Contents Category: Plant Viruses – RNA

Estimation of the *in vivo* recombination rate for a plant RNA virus

Nicolas Tromas,[1] Mark P. Zwart,[1] Maïté Poulain,[2] and Santiago F. Elena[1,3]

[1]Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-UPV, 46022 Valencia, Spain

[2]Genoscreen, 1 Rue du Professeur Calmette, 59000 Lille, France

[3]The Santa Fe Institute, Santa Fe, New Mexico 87501, USA

Running title: TEV recombination rate

Keywords: experimental evolution; plant virus; *Tobacco etch potyvirus*; virus evolution; virus replication

Word count: Summary: 214. Full manuscript: 5099. Number of figures: 2. Number of tables: 3.

Correspondence:

Prof. Santiago F. Elena, E-mail: santiago.elena@csic.es

Dr. Nicolas Tromas, E-mail: ntromas@ibmcp.upv.es

IBMCP (CSIC-UPV), Campus UPV CPI 8E, Ingeniero Fausto Elio s/n, 46022 Valencia, Spain; Phone: +34 963 877 895; Fax: +34 963 877 859.

ABSTRACT

Phylogenomic evidence suggests that recombination is an important evolutionary force for potyviruses, one of the larger families of plant RNA viruses. However, mixed-genotype potyvirus infections are marked by low levels of cellular coinfection, precluding template switching and recombination events between virus genotypes during genomic RNA replication. To reconcile these conflicting observations, we have evaluated the *in vivo* recombination rate (r_g) of *Tobacco etch potyvirus* by coinfecting plants with pairs of genotypes marked with engineered restriction sites as neutral markers. The recombination rate was then estimated using two different approaches: (i) a classical approach that assumes recombination between marked genotypes can occur in the whole virus population, rendering an estimate of $r_g = 7.762\times10^{-8}$ events per nucleotide site per generation; (ii) an alternative method that assumes recombination between marked genotypes can occur only in coinfected cells, rendering a much higher estimate of $r_g = 3.427\times10^{-5}$ events per nucleotide site per generation. This last estimate is similar to the TEV mutation rate, suggesting that recombination should be at least as important as point mutation in creating variability. Finally, we compared our mutation and recombination rate estimates to those reported for animal RNA viruses. Our analysis suggests that high recombination rates may be an unavoidable consequence of selection for fast replication at the cost of low fidelity.

45 **INTRODUCTION**

46 Three different mechanisms are ultimately responsible for the observed high genetic
47 variability of plant RNA viruses: mutation, recombination and segment reassortment.
48 Both the roles of mutation (Aranda *et al*., 1997; Malpica *et al*., 2002; Sanjuán *et al*.,
49 2009; Tromas & Elena, 2010) and recombination (Aranda *et al*., 1997; Bonnet *et al*.,
50 2005; Chen *et al*., 2002; Fernández-Cuartero *et al*., 1994; Froissart *et al*., 2005; Martín
51 *et al*., 2009; van der Walt *et al*., 2009) in the evolution of plant RNA viruses have been
52 extensively studied. Although it has been widely suggested that both high mutation and
53 recombination rates are beneficial *per se*, they could also be byproducts of the parasitic
54 lifestyle of viruses that favors fast replication over high fidelity (Belshaw *et al*., 2007;
55 Elena & Sanjuán 2005) and of the modularity of viral RNA genomes (Martin *et al*.,
56 2005; Simon-Loriere & Holmes, 2011).

57 Recombination in RNA viruses can be defined mechanistically as an exchange of
58 genetic material between at least two different viral genomes caused by replicase-driven
59 template switching. Three classes of recombination have been proposed, and are
60 distinguished by the precise mechanism of template switching (Nagy & Simon, 1997;
61 Simon-Loriere & Holmes, 2011; Sztuba-Solińska *et al*., 2011). The first class is base-
62 pairing dependent and therefore requires a perfect alignment of the donor and acceptor
63 RNA molecules. The second recombination class does not require similarity between
64 sequences, but instead requires similarity between RNA secondary structures or *cis*-
65 acting replication elements. The last recombination class combines characteristics of
66 the two first classes. RNA virus recombination frequencies and rates have been
67 estimated with a wide battery of techniques, including assays based on combinations of
68 viruses with genetic markers in cell culture (Kirkegaard & Baltimore, 1986; Levy *et al*.,
69 2004; Reiter *et al*., 2011) and *in vivo* (Bruyere *et al*., 2000; Froissart *et al*., 2005; Pita &

70 Roossinck, 2013; Urbanowicz *et al*., 2005), and by phylogenetic methods (Chare &
71 Holmes, 2005; Codoñer & Elena, 2008; Martín *et al*., 2009; Ohshima *et al*., 2007;
72 Revers *et al*., 1996). Overall, quantitative estimates are highly variable, ranging from
73 1.4×10^{-5} recombination events per site per generation for *Human immunodeficiency*
74 *virus* type 1 (HIV-1) within a host (Neher & Leitner, 2010) to 4×10^{-8} in the case of
75 *Hepatitis C virus* (HCV) (Reiter *et al*., 2011). Froissart *et al.* (2005) reported the first *in*
76 *planta* recombination rate for a plant virus, *Cauliflower mosaic virus* (CaMV). They
77 found that recombination for this pararetrovirus was frequent and estimated its rate to be
78 4×10^{-5} events per nucleotide site and per replication cycle. In a recent report, Pita &
79 Roossinck (2013) described frequent recombination events for *Cucumber mosaic virus*
80 (CMV), although their experimental design did not allow for estimation of
81 recombination rates. Unfortunately, *in planta* estimates for recombination rate of
82 single-stranded positive-sense RNA viruses, the most common amongst known plant
83 viruses, are still missing.
84 Potyviruses represent a particularly interesting model system for studying
85 recombination, since two apparently conflicting observations have been made. On the
86 one hand, phylogenetic evidence suggests that unusually high frequencies of
87 recombination occur (Chare & Holmes, 2005; Revers *et al*., 1996). On the other hand,
88 during mixed-genotype potyvirus infections low levels of cellular coinfection have been
89 observed (Dietrich & Maiss, 2003; N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena,
90 unpublished manuscript). Template switching between virus genotypes can only occur
91 if the genomic RNA of two virus genotypes is being replicated in the same cell, and
92 hence these low levels of cellular coinfection form an impediment to recombination
93 between virus genotypes. These conflicting observations therefore call for an
94 experimental evaluation of the potyvirus recombination rate. Moreover, they raise the

question of what the effects of cellular coinfection, or paucity thereof, may have on estimates of the recombination rate.

In this study, we provide an estimate of the recombination rate of *Tobacco etch virus* (TEV; genus *Potyvirus*, family *Potyviridae*) during a single infection cycle in its primary host *Nicotiana tabacum*. TEV is a prototypical single-strand positive-sense RNA virus that encodes a 346-kDa polypeptide that self-processes into ten mature proteins (Riechmann *et al*., 1992) plus an additional peptide resulting from a +2 frame shift within the third cistron during translation (Chung *et al*., 2008). Our strategy consisted of inoculating equimolar mixtures of pairs of engineered genotypes carrying different neutral markers, characterizing the virus populations resulting from systemic infection and subsequently using different approaches to estimate the recombination rate. We have hereby provided the first estimate of a potyvirus recombination rate, and considered in detail the effects of cellular coinfection on recombination between virus genotypes.

RESULTS AND DISCUSSION

We used the pTEV-7DA (GenBank DQ986288) infectious clone (Dolja *et al*., 1992) as a source for TEV. To analyze the TEV recombination rate, we introduced four neutral genetic markers, in the form of artificial restriction sites, along the TEV genome. New *Asc*I and *Pme*I restriction sites were created at positions 402 and 3735, respectively, whereas natural restrictions sites *Eco*47III and *Sal*I were removed from positions 4969 and 7166, respectively (Fig. 1A). *N. tabacum* plants were then infected with *in vitro* synthesized RNA of each virus variant. We observed no differences in the time until the onset of TEV symptoms among the marked viruses; in all cases symptoms appeared 6 - 7 days post inoculation (dpi). Furthermore, the accumulation of TEV genomes was

measured by RT-qPCR 7 dpi for each engineered genotype (Fig. S1, supplementary material). No differences were observed between the marked viruses (Model II nested ANOVA, $F_{3,10} = 0.216$, $P = 0.883$), thus confirming the neutrality of the markers.

For the actual experiments to measure the recombination frequency, four-week-old *N. tabacum* plants were mechanically inoculated on the third true leaf with 7 µg of an equimolecular mixture of RNA transcripts. The three combinations assayed were *Asc*I/*Pme*I, *Pme*I/*Eco*47III, and *Eco*47III/*Sal*I (Fig. 1). Each combination was inoculated on 5 plants, and after 15 days we harvested all systemically infected leaves. RT-PCR was used to amplify a region containing the two markers (see Materials and Methods). We adapted our PCR protocol to use relatively small amounts of template cDNA and a low number of cycles to avoid the formation of recombinants during this step. In 140 control reactions, we did not find any false positives (see Materials and Methods).

We sequenced individual clones originating from each combination in order to check for recombinants (Table 1). For the *Pme*I/*Eco47*I mixture, one out of the five plants was not infected. The average observed frequency of recombinant genotypes per marker combination varied from zero for the *Eco47*III/*Sal*I combination, where no recombinant genotypes were detected, to 1.85% for the *Asc*I/*Pme*I combination (Table 1). Contrary to our expectations, we did not find a significant relationship between the distance between markers and the frequency of recombinant genotypes (Spearman correlation coefficient, $r = 0.243$, 12 d.f., $P = 0.402$). This result is probably due to low statistical power, given that we sampled few recombinants per marker combination. When frequencies of recombinant genotypes were transformed into recombination rates per site (Kosambi, 1944), the grand mean recombination rate (±1 SEM) for all three combinations was estimated to be $r = (3.388 \pm 2.973) \times 10^{-6}$ recombination events per

nucleotide site (r/s). Marker combination had no effect on recombination rate (Kruskal-Wallis test, $\chi^2 = 4.735$, 3 d.f., $P = 0.094$), suggesting that the frequency of recombinant genotypes did not vary along the TEV genome. Similarly, distance between markers did not have a significant effect on recombination rate (Spearman correlation coefficient, $r = 0.264$, 12 d.f., $P = 0.362$), as has been observed previously (Anderson *et al*., 1998; Froissart *et al*., 2005). Finally, we rescaled the estimate of recombination rate to the more biologically meaningful units of recombination events per nucleotide site and per generation (r/s/g). Given a generation time of 2.91±0.58 (±1 SEM) generations per day (Martínez *et al*., 2011), 15 dpi are equivalent to 43.65±8.70 generations and therefore $r_g = (7.762 \pm 6.985) \times 10^{-8}$ r/s/g.

Factors biasing recombination rate estimation for TEV: Unbalanced mixtures and cellular coinfection

The classic method we used to estimate recombination rate assumes that the two marked virus genotypes are present at the same frequency during infection (i.e., a balanced mixture). For an unbalanced mixture of parentals, it is likely that there will be fewer opportunities for recombination between the two parental genotypes, as has been observed for *Murine leukemia virus* (MuLV) (Anderson *et al.*, 1998). As we can only detect recombination events between the two parental genomes, this may result in a lower observed recombination frequency and therefore considerably lower the estimated recombination rate. Although we carefully quantified and mixed both genotypes to ensure the 1:1 initial ratio, in most individual plants the frequency of the genotypes changed considerably after 15 dpi. However, for the three marker combinations included in the analysis the mean frequency of the genotypes across plants did not deviate significantly from the inoculum mixture (one-sample *t*-test against a value of

0.5: *Asc*I/*Pme*I: $t = 0.638$, 4 d.f., $P = 0.558$; *Pme*I/*Eco*47III: $t = 1.155$, 3 d.f., $P = 0.332$; *Eco*47III/*Sal*I: $t = 0.408$, 4 d.f., $P = 0.704$), which suggests variation may be due to genetic drift rather than to differences in fitness.

To further evaluate the effect of the marker composition we explored the dependence of the estimates of recombination rate on the observed ratio of the less abundant to the most abundant parental genotype. A linear regression showed a significant relationship between this ratio and the recombination rate ($R^2 = 0.395$; $F_{1,12} = 7.839$, $P = 0.016$), confirming that the estimated recombination rate decreased as the ratio of parental genotypes departed from the hypothetical 1:1. Indeed, the regression equation $r = a + b \times RATIO$ can be used to correct for the effect of the unbiased mixtures. In the worse scenario, that is when *RATIO* has the largest possible departure from the 1:1 expectation, the (underestimated) expected recombination rate is simply the intercept of the equation $a = (0.852 \pm 1.578) \times 10^{-6}$ r/s. By contrast, in the optimal situation, namely when the mixture is well balanced (e.g., *RATIO* = 1), then the expected recombination rate is $a + b$. Hence, $r = (2.101 \pm 0.878) \times 10^{-5}$ r/s, or $r_g = (4.813 \pm 2.228) \times 10^{-7}$ r/s/g, a value that is 6.2 times larger than the value reported in the previous section.

A second factor that could bias recombination rate estimates for TEV is the low frequency of cellular coinfection. Recombination between the marked parental viruses can only occur in those cells that are coinfected by both viruses. Therefore detectable recombination will only occur in a fraction of infected cells, even if the parental ratio is perfectly balanced. If the parental ratio is not balanced, there will be fewer coinfected cells and consequently even less detectable recombination events. However, only one virus genotype will be present in most infected cells (Dietrich & Maiss, 2003). In another study, the rate of cellular coinfection was measured by flow cytometry on protoplasts extract from *N. tabacum* plants infected with two TEV variants carrying

195 different fluorescent markers (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena,
196 unpublished manuscript). Cellular coinfection was highest at 10 dpi, being 0.138±0.029
197 % of infected cells. Low levels of cellular coinfection are therefore probably a general
198 characteristic of potyvirus infection, and are probably a common impediment to
199 recombination between different potyvirus variants. In the next section, we develop an
200 approach to incorporate the level of coinfection into the estimation of r_g.
201
202 **Maximum-likelihood estimate of the recombination rate**
203 We developed a more sophisticated maximum-likelihood-based method to estimate
204 recombination rates that takes into account details of the infection process. First, we
205 incorporate a time-varying cellular multiplicity of infection (*MOI*). The *MOI* is the
206 number of virions infecting a cell, which changes over the course of plant virus
207 infection (González-Jara *et al*., 2008; Gutiérrez *et al*., 2010; Zwart *et al*., 2013). A
208 time-dependent mathematical function for *MOI* was determined based on empirical
209 estimates of *MOI* over time (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena,
210 unpublished manuscript) (see Materials and Methods). If the *MOI* and the frequency of
211 parental genotypes are known, these two variables can be used to predict the expected
212 frequency of cellular coinfection for each time point in each plant.
213 Second, we considered the effects of virus expansion in the host plant on the expected
214 frequency of recombinants, for two reasons: (i) recombinants that are generated early in
215 an expanding population can have a large number of descendants and therefore reach
216 high frequencies, (ii) recombinants will probably only reach appreciable frequencies if
217 they occur in cells were replication occurs, since they can then be replicated within the
218 cell and possibly be transmitted to other cells. Plant viruses move locally by cell-to-cell
219 movement, and each cell can only infect those cells adjacent to it (Dietrich & Maiss,

2003; González-Jara *et al*., 2008), and hence a virus variant can be "trapped" if it is surrounded by other variants (Zwart *et al*., 2011). Hence, observable recombination events in this setup can only occur a short time after initial infection of coinfected cells. One approach for capturing the dynamics of virus colonization of the host is to estimate the time-varying cellular contagion rate (*C*): the number of cells infected per infected cell per day. Time-varying estimates of *C* were made for TEV infection of *N. tabacum* based on flow cytometry data on protoplasts (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript), and we used these estimates here to obtain a mathematical function predicting changes in *C* over time (see Materials and Methods). We then developed an infection model incorporating both changes in cellular coinfection (as predicted by the *MOI*) and virus expansion (as embodied by *C*), which predicts the frequency of recombinants. We then used a maximum-likelihood-based method to fit the model to the data by comparing the predicted and observed frequency of recombinants, and hereby estimate the recombination rate (see Materials and Methods). Bootstrapping was used to estimate its 95% confidence interval (CI). Using this approach we estimated the recombination rate to be $r_g = 3.427\times10^{-5}$ r/s/g (95% CI: 1.346×10^{-5} - 5.998×10^{-5}).

Note that the cellular contagion rate is highest early in infection when the *MOI* is still low, further limiting cellular coinfection to a small time window and hereby severely limiting opportunities for recombination between virus genotypes. The effects of such details of the infection process can only be captured by the fitted model, and not by simple corrections for unbalanced mixtures of the rate of coinfection. This estimate of r_g is ca. 71-fold higher than the estimate obtained above ignoring the details of TEV colonization of plant tissues. It should be noted, however, that even the maximum-likelihood-based estimate that considers details of the infection process is probably best

seen as a lower limit of the recombination rate. The model corrects for the effects of the segregation at the cellular level, but any further segregation of the genotypes at lower levels of organization (i.e, replication complexes within the cell) is not considered. Moreover, we have tailored this experimental system for considering homologous recombination, whereas non-homologous recombination events are also known to occur.

Comparison with estimates of recombination rate for other plant RNA viruses

The only previous estimate of recombination rate for a plant virus during real infection conditions was reported by Froissart *et al*. (2005) for CaMV, with r_g have an estimated range of $(2 - 4)\times10^{-5}$ r/s/g. This estimate is very similar to our maximum-likelihood-based estimate using an infection model $(1 - 6)\times10^{-5}$ r/s/g, although Froissart *et al*. (2005) incorporated no corrections for levels of cellular coinfection. However, CaMV reaches high cellular coinfection and *MOI* values (Gutiérrez *et al*., 2010), unlike TEV and other potyviruses (Dietrich & Maiss, 2003; N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript). Such a correction is therefore unlikely to appreciably alter estimates of the CaMV recombination rate, and we propose that a comparison of these two estimates is meaningful and suggests recombination rates are similar.

The recombination rate for the tripartite *Brome mosaic virus* (BMV) has been evaluated in several studies (Bruyere *et al*., 2000; Olsthoorn *et al*., 2002; Urbanowicz *et al*., 2005). Unfortunately, comparison of these studies and our own one is not straightforward for several reasons: (i) none of these studies made a rigorous statistical data analyses and just reported counts of recombinant and parental genomes, (ii) each study focused on a particular genomic region, which may or may not be representative

for the whole genome, and (iii) there are no data from which to estimate the number of generations per day for BMV. Nonetheless, it is still possible to compute the frequency of recombinants and r from the numbers provided in different tables and figures of these studies. Bruyere *et al*. (2000) introduced several restriction site markers in BMV RNA3, and inoculated the non-natural host *Chenopodium quinoa* with balanced mixtures. Variable numbers of local lesions were analyzed for the presence of parental and recombinant genomes. Averaging across the four experimental replicates described in their Table 1, we estimated $r = (3.388 \pm 2.973) \times 10^{-6}$ r/s. In a follow-up study Urbanowicz *et al*. (2005) used a highly similar method to estimate the recombination frequencies for RNA1 and RNA2. Using the data contained in their Figs. 2 and 3, we estimated $r = (1.739 \pm 0.433) \times 10^{-4}$ r/s for RNA1 and $r = (2.490 \pm 0.400) \times 10^{-4}$ r/s for RNA2, averaging across experiments for RNA2. The results for the three segments are homogeneous (Kruskal-Wallis test, $\chi^2 = 2.444$, 2 d.f., $P = 0.295$) and thus we can estimate an average genome-wide recombination rate per site for BMV of $r = (2.104 \pm 0.268) \times 10^{-4}$ r/s. This value is far higher than our estimate of r for TEV, and suggests that TEV recombines less than BMV, a multipartite positive-stranded RNA virus. Unfortunately, to our knowledge, the frequency of cellular coinfection for BMV is still unknown as it is the number of generations per day.

Relationship between mutation and recombination rates for RNA viruses

Tromas & Elena (2010) estimated the TEV point mutation rate to be in the range $(0.475 - 6.299) \times 10^{-5}$ mutations per site per generation. The mutation rate range is therefore, given the uncertainties associated to both estimates, similar to the TEV recombination rate estimated with the infection model. This similarity suggests that mutation and

294 recombination may have a similar impact on the generation of genetic diversity for
295 TEV.
296 If mutation and recombination rates are indeed similar, this has important evolutionary
297 implications. For example, high mutation rates in combination with small population
298 sizes during viral transmission turn on Muller's ratchet, a phenomenon already shown
299 to operate in experimental TEV populations transmitted throughout very dramatic
300 bottlenecks (De la Iglesia & Elena, 2007). A high recombination rate could counteract
301 the effect of Muller's ratchet by recreating mutation-free genomes. The main
302 evolutionary advantage of recombination may be as a mechanism of sex that helps
303 purge deleterious mutations (Muller, 1964; Simon-Loriere & Holmes, 2011). In a
304 similar vein, high recombination rates may have also be advantageous because they
305 speed up the rate of adaptation by bringing together beneficial mutations that otherwise
306 would exist in different genomes, thus minimizing the effect of clonal interference
307 (Fisher, 1930; Muller, 1932; Simon-Loriere & Holmes, 2011). In addition to these
308 "fitness advantage" theories, other models have been brought forward to explain viral
309 recombination. One intriguing possibility is that RNA virus recombination may have
310 evolved as a byproduct of the high nucleotide incorporation rate of viral RNA
311 polymerases (Simon-Loriere & Holmes, 2011): the faster the action of the replicase, the
312 more slippery it becomes. RNA virus high mutation rates are also a likely consequence
313 of the tradeoff between fast replication and accuracy (Belshaw *et al*., 2007; Elena &
314 Sanjuán, 2005), and one would then expect a positive correlation between mutation and
315 recombination rates across RNA viruses.
316 To test this prediction, we searched the literature on RNA viruses for cases in which
317 estimates of both the mutation and recombination rates are available. Unfortunately, the
318 set is just limited to the following seven cases: HIV-1 (Batorsky *et al*., 2011; Jezt *et al*.,

2000), HCV (Reiter *et al*., 2011), *Mouse hepatitis virus* (MHV; Baric *et al*., 1990), MuLV (Anderson *et al*., 1997; Zhuang *et al*., 2006), *Poliovirus* (PV; Duggal *et al.*, 1997; Jarvis & Kirkegaard, 1992; King, 1988), *Spleen necrosis virus* (SNV; Hu & Temin, 1989), and TEV (this study). When more than one estimate existed for one virus, the average was taken. Mutation rates were taken from Sanjuán *et al*. (2010). Note that the low recombination rate of HCV measured by Reiter *et al*. (2011) has been questioned by others (González-Candelas *et al*., 2011). Fig. 2 illustrates the relationship between mutation and recombination rates. A positive correlation exists between both traits, which becomes highly significant if the discordant data point for HCV is removed from the computation (Pearson's $r = 0.963$, 4 d.f., $P = 0.002$). Therefore, these data support the hypothesis that TEV recombination rate may be a side effect of selection for fast but error prone replication, rather than being selected for the fitness advantages it may provide in the long run.

METHODS

Generation of restriction sites as genetic markers. All mutations necessary to create or remove restriction sites were introduced by PCR-directed mutagenesis using the Quickchange® II XL kit (Stratagene) and following the indications given by the manufacturer. Primer pairs (Table 2) for mutagenesis were designed following Stratagene's recommendations. To minimize unwanted errors during the mutagenesis process, the kit incorporates the *Pfu*Ultra™ high fidelity DNA polymerase (Stratagene). The amplification conditions were 1 min at 95ºC (initial denaturation), followed by 18 cycles consisting of 30 s at 95 ºC, 45 s at 65 ºC and 18 min at 68 ºC, and a final extension step of 28 min at 68ºC. PCR products were digested with *Dpn*I (New England Biolabs) to remove the parental methylated strands and transformed into

344 electrocompetent *Escherichia coli* DH5α. At least 15 clones were sequenced to
345 confirm the successful incorporation of desired mutations.
346
347 **Neutrality of restriction markers.** Sequence-validated plasmids containing the
348 corresponding restriction site maker were linearized with *Bgl*II (Takara) and transcribed
349 into 5'-capped RNAs using the SP6 mMESSAGE mMACHINE kit (Ambion Inc).
350 Transcripts were precipitated (1.5 volumes of DEPC-treated water, 1.5 volumes of 7.5
351 M LiCl, 50 mM EDTA), collected and resuspended in DEPC-treated water (Carrasco *et*
352 *al*. 2007). RNA integrity was assessed by gel electrophoresis and concentration was
353 spectrophotometrically determined using a Biophotometer (Eppendorf).
354 To evaluate the neutrality of the four markers, we proceeded as follows. Three four-
355 week-old *N. tabacum* plants were inoculated by abrasion on the third true leaf with 7 µg
356 of transcribed RNA from each individual marker as described elsewhere (Carrasco *et*
357 *al*., 2007). Inoculated plants were placed in a BSL-2 greenhouse at 25 ºC and 16 h
358 light/8 h dark period. RT-qPCR was performed as described elsewhere (Lalić *et al*.,
359 2011).
360
361 **Coinoculation experiments and restriction analysis.** Total RNA was extracted using
362 InviTrap® Spin Plant RNA Mini Kit (Invitek) from the symptomatic leafs for each of
363 five plants 15 dpi. Each plant was analyzed separately, thus providing independent
364 replicates of the recombination rate among pairs of markers. For each combination of
365 restriction makers, the regions of interest was reverse transcribed using the following
366 reaction mixture: 1× RT reaction buffer (Fermentas), 0.2 mM each dNTP, 0.25 µM
367 forward primer (Table 3), 0.2 µL RNase inhibitor, 40 units of Moloney MuLV reverse
368 transcriptase (Fermentas), 10 ng total RNA, and DEPC-treated water to complete 20 µL

reaction volume. Five PCR reactions for each RT reaction were then done using the following reaction mixture: 1× HF buffer (Finzymes), 0.2 mM each dNTP, 0.25 µM each primer (Table 3), 0.5 units of ultra high-fidelity Phusion DNA polymerase (Finzymes), 1 µL DMSO, 3 µL from the reverse-transcription reaction, and DEPC-treated water to complete a reaction volume of 25 µL. The cycling conditions were optimized to limit false-positive recombination events: 1 min at 98 ºC; followed by 25 cycles consisting of 8 s at 98 ºC, 25 s at 57 ºC and 25 s/kb at 72 ºC; and then a final extension of 5 min at 72 ºC. PCR products were gel purified with GeneJET™ Gel Extraction Kit (Fermentas), cloned into the plasmid pUC19/*Sma*I (Fermentas), and used to transform *E. coli*. The analyses of a large number of clones were performed by amplifying the region of interest using colonies-PCRs: 1× Taq buffer (Roche), 0.2 mM each dNTP, 0.25 µM each primer (Table 3), 2 units of Taq DNA polymerase (Roche) and sequencing by Genoscreen (http://www.genoscreen.fr) using BIGDYE 3.1 and a 96-capillars ABI3730XL sequencing system (Applied Biosystems). Sequences were analyzed using GENEIOUS version 4.8 (http://www.geneious.com). The number of clones that rendered useful sequences was 525 (instead of the 672 submitted for sequencing). The number of sequenced clones per plant ranged between 10 and 47, with a median value of 43.

For each pair of markers, four progeny genotypes are expected, the two parentals (Fig. 1B, left column) and the two recombinants (Figure 1B, right column).

Minimizing the formation of recombinant molecules during RT-PCR. A worrisome aspect of PCR based studies of virus variability is the phenomenon of PCR-mediated recombination, or chimera formation (Meyerhans *et al*., 1990). In a recent study, Lahr & Katz (2009) have shown that using the Phusion DNA polymerase, no

394 more than 30 amplification cycles and a low initial template concentration minimized
395 chimera formation. To determine whether the RT-PCR conditions used in our
396 experiments may have favored the formation of chimeras, we first inoculated two four-
397 week-old *N. tabacum* plants with respectively TEV RNAs containing *Pme*I and
398 *Eco*47III markers. After 15 dpi, total RNA from each plant was extracted and mixed in
399 a 1:1 (w:w) ratio. Three serial dilutions of this mixture were made: 50 ng/µL, 5 ng/µL
400 and 0.5 ng/µL. Each of these dilutions was then used as template for an RT-PCR
401 experiment (25 cycles, Phusion DNA polymerase). We failed to get an amplification
402 product for the highest dilution (0.5 ng/µL). For the two others dilutions, PCR products
403 were purified, cloned and transformed into *E. coli*. Forty-eight clones were genotyped
404 by restriction analysis with *Pme*I and *Eco*47III for the 50 ng/µL dilution, finding two
405 false positives. One hundred and seven clones were likewise genotyped for the 5 ng/µL
406 dilution without observing false positives; 33 additional clones were genotyped by
407 sequencing for this dilution, with no false positives observed. No homologous or non-
408 homologous recombination events were observed, confirming these conditions avoid
409 false positive results. These results also confirm that the concentration of template
410 RNA molecules and the number of PCR cycles determine the chances of generating
411 chimera molecules during RT-PCR.
412
413 **Estimation of *r* and r_g.** The frequency of recombinant genotypes (*f*) on each analyzed
414 plant was estimated as *f* = *Recombinants*/(*Recombinants* + *Parentals*). The
415 recombination rate between restriction markers was computed according to Kosambi
416 (1944) equation:
417 $$r = \frac{1}{4L} \ln \frac{1+2f}{1-2f},$$

where L is the physical distance separating the two markers (in nucleotides). The units of the resulting estimate are recombination events per site (r/s). The Kosambi method was chosen to minimize the potential effect that multiple crossovers may have in the inference of r. The independent estimates for each pairs of markers were averaged and the corresponding SEM computed. A genome-wide r was computed averaging all estimates obtained from different combinations of restriction markers and experimental replicates. Note that r was divided by the number of generations (15 days × 2.91 generations per day; Martínez *et al*., 2011) to obtain r_g.

Maximum-likelihood estimate of r_g. To obtain a function for *MOI* over time (t in days), m_t, we fitted a logistic model to *MOI* estimates for TEV infection of *N. tabacum* leaves 3, 5, 6, and 7 on days 3, 5, 7, and 10 dpi (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript), such that $m_t = \kappa \Big/ \left[1 - \left(1 - \frac{\kappa}{m_0}\right) e^{-\upsilon t}\right]$. Here $m_t = m_0$ is the *MOI* at $t = 0$, κ is the maximum attainable value for *MOI* and υ is the initial rate of increase in *MOI*. A grid search was used to estimate $m_0 = 1.12 \times 10^{-3}$, $\kappa = 0.47$ and $\upsilon = 1.02$, by minimizing the residual sum of squares. From m_t the frequency of coinfected cells can be calculated as $c_t = \left(1 - e^{-m_t \alpha}\right)\left(1 - e^{-m_t(1-\alpha)}\right) \Big/ \left(1 - e^{-m_t}\right)$, where α is the LaPlace binomial point estimator of the frequency of a parental genotype. Note that m_t is the *MOI* in all cells, including those that are not infected, and we must therefore divide the fraction of coinfected cells by the fraction of infected cells $\left(1 - e^{-m_t}\right)$.

To obtain a function for the change in C (the cellular contagion rate), we fitted an exponential function to estimates of C for TEV infection of pooled data of different leaves of *N. tabacum* on days 3, 5, 7, and 10 dpi (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript), such that $C_t = C_0 e^{-\gamma t}$. Here, $C = C_0$ if t

442 = 0 and γ is the decay rate. A grid search was used to estimate C_0 = 45.0 and γ = 0.84,
443 by minimizing the residual sum of squares. The difference equation for the number of
444 infected cells is $n_{t+1} = (1 + C_t)n_t$.
445 To estimate the expected final frequency of recombinants, we first estimate the
446 frequency of *de novo* recombinants each day (y) as determined by Kosambi (1944):
447 $y = \frac{\theta}{2}\tanh\left(2Lr_g\right)$, where θ is the number of virus generations per day of infection
448 (2.91). The frequency of recombinants on a given day (f_t) is then $f_t = f_{t-1} + yc_t$. The
449 frequency of recombinants from the previous day is included, since we are considering
450 only those recombinants that occur in the newly infected cells, but assume that newly
451 occurring recombinants will be maintained in the virus population. The window of one
452 day for allowing recombination to occur in newly infected cells is probably
453 conservative given the rapid advance of TEV infection (Dolja *et al*., 1993) and our
454 estimate of θ = 2.91 generations per day (Martinez *et al*., 2011). We can then estimate
455 the expected frequency of recombinants in the final population $f_{15} = \frac{1}{n_{15}}\sum_{t=1}^{15} n_t C_t f_t$, where
456 n_t is the number of infected cells after t days. This equation in essence estimates the
457 mean frequency of recombinants occurring over days, weighted by the amount of
458 expansion occurring on a particular day ($n_t C_t$). In order to estimate r_g, we minimized
459 the negative log-likelihood by a grid search. The likelihood of the corresponding f_{15}
460 value is given by: $L\left(f_{15}\middle|g,h\right) = \binom{g}{h} f^h\left(1-f\right)^{g-h}$, where g is the total number of clones
461 sequenced, and h is the number of sequenced clones that were recombinant.
462

463 **ACKNOWLEDGMENTS**

We thank Francisca de la Iglesia and Àngels Pròsper for excellent technical assistance, José A. Daròs for methodological advice, José M. Cuevas for critical reading of the manuscript, and other lab members for helpful discussions. This work was supported by grants BFU2009-06993 and BFU2012-30805 from the Spanish Secretaría de Estado de Investigación, Desarrollo e Innovación. N.T. was supported by a pre-doctoral fellowship from the former Spanish Ministerio de Ciencia e Innovación.

REFERENCES

Anderson, J. A., Bowman, E. H. & Hu, W. S. (1998). Retroviral recombination rates do not increase linearly with marker distance and are limited by the size of the recombining subpopulation. *J Virol* **72**, 1195-1202.

Aranda, M. A., Fraile, A., Dopazo, J., Malpica, J. M. & García-Arenal, F. (1997). Contribution of mutation and RNA recombination to the evolution of a plant pathogenic RNA. *J Mol Evol* **44**, 81-88.

Baric, R. S., Fu, K., Schaad, M. C. & Stohlman, S. A. (1990). Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. *Virology* **177**, 646-656.

Batorsky, R., Kearney, M. F., Palmer, S. E., Maldarelli, F., Rouzine, I. M. & Coffin, J. M. (2011). Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc Natl Acad Sci USA* **108**, 5661-5666.

Belshaw, R., Gardner, A., Rambaut, A. & Pybus, O. G. (2007). Pacing a small cage: mutation and RNA viruses. *Trends Ecol Evol* **23**, 188-193.

Bonnet, J., Fraile, A., Sacristán, S., Malpica, J. M. & García-Arenal, F. (2005). Role of recombination in the evolution of natural populations of *Cucumber mosaic virus*, a tripartite RNA plant virus. *Virology* **332**, 359-368.

Bruyere, A., Wantroba, M., Flasinski, S., Dzianott, A. & Bujarski, J. J. (2000). Frequent homologous recombination events between molecules of one RNA component in a multipartite RNA virus. *J Virol* **74**, 4214-4219.

Carrasco, P., Daròs, J. A., Agudelo-Romero, P. & Elena, S. F. (2007). A real-time RT-PCR assay for quantifying the fitness of *Tobacco etch virus* in competition experiments. *J Virol Meth* **139**, 181-188.

Chare, E. R. & Holmes, E. C. (2005). A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch Virol* **151**, 933-946.

Chen, Y. K., Goldbach, R. & Prins, M. (2002). Inter- and intramolecular recombinations in the *Cucumber mosaic virus* genome related to adaptation to *Alstroemeria*. *J Virol* **76**, 4119-4124.

Chung, B. Y. W., Miller, W. A., Atkins, J. F. & Firth, A. E. (2008). An overlapping essential gene in the *Potyviridae*. *Proc Natl Acad Sci USA* **15**, 5897-5902.

Codoñer, F. M. & Elena, S. F. (2008). The promiscuous evolutionary history of the family *Bromoviridae*. *J Gen Virol* **89**, 1739-1747.

De la Iglesia, F. & Elena, S. F. (2007). Fitness declines in *Tobacco etch virus* upon serial bottleneck transfers. *J Virol* **10**, 4941-4947.

Dietrich, C. & Maiss, E. (2003). Fluorescent labelling reveals spatial separation of potyvirus populations in mixed infected *Nicotiana benthamiana* plants. *J Gen Virol* **84**, 2871-2876.

Dolja, V. V., Herndon, K. L., Pirone, T. P. & Carrington, J. C. (1993). Spontaneous mutagenesis of a plant potyvirus genome after insertion of a foreign gene. *J Virol* **67**, 5967-5975.

Dolja, V. V., McBride, H. J. & Carrington, J. C. (1992). Tagging of plant potyvirus replication and movement by insertion of β-glucoronidase into the viral polyprotein. *Proc Natl Acad Sci USA* **89**, 10208-10212.

Duggal, R., Cuconati, A., Gromeier, M. & Wimmer, E. (1997). Genetic recombination of poliovirus in a cell-free system. *Proc Natl Acad Sci USA* **94**, 13786-13791.

Elena, S. F. & Sanjuán, R. (2005). Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. *J Virol* **79**, 11555-11558.

Fernández-Cuartero, B., Burgyán, J., Aranda, M. A., Salánki, K., Moriones, E. & García-Arenal, F. (1994). Increase in relative fitness of a plant virus RNA associated with its recombinant nature. *Virology* **203**, 373-377.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. University Press, Oxford, UK.

Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S. & Michalakis, Y. (2005). Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. *PLoS Biol* **3**, e89.

González-Candelas, F., López-Labrador, F. X. & Bracho, M. A. (2011). Recombination in *Hepatitis C virus*. *Viruses* **3**, 2006-2024

González-Jara, P., Fraile, A., Cantó, 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.

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.

Hu, W. S. & Temin, H. M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particles: pseudodiploidy and high rate of genetic recombination. *Proc Natl Acad Sci USA* **87,** 1556-1560.

Jarvis, T. C. & Kirkegaard, K. (1992). Poliovirus RNA recombination: mechanistic studies in the absence of selection. *EMBO J* **11**, 135-145.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D. & Dougherty, J. P. (2000). High rate of recombination throughout the *Human immunodeficiency virus* type 1 genome. *J Virol* **74**, 1234-1240.

King, A. M. (1988). Preferred sites of recombination in poliovirus RNA: an analysis of 40 intertypic cross-over sequences. *Nucl Acids Res* **16**, 11705-11723.

Kirkegaard, K. & Baltimore, D. (1986). The mechanism of RNA recombination in poliovirus. *Cell* **47**, 433-443.

Kosambi, D. D. (1944). The estimation of map distance from recombination values. *Ann Eugen* **12**, 172-175.

Lahr, D. J. G. & Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* **47**, 857-866.

Lalić, J., Cuevas, J. M. & Elena, S. F. (2011). Effect of host species on the distribution of mutational fitness effects for an RNA virus. *PLoS Genet* **7**, e1002378.

Levy, D. N., Aldrovandi, G. M., Kutsch, O. & Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. *Proc Natl Acad Sci USA* **101**, 4204-4209.

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W. & García-Arenal, F. (2002). The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505-1511.

Martin, D. P., van der Walt, E., Posada, D. & Rybicki, E. P. (2005). The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet* **1,** e51.

Martín, S., Sambade, A. Rubio, L., Vives, M. C., Moya, P., Guerri, J., Elena, S. F. & Moreno, P. (2009). Contribution of recombination and selection to molecular evolution of *Citrus tristeza virus*. *J Gen Virol* **90,** 1527-1538.

Martínez, F., Sardanyés, J., Elena, S. F. & Daròs, J. A. (2011). Dynamics of a plant RNA virus intracellular accumulation: stamping machine vs. geometric replication. *Genetics* **188**, 637-464.

Meyerhans, A., Vartanian, J. P. & Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucl Acids Res* **18,** 1687-1691.

Muller, H. J. (1932). Some genetic aspects of sex. *Am Nat* **66,** 118-138.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mut Res* **1,** 2-9.

Nagy, P. D. & Simon, A. E. (1997). New insights into the mechanisms of RNA recombination. *Virology* **235**, 1-9.

Neher, R. A. & Leitner, T. (2010). Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* **6**, e1000660.

Ohshima, K., Tomitaka, Y., Wood, J. T., Minematsu, Y., Kajiyama, H., Tomimura, K. & Gibbs, A. J. (2007). Pattern of recombination in *Turnip mosaic virus* genomic sequences indicate hotspots of recombination. *J. Gen. Virol*. **88**, 298-315.

Olsthoorn, R. C. L., Bruyere, A., Dzianott, A. & Bujarski, J. J. (2002). RNA recombination in *Brome mosaic virus*: effects of strand-specific stem-loop inserts. *J Virol* **76**, 12654-12662.

Pita, J. S. & Roossinck, M. J. (2013). Fixation of emerging interviral recombinants in *Cucumber mosaic virus* popualtions. *J Virol* **87**, 1264-1269.

Reiter, J., Pérez-Vilaró, G., Scheller, N., Mina, L. B., Díez, J. & Meyerhans, A. (2011). *Hepatitis C virus* RNA recombination in cell culture. *J Hepatol* **55**, 777-783.

Revers, F., Le Gall, O., Candresse, T., Le Romancer, M. & Dunez, J. (1996). Frequent occurrence of recombinant potyvirus isolates. *J Gen Virol* **77**, 1953-1965.

Riechmann, J. L., Laín, S. & García, J. A. (1992). Highlights and prospects of potyvirus molecular biology. *J Gen Virol* **73**, 1-16.

Sanjuán, R., Agudelo-Romero, P. & Elena, S. F. (2009). Upper-limit mutation rate for a plant RNA virus. *Biol Lett* **5**, 394-396.

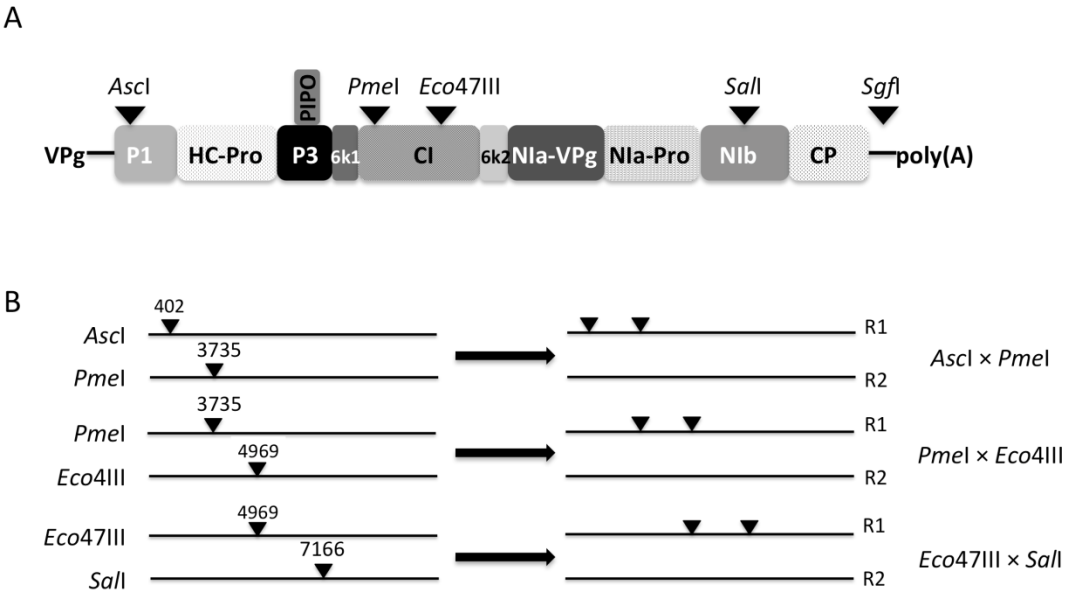
Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. (2010). Viral mutation rates. *J Virol* **84**, 9733-9748.

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

Sztuba-Solińska, J., Urbanowicz, A., Figlerowiz, M. & Bujarski, J. J. (2011). RNA-RNA recombination in plant virus replication and evolution. *Annu Rev Phytopathol* **49**, 415-443.

606 **Tromas, N. & Elena, S. F. (2010).** The rate and spectrum of spontaneous mutations in
607 a plant RNA virus. *Genetics* **185**, 983-989.
608 **Urbanowicz, A., Alejska, M., Formanowicz, P., Błażewicz, J., Figlerowicz, M. &**
609 **Bujarski, J. J. (2005).** Homologous crossover among molecules of *Brome mosaic*
610 *bromovirus* RNA1 or RNA2 segments *in vivo*. *J Virol* **79**, 5732-5742.
611 **Van der Walt, E., Rybicki, E. P., Varsani, A., Polston, J. E., Billharz, R.,**
612 **Donaldson, L., Monjane, A. L. & Martin, D. P. (2009).** Rapid host adaptation by
613 extensive recombination. *J Gen Virol* **90**, 734-746.
614 **Zhuang, J., Mukherjee, S., Ron, Y. & Dougherty, J. P. (2006).** High rate of genetic
615 recombination in *Murine leukemia virus*: implications for influencing proviral
616 ploidy. *J Virol* **80**, 6706-6711.
617 **Zwart, M. P., Daròs, J. A. & Elena, S. F. (2011).** One is enough: *in vivo* effective
618 population size is dose-dependent for a plant RNA virus. *PLoS Pathog* **7,**
619 e1002122.
620 **Zwart, M. P., Tromas, N. & Elena, S. F. (2013).** Model-selection-based approach for
621 calculating cellular multiplicity of infection during virus colonization of multi-
622 cellular hosts. *PLoS ONE* **8**, e64657.
623

624 **Fig. 1.** (A) Location of the different restriction site markers in TEV genome. (B)
625 Expected restriction profile for each pair of markers. The left column shows the two
626 parental genotypes and the right column the two recombinant ones.



627

628

Fig. 2. Relationship between mutation and recombination rates for seven RNA viruses and retroviruses.

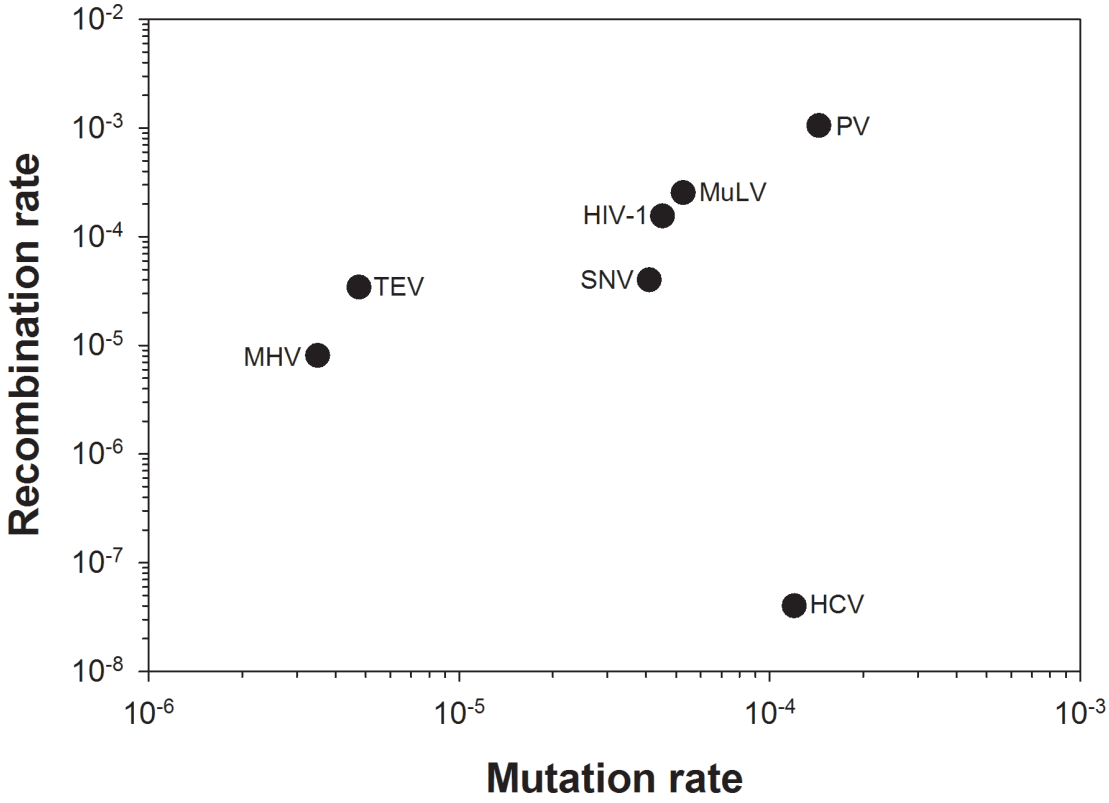


Table 1. Observed frequency of homologous recombinants (*f*) and estimated recombination rate per site (*r*).

Combination	Fragment size (nt)	Plants analyzed	*f* (±1 SEM) (%)	*r* (±1 SEM) ($\times 10^{-6}$)
*Asc*I/*Pme*I	3334	5	1.854±0.125	5.561±0.376
*Pme*I/*Eco*47III	1234	4	0.568±0.084	4.604±0.684
*Eco*47III/*Sal*I	2197	5	0	0

Table 2. Artificial restriction sites engineered as genetic markers for this study

Restriction enzyme	Cistron	Genomic position for the cut	Mutagenesis primer (5' to 3')*
*Asc*I	P1	402	TTATCTTGGTCG***G***CGCGC***C***CTCACCCATGGC
*Pme*I	CI	3735	AGCCTTCCTGGAGTCAC***G***TTTAA***A***CAATGGTGGAACAACCA
*Eco*47III	CI	4969	AGTCATACATGACAAGCTGAA***A***CG***T***TTTAAGCTACACACTTGTGAG
*Sal*I	NIb	7166	GATGGGAGCATATAAGCCAA***CC***CGACTTAATAGAGAGGCG

*Restriction sites are underlined. Mutagenized sites are indicated with bold cursive type.

Table 3. Primers used to amplify the region containing the pair of restriction sites

Combination	TEV genome position (5')	Sequence (5' to 3')
*Asc*I/*Pme*I	46	GCAATCAAGCATTCTACTTC
	3894	ATCCAACAGCACCTCTCAC
*Pme*I/*Eco*47III	3541	TTGACGCTGAGCGGAGTGATGG
	5275	CTATTGATGCATGCTAGAGTC
*Eco*47III/*Sal*I	4972	TTAAGCTACACACTTGTGAGAC
	7394	TTCTTTCTTCTTGCCTTTG

Supplementary Material for "Estimation of the *in vivo* recombination rate for a plant RNA virus"

Nicolas Tromas, Mark P. Zwart, Maïté Poulain, and Santiago F. Elena

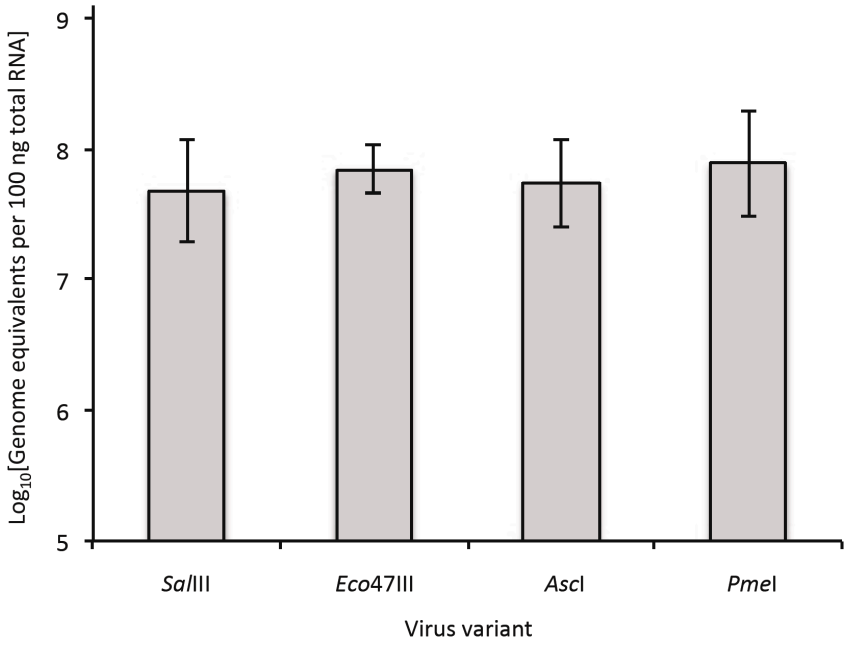


Fig. S1. Mean $\log_{10}$-transformed accumulation levels of the four marked virus variants, as measured by RT-qPCR. Error bars represent 1 standard deviation.