Estimation of the \textit{in vivo} recombination rate for a plant RNA virus

Nicolas Tromas,\textsuperscript{1} Mark P. Zwart,\textsuperscript{1} Maïté Poulain,\textsuperscript{2} and Santiago F. Elena\textsuperscript{1,3}

\textsuperscript{1}Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-UPV, 46022 Valencia, Spain
\textsuperscript{2}Genoscreen, 1 Rue du Professeur Calmette, 59000 Lille, France
\textsuperscript{3}The Santa Fe Institute, Santa Fe, New Mexico 87501, USA

\textbf{Running title}: TEV recombination rate

\textbf{Keywords}: experimental evolution; plant virus; \textit{Tobacco etch potyvirus}; virus evolution; virus replication


\textbf{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.
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 \( r_g = 7.762 \times 10^{-8} \) 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 \( r_g = 3.427 \times 10^{-5} \) 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.
INTRODUCTION

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

Recombination in RNA viruses can be defined mechanistically as an exchange of genetic material between at least two different viral genomes caused by replicase-driven template switching. Three classes of recombination have been proposed, and are distinguished by the precise mechanism of template switching (Nagy & Simon, 1997; Simon-Loriere & Holmes, 2011; Sztuba-Solińska et al., 2011). The first class is base-pairing dependent and therefore requires a perfect alignment of the donor and acceptor RNA molecules. The second recombination class does not require similarity between sequences, but instead requires similarity between RNA secondary structures or cis-acting replication elements. The last recombination class combines characteristics of the two first classes. RNA virus recombination frequencies and rates have been estimated with a wide battery of techniques, including assays based on combinations of viruses with genetic markers in cell culture (Kirkegaard & Baltimore, 1986; Levy et al., 2004; Reiter et al., 2011) and in vivo (Bruyere et al., 2000; Froissart et al., 2005; Pita &
Roossinck, 2013; Urbanowicz et al., 2005), and by phylogenetic methods (Chare & Holmes, 2005; Codoñer & Elena, 2008; Martin et al., 2009; Ohshima et al., 2007; Revers et al., 1996). Overall, quantitative estimates are highly variable, ranging from $1.4 \times 10^{-5}$ recombination events per site per generation for Human immunodeficiency virus type 1 (HIV-1) within a host (Neher & Leitner, 2010) to $4 \times 10^{-8}$ in the case of Hepatitis C virus (HCV) (Reiter et al., 2011). Froissart et al. (2005) reported the first in planta recombination rate for a plant virus, Cauliflower mosaic virus (CaMV). They found that recombination for this pararetrovirus was frequent and estimated its rate to be $4 \times 10^{-5}$ events per nucleotide site and per replication cycle. In a recent report, Pita & Roossinck (2013) described frequent recombination events for Cucumber mosaic virus (CMV), although their experimental design did not allow for estimation of recombination rates. Unfortunately, in planta estimates for recombination rate of single-stranded positive-sense RNA viruses, the most common amongst known plant viruses, are still missing.

Potyviruses represent a particularly interesting model system for studying recombination, since two apparently conflicting observations have been made. On the one hand, phylogenetic evidence suggests that unusually high frequencies of recombination occur (Chare & Holmes, 2005; Revers et al., 1996). On the other hand, during mixed-genotype potyvirus infections low levels of cellular coinfection have been observed (Dietrich & Maiss, 2003; N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript). Template switching between virus genotypes can only occur if the genomic RNA of two virus genotypes is being replicated in the same cell, and hence these low levels of cellular coinfection form an impediment to recombination between virus genotypes. These conflicting observations therefore call for an 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 *AscI* and *PmeI* restriction sites were created at positions 402 and 3735, respectively, whereas natural restrictions sites *Eco47III* and *SalI* 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 *AscI/PmeI*, *PmeI/Eco47III*, and *Eco47III/SalI* (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 *PmeI/Eco47I* 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 *Eco47III/SalI* combination, where no recombinant genotypes were detected, to 1.85% for the *AscI/PmeI* 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 ($\pm 1$ 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 \( r_g = (7.762±6.985)\times10^{-8} \) 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: *Ascl/Pmel*: $t = 0.638$, 4 d.f., $P = 0.558$; *Pmel/Eco47III*: $t = 1.155$, 3 d.f., $P = 0.332$;
*Eco47III/SalI*: $t = 0.408$, 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

8
different fluorescent markers (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript). Cellular coinfection was highest at 10 dpi, being 0.138±0.029 % of infected cells. Low levels of cellular coinfection are therefore probably a general characteristic of potyvirus infection, and are probably a common impediment to recombination between different potyvirus variants. In the next section, we develop an approach to incorporate the level of coinfection into the estimation of $r_g$.

**Maximum-likelihood estimate of the recombination rate**

We developed a more sophisticated maximum-likelihood-based method to estimate recombination rates that takes into account details of the infection process. First, we incorporate a time-varying cellular multiplicity of infection (MOI). The MOI is the number of virions infecting a cell, which changes over the course of plant virus infection (González-Jara *et al*., 2008; Gutiérrez *et al*., 2010; Zwart *et al*., 2013). A time-dependent mathematical function for MOI was determined based on empirical estimates of MOI over time (N.T. Tromas, M.P. Zwart, G. Lafforgue, S.F. Elena, unpublished manuscript) (see Materials and Methods). If the MOI and the frequency of parental genotypes are known, these two variables can be used to predict the expected frequency of cellular coinfection for each time point in each plant.

Second, we considered the effects of virus expansion in the host plant on the expected frequency of recombinants, for two reasons: (i) recombinants that are generated early in an expanding population can have a large number of descendants and therefore reach high frequencies, (ii) recombinants will probably only reach appreciable frequencies if they occur in cells were replication occurs, since they can then be replicated within the cell and possibly be transmitted to other cells. Plant viruses move locally by cell-to-cell movement, and each cell can only infect those cells adjacent to it (Dietrich & Maiss,
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 $r_g = 3.427 \times 10^{-5}$ r/s/g (95% CI: $1.346 \times 10^{-5} - 5.998 \times 10^{-5}$).

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 $(2 - 4) \times 10^{-5} \text{ r/s/g}$. This estimate is very similar to our maximum-likelihood-based estimate using an infection model $(1 - 6) \times 10^{-5} \text{ r/s/g}$, 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
recombination may have a similar impact on the generation of genetic diversity for TEV.

If mutation and recombination rates are indeed similar, this has important evolutionary implications. For example, high mutation rates in combination with small population sizes during viral transmission turn on Muller’s ratchet, a phenomenon already shown to operate in experimental TEV populations transmitted throughout very dramatic bottlenecks (De la Iglesia & Elena, 2007). A high recombination rate could counteract the effect of Muller’s ratchet by recreating mutation-free genomes. The main evolutionary advantage of recombination may be as a mechanism of sex that helps purge deleterious mutations (Muller, 1964; Simon-Loriere & Holmes, 2011). In a similar vein, high recombination rates may have also be advantageous because they speed up the rate of adaptation by bringing together beneficial mutations that otherwise would exist in different genomes, thus minimizing the effect of clonal interference (Fisher, 1930; Muller, 1932; Simon-Loriere & Holmes, 2011). In addition to these “fitness advantage” theories, other models have been brought forward to explain viral recombination. One intriguing possibility is that RNA virus recombination may have evolved as a byproduct of the high nucleotide incorporation rate of viral RNA polymerases (Simon-Loriere & Holmes, 2011): the faster the action of the replicase, the more slippery it becomes. RNA virus high mutation rates are also a likely consequence of the tradeoff between fast replication and accuracy (Belshaw et al., 2007; Elena & Sanjuán, 2005), and one would then expect a positive correlation between mutation and recombination rates across RNA viruses.

To test this prediction, we searched the literature on RNA viruses for cases in which estimates of both the mutation and recombination rates are available. Unfortunately, the 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 PfuiUltra™ high fidelity DNA polymerase (Stratagene). The amplification conditions were 1 min at 95°C (initial denaturation), followed by 18 cycles consisting of 30 s at 95 °C, 45 s at 65 °C and 18 min at 68 °C, and a final extension step of 28 min at 68°C. PCR products were digested with DpnI (New England Biolabs) to remove the parental methylated strands and transformed into
electrocompetent *Escherichia coli* DH5α. At least 15 clones were sequenced to confirm the successful incorporation of desired mutations.

**Neutrality of restriction markers.** Sequence-validated plasmids containing the corresponding restriction site maker were linearized with *Bgl*II (Takara) and transcribed into 5’-capped RNAs using the SP6 mMESSAGE mMACHINE kit (Ambion Inc). Transcripts were precipitated (1.5 volumes of DEPC-treated water, 1.5 volumes of 7.5 M LiCl, 50 mM EDTA), collected and resuspended in DEPC-treated water (Carrasco *et al.*, 2007). RNA integrity was assessed by gel electrophoresis and concentration was spectrophotometrically determined using a Biophotometer (Eppendorf).

To evaluate the neutrality of the four markers, we proceeded as follows. Three four-week-old *N. tabacum* plants were inoculated by abrasion on the third true leaf with 7 µg of transcribed RNA from each individual marker as described elsewhere (Carrasco *et al.*, 2007). Inoculated plants were placed in a BSL-2 greenhouse at 25 °C and 16 h light/8 h dark period. RT-qPCR was performed as described elsewhere (Lalić *et al.*, 2011).

**Coinoculation experiments and restriction analysis.** Total RNA was extracted using InviTrap® Spin Plant RNA Mini Kit (Invitek) from the symptomatic leafs for each of five plants 15 dpi. Each plant was analyzed separately, thus providing independent replicates of the recombination rate among pairs of markers. For each combination of restriction makers, the regions of interest was reverse transcribed using the following reaction mixture: 1× RT reaction buffer (Fermentas), 0.2 mM each dNTP, 0.25 µM forward primer (Table 3), 0.2 µL RNase inhibitor, 40 units of Moloney MuLV reverse 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/SmaI (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
more than 30 amplification cycles and a low initial template concentration minimized chimera formation. To determine whether the RT-PCR conditions used in our experiments may have favored the formation of chimeras, we first inoculated two four-week-old *N. tabacum* plants with respectively TEV RNAs containing *Pme*I and *Eco47III* markers. After 15 dpi, total RNA from each plant was extracted and mixed in a 1:1 (w:w) ratio. Three serial dilutions of this mixture were made: 50 ng/µL, 5 ng/µL and 0.5 ng/µL. Each of these dilutions was then used as template for an RT-PCR experiment (25 cycles, Phusion DNA polymerase). We failed to get an amplification product for the highest dilution (0.5 ng/µL). For the two others dilutions, PCR products were purified, cloned and transformed into *E. coli*. Forty-eight clones were genotyped by restriction analysis with *Pme*I and *Eco47III* for the 50 ng/µL dilution, finding two false positives. One hundred and seven clones were likewise genotyped for the 5 ng/µL dilution without observing false positives; 33 additional clones were genotyped by sequencing for this dilution, with no false positives observed. No homologous or non-homologous recombination events were observed, confirming these conditions avoid false positive results. These results also confirm that the concentration of template RNA molecules and the number of PCR cycles determine the chances of generating chimera molecules during RT-PCR.

**Estimation of *r* and *r*.* The frequency of recombinant genotypes (*f*) on each analyzed plant was estimated as *f* = *Recombinants*/(*Recombinants* + *Parentals*). The recombination rate between restriction markers was computed according to Kosambi (1944) equation:

\[
    r = \frac{1}{4L} \ln \frac{1+2f}{1-2f},
\]
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 $\times$ 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), $m_0$, 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 $m_t = \kappa \sqrt{1 - \left(1 - \frac{\kappa}{m_0}\right)e^{-\nu t}}$.

Here $m_t = m_0$ is the MOI at $t = 0$, $\kappa$ is the maximum attainable value for MOI and $\nu$ 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 $\nu = 1.02$, by minimizing the residual sum of squares. From $m_t$ the frequency of coinfected cells can be calculated as $c_t = \left(1 - e^{-m_t}\right)\left(1 - e^{-m_t}e^{-\nu t}\right)\left(1 - e^{-m_t}\right)$, where $\alpha$ 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 $t$
= 0 and γ is the decay rate. A grid search was used to estimate $C_0 = 45.0$ and $γ = 0.84$, by minimizing the residual sum of squares. The difference equation for the number of infected cells is $n_{t+1} = (1 + C_t)n_t$.

To estimate the expected final frequency of recombinants, we first estimate the frequency of de novo recombinants each day ($y$) as determined by Kosambi (1944): $y = \frac{θ}{2}\tanh\left(2Lr\right)$, where $θ$ is the number of virus generations per day of infection (2.91). The frequency of recombinants on a given day ($f_t$) is then $f_t = f_{t-1} + yc_t$. The frequency of recombinants from the previous day is included, since we are considering only those recombinants that occur in the newly infected cells, but assume that newly occurring recombinants will be maintained in the virus population. The window of one day for allowing recombination to occur in newly infected cells is probably conservative given the rapid advance of TEV infection (Dolja et al., 1993) and our estimate of $θ = 2.91$ generations per day (Martinez et al., 2011). We can then estimate the expected frequency of recombinants in the final population $f_{15} = \frac{1}{n_{15}} \sum_{t=1}^{15} n_tC_t f_t$, where $n_t$ is the number of infected cells after $t$ days. This equation in essence estimates the mean frequency of recombinants occurring over days, weighted by the amount of expansion occurring on a particular day ($n_tC_t$). In order to estimate $r_g$, we minimized the negative log-likelihood by a grid search. The likelihood of the corresponding $f_{15}$ value is given by: $L(f_{15} \mid g, h) = \left(\begin{array}{c} g \\ h \end{array}\right) f^h (1 - f)^{g-h}$, where $g$ is the total number of clones sequenced, and $h$ is the number of sequenced clones that were recombinant.

 acknowledgment
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


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

<table>
<thead>
<tr>
<th>Combination</th>
<th>Fragment size (nt)</th>
<th>Plants analyzed</th>
<th>$f$ (±1 SEM) (%)</th>
<th>$r$ (±1 SEM) ($\times 10^{-6}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AscI/PmeI</td>
<td>3334</td>
<td>5</td>
<td>1.854±0.125</td>
<td>5.561±0.376</td>
</tr>
<tr>
<td>PmeI/Eco47III</td>
<td>1234</td>
<td>4</td>
<td>0.568±0.084</td>
<td>4.604±0.684</td>
</tr>
<tr>
<td>Eco47III/SalI</td>
<td>2197</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2. Artificial restriction sites engineered as genetic markers for this study

<table>
<thead>
<tr>
<th>Restriction enzyme</th>
<th>Cistron</th>
<th>Genomic position</th>
<th>Mutagenesis primer (5’ to 3’)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>AscI P1</td>
<td>402</td>
<td>TTATCTTGTCGCGCCGCCTCAACCACATGGC</td>
<td></td>
</tr>
<tr>
<td>PmeI CI</td>
<td>3735</td>
<td>AGCCTTCCTGGAGTCACGTATAAAACAATGGGAAACAACCA</td>
<td></td>
</tr>
<tr>
<td>Eco47III CI</td>
<td>4969</td>
<td>AGTCATAACATGACAGCTGAAACGTATTAAGCTACACACTTGTGAG</td>
<td></td>
</tr>
<tr>
<td>SalI Nib</td>
<td>7166</td>
<td>GATGGGACATATAAGCCAAACCGACTTTAATAGAGAGGCG</td>
<td></td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Combination</th>
<th>TEV genome position (5’)</th>
<th>Sequence (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ascl/Pmel</em></td>
<td>46</td>
<td>GCAATCAAGCATTCTACTTC</td>
</tr>
<tr>
<td></td>
<td>3894</td>
<td>ATCCAAACGACCTCTCAC</td>
</tr>
<tr>
<td><em>Pmel/Eco47III</em></td>
<td>3541</td>
<td>TTGACGCTGAGCGGAGTGATGG</td>
</tr>
<tr>
<td></td>
<td>5275</td>
<td>CTATTGATGCATGCTAGAGTC</td>
</tr>
<tr>
<td><em>Eco47III/SalI</em></td>
<td>4972</td>
<td>TTAAGCTACACACTTGAGAC</td>
</tr>
<tr>
<td></td>
<td>7394</td>
<td>TTCTTTCTTCTTGCTTTT</td>
</tr>
</tbody>
</table>
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.