Submission type: Article
Section: Discoveries

Experimental evolution of pseudogenization and gene loss in a plant RNA virus

Mark P. Zwart,*[,1] Anouk Willemsen,[1] José-Antonio Daròs,[1] and Santiago F. Elena[1,2]

[1] Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-UPV, 46022 València, Spain
[2] The Santa Fe Institute, Santa Fe, NM 87501, USA

* Corresponding author: Email: marzwa@ibmcp.upv.es

Running title: Experimental evolution of pseudogenization

Abstract

Viruses have evolved highly streamlined genomes and a variety of mechanisms to compress them, suggesting that genome size is under strong selection. Horizontal gene transfer has, on the other hand, played an important role in virus evolution. However, evolution cannot integrate initially nonfunctional sequences into the viral genome if they are rapidly purged by selection. Here we report on the experimental evolution of pseudogenization in virus genomes using a plant RNA virus expressing a heterologous gene. When long 9-week passages were performed the added gene was lost in all lineages, whereas viruses with large genomic deletions were fixed in only two out of ten 3-week lineages and none in 1-week lineages. Illumina NGS revealed considerable convergent evolution in the 9- and 3-week lineages with genomic deletions. Genome size was correlated to within-host competitive fitness, although there was no correlation with virus accumulation or virulence. Within-host competitive fitness of the 3-week virus lineages without genomic deletions was higher than for the 1-week lineages. Our results show that the strength of selection for a reduced genome size and the rate of pseudogenization depend on demographic conditions. Moreover, for the 3-week passage condition, we observed increases in within-host fitness whilst selection was not strong enough to quickly remove the nonfunctional heterologous gene. These results suggest a demographically determined “sweet spot” might exist, where heterologous insertions are not immediately lost while at the same time evolution can act to integrate them into the viral genome.

Introduction

Virus genomes are highly streamlined. Compared to more complex organisms, viruses tend to have small genomes with (i) a high percentage of coding sequences, (ii) none or little intronic sequences, and (iii) only short stretches of intergenic sequences (Belshaw et al. 2007; Lynch 2006). Moreover, viruses have evolved strategies to further compress their genomes, such frameshifts and overlapping ORFs (e.g., Belshaw et al. 2007; Chung et al. 2008). Field observations suggest that genome shrinkage sometimes occurs during epidemic spread, and might be linked to increased within-host fitness and be adaptive for *White spot syndrome virus* (WSSV), a large DNA virus (Marks et al. 2005; Zwart et al. 2010a). Moreover, it appears to be a very general observation that viruses expressing heterologous genes tend to be unstable (Chapman et al. 1992; Chung et al. 2007; Dolja et al. 1993; Guo et al. 1998; Lee et al. 2007, Paar et al. 2007, Pijlman et al. 2001). Furthermore, under conditions maximizing selection for competitive fitness – exemplified by undiluted serial passage in cultured cells – viruses tend to rapidly evolve defective interfering particles (DIPs): viruses with large genomic deletions unable to replicate autonomously, but with a replicative advantage at high multiplicities of infection (Huang 1973; Pathak and Nagy 2009; Simon et al. 2004; Zwart et al. 2008). All these observations suggest that genome size is under strong selection for viruses, and that having unnecessary genomic sequences has fitness costs. In contrast, for bacteria striking cases of genome shrinkage have been found for obligate host-dependent species, but this shrinkage appears to be the result of a mutational bias towards deletions and genetic drift (Kuo and Ochman 2009; Ochman and Davalos 2006).

Viruses play an important evolutionary role as vectors for horizontal gene transfer (HGT) in the genomes of their hosts (Belshaw et al. 2004; Canchaya et al. 2003; Routh et al. 2012). It is moreover becoming increasingly apparent that HGT is also widespread in most viruses and is a key mechanism in their evolution (Dolja et al. 2011; Hughes and Friedman

2005; Koonin and Dolja 2012; Liu et al. 2011, 2012; Yutin and Koonin 2012). Striking innovations such as the DNA-RNA virus hybrid (Diemer and Stedeman 2012) and the cooption of an entire host immunity mechanism by a phage (Seed et al. 2013) exemplify how HGT can empower the evolutionary process. However, strong selection for genome size would in principle be an impediment to HGT. In order for a heterologous sequence to be beneficial to the recipient virus, it must be accommodated into the virus genome, transcriptome and proteome. Mutation and selection must therefore act on newly transferred heterologous sequences, but in order to do so these sequences must not be purged right upon acquisition because of selection for a streamlined genome. The mutational supply may favor the deletion of heterologous sequences; deletion of a sequence by recombination probably has a greater likelihood than the occurrence of beneficial mutations functionally integrating this element in the virus. If the effects of deletions of heterologous sequences are beneficial on the short term, how can HGT be common in viruses?

One possible answer to this question is that selection for genome size is not a very strong force. First, it is not at all clear that the metabolic cost of additional genetic material is sufficient to account for the expected fitness costs (Lynch 2007). Second, experimental results on the relationship between genome size and replicative fitness are ambiguous. When comparing phages with different genome organizations adapted for fast replication, the expected relationship was not found (Bull et al. 2004). When expressing different-size marker proteins in *Sendai virus*, an inverse relationship between insert size and replication was found in cultured cells (Sakai et al. 1999). However, this relationship was not observed *in vivo* (Sakai et al. 1999) and, furthermore, the use of sequences coding for different marker proteins makes the comparison troublesome (Majer et al. 2013). Moreover, instability of viruses expressing heterologous sequences (Chapman et al. 1992; Chung et al. 2007; Dolja et al. 1993; Guo et al. 1998) may depend on many environmental factors (Paar et al. 2007), and even subtleties of the heterologous sequence, such as G/U content

(Lee et al. 2002). Finally, experiments corroborating the relationship between genome size and fitness for WSSV were performed with field isolates (Marks et al. 2005; Zwart et al. 2010a), and hence other genetic variation could be a confounding factor. In considering whether HGT is really implausible, we therefore need to ask whether increases in genome size really have appreciable fitness costs, and what fitness components might be affected.

Here we explore the process of pseudogenization in virus genomes by means of experimental evolution. As a model system, we consider a plant RNA virus expressing a non-toxic heterologous gene, whose expression has been engineered to minimally disrupt the viral genome. We first looked for conditions under which the heterologous gene would be maintained in the genome. It has been shown that the time period between two consecutive transmission events, that is, the time a viral population has to expand between two consecutive bottlenecks, can be instrumental in determining genome stability in plant RNA viruses (Dolja et al. 1993). However, in this study the heterologous gene was expressed as a fusion with one of the viral cistrons, and given the strong effects on viral accumulation (Dolja et al. 1993) must therefore be seen as a deleterious, rather than merely non-functional, addition. We then consider whether the deletion of the heterologous gene was adaptive, and what virus characteristics it modifies. Finally, we consider whether there are conditions that fulfill two requirements: (i) the heterologous gene has a high probability of being maintained in the evolved virus population, and (ii) there is evidence that the virus population is under positive selection and experiences increases in fitness. We think the combination of conditions is relevant to the context of HGT in viruses. If these two conditions are fulfilled, then in principle a heterologous sequence can persist for long periods of time in the virus population, whilst increases in fitness imply that natural selection is acting on the populations and could “tinker” with the heterologous gene, sequences regulating its expression, and other loci interacting with the heterologous gene, potentially functionally integrating it into the viral genome. On the other hand, a heterologous gene may not be lost in a virus population subject to high levels of genetic drift, but it is then also

unlikely that natural selection acts to functionally integrate it. Similarly, a virus population under strong positive selection in which the heterologous gene is lost is also a dead end for HGT. However, simultaneously having maintenance of the heterologous gene and increases in viral fitness suggests the occurrence of a “sweet spot” that could help explain how HGT occurs in viruses.

Results and Discussion

Results of serial passage experiments

As a model system for virus expressing a heterologous gene without appreciable toxicity or disruption of viral replication, we used a variant of *Tobacco etch virus* (TEV; genus *Potyvirus*, family *Potyviridae*). TEV is a positive-sense single-stranded RNA virus that encodes a polyprotein auto-catalytically cleaved into ten mature proteins (Riechmann et al. 1992), and a partially overlapping ORF with a +2 frameshift (Chung et al. 2008). The TEV variant we used expresses eGFP as a separate cistron between P1 and HC-Pro by introducing a second NIa-Pro proteolytic site downstream of the eGFP sequence, whilst retaining the existing C-terminal site in P1 (fig. 1) (Zwart et al. 2011). Evolution experiments were performed in *Nicotiana tabacum* L. cv. Xanthi plants. In brief, four-week-old plants were inoculated with high virus doses, and passages lasting either one week (for a total of 27 consecutive passages), three weeks (nine passages) or nine weeks (three passages) were performed. Each “passage” is the infection of a single plant, and harvesting of tissues at the designated time (i.e., the “passage duration”). Therefore, although the passage duration varied among treatments, each lineage evolved for the same total time (27 weeks) in *N. tabacum.* At the end of a passage, all the leaves above the inoculated leaf were collected, pooled and used to obtain the inoculum for the next round of serial passaging. Ten independent lineages were generated and maintained for each passage duration (1, 3 or 9 weeks). An overview of the experimental setup used is

given in fig. 2, and further details are given in the Materials and Methods section.

eGFP expression was readily apparent in infected plants (fig. 3A) and was used as a first indication of whether the heterologous gene was intact. Partial losses of fluorescence (fig. 3B) almost always preceded complete losses of fluorescence (fig. 3C). One out of ten 1-week lineages showed a partial loss of fluorescence, first observed at passage 7 and maintained until passage 27 (fig. 3E-F). Two out of ten 3-week lineages showed a loss of fluorescence. All but one 9-week passage showed decreased levels of fluorescence after a single passage, and all lineages showed a complete loss of fluorescence after two passages. RT-PCR with primers flanking eGFP (Materials and Methods) confirmed the occurrence of genomic deletions in all lineages with a partial or complete loss of fluorescence (fig. 3G). These results are congruent with previous work with TEV (Dolja et al. 1993) except that the deletion of the heterologous gene occurs much more slowly here, as anticipated.

We then inoculated *N. tabacum* with TEV-eGFP and, at 9 weeks of infection, harvested every 5th leaf up the stem. Since the infection progresses linearly as the plant grows, these leaves enable us to monitor a qualitative time course of evolution within the plant. All plants had by then reached the 45-leaf stage, except for one that had only 40 leaves. We performed RT-qPCR on individual leaves to ascertain at what leaf level deletions occurred, and if they were subsequently maintained in the population. This analysis can be performed since the virus moves mainly upwards in the plant (Dolja et al. 1992). In most cases once a deletion was detected in one leaf, it was maintained and fixed in the superior leaves (fig. 4). This observation suggests that selection for deletion variants is very strong in this experiment, being a stronger evolutionary force than genetic drift within the host. The first leaf in which a novel deletion was detected was not uniformly distributed over all tested leaves (Leaves 5-45; one-sample Kolmogorov-Smirnoff test: $n = 13$; $P = 0.019$); new deletion variants were usually first detected in higher leaves (mean ±SD = 32.31 ±9.92). This result suggests that passage duration is important to the dynamics of heterologous

156 gene deletion because it regulates the amount of expansion that occurs between
157 bottlenecks.

158 During virus infection of mechanically inoculated plants, genetic bottlenecks in the virus
159 population can occur during primary infection of the inoculated leaf and the subsequent
160 entry into systemically infected leaves (Gutiérrez et al. 2012; Hall et al. 2001; Sacristán et
161 al. 2003). For TEV infection of tobacco plants, the number of primary infection foci in the
162 inoculated leaf is a good estimator of the number of founders, whereas the subsequent
163 bottlenecks during entry into systemically infected leaves do not appear to be severe (Zwart
164 et al. 2011). We therefore inoculated eight plants with TEV-eGFP using the same
165 procedure and conditions as during serial passaging (see Materials and Methods;
166 homogenized tissue of plants infected with TEV-eGFP was used as an inoculum), and
167 counted the number of primary infection foci (Zwart et al. 2011). The mean number of foci
168 observed ±SD was 417 ±140, suggesting that although there is a bottleneck at the start of
169 infection, it is not too severe. Nevertheless, this bottleneck could remove variation
170 generated *de novo* during the previous passage and hereby limit the variation upon which
171 selection can act. Longer passage duration would allow beneficial variation to increase in
172 frequency, thus making it less likely to lose them due to genetic drift at the next
173 transmission event. The occurrence of genetic bottlenecks therefore reinforces the idea
174 that passage duration might be important to the evolutionary dynamics in this system.
175

176 **Genome sequences of evolved lines with deletions**

177 All evolved lineages in which deletions had been detected by RT-PCR were fully
178 sequenced by Illumina. We developed an approach (Materials and Methods) for mapping
179 large genomic deletions (i.e., deletions larger than the read size). We consistently saw
180 pseudogenization or complete loss of eGFP in all these lineages (fig. 5A). None of these
181 deletions included the C-terminus of P1, whilst for seven out of 13 lineages these deletions
182 included N-terminal regions of HC-Pro, similar to the previous results (Dolja et al. 1993).

The N-terminal region of HC-Pro is not essential for replication and movement (Cronin et al. 1995; Dolja et al. 1993) but has been implicated in vector-borne transmission (Atreya et al. 1992; Thornbury et al. 1990), which is not a selective force in our mechanical transmission passage experiments. In the seven lineages with deletions extending into HC-Pro, the remains of eGFP were fused with HC-Pro, whilst the proteolytic site between P1 and eGFP remained intact. The start of genomic deletion (5' end) was not uniformly distributed (one-sample Kolmogorov-Smirnoff test: $n = 13$; $P = 0.006$), and there is clustering at the 5' of the eGFP cistron (fig. 5B-C), suggesting the existence of a hotspot for recombination or the unviability of any deletions in the P1-eGFP proteolytic site. On the other hand, the 3' end of the genomic deletion was uniformly distributed (one-sample Kolmogorov-Smirnoff test: $n = 13$; $P = 0.130$).

We performed additional analyses to detect minority variants with different deletion sizes in sequenced lineages (Materials and Methods). Although minority variants were sometimes detected in the 3- and 9-week lineages, these were always present at low frequencies (< 1.5%). Only in the case of the 1-week lineage with a partial eGFP loss was a minority variant present: 47.1% of the population was composed of a variant with intact eGFP. The deletion in the majority variant extends beyond the HC-Pro N-terminal regions nonessential for replication and movement (Cronin et al. 1995), suggesting this majority variant is not able to replicate without being complemented by the full-length variant. When plants were inoculated with low virus doses of this evolved lineage, we were indeed unable to *in vivo* clone the majority variant without intact eGFP (table 1).

When single-base substitutions were detected in sequenced lineages, there appeared to be convergent evolution in the 3- and 9-week lineages (fig. 5A). All these lineages contained at least one substitution present in another lineage, two substitutions were present in eight out of 12 lineages, and one substitution was present in nine out of 12 lineages. Overall, more than half of the substitutions found were present in other lineages, and some lineages contained only substitutions also present in other lineages (9-week

lineages 1 and 2). In the single 1-week passage sequenced, none of the repeated substitutions were present, suggesting that convergent evolution did not occur under these conditions.

Of the substitutions found here, 28 were nonsynonymous and 53 were synonymous. Eleven nonsynonymous substitutions were convergent, whilst 29 synonymous substitutions were convergent. Synonymous substitutions were therefore more common than nonsynonymous substitutions, although both were equally likely among cases of convergence (Fisher's exact test $P = 0.244$). Convergent, nonsynonymous substitutions were always found in the P1 cistron (A872G, N → S in 9 out of 13 lineages; all positions given relative to the original TEV-eGFP genome, GenBank KC918545), or the remaining 5' end of the eGFP cistron (A1085U, E → V in 2 lineages). Convergent synonymous substitutions were found in P1 (C795U in 6 lineages), HC-Pro (C1927A in 5 lineages), NIa-Pro (U7092C in 2 lineages and A7479C in 8 lineages), and NIb cistrons (A8253C in 8 lineages). The A7479C and A8253C substitutions always occurred together, suggesting synergistic epistasis or possibly even reciprocal sign epistasis, whereas the A7479C and A8253C never occurred together with U7092C, suggesting antagonistic epistasis. However, neither of these two effects was significant given the number of observations (table 2). Not a single mutation was detected in the CP. The overall d_N/d_S ratio was significantly smaller than one (mean ±SD = 0.058 ±0.002; *z*-test $P < 0.001$), suggesting the polyprotein sequence is under purifying selection. Within-population single nucleotide polymorphisms (SNPs) were analyzed for every evolved lineage against their corresponding consensus sequence. Of the SNPs found, 10 were synonymous and six were nonsynonymous. Five out of 13 evolved lineages contained the same synonymous SNP in the HC-Pro cistron (C1879A). None of other synonymous and nonsynonymous SNPs were repeated in the evolved lineages.

Previous evolution experiments with TEV have also shown no genomic convergences

for 1-week passages in *N. tabacum*: after 15 weeks of evolution no substitutions were repeated in different lineages (Bedhomme et al. 2012). These observations suggest that little adaptive evolution might occur when short 1-week serial passages are performed. Furthermore, the specific convergent mutations observed here were not observed in other 1-week passage experiments (Bedhomme et al. 2011; N. Tromas, M.P. Zwart and S.F. Elena, unpublished manuscript). This suggests these convergent mutations may be linked to the insertion of the eGFP gene. To test this possibility, we considered whether the most common mutations (C795U, A872G, A7479C, and A8253C) occurred in lineages of the wild-type TEV put through three 9-week passages in *N. tabacum* (Materials and Methods). In none of such lineages were any of these mutations found, strongly suggesting they are linked to the presence of eGFP. Although we sequenced only a small part of these evolved TEV genomes, we did find one mutation repeated in 3 out of 10 lineages (A6806G, K → E), suggesting there is at least some convergent evolution when the wild-type TEV is put through long passages.

Accumulation, virulence and within-host competitive fitness of evolved lineages

We biologically characterized all evolved lineages, in terms of their virulence and viral accumulation, and measured within-host competitive fitness (*W*; Materials and Methods). There was no effect of passage duration on either viral accumulation at 7 dpi or virulence (fig. 6A-B; table 3). On the other hand, there was a highly significant effect of passage duration on within-host competitive fitness, which increased significantly with passage duration (fig. 6C; table 3).

For the current experimental setup, we *a priori* expect that within-host competitive fitness will be under selection, because we are passaging a virus within a single host at high inoculation doses. Those virus variants that exist at the highest frequency at the end of infection are therefore most likely to be transferred, irrespective of accumulation levels of

the entire population. We did not expect virulence to change, since we do not expect it to be under selection and it is probably not linked to within-host fitness in our model system (Carrasco et al. 2007b). Given that high inoculation doses are used, virus accumulation is not likely to be very important either, so long as it is enough to maintain infection in the next round of passaging. Note that a 1000-fold dilution of an inoculum still causes moderate levels of infection, as shown by the *in vivo* cloning results (table 1). Published data (Carrasco et al. 2007b; Lalić et al. 2011) show that the mutational effects on within-host fitness and virus accumulation at 7 dpi, expressed as the Malthusian growth rate per day (Lalić et al. 2011), are not correlated (Spearman correlation: ρ = 0.034, 19 d.f., P = 0.884; see also fig. S1; Supplementary Material). Therefore, accumulation was not expected to increase as a pleiotropic effect of increases in within-host fitness either. Accumulation will be an important parameter when each virus lineage is evolved in multiple host organisms, as higher accumulation can then lead to a higher frequency of a virus variant in the final population (e.g., Zwart et al. 2010b).

For a plant virus, it is plausible that within-host fitness and accumulation are largely decoupled. Local infection by cell-to-cell movement can be achieved by a small number of virions transported to an adjacent cell (Miyashita and Kishino 2010). Therefore, those virus variants that spread rapidly need not necessarily accumulate a high number of virions per cell. Exclusion is moreover thought to play an important role in infection (Dietrich and Maiss 2004, Folimonova 2012), potentially allowing viruses that spread quickly to reach a high frequency and yet have relatively low accumulation. Furthermore, it should be noted that even if two virus variants have the same level of accumulation late in infection (e.g. 7 dpi), viruses with a high within-host competitive fitness may reach higher levels of accumulation early in infection (e.g. 3 dpi). When a virus rapidly exits the inoculated leaf, this can lead to higher levels of infection before infection levels saturate (Lafforgue et al. 2012; Zwart et al. 2012). Rapid replication and movement might therefore be the mechanisms by which within-host fitness is increased in the evolved lineages, especially if

289 genome size were directly linked to replication. Nevertheless, the key trait to measure from
290 an evolutionary perspective – because it is expected to be under selection in this
291 experimental setup – is competitive within-host fitness.

292 We then considered the relationship between genome size and within-host fitness for the
293 evolved lineages (fig. 7), and also found a highly significant relationship (Spearman
294 correlation: $\rho = -0.877$, 28 d.f., $P < 0.001$). The mean fitness of evolved lineages with
295 genomic deletions was higher than that of the ancestral virus without the heterologous gene
296 (TEV) for ten out of 12 lineages. However, when we performed pair-wise comparisons
297 between TEV and the evolved strains that fixed deletions, no significant differences were
298 found (*t*-test with Holm-Bonferroni correction on the log-transformed *W* values). Statistical
299 power is low when comparing individual lineages because individual-plant level variation is
300 high, a limitation of our experimental system. We must therefore conclude that fitness of
301 the wild-type TEV and evolved strains is similar. This result is, however, congruent with the
302 observation that the convergent single-nucleotide mutations observed in the evolved TEV-
303 eGFP lineages were not observed in wild-type TEV in this study and others (Bedhomme et
304 al. 2011; N. Tromas, M.P. Zwart and S.F. Elena, unpublished manuscript); it supports the
305 suggestion that these mutations are specific for accommodating changes in the TEV-eGFP
306 background and will probably not be beneficial in the wild-type virus background.

307 We then considered the within-host competitive fitness of those lineages without
308 genomic deletions (fig. 6C). We found that 3-week lineages without genomic deletions had
309 a significantly higher fitness than the 1-week lineages (table 3). The 3-week lineages
310 appear to be at the “sweet spot” where the heterologous gene is maintained in many viral
311 lineages while there are concomitantly significant increases in viral fitness. This
312 observation suggests that demography can play an important role in modulating the
313 evolutionary outcome of HGT.

314 Although we have shown simultaneous maintenance of the heterologous gene and
315 increases in fitness, the increases in fitness are probably not related to retention of the

316 eGFP sequence, since it is most unlikely to evolve any beneficial function to the virus. On
317 the other hand, convergent single-nucleotide mutations occurring in TEV-eGFP did not
318 occur in the wild-type virus, suggesting that these mutations are linked to the insertion of
319 eGFP or the additional polyprotein cleavage site. The convergent evolution seen is
320 therefore probably related to the heterologous sequence. Nevertheless, the next step is to
321 perform evolution experiments with viruses carrying functional sequences. Whereas
322 previous reports show heterologous sequences were generally unstable (e.g., Chapman et
323 al. 1992; Chung et al. 2007; Dolja et al. 1993; Guo et al. 1998), here we have identified a
324 demographic condition under which such sequences can be maintained, and under which
325 within-host fitness is still likely to be under selection. Indeed, preliminary results show that
326 the *Cucumber mosaic virus* 2b silencing suppressor is stably maintained in TEV genome
327 when 3-week passages are made (A. Willemsen, M.P. Zwart, S.F. Elena; unpublished
328 data). On the other hand, the use of a heterologous sequence not expected to acquire any
329 function has the advantage of focusing entirely on the process of pseudogenization and
330 gene loss.
331
332 **Alternative models that explain the data**
333 Fixation of genomic deletions and increases in within-host fitness were not observed in the
334 1-week lineages. Moreover, we did not observe the convergent single-nucleotide mutations
335 in the single 1-week lineage fully sequenced, and in another report convergent evolution
336 was not found after 1-week passages in *N. tabacum* (Bedhomme et al. 2012). Why is the
337 outcome of short (1-week) and longer (3- and 9-week) passages so different? We discuss
338 two conceptual models that may explain the data.
339 The first model focuses on the genetic bottlenecks during passaging, and asserts that
340 there may be too much drift for selection to operate effectively during the short-duration
341 passages. Beneficial variation that arises *de novo* can be maintained in the virus
342 population if its frequency increases to levels where it is likely to be sampled in the next

round of infection. Therefore, the time between bottleneck events at the start of infection (i.e. passage duration) might be critical to the outcome of the evolutionary process. For the short 1-week passages, the passage duration is too short to allow any beneficial variation to increase to frequencies where it is likely to be sampled. For intermediate 3-week passages, this model necessarily postulates that the rate at which beneficial single-nucleotide mutations occur is higher than the rate at which recombination leading to the deletion of eGFP occurs, contrary to our expectations. Hence eGFP is lost in few intermediate-duration lineages, while beneficial mutations are fixed in many lineages. This leads to an overall increase in fitness, also for those lineages without deletions. Finally, for the long 9-week lineages, the passage duration is sufficiently long for both beneficial single-nucleotide mutations and recombinations to be selected to frequencies where they are maintained, and in most lineages variants incorporating both predominate. This model predicts a "sweet spot": functional integration of a transgene will be most likely during intermediate-duration passages, where the heterologous sequence is maintained and passages are long enough for selection to act effectively on *de novo* variation. During longer passages, the heterologous sequence will simply be deleted. During shorter passages, the heterologous sequence will be maintained because selection cannot get rid of it. On the other hand, it will be doomed to remain nonfunctional because of the absence of effective selection.

The second model asserts that demographic conditions determine whether the heterologous sequence is detrimental or not. Early in infection, virus titers are low whilst the virus rapidly colonizes most plant tissues (Dolja et al. 1992; Zwart et al. 2012). For longer infections, virus titers will consistently be higher, and the virus continually expands into newly generated apical tissues. As a consequence, the cellular multiplicity of infection (MOI) is also likely to increase over the course of infection (González-Jara et al. 2009; Gutiérrez et al. 2010; Zwart et al. 2013). Different MOIs may lead to different selection pressures acting on the virus population, as exemplified by DIP viruses (e.g. Zwart et al.

2008). This second model predicts conditions under which the heterologous sequence can be maintained in the virus population. Integration of the heterologous sequence will depend on if and how demography affects selection for improvement or new functions that can be re-coded for by the heterologous sequence. However, in the absence of demographic effects on selection for transgene function, a "sweet spot" may still exist given that the strength of genetic drift will increase with shorter passages. Indeed, this may explain the lack of any convergent evolution in 1-week passages in tobacco (Bedhomme et al. 2012).

We think the first model is better supported by our data, because direct competition experiments suggest the fitness of wild-type TEV is higher than that of the TEV-GFP in 1-week passage (*t*-test on the log-transformed *W* values: $t = 3.616$, 4 d.f., $P = 0.022$; see also fig. 6C). Hence, the occurrence of selection for smaller genome size appears to be independent of demographic conditions, although its intensity may not. Under both models, the proposed occurrence of a "sweet spot" for the evolutionary integration of heterologous genes will probably be pathosystem-specific and, moreover, it may be sensitive to the exact virus genotype and environment used. E.g., the stability of a heterologous gene and evolvability of the virus may both change with replication levels. Moreover, in some viruses in which inserts are highly unstable (Chung et al. 2007) there may always be a high number of lineages in which a heterologous gene is deleted, regardless of demography.

Concluding remarks

We found a clear relationship between genome size and within-host fitness for TEV, although other traits such as virus accumulation and virulence appear to be unaffected. This means that within-host fitness is sometimes under strong selection, apparently depending on the duration of infection between consecutive transmission events. This was exemplified by considering leaf-to-leaf evolution during a 9-week infection period, where in some cases deletions had already been fixed early in infection (e.g., leaf 15 of 45) and in

most cases once a deletion occurred, it was maintained and fixed. We then asked whether there are conditions under which viral lineages evolved higher within-host fitness (suggesting that deterministic forces predominate over random forces in evolution) and the heterologous gene is concurrently maintained in most lineages. We found that this is the case for the 3-week passage condition. The within-host fitness of these lineages that had not lost the heterologous gene significantly increased, whilst eight out of 10 lineages did not lose the heterologous gene after 27 weeks of evolution.

We therefore conclude that demographic conditions, in this case the duration of the infection period, can modulate the evolutionary process and possibly create conditions which one would expect to be more conductive for HGT. This suggested demographic "sweet spot" may help to resolve the paradox that viruses can have both streamlined genomes (Belshaw et al. 2007; Lynch 2006) and high levels of HGT (Dolja and Koonin 2011; Koonin and Dolja 2012). Moreover, our results suggest that demography may be critical for evolutionary innovation. However, observing the real-time integration of functional elements in the viral genome by experimental evolution would provide stronger support for these ideas.

These results also have implications for synthetic biology, in particular the generation and employment of viral expression constructs. First, the significant difference in within-host competitive fitness between the 1-week and 3-week lineages without genomic deletions suggests that an evolutionary approach could be used to optimize the fitness of expression vectors. All these lineages have intact eGFP expression (fig. 3), whilst the within-host fitness of the 3-week lineages is higher (table 3). A key question, in this respect, that we have not addressed here is whether the evolved lineages with high fitness also have a higher stability of the eGFP insert. Second, the results suggest that intermediate-duration infections (i.e., approximately 3 weeks) can be suitable for expression of heterologous genes. This result contrasts with those of Dolja et al. (1993), who reported high instability of the β-glucuronidase marker in TEV. Both markers are approximately the

same size. This difference might be explained by the expression strategies: eGFP was cleaved from both P1 and HC-Pro (Zwart et al. 2011) whereas β-glucuronidase was fused to HC-Pro (Dolja et al. 1992), possibly affecting HC-Pro functions and hereby lowering viral fitness. On the other hand, for many of the evolved TEV-eGFP variants, the N-terminal remains of eGFP were fused to HC-Pro (fig. 5a). Another possibility is therefore that expression of β-glucuronidase is harmful for viral replication.

Previous work has demonstrated that passage duration is of crucial importance to evolutionary outcomes for a plant virus (Dolja et al. 1993). Our results confirm the importance of passage duration for the evolution of genomic deletions, but we have also shown high levels of convergent evolution for single-nucleotide substitutions. In some lineages there were no unique mutations – although between-lineages variance was high – making our results qualitatively comparable to other cases of convergent evolution in viruses (Bull et al. 1997). Our results therefore underscore that passage duration is crucial to the outcome of plant virus evolution, and that adaptive and convergent evolution can both be observed in real time during long passages.

Materials and Methods

Plants, virus stocks and infections

N. tabacum were kept in a greenhouse at 24 °C with 16 h light. For the within-host competitive fitness assays and *in vivo* cloning experiment, plants were transferred to a growth chamber at 24 °C with 16 h light after inoculation. To generate a virus stock of the ancestral TEV-eGFP, we first transcribed RNA from the pMTEV-eGFP plasmid (Zwart et al. 2011) as described elsewhere (Carrasco et al. 2007a). The third true leaf of four-week-old *N. tabacum* plants was inoculated with 5 μg of RNA. All systemically infected tissues were harvested nine days post inoculation and virions were purified and stored as described elsewhere (Carrasco et al. 2007a; Zwart et al. 2011).

Serial passage experiments were initiated by inoculating plants with 10 µl of the purified virion stock, by rub inoculating the third true leaf using Carborundum. For passages two and onwards, approximately 500 mg of homogenized infected tissue from the previous passage was diluted in 500 µl phosphate buffer (50 mM potassium phosphate pH 7.0, 3% polyethylene glycol 6000). Fifty µl were then rub-inoculated to the third true leaf. For *in vivo* cloning, plants were inoculated with a 1:1000 dilution in phosphate buffer of 1:1 mixture of infected tissue and phosphate buffer. eGFP fluorescence was observed with a Leica MZ16F stereomicroscope, using a 0.5× objective and GFP2 filters (Leica).

RT-PCR

To determine whether deletions had occurred at the eGFP locus, RNA was extracted from infected tissue using the RNeasy Plant kit (Qiagen). RT was performed using M-MuLV (Fermentas) and random hexamers. PCR was then performed with Taq DNA polymerase (Roche) and primers flanking the eGFP gene: forward primer 5'-CAATTGTTCGCAAGTGTGC-3', reverse primer 5'-ATGGTATGAAGAATGCCTC-3'. 1% agarose gels were used to resolve PCR products. Note that this PCR assay was shown to have a high sensitivity for variants missing the eGFP gene in previous work (Zwart et al. 2011).

Illumina NGS, SNP calling and mapping of genomic deletions

For the sequencing of the experimentally evolved lineages containing deletions at the eGFP site, the virus genome was RT-PCR amplified using Accuscript RT (Agilent Technologies) and Phusion DNA polymerase (Thermo Scientific), with seven independent replicates that were pooled. The virus genome was amplified using three primer sets (set 1: 5'-GCAATCAAGCATTCTACTTC-3' and 5'-ATCCAACAGCACCTCTCAC-3'; set 2: 5'-TTGACGCTGAGCGGAGTGATGG-3' and 5'-AATGCTTCCAGAATATGCC-3'; set 3: 5'-TCATTACAAACAAGCACTTG-3' and 5'-CGCACTACATAGGAGAATTAG-3'), and

equimolar mixtures of PCR products were made. Sequencing was performed at GenoScreen (www.genoscreen.com). Illumina HiSeq2000 2×100bp paired-end libraries with multiplex adaptors were prepared along with an internal PhiX control. Sequencing quality control was performed by GenoScreen, based on PhiX error rate and Q30 values. Artifact filtering and read quality trimming (3' minimum Q20 and minimum read-length of 50 bp) was done using FASTX-Toolkit v0.0.13.2 (hannonlab.cshl.edu/fastx_toolkit/index.html). De-replication of the reads and 5' quality trimming requiring a minimum of Q20 was done using PRINSEQ-lite v0.20.3 (Schmieder and Edwards 2011). Reads containing undefined nucleotides (N) were discarded. Mapping was done against the reference genome TEV-eGFP (GenBank KC918545) with Bowtie 2 v2.1.0 (Langmead and Salzberg 2012), which allows for gapped-read alignments. For every evolved lineage, SNPs were identified using SAMtools' mpileup (Li et al. 2009).

After the pre-mapping step the most common deletions observed were defined manually and for every lineage a new reference sequence was constructed masking each position of the defined deletion with the symbol N. These new reference genomes, together with the cleaned reads, were used as input for the program GapFiller v1.9 (Boetzer and Pirovano 2012), which reliably closes gaps within pre-assembled scaffolds using paired reads. GapFiller fills the gap from each edge in an iterative manner. In our case it partially closed the gaps, base by base, until it could not extend any further given the difference between the *a priori* estimated deletion size and the actual size encountered. At both sides overlapping sequences were manually identified and the ends were joined to reconstruct the new consensus sequences. These were error corrected using the software package Polisher v2.0.8 (available for academic use from the Joint Genome Institute). Accession numbers for the new consensus sequences are GenBank KC918546-KC918555 for 9-week lineages 1-10, GenBank KC918556 for 3-week lineage 3, GenBank KC918557 for 3-week lineage 6 and GenBank KC918558 for 1-week lineage 7.

The consensus sequences for the evolved TEV-eGFP lineages were reconstructed and

504 the cleaned reads were re-mapped against the corresponding consensus for every lineage
505 (Materials and Methods). This re-mapping was efficient with about 93%-98% of the reads
506 without a mate and about 96%-98% of the paired reads mapping exactly one time, for the
507 9-week lineages. For the 3-week lineages this was 91%-96% and 94%-97% respectively.
508 For the 1-week lineage this was 85% and 87% respectively with 15% and 13% of the reads
509 aligning zero times, for which most of these reads probably belong to the eGFP region of
510 full-length TEV-eGFP variant present in the population. To detect other populations with a
511 different deletion in the evolved lineages we extracted all the reads that did not map exactly
512 end-to-end, 100 bp around the deletion site, and mapped these reads against the original
513 reference genome TEV-eGFP. In some regions we found other deletions, but these
514 deletions occurred at very low frequencies ranging from 0.04% to 1.42%. The fact these
515 frequencies are so low, and moreover that some of these deletions start at the same
516 position, suggests that at least some of them could arise due to sequencing error or low-
517 level contamination.
518 For each new consensus, SNPs were re-identified for every lineage using SAMtools'
519 mpileup. Coverage by 35 bp windows distributions was generated for each new
520 consensus. Statistically low-covered regions were searched for approximating the
521 distribution to a normal and calculating a *P*-value per window for a two-sided normal test.
522 However, no regions with significantly lower coverage were identified, confirming that if
523 other deletion variants exist, they are present at low frequencies.
524
525 **Evolution and sequencing of wild-type TEV**
526 In order to obtain wild-type TEV, plants were agroinoculated with pGTEVa (Bedoya et al.
527 2012). Three 9-week passages were performed as described for TEV-eGFP, with 10
528 independent lineages. RNA was extracted from ground plant tissue of the evolved lineages
529 and the 46-2466 and 6376-8779 regions of the TEV genome (GenBank DQ986288) were
530 RT-PCR amplified. The PCR products were sequenced with the 5´-

GCAATCAAGCATTCTACTTC-3′ and 5′-CCTGATATGTTTCCTGATAAC-3′ primers, and the 5′-TCATTACAAACAAGCACTTG-3′ and 5′-AGGCCCAACTCTCCGAAAG-3′ primers, respectively.

Virus accumulation and virulence assay and within-host competitive fitness assays

Four-week-old *N. tabacum* plants were infected with purified virions of the wild-type virus without the heterologous gene (TEV; derived from the pMTEV plasmid (Bedoya and Daròs 2010)), the ancestral virus for the evolution experiments (TEV-eGFP), and TEV marked with the mCherry reporter gene (TEV-mCherry, derived from the pMTEV-mCherry plasmid (Zwart et al. 2011)). Infected tissues were harvested after one week. Virus accumulation for these stocks of TEV, TEV-eGFP and all evolved lineages were then determined. The Invitrap Spin Plant RNA Mini Kit (Stratec Molecular) was used to isolate total RNA. One-step RT-qPCR was then performed using the Primescript RT-PCR Kit II (Takara), in accordance with manufacturer instructions, and a PRISM Sequence Analyser 7500 (Applied Biosystems). Specific primers for the coat protein (CP) were used: 5'-TTGGTCTTGATGGCAACGTG-3' and reverse primer 5'-TGTGCCGTTCAGTGTCTTCCT-3'. 7500 Software version 2.0.4 (Applied Biosystems) was used to analyze the data. The concentration of genome equivalents could then be normalized to that of the sample with the lowest concentration, using phosphate buffer. Ten four-week-old *N. tabacum* plants were then inoculated with 50 µl of these dilutions. Five randomly selected plants were harvested after one week, and virus accumulation was again determined using RT-qPCR, and these values were used as the viral accumulation of TEV, TEV-eGFP, and the evolved lineages. The height of the remaining five plants was measured 3 weeks post inoculation. To measure competitive within-host fitness, TEV, TEV-eGFP and all evolved lineages were again normalized to the sample of the lowest concentration, and 1:1 mixture of genome equivalents was made with TEV-mCherry. This virus has a similar insert size and within-host fitness compared to TEV-eGFP (Zwart et al. 2011), and was used as a common

competitor for all virus strains. The 1:1 mixture of genome equivalents was rub-inoculated in *N. tabacum* plants and infected tissues were harvested after 1 week. The mixture of TEV-eGFP and TEV-mCherry gave a 1:1 ratio of foci of primary infection (Zwart et al. 2011), validating the procedure to quantify genome equivalents and make mixtures. RT-qPCR for the CP was used to determine viral accumulation, whilst independent one-step RT-qPCR reactions were also performed for the mCherry sequence, using specific primers: 5'-CGGCGAGTTCATCTACAAGG-3' and 5'-TGGTCTTCTTCTGCATTACGG-3'. The RT-qPCR method was otherwise identical to that used for the CP. The ratio of the evolved lineage to TEV-mCherry (R) is then: $R = \left(n_{CP} - n_{mCherry}\right)/n_{mCherry}$, where n_{CP} and $n_{mCherry}$ are the RT-qPCR-measured copy numbers of the CP and mCherry, respectively. From the ratio at the start of the experiment (R_0 = 1), we can then estimate the replicative advantage (W, see (Carrasco et al. 2007a)) as: $W = (R_t/R_0)^{1/t}$, where t is the time in days and R_t the virus ratio at the end of the experiment. We consider the replicative advantage as a measure of within-host competitive fitness.

Acknowledgements

The authors thank Alejandro Manzano Marín for his bioinformatics guidance with the Illumina analysis and Francisca de la Iglesia, Paula Agudo and Àngels Pròsper for technical support. This project was made possible through the support of grant 22371 from the John Templeton Foundation to S.F.E. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. Additional support was received from the Spanish Dirección General de Investigación Científica y Técnica grants BFU2012-30805 to S.F.E, JCI2011-10379 to M.P.Z and BIO2011-26741 to J.A.D., and by a Rubicon grant from the Netherlands Organization for Scientific Research (www.nwo.nl) to M.P.Z.

References

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.

Bedhomme S, Lafforgue G, Elena SF. 2012. Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. *Mol Biol Evol* 29:1481-1492.

Bedoya LC, Daròs JA. 2010. Stability of *Tobacco etch virus* infectious clones in plasmid vectors. *Virus Res* 149:234-240.

Bedoya LC, Martínez F, Orzáez D, Daròs JA. 2012. Visual tracking of plant virus infection and movement using a reporter MYB transcription factor that activates anthocyanin biosynthesis. *Plant Physiol* 158:1130-1138.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci USA* 101:4894-4899.

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

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol* 13:R56.

Bull JJ, Badgett MR, Springman R, Molineux IJ. 2004. Genome properties and the limits of adaptation in bacteriophages. *Evolution* 58:692-701.

Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997. Exceptional convergent evolution in a virus. *Genetics* 147:1497-1507.

Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. 2003. Prophage genomics. *Microbiol Mol Biol Rev* 67:238-276.

Carrasco P, Daròs JA, Agudelo-Romero P, Elena SF. 2007a. A real-time RT-PCR assay for

610 quantifying the fitness of *Tobacco etch virus* in competition experiments. *J Virol Methods*
611 139:181-188.
612 Carrasco P, de la Iglesia F, Elena SF. 2007b. Distribution of fitness and virulence effects
613 caused by single-nucleotide substitutions in *Tobacco etch virus*. *J Virol* 81:12979-12984.
614 Chapman S, Kavanagh T, Baulcombe D. 1992. *Potato virus X* as a vector for gene
615 expression in plants. *Plant J* 2:549-557.
616 Chung BN, Canto T, Palukaitis P. 2007. Stability of recombinant plant viruses containing
617 genes of unrelated plant viruses. *J Gen Virol* 88:1347-1355.
618 Chung BYW, Miller WA, Atkins JF, Firth AE. 2008. An overlapping essential gene in the
619 *Potyviridae*. *Proc Natl Acad Sci USA* 105:5897-5902.
620 Cronin S, Verchot J, Haldemancahill R, Schaad MC, Carrington JC. 1995. Long-distance
621 movement factor - a transport function of the potyvirus helper component proteinase. *Plant*
622 *Cell* 7:549-559.
623 Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme
624 environment suggests recombination between unrelated groups of RNA and DNA viruses.
625 *Biol Direct* 7:13.
626 Dietrich C, Maiss E. 2003. Fluorescent labelling reveals spatial separation of potyvirus
627 populations in mixed infected *Nicotiana benthamiana* plants. *J Gen Virol* 84: 2871-2876.
628 Dolja VV, Herndon KL, Pirone TP, Carrington JC. 1993. Spontanceous mutagenesis of a
629 plant potyvirus genome after insertion of a foreign gene. *J Virol* 67:5968-5975.
630 Dolja VV, Koonin EV. 2011. Common origins and host-dependent diversity of plant and
631 animal viromes. *Curr Opin Virol* 1:322-331.
632 Dolja VV, McBride HJ, Carrington JC. 1992. Tagging of plant potyvirus replication and
633 movement by insertion of β-glucuronidase into the viral polyprotein. *Proc Natl Acad Sci*
634 *USA* 89:10208-10212.
635 Folimonova SY. 2012. Superinfection exclusion is an active virus-controlled function that
636 requires a specific viral protein. *J Virol* 86:5554-5561.

González-Jara P, Fraile A, Canto T, García-Arenal F. 2009. The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. *J Virol* 83: 7487-7494.

Guo HS, López-Moya JJ, García JA. 1998. Susceptibility to recombination rearrangement of a chimeric *Plum pox potyvirus* genome after insertion of a foreign gene. *Virus Res* 57:183-195.

Gutiérrez S, Michalakis Y, Blanc S. 2012. Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr Opin Virol* 2:546–555.

Gutiérrez S, Yvon M, Thébaud G, Monsion B, Michalakis Y, Blanc S. 2010. Dynamics of the multiplicity of cellular infection in a plant virus. *PLoS Pathog* 6: e1001113.

Hall JS, French R, Hein GL, Morris TJ, Stenger DC 2001. Three distinct mechanisms facilitate genetic isolation of sympatric *Wheat streak mosaic virus* lineages. *Virology* 282: 230-236.

Huang AS. 1973. Defective interfering viruses. *Annu Rev Microbiol* 27:101-117.

Hughes AL, Friedman R. 2005. Poxvirus genome evolution by gene gain and loss. *Mol Phylogen Evol* 35:186-195.

Koonin EV, Dolja VV. 2012. Expanding networks of RNA virus evolution. *BMC Biol* 10:54.

Lafforgue G, Tromas N, Elena SF, Zwart MP. 2012. Dynamics of the establishment of systemic potyvirus infection: Independent yet cumulative action of primary infection sites. *J Virol* 86:12912-12922.

Lalić J, Cuevas JM, Elena SF. 2011. Effect of host species on the distribution of mutational fitness effects for an RNA virus. *PLoS Genet* 7:e1002378.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-359.

Lee SG, Kim DY, Hyun BH, Bae YS. 2002. Novel design architecture for genetic stability of recombinant poliovirus: the manipulation of G/C contents and their distribution patterns increases the genetic stability of inserts in a poliovirus-based RPS-Vax vector system. *J Virol* 76:1649-1662.

664 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
665 R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map
666 format and SAMtools. *Bioinformatics* 25: 2078-2079.
667 Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, Ghabrial SA, Li G, Yi X, Jiang D. 2011. Widespread
668 horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes.
669 *BMC Evol Biol* 11:91.
670 Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D. 2012. Evolutionary
671 genomics of mycovirus-related dsRNA viruses reveals cross-family horizontal gene transfer
672 and evolution of diverse viral lineages. *BMC Evol Biol* 12:276.
673 Lynch M. 2006. Streamlining and simplification of microbial genome architecture. *Annu Rev*
674 *Microbiol* 60:327-349.
675 Kuo CH and Ochman H. 2009. Deletional bias across the three domains of life. *Genome*
676 *Biol Evol* 1:142-152.
677 Lynch M. 2007. *The origins of genome architecture.* Sunderland, MA: Sinauer Associates,
678 Inc.
679 Majer E, Daròs JA, Zwart MP. 2013. Stability and fitness impact of the visually discernible
680 Rosea1 marker in the *Tobacco etch virus* genome. *Viruses* 5:2153-2168.
681 Marks H, van Duijse JJA, Zuidema D, van Hulten MCW, Vlak JM. 2005. Fitness and
682 virulence of an ancestral white spot syndrome virus isolate from shrimp. *Virus Res* 110:9-
683 20.
684 Miyashita S, Kishino H. 2010. Estimation of the size of genetic bottlenecks in cell-to-cell
685 movement of *Soil-borne wheat mosaic virus* and the possible role of the bottlenecks in
686 speeding up selection of variations in trans-acting genes or elements. *J Virol* 84:1828-1837.
687 Ochman H and Davalos LM. 2006. The nature and dynamics of bacterial genomes. *Science*
688 311:1730-1733.
689 Paar M, Schwab S, Rosenfellner D, Salmons B, Günzburg WH, Renner M, Portsmouth D.
690 2007. Effects of viral strain, transgene position, and target cell type on replication kinetics,

691 genomic stability, and transgene expression of replication-competent *Murine leukemia*
692 *virus*-based vectors. *J Virol* 81:6973-6983.
693 Pathak KB, Nagy PD. 2009. Defective interfering RNAs: Foes of viruses and friend of
694 virologists. *Viruses* 1:895-919.
695 Pijlman GP, van den Born E, Martens DE, Vlak JM. 2001. *Autographa californica*
696 baculoviruses with large genomic deletions are rapidly generated in infected insect cells.
697 *Virology* 283:132-138.
698 Riechmann JL, Lain S, Garcia JA. 1992. Highlights and prospects of potyvirus molecular
699 biology. *J Gen Virol* 73: 1-16.
700 Routh A, Domitrovic T, Johnson JE 2012. Host RNAs, including transposons, are
701 encapsidated by eukaryotic single-stranded RNA virus. *Proc Natl Acad Sci USA* 109:1907-
702 1912.
703 Sacristán S, Malpica JM, Fraile A, García-Arenal F. 2003. Estimation of population
704 bottlenecks during systemic movement of *Tobacco mosaic virus* in tobacco plants. *J Virol*
705 77: 9906-9911.
706 Sakai Y, Kiyotani K, Fukumura M, Asakawa M, Kato A, Shioda T, Yoshida T, Tanaka A,
707 Hasegawa M, Nagai Y. 1999. Accommodation of foreign genes into the *Sendai virus*
708 genome: sizes of inserted genes and viral replication. *FEBS Lett* 456: 221-226.
709 Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic
710 datasets. *Bioinformatics* 27:863-864.
711 Seed KD, Lazinski DW, Calderwood SB, Camilli A. 2013. A bacteriophage encodes its own
712 CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494:489-491.
713 Simon AE, Roossinck MJ, Havelda Z. 2004. Plant virus satellite and defective interfering
714 RNAs: New paradigms for a new century. *Annu Rev Phytopathol* 42:415-437.
715 Thornbury DW, Patterson CA, Dessens JT, Pirone TP. 1990. Comparative sequence of the
716 helper component (HC) region of *Potato virus Y* and a HC-defective strain, *Potato virus C*.
717 *Virology* 178:573-578.

718 Yutin N, Koonin EV. 2012. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large
719 DNA viruses of eukaryotes. *Virol J* 9:161.
720 Zwart MP, Daròs JA, Elena SF. 2011. One is enough: *In vivo* effective population size is
721 dose-dependent for a plant RNA virus. *PLoS Pathog* 7:e1002122.
722 Zwart MP, Daròs JA, Elena SF. 2012. Effects of *Potyvirus* effective population size in
723 inoculated leaves on viral accumulation and the onset of symptoms. *J Virol* 86:9737-9747.
724 Zwart MP, Dieu BTM, Hemerik L, Vlak JM. 2010a. Evolutionary trajectory of *White spot*
725 *syndrome virus* (WSSV) genome shrinkage during spread in Asia. *PLoS ONE* 5:e13400.
726 Zwart MP, Van der Werf W, Georgievska L, Van Oers MM, Vlak JM, Cory JS. 2010b.
727 Mixed-genotype infections of *Trichoplusia ni* larvae with *Autographa californica* multicapsid
728 nucleopolyhedrovirus: Speed of action and persistence of a recombinant in serial passage.
729 *Biol Control* 52:77-83.
730 Zwart MP, Erro E, Van Oers MM, De Visser JAGM, Vlak JM. 2008. Low multiplicity of
731 infection *in vivo* results in purifying selection against baculovirus deletion mutants. *J Gen*
732 *Virol* 89:1220-1224.
733 Zwart MP, Tromas N, Elena SF. 2013. Model-selection-based approach for calculating
734 cellular multiplicity of infection during virus colonization of multi-cellular hosts. *PLoS ONE* 5:
735 e64657.
736

Figures

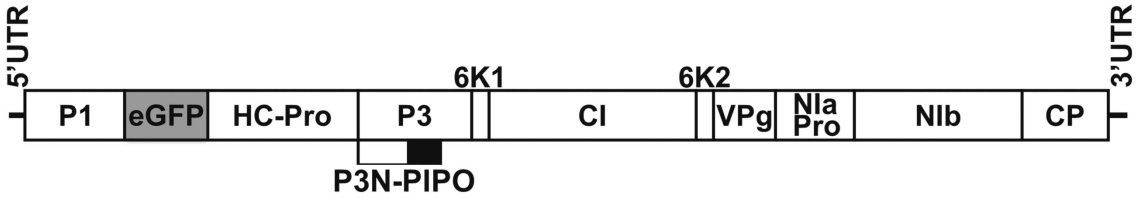
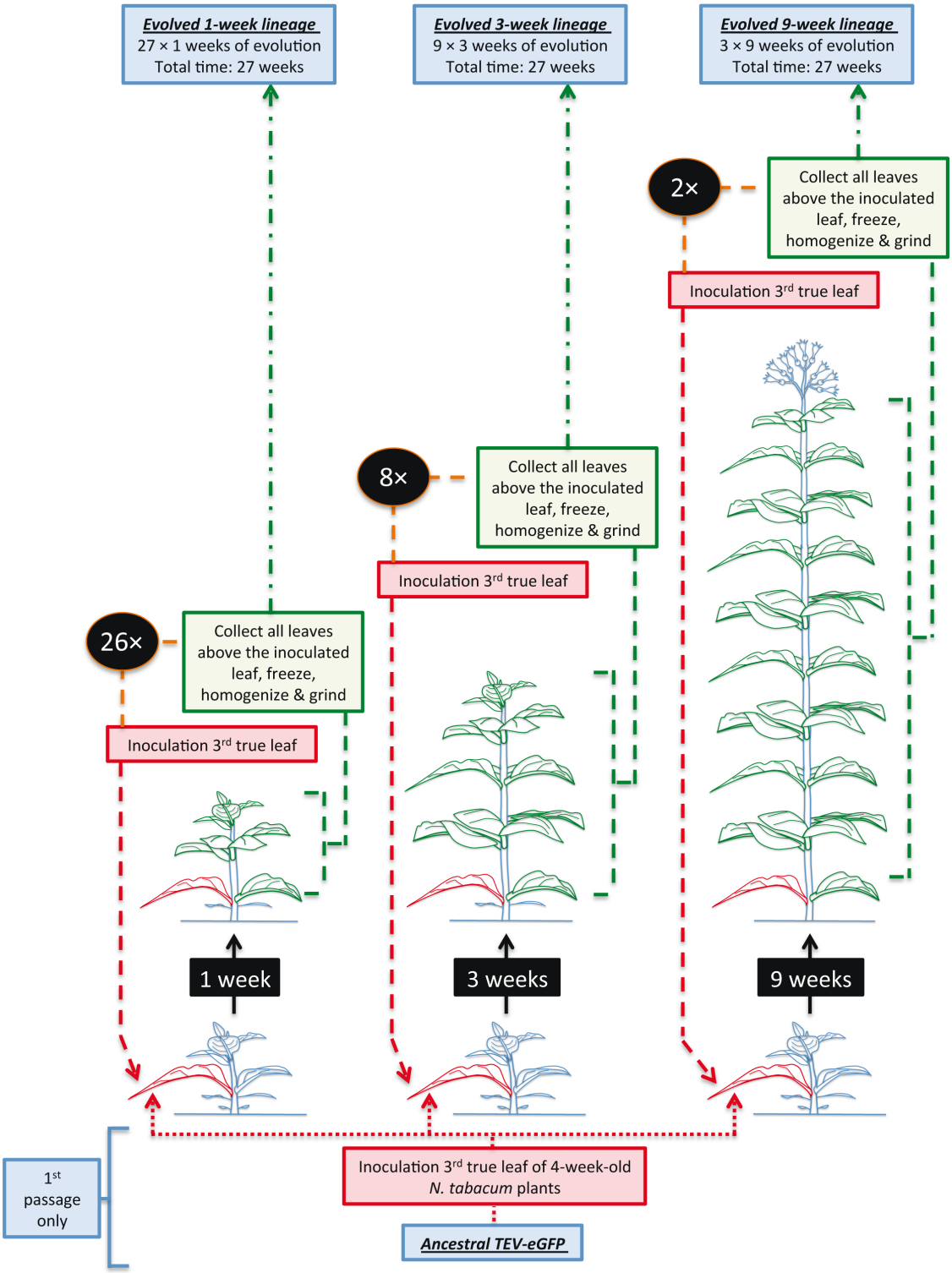


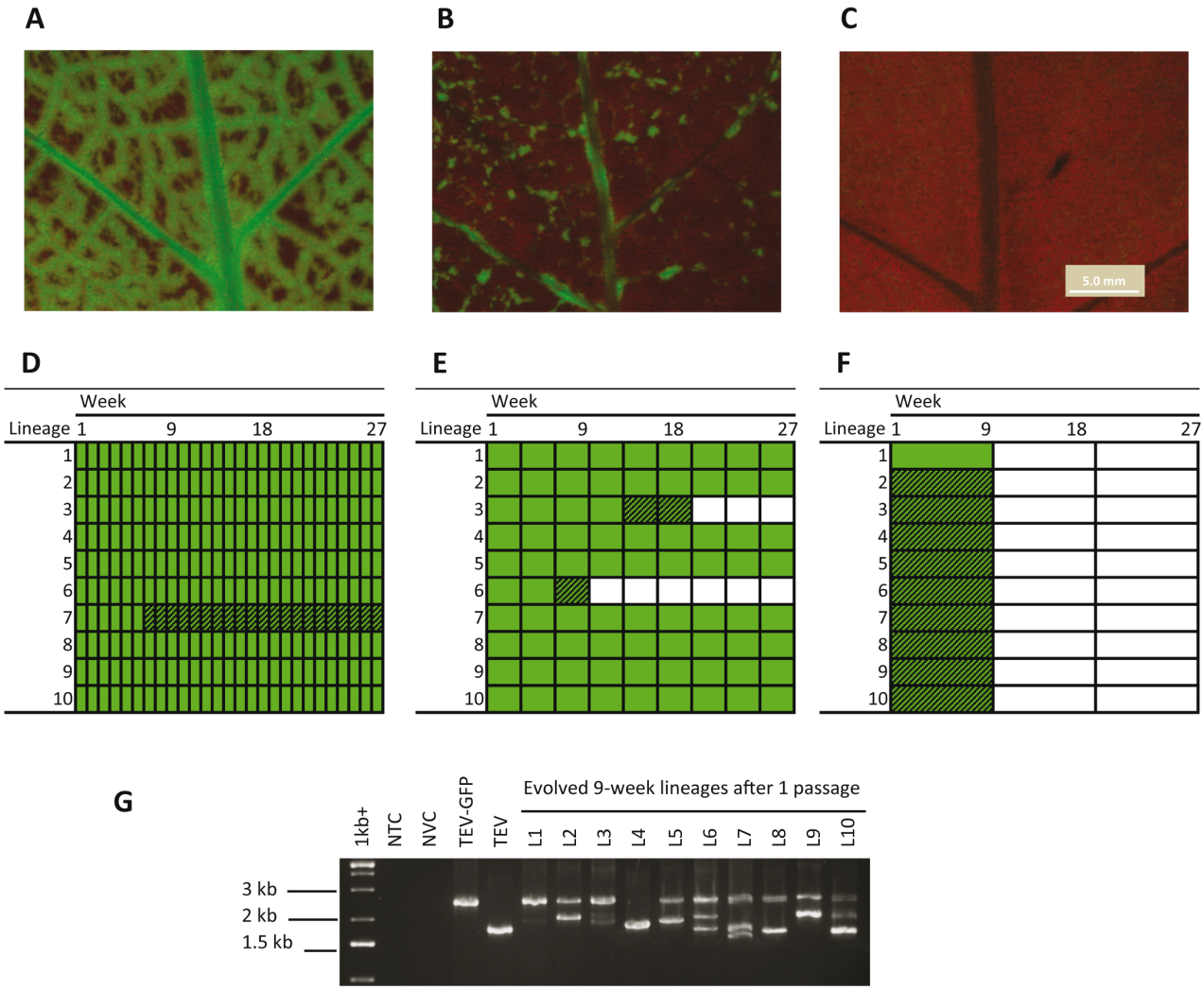
Fig. 1. Scheme of TEV-eGFP. Lines represent the viral 5' and 3' untranslated regions (5' UTR and 3' UTR), the grey box represents eGFP, open boxes represent the viral cistrons P1, HC-Pro, P3, 6K1, CI, 6K2, VPg, NIa-Pro, NIb, and CP, whereas P3N-PIPO is indicated by the lower box.



744

745 **Fig. 2.** Overview of the experimental setup employed in the study. At the start of the serial
746 passage experiment, 4-week-old *N. tabacum* plants were mechanically inoculated with
747 TEV-eGFP in the third true leaf (indicated in red above). At the end of the designated
748 passage duration (1, 3 or 9 weeks), all leaves above the inoculated leaf, which are

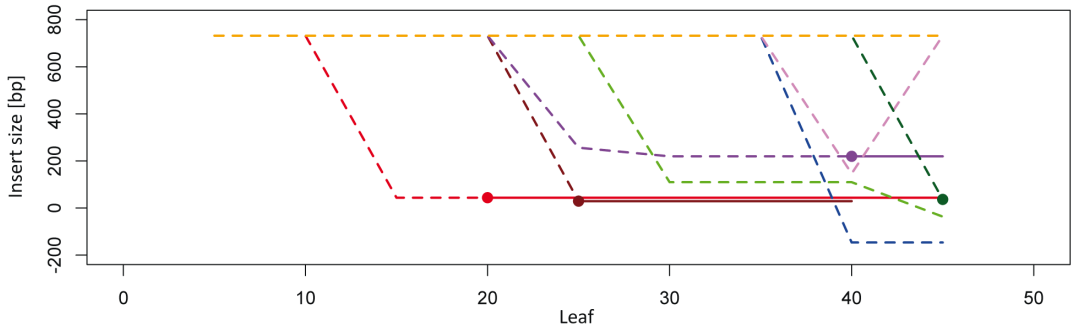
749 indicated in green above, were collected and stored at −80 °C. The frozen tissue was then
750 homogenized, and a sample of the homogenized tissue was ground to a fine powder. For
751 inoculation of subsequent passages, powder was resuspended in inoculation buffer and
752 new *N. tabacum* plants were inoculated. Although the duration of the passages varied (1, 3
753 and 9 weeks), the number of passages was set so that the total time each lineage was
754 present in plants was the same, being 27 weeks for all lineages. For each passage
755 duration used, 10 independent lineages were taken. Note that the figure is only schematic:
756 the distance between leaf layers has been exaggerated, and after 9 weeks of infection *N.*
757 *tabacum* plants are in reality relatively taller than depicted here.



758

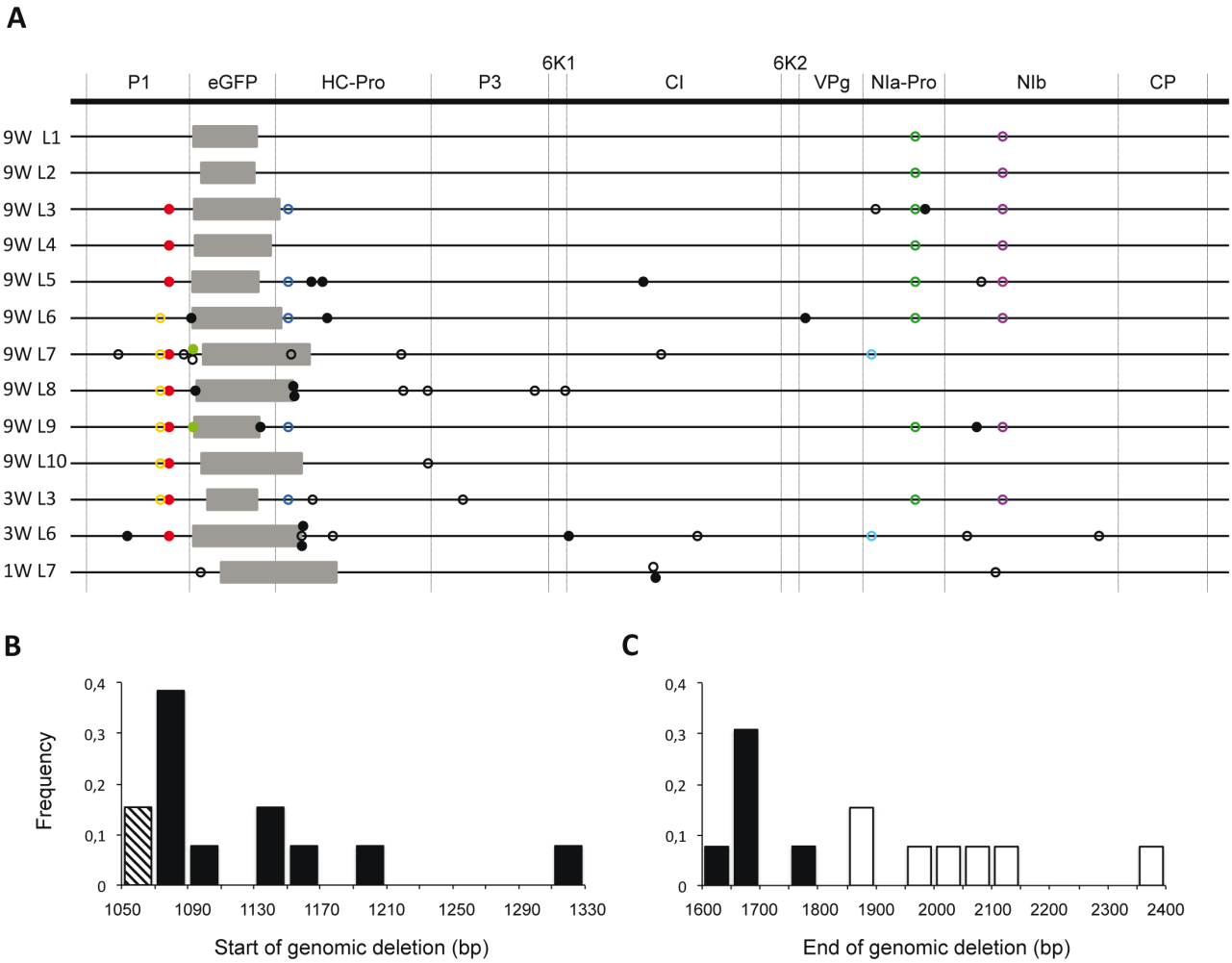
759 **Fig. 3.** Loss of eGFP fluorescence and sequence during serial passages: (A) A *N. tabacum*
760 leaf that is completely symptomatic, indicating heavy virus infection, shown one week post
761 inoculation with an evolved TEV-eGFP lineage with no loss of fluorescence. (B) A
762 completely symptomatic leaf shown one week post inoculation with an evolved TEV-eGFP
763 lineage with a partial loss of fluorescence (1-week passage Lineage 7). (C) A completely
764 symptomatic leaf shown one week post inoculation with an evolved TEV-GFP lineage with
765 a complete loss of fluorescence. (D) Observed fluorescence during serial passage of TEV-
766 eGFP for 1-week passages. For panels D-E, green squares indicate no loss of eGFP
767 fluorescence (as in panel A), hatched green squares indicate a partial loss of fluorescence
768 (as in panel B) in parts or all of the plant, and white squares indicate no fluorescence was

observed in the whole plant (as in panel C). (E) Observed fluorescence for 3-week passages. (F) Observed fluorescence for 9-week passages. (G) An agarose gel with RT-PCR products for deletions in the eGFP locus of TEV-eGFP is shown, for the first passage of the 9-week lineages. 1kb+ indicates the lane with a 1kb+ DNA ladder, NTC is the non-template control and NVC is the non-virus control, a mock-inoculated healthy plant. TEV-eGFP, the ancestral virus for the evolution experiments, and TEV, the wild type virus without the heterologous gene inserted, are included for comparison. Note that in each evolved lineage deletions of eGFP are visible, although their frequency appears to vary. In Lineage 1 the band corresponding to the ancestral virus is still very strong, whereas for Lineage 4 it is not longer visible.



780

781 **Fig. 4.** A single 9-week passage of TEV-eGFP was performed in eight plants, and the
782 smallest observed insert size at the eGFP locus (ordinate) was measured every fifth leaf
783 (abscissae). Dotted lines indicate the full-length eGFP sequence was still detected, solid
784 lines indicate it is no longer detected, and a circle indicates the point at which the full-length
785 sequence is first no longer detected, and each replicate has a different color. Deletions
786 were fixed in four out of eight replicates, and only in single replicate were no deletions
787 detected throughout infection (orange). Only in one case (pink) was a deletion not
788 maintained after it has first been observed (a deletion is observed in leaf 40 but not in leaf
789 45).



790

791 **Fig. 5.** Genome sequences of evolved lineages: (A) NGS data for the evolved lineages
792 with deletions from the serial passage experiment (fig. 3D-F) are given. The names on the
793 left identify lineages (e.g., 9W L1 is the final population of 9-week passage, lineage 1).
794 Grey boxes indicate genomic deletions in the majority variant. Full circles and open circles
795 are non-synonymous and synonymous substitutions, respectively. Black substitutions
796 occur in only one lineage, whereas color-coded substitutions are repeated in two or more
797 lineages. For lineage 1W L7, only a single sequence is represented. However, note that in
798 this lineage another variant with the full-length TEV-GFP genome is present at a frequency
799 47.1%. None of the single-nucleotide mutations in 1W L7 were fixed, suggesting they
800 occurred after the genomic deletion and are present in only one of the two variants present.
801 However, the sequence variation is minimal and that the two variants are present at

approximately same frequencies in this lineage. The single-nucleotide mutations can therefore not be assigned to the full-length or the deleted variant, although they are represented on the deletion variant in the figure. In all other lineages, the full-length virus was not detected and variants with other genomic deletions were present only at very low frequencies (< 1.5%). (B) Histogram of the position of the start of genomic deletions in the evolved lines. For panels B and C, dark lines indicate regions in the eGFP cistron, white lines indicate regions in the viral genome, and hatched lines indicate the regions encompassing two cistrons (i.e., P1 and eGFP in panel B). (C) Histogram of the position of the end of the genomic deletion in evolved lines.

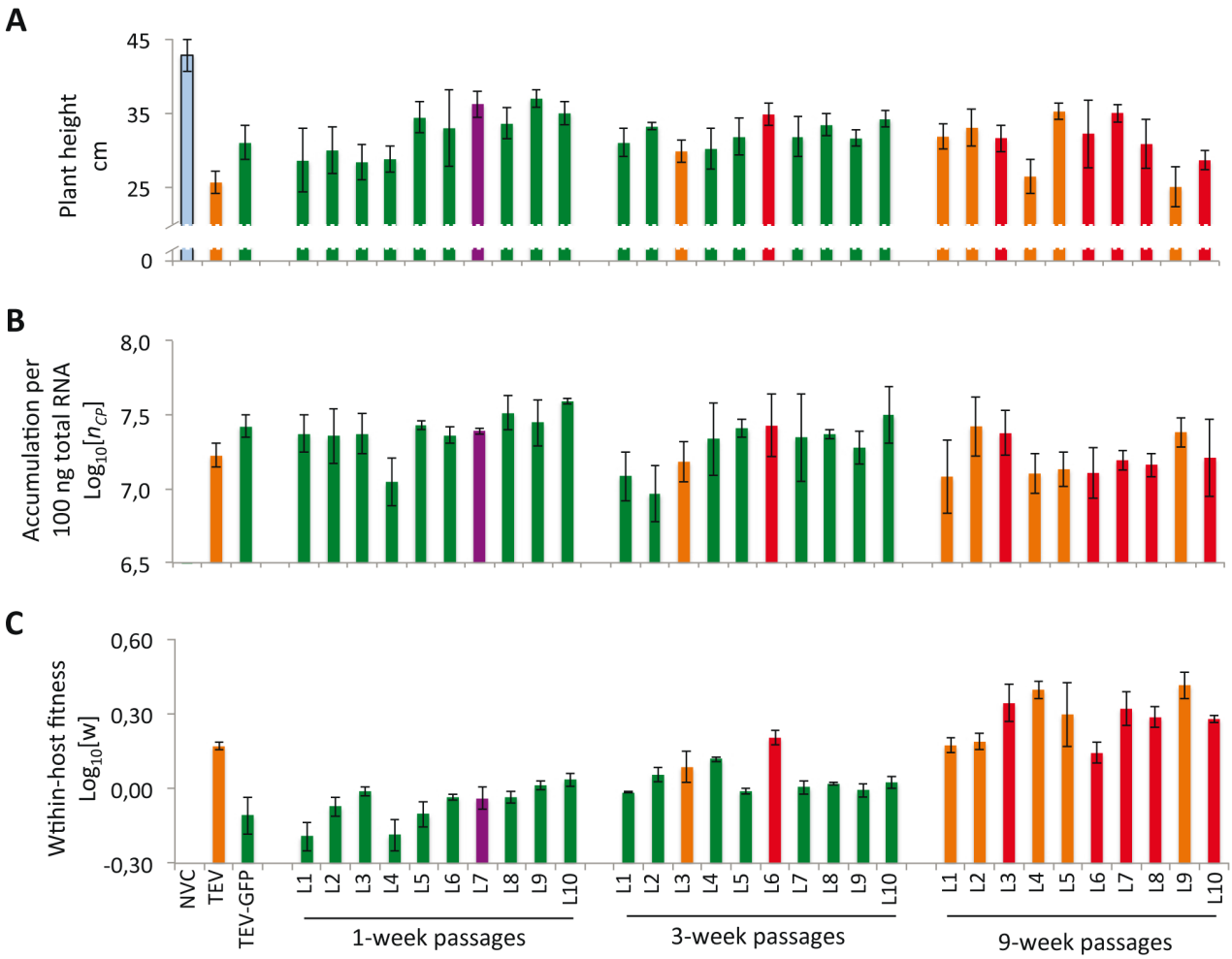
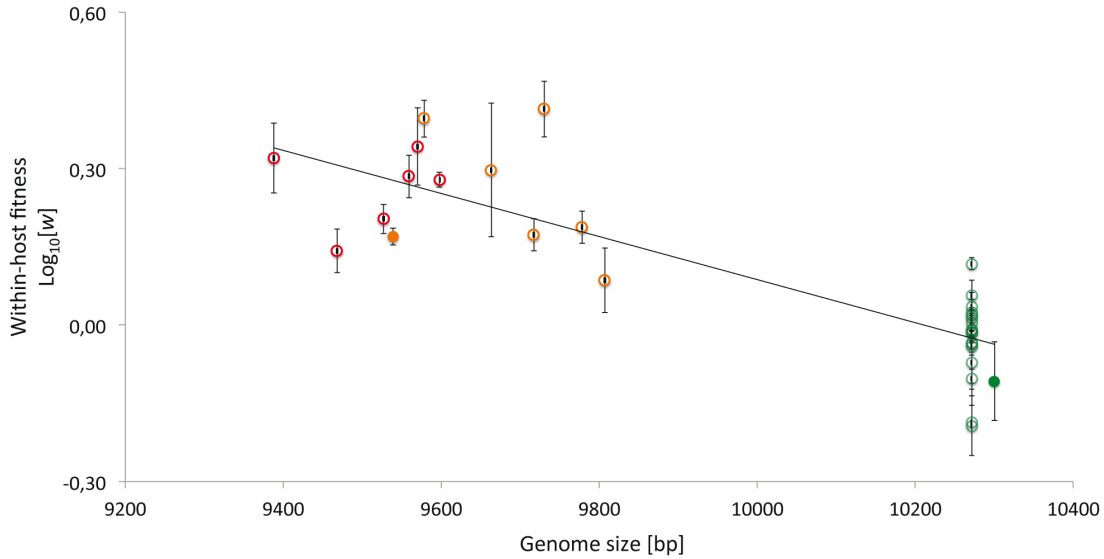


Fig. 6. Virulence, accumulation and within-host fitness of evolved strains are given. (A) The height of control plants (NVC), and plants infected with TEV, TEV-eGFP (the ancestral virus for evolved lineages), and all lineages of evolved viruses is given. We consider the inverse of plant height as a proxy for virulence, though in the absence of significant differences in the data we simply present the raw data. For all panels green columns indicate no deletions in the heterologous gene (eGFP) were detected, orange indicates part or all of the eGFP was not present, and red indicates part of eGFP and the viral HC-Pro cistron are not present. 1W L7 population is marked magenta because it contains a large deletion resulting in a virus apparently unable to infect on its own. (B) Virus accumulation, as measured by RT-qPCR, (C) The replicative advantage (*W*) of the tested virus with respect to a common competitor, TEV-mCherry, is given, as determined by competition

824 experiments and RT-qPCR (Materials and Methods). We consider *W* as an indicator of the
825 within-host competitive fitness.
826



827

828 **Fig. 7.** The relationship between genome size (abscissa) and within-host competitive
829 fitness (ordinate) is given. Green data points indicate no deletions in the heterologous
830 gene (eGFP) were detected, orange indicates part or all of the eGFP was not present, and
831 red indicates part of eGFP and of the viral HC-Pro cistron are not present. The data points
832 for the ancestral TEV-eGFP and TEV have been filled, and TEV-eGFP has been shifted to
833 the right so that it can be easily identified, even though its genome size is the same as
834 other lineages without deletions. A linear regression line has been added only to
835 emphasize the trend in the data. Note that most of the evolved viruses with deletions have
836 a higher fitness than TEV, implicating the observed substitutions with increased fitness.

837

Tables

Table 1. *In vivo* cloning of 1-week passage lineage 7

		Plants			
Experiment	Mean foci	Uninfected	eGFP only	Mixture	No eGFP
1	1.022	9 (0.360)	12 (0.480)	4 (0.160)	0 (0.000)
2	1.514	18 (0.243)	36 (0.486)	20 (0.270)	0 (0.000)

Note.–Two replicate experiments were performed in which *N. tabacum* plants were infected with a 1:1000 dilution of infectious sap of the final evolved population of 1-week passage lineage 7. This population had apparently harbored a virus variant that had lost fluorescence (fig. 3B), although this variant never went to fixation (fig. 3D). Even at low doses, all infected plants (as determined by symptoms 2 weeks after inoculation) contained only the eGFP-expressing virus ("eGFP only"; fluorescence patterns similar to fig. 3A), or both virus variants ("Mixture"; fig. 3B). A symptomatic plant without eGFP expression was never observed ("No eGFP"; fig. 3C), although the low dose resulted in a low mean number of primary infection foci of TEV-eGFP ("mean foci") and many uninfected plants. RT-PCR was performed on five plants judged as being uninfected, infected only with the eGFP variant, or infected with a mixture of the two viruses, and the RT-PCR results were always congruent with microscopic observations. We therefore conclude the virus variant in the population which does not express eGFP has therefore probably lost infectivity, or it has a very low infectivity. This variant is nevertheless surprisingly stable. The 1-week passage lineage 7 (1W L7) population was put through three 1-week passages or one 3-week passage, with ten replicates each. The presence of the virus variant not expressing eGFP could always be deduced from eGFP expression patterns (i.e., fig. 3B).

Table 2. Co-occurrence of single-nucleotide substitutions.

Substitution Combination	$\Pr(A)\Pr(B) = \Pr(A \cap B)$.	$Obs(A \cap B)$.	*P*-value
$A7479C \cap A8253C$.	(8/12)(8/12) = 0.444	8/12 = 0.667	0.1501
$(A7479C \cup A8253C) \cap$ U7092C.	(8/12)(2/12) = 0.111	0/12 = 0	0.3841

Note.–Here we test whether the co-occurrence, or lack thereof, is statistically significant for two groups of single-nucleotide substitution (i.e., substitution combination). $\Pr(A)\Pr(B) = \Pr(A \cap B)$ is the product of the frequency of occurrence of the substitutions, the expected frequency at which we expect to see both substitutions in the absence of any epistatic interactions. We only considered the 3- and 9-week passages, since we do not think selection is acting on the 1-week lineage. $Obs(A \cap B)$ gives the frequency at which the combination was observed, and *P*-value is the significance as determined by comparison of predicted and observed values with an exact Binomial test.

868 **Table 3.** Nested ANOVAs on plant height, accumulation and within-host fitness of evolved
869 lineages.

Trait	Source of variation	SS	d.f.	MS	*F*	*p*
Plant height	Treatment	66.040	2	33.020	0.791	0.463
	Lineage within Treatment	1126.600	27	41.726	7.120	< 0.001
	Error	703.200	120	5.826		
Accumulation	Treatment	41.440	2	20.720	0.391	0.680
	Lineage within Treatment	1431.520	27	53.019	6.998	< 0.001
	Error	909.200	120	7.577		
Fitness[a]	Treatment	1.881	2	0.940	48.534	< 0.001
	Lineage within Treatment	0.523	27	0.019	3.267	< 0.001
	Error	0.356	60	0.006		
Fitness[b]	Treatment	0.100	1	0.100	7.868	0.013
	Lineage within Treatment	0.203	16	0.013	4.096	< 0.001
	Error	0.112	36	0.003		

870 Note.–Fitness[a] is a comparison of all lineages. Fitness[b] compares only those lineages that
871 have not fixed genomic deletions, ten of which are from the 1-week treatment and eight of
872 which are from the 3-week treatment (see fig. 5). ANOVA was used for this comparison of
873 two groups to allow lineage to be nested within treatment. Treatment is the passage
874 duration (1, 3 or 9 weeks).