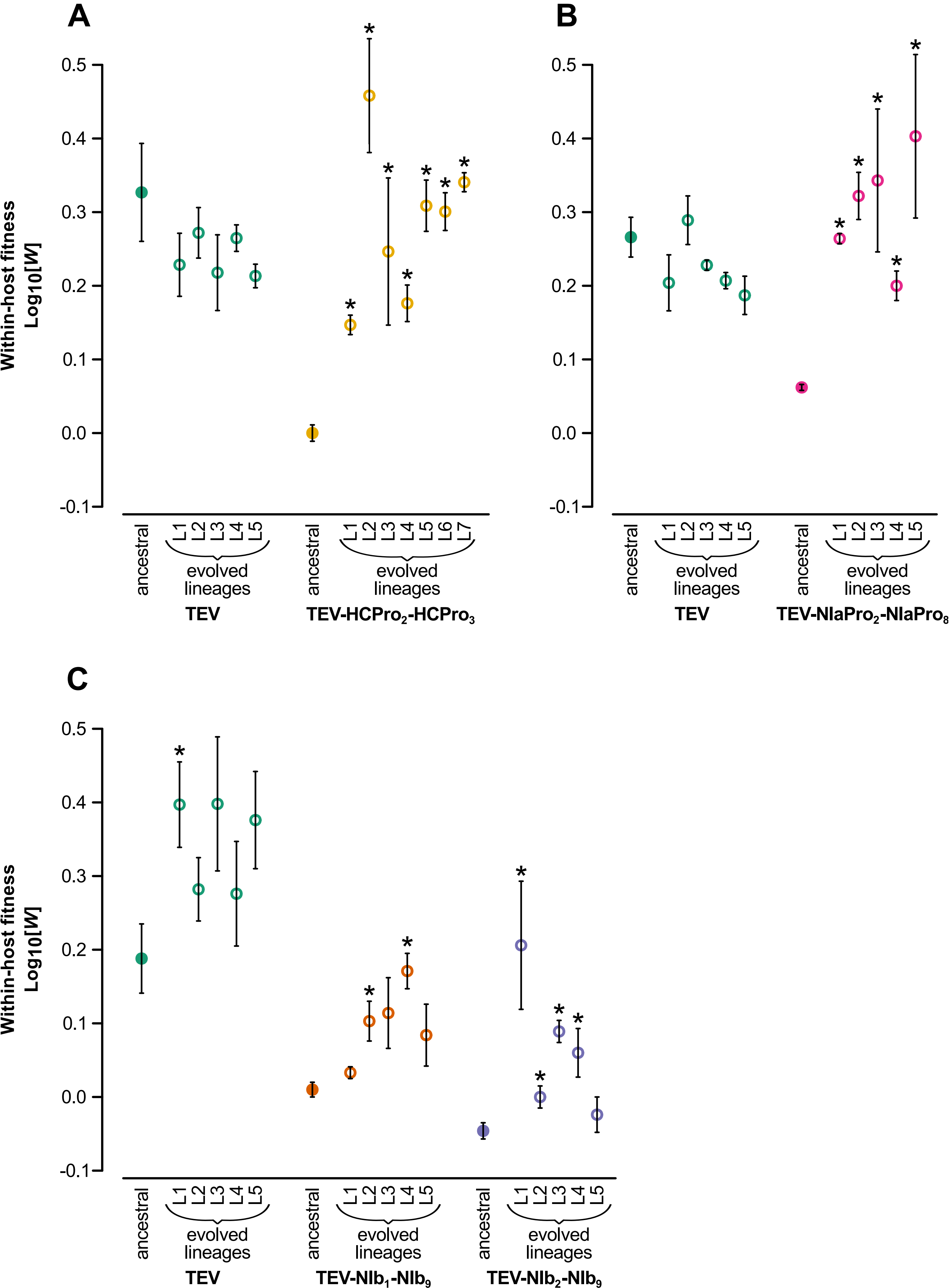

A
B
C
Within-host fitness
Log10[*W*]
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
*
ancestral
L1 L2 L3 L4 L5
L1 L2 L3 L4 L5 L6 L7
evolved lineages
TEV
TEV-HCPro$_2$-HCPro$_3$
TEV-NIaPro$_2$-NIaPro$_8$
TEV-NIb$_1$-NIb$_9$
TEV-NIb$_2$-NIb$_9$

Figure 5

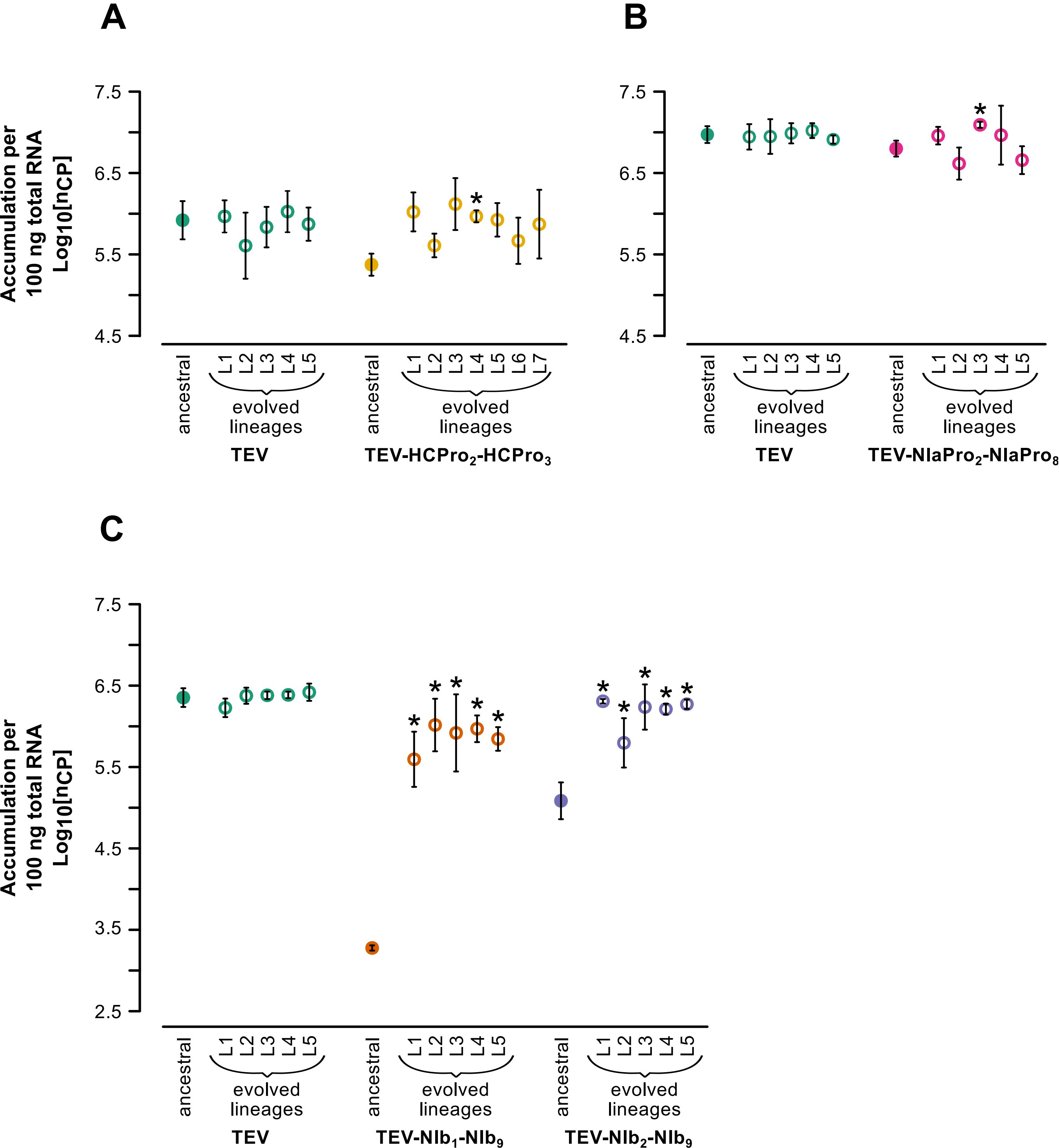

A
B
C
Accumulation per 100 ng total RNA Log10[nCP]
7.5
6.5
5.5
4.5
3.5
2.5
ancestral
L1 L2 L3 L4 L5
L1 L2 L3 L4 L5 L6 L7
evolved lineages
TEV
TEV-HCPro$_2$-HCPro$_3$
TEV-NIaPro$_2$-NIaPro$_8$
TEV-NIb$_1$-NIb$_9$
TEV-NIb$_2$-NIb$_9$
*

Figure 6

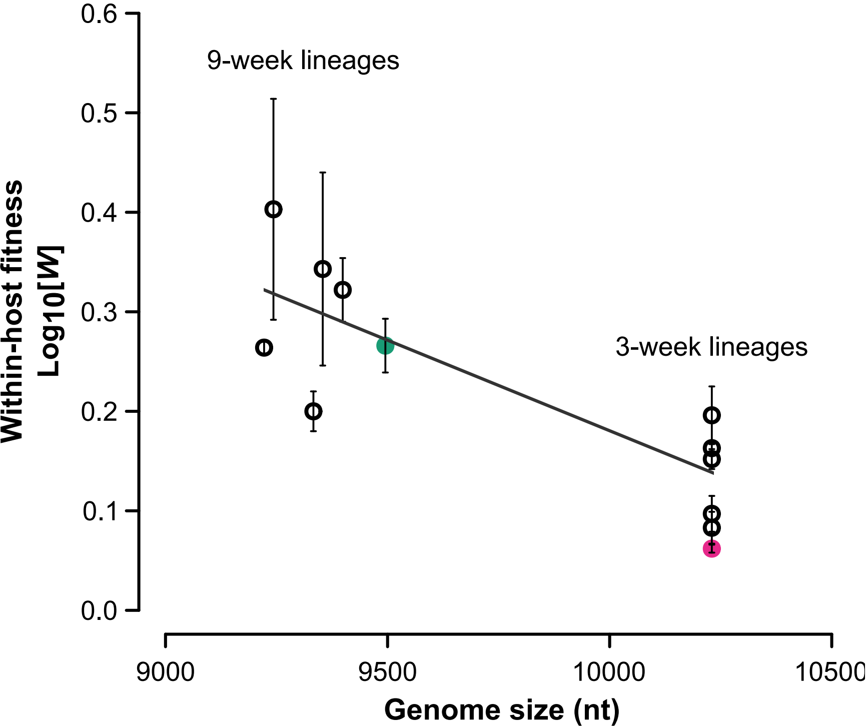
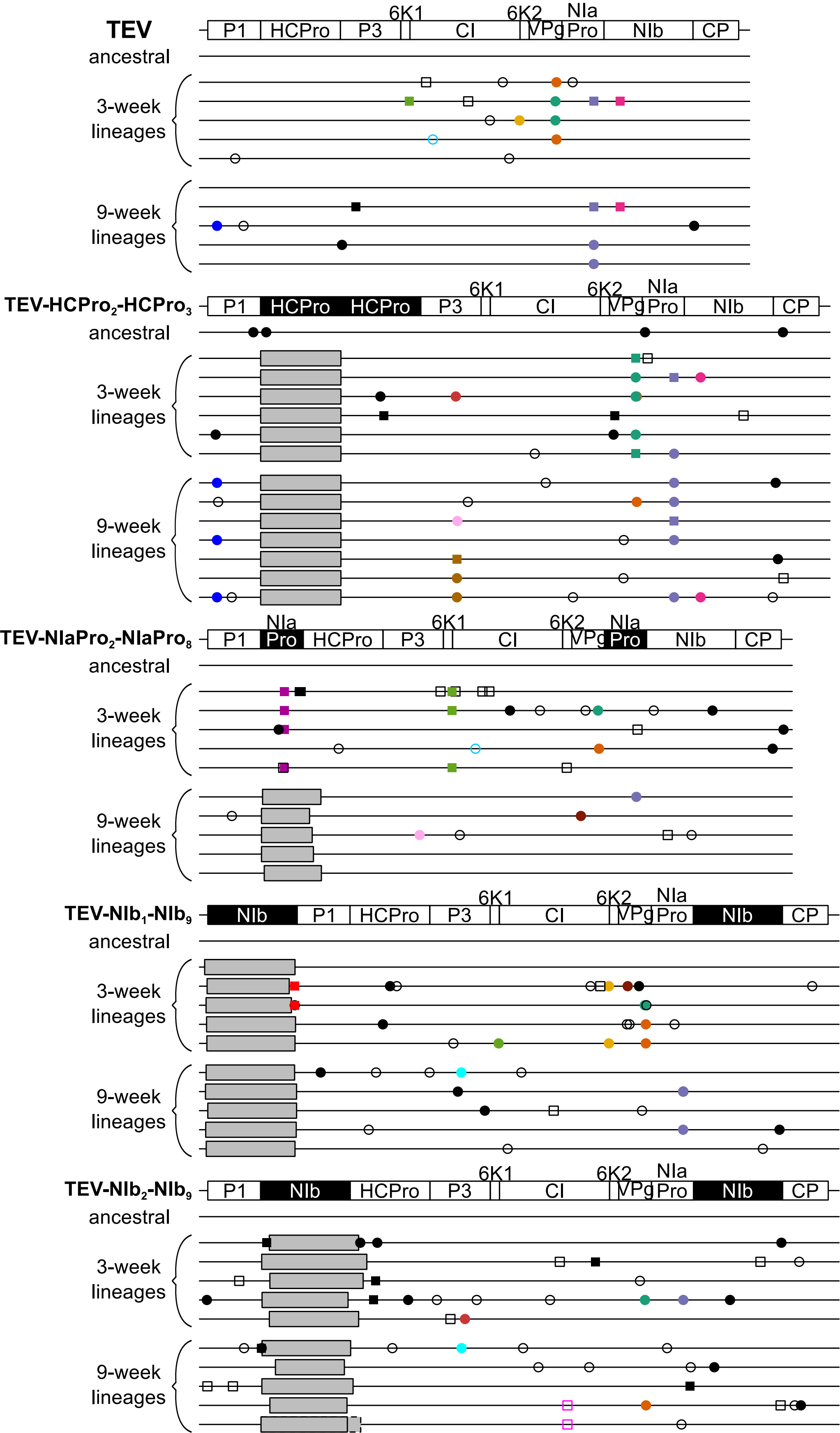

9-week lineages
3-week lineages
Within-host fitness
Log10[*W*]
0.6
0.5
0.4
0.3
0.2
0.1
0.0
9000
9500
10000
10500
Genome size (nt)

Figure 7



TEV
P1
HCPro
P3
6K1
CI
6K2
VPg
NIa
Pro
NIb
CP
ancestral
3-week lineages
9-week lineages
TEV-HCPro$_2$-HCPro$_3$
P1
HCPro
HCPro
P3
6K1
CI
6K2
VPg
NIa
Pro
NIb
CP
ancestral
3-week lineages
9-week lineages
TEV-NIaPro$_2$-NIaPro$_8$
P1
NIa
Pro
HCPro
P3
6K1
CI
6K2
VPg
NIa
Pro
NIb
CP
ancestral
3-week lineages
9-week lineages
TEV-NIb$_1$-NIb$_9$
NIb
P1
HCPro
P3
6K1
CI
6K2
VPg
NIa
Pro
NIb
CP
ancestral
3-week lineages
9-week lineages
TEV-NIb$_2$-NIb$_9$
P1
NIb
HCPro
P3
6K1
CI
6K2
VPg
NIa
Pro
NIb
CP
ancestral
3-week lineages
9-week lineages