Estimating virus effective population size and selection without neutral markers

Elsa Rousseau, Benoît Moury, Ludovic Mailleret, Rachid Senoussi, Alain Palloix, Vincent Simon, Sophie Valière, Frédéric Grognard, Frédéric Fabre

PLOS | PATHOGENS

RESEARCH ARTICLE

Estimating virus effective population size and selection without neutral markers

Elsa Rousseau[1,2,3☯]*, Benoît Moury[3], Ludovic Mailleret[1,2], Rachid Senoussi[4], Alain Palloix[5†], Vincent Simon[3,6], Sophie Valière[7,8], Frédéric Grognard[1], Frédéric Fabre[9☯]*

1 Université Côte d'Azur, Inria, INRA, CNRS, UPMC Univ Paris 06, Biocore team, Sophia Antipolis, France, **2** Université Côte d'Azur, INRA, CNRS, ISA, Sophia Antipolis, France, **3** Pathologie Végétale, INRA, 84140 Montfavet, France, **4** UR BioSp, INRA, Avignon, France, **5** UR GAFL, INRA, Montfavet, France, **6** UMR BFP, INRA, Villenave d'Ornon, France, **7** GeT-PlaGe, INRA, Genotoul, Castanet-tolosan, France, **8** UAR DEPT GA, INRA, Castanet-Tolosan, France, **9** UMR SAVE, INRA, Villenave d'Ornon, France

☯ These authors contributed equally to this work.
† Deceased.
* elsa.rousseau7@gmail.com (ER); frederic.fabre@inra.fr (FF)

Citation: Rousseau E, Moury B, Mailleret L, Senoussi R, Palloix A, Simon V, et al. (2017) Estimating virus effective population size and selection without neutral markers. PLoS Pathog 13(11): e1006702. https://doi.org/10.1371/journal.ppat.1006702

Editor: Fernando Garcia-Arenal, Universidad Politecnica de Madrid, SPAIN

Received: June 23, 2017

Accepted: October 19, 2017

Published: November 20, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the SMaCH (Sustainable Management of Crop Health) metaprogramme of INRA and by a grant over-seen by the French National Research Agency (ANR) as part of the "Blanc2013" program (ANR-13-BSV7-0011, FunFit project). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

By combining high-throughput sequencing (HTS) with experimental evolution, we can observe the within-host dynamics of pathogen variants of biomedical or ecological interest. We studied the evolutionary dynamics of five variants of *Potato virus Y* (PVY) in 15 doubled-haploid lines of pepper. All plants were inoculated with the same mixture of virus variants and variant frequencies were determined by HTS in eight plants of each pepper line at each of six sampling dates. We developed a method for estimating the intensities of selection and genetic drift in a multi-allelic Wright-Fisher model, applicable whether these forces are strong or weak, and in the absence of neutral markers. This method requires variant frequency determination at several time points, in independent hosts. The parameters are the selection coefficients for each PVY variant and four effective population sizes N_e at different time-points of the experiment. Numerical simulations of asexual haploid Wright-Fisher populations subjected to contrasting genetic drift ($N_e \in [10, 2000]$) and selection ($|s| \in [0, 0.15]$) regimes were used to validate the method proposed. The experiment in closely related pepper host genotypes revealed that viruses experienced a considerable diversity of selection and genetic drift regimes. The resulting variant dynamics were accurately described by Wright-Fisher models. The fitness ranks of the variants were almost identical between host genotypes. By contrast, the dynamics of N_e were highly variable, although a bottleneck was often identified during the systemic movement of the virus. We demonstrated that, for a fixed initial PVY population, virus effective population size is a heritable trait in plants. These findings pave the way for the breeding of plant varieties exposing viruses to stronger genetic drift, thereby slowing virus adaptation.

Competing interests: The authors have declared that no competing interests exist.

Author summary

A growing number of experimental evolution studies are using an "evolve-and-resequence" approach to observe the within-host dynamics of pathogen variants of biomedical or ecological interest. The resulting data are particularly appropriate for studying the effects of evolutionary forces, such as selection and genetic drift, on the emergence of new pathogen variants. However, it remains challenging to unravel the effects of selection and genetic drift in the absence of neutral markers, a situation frequently encountered for microbes, such as viruses, due to their small constrained genomes. Using such an approach on a plant virus, we observed that the same set of virus variants displayed highly diverse dynamics in closely related plant genotypes. We developed and validated a method that does not require neutral markers, for estimating selection coefficients and effective population sizes from these experimental evolution data. We found that the viruses experienced considerable diversity in genetic drift regimes, depending on host genotype. Importantly, genetic drift experienced by virus populations was shown to be a heritable plant trait. These findings pave the way for the breeding of plant varieties exposing viruses to strong genetic drift, thereby slowing virus adaptation.

Introduction

Evolution in isolated populations results from the interplay between several forces, including mutation, selection, and genetic drift. Mutation creates genetic diversity within a population. Subsequent selection and genetic drift drive the evolution of diversity within the population. Selection is a deterministic force that increases the frequency of the fittest variants at the expense of the weakest ones. It can be characterized by the selection coefficient s, commonly calculated, at a specific locus, as the relative difference in fitness conferred by two alleles. Genetic drift, unlike selection, acts equally on all variants. It is the outcome of random sampling effects between generations, resulting in stochastic fluctuations in variant frequencies [1]. The strength of genetic drift is frequently evaluated by determining the effective population size N_e [1]. N_e is defined as the size of an ideal panmictic population of constant size with non-overlapping generations that would display the same degree of randomness in allele frequencies as the population studied [2]. N_e is often much lower than the census population size [3, 4], but it can be seen as its evolutionary analog [5]. When N_e is small, sampling effects are magnified between generations, and allele frequencies therefore fluctuate strongly. For populations varying in size over time, the effective population size over a given number of generations can be approximated by the harmonic mean $\bar{N}_e$ of effective population sizes at each generation. This approximation holds provided that the number of generations is much smaller than $\bar{N}_e$ [6–8] and that mutation can be neglected [9]. Population size may vary over time due to bottlenecks, which are common in natural populations. As they greatly decrease population size, they have a disproportionate effect on the overall value of $\bar{N}_e$ [1].

When selection and genetic drift act simultaneously, the probability of fixation of a new mutation (with a selection coefficient s), and, more generally, its evolutionary dynamics, is controlled by the product $N_e \times |s|$ [1, 10]. If $N_e \times |s| \ll 1$, then genetic drift predominates over selection and evolution is mostly stochastic. If $N_e \times |s| \gg 1$, then selection becomes effective and evolution is mostly deterministic [10]. This rule of thumb can be applied to the evolutionary dynamics of pathogen variants of biomedical or ecological interest, during the course of infection of a single host, for microbe variants escaping the immune response of their host, or

becoming resistant to drug therapy (e.g. [11]) or, in the case of plant pathogens, for variants adapting to host resistance genes (e.g. [12]). In this study, we combined high-throughput sequencing (HTS) with experimental evolution to measure the within-host dynamics of five variants of *Potato virus Y* (PVY, genus *Potyvirus*, family Potyviridae) in closely related plant genotypes [13].

It remains challenging to unravel the effects of genetic drift and selection in the absence of neutral markers, in studies of the adaptation dynamics of pathogens. This situation is frequently encountered for pathogens with small genomes, especially viruses [14, 15]. Various approaches based on moment [16, 17] or likelihood [18–20] methods have been proposed for estimating N_e, but all require the genetic markers studied to be neutral. Various methods have also been proposed for detecting selection and estimating selection coefficients. These methods require at least some prior information about N_e (e.g. [21]) or assume that genetic drift is negligible (e.g. [22]). However, in the absence of neutral markers and without prior estimates of N_e, both selection and genetic drift must be taken into account, as these two forces act simultaneously on evolution. This greatly complicates the estimation of N_e and s. Only a few methods have been proposed for the joint estimation of N_e and s from time-sampled data (see [11] and [23] for a review). For large effective population sizes (typically $N_e > 5000$) and small selection coefficients (typically $|s| < 0.01$), several likelihood methods based on diffusion approximations of the Wright-Fisher model [1, 11] are available [24–27]. In the situations in which these methods are valid, the ranges of N_e and s values obtained are rather restrictive for many microorganisms, particularly viruses [28–31]. Foll *et al.* [32] recently proposed the use of approximate Bayesian computation (ABC) for the joint estimation of N_e and s in a Wright-Fisher model. Their method can deal with both weak and strong selection regimes, but still requires multilocus genome-wide data with mostly neutral loci to estimate N_e accurately.

In this study, we investigated the evolutionary dynamics of five variants of PVY in 15 closely related pepper genotypes. All plants were inoculated with the same mixture of virus variants and variant frequencies were determined with HTS in eight plants of each genotype at each of six sampling dates after inoculation. A diverse range of evolutionary patterns was observed. We developed a method for estimating the parameters of a multi-allelic Wright-Fisher model with selection and genetic drift, to investigate the underlying evolutionary processes. This method has two main advantages: it applies to a large range of selection and genetic drift intensities and it works efficiently in the absence of neutral markers. The parameters of the Wright-Fisher model (i.e. selection coefficients for each virus variant and effective population sizes at given time points) can be estimated by coupling maximum likelihood and ABC methods and applying them to a set of variant frequencies determined at several time points in independent hosts. We tested the method with numerical simulations mimicking the datasets obtained with HTS in evolve-and-resequence experiments [33]. The simulations covered an extensive range of N_e and s values. We were then able to estimate the selection coefficient of each PVY variant in each pepper genotype and the changes in effective population size over time during the colonization of the plant by the virus. Finally, by varying pepper genotypes and fixing the initial PVY population, we provided evidence that the effective population size of PVY is a heritable plant trait. This finding paves the way for the breeding of plant cultivars exposing viruses to greater genetic drift and/or smaller selection effects.

Materials and methods

Biological experiment

Plant and virus material. We used 15 doubled-haploid (DH) lines of pepper (*Capsicum annuum*, family Solanaceae). All the plants of a given genotype were thus genetically identical.

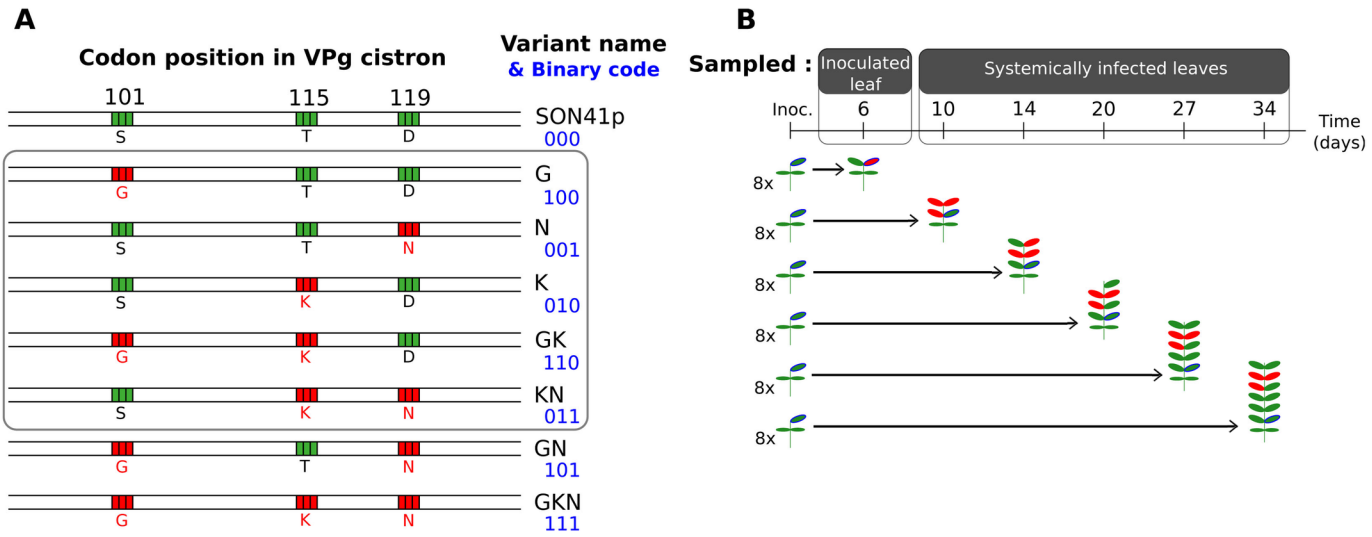

Fig 1. Virus variants inoculated to pepper plants and sampling protocol. (A) The five virus variants (in the gray box) were derived from the SON41p PVY clone and differed only at codon positions 101, 115 and 119 of the VPg cistron. These positions are shown in green if they correspond to the SON41p clone and in red if a non-synonymous substitution was introduced by site-directed mutagenesis. Single-letter amino acid abbreviations are presented below each position and PVY variant. Variant names and the corresponding binary code for the three point mutations of interest are given on the right of the sequences, with the binary code of the SON41p variant set to 000. The two additional possible variants, based on the three-digit binary code, are also shown at the bottom. (B) Sampling protocol for one pepper genotype. We inoculated 48 plants with the virus. Eight plants were sampled at each sampling time, from 6 to 34 days post-inoculation. The leaf circled in blue is the leaf inoculated with the virus. The leaves sampled are shown in red.

https://doi.org/10.1371/journal.ppat.1006702.g001

All DH lines carried the major resistance gene *pvr2*^{3} and differed in terms of their genetic background [12]. They are issued from the F_1 hybrid between two pepper lines, Perennial and Yolo Wonder. Thus, on average, any pair of DH lines had 50% of alleles in common, at markers distinguishing between Perennial and Yolo Wonder. Each DH line therefore constituted a different host environment for plant colonization by PVY. These lines were chosen for study on the basis of quantitative differences in three previously measured factors, so as to generate different intensities of genetic drift and selection acting on PVY populations: (i) relative within-plant viral accumulation, (ii) resistance breakdown (RB) frequency [12] and (iii) the number of primary infection foci after mechanical inoculation with the virus [34] (S1 Fig).

All plants were mechanically inoculated with the same equimolar mixture, based on quantitative double-antibody sandwich enzyme-linked immunosorbent assay (DAS-ELISA), of the five PVY variants G, N, K, GK and KN [35]. Single- and double-letter names indicate single and double mutants, respectively, of the infectious clone SON41p (Fig 1A). Three mutations located close together in the PVY genome differentiate the five variants, and these mutations are named after the amino-acid substitutions observed at positions 101 for the S (serine) to G (glycine) substitution, 115 for the T (threonine) to K (lysine) substitution, and 119 for the D (aspartic acid) to N (asparagine) substitution, in the VPg (viral protein genome-linked). The G and N variants displayed a low level of adaptation to the major resistance gene *pvr2*^{3} carried by all plant genotypes, whereas variants K, GK and KN displayed higher levels of adaptation [12].

Experimental set-up and plant sampling. For each pepper genotype, 48 plants were arranged in randomized blocks, to minimize environmental effects. The first true leaf of each plant was inoculated 29 days after sowing. We then analyzed eight plants per DH line at 6, 10, 14, 20, 27 and 34 days post-inoculation (dpi) (Fig 1B). The inoculated leaf was sampled at 6 dpi, and, on subsequent sampling dates, three uninoculated leaves, corresponding to the three

youngest unfolded leaves, were sampled and pooled together. As the plants were removed after sampling, the virus populations obtained from each plant sample were independent. This prevented possible effects on virus population dynamics due to the removal of infected leaves and subsequent re-sampling. Each leaf sample was ground in four volumes of 0.03 M phosphate buffer (pH 7.0) supplemented with 2% (w/v) diethyldithiocarbamate, as previously described [35].

High-throughput sequencing and determination of PVY variant frequencies. Total RNA was purified from individual plant samples with the Tri-Reagent kit (Sigma-Aldrich). It was subjected to reverse transcription-polymerase chain reaction (RT-PCR) with tagged primers for the amplification, over 35 cycles, of a 104-nucleotide region encompassing the polymorphic region of the PVY VPg cistron. Eight differently tagged primers were used, corresponding to the eight different plant replicates of the same plant genotype for each sampling date (S1 Table). Amplified DNAs corresponding to the eight plant replicates were pooled together on the basis of their intensity on electrophoresis gels.

HTS was performed at the Genomic Platform of INRA Toulouse. For this purpose, 2 × 150 base-pair (bp) libraries with multiplex adapters were prepared, and all the RT-PCR-amplified products were pooled into a single large sample (12 cycles). This sample was run on a MiSeq Illumina paired-end sequencer with the MiSeq Reagent Kit v2, for 500 cycles. We chose to use MiSeq Illumina sequencing because this technology has a much lower error rate than other high-throughput sequencing technologies, such as 454 sequencing [36]. By using tagged primers and subsequent multiplex adapters, we were able to assign a plant genotype and a sampling date to each sequence.

In the initial sequence analysis, we used FLASH software to obtain the consensus sequence from reads 1 and 2 with a minimum overlap length between the two reads of 63, a maximum overlap length of 153 and a maximum allowed ratio of the number of mismatched base pairs to overlap length of 0.2 [37]. The sequences were then sorted by adapter and by tag. Finally, the sequences corresponding to each PVY variant were determined with the help of 'agrep' function in R software [38], and sequence counts were used to assess the composition of the virus population in each sample. After sequence sorting, we had 374 to 14141 sequences per sample, with a mean of 3295 sequences per sample.

We carried out two complementary sequence analyses to detect PVY mutations (see S1 Text for details). It was important to perform these analyses as the presence of mutants might have affected virus population dynamics and the intensities of the evolutionary forces studied. In the first analysis, we looked for all eight possible variants based on the three codon positions of interest in the VPg cistron, *i.e.* the five analyzed variants, G, N, K, GK and KN, together with the SON41p, GN and GKN variants (corresponding to all the possible binary codes in Fig 1A). The sum of the frequencies of all three additional variants, SON41p, GN and GKN, remained below 5%, and these variants were not, therefore, considered in N_e and s estimations. The raw data from this analysis (*i.e.* number of sequences of each variant in each of the 677 samples analyzed) are available from S2 Table. We then calculated the frequencies of *de novo* substitutions in each sample and at each nucleotide position of all sequences, by comparison with the sequence of the SON41p reference clone (equivalent to comparison with sequences of the G, N, K, GK and KN clones). In all, only PVY populations sampled from six of the 677 plants studied presented a *de novo* substitution with a frequency exceeding 5% (S1 Text). These PVY populations were removed for subsequent analyses. Furthermore, numerical simulations showed that a sixth unaccounted for variant present at a mean frequency of 7% in virus populations had no significant impact on estimates of N_e and s (see below), justifying our use of a 5% threshold.

Genetic analysis and heritability estimation. With this experimental design, we studied the phenotype of each pepper DH line in terms of its effect on PVY populations. The plants within each pepper DH line were genetically identical and experimental conditions were set up so as to ensure an absence of differences in environmental effects between DH lines. We were therefore able to estimate the heritability of plant traits of interest, corresponding to the evolutionary forces exerted by the plant on PVY populations, by constituting two replicates for each DH line dataset. More precisely, we assessed the heritability of the intrinsic rates of increase for the five PVY variants and of the effective population sizes of PVY. As the initial population of PVY was fixed and identical for all plant genotypes, we did not consider the effect of PVY population composition on the heritability of effective PVY population size. To estimate heritabilities, we split the dataset for the 48 plants for each DH line in two, by randomly selecting four of the eight plants at each sampling date. The broad-sense heritability of any plant trait of interest can be estimated as $h^2 = \sigma_G^2/(\sigma_G^2 + \sigma_e^2/n)$, where σ_G^2 corresponds to the genotypic variance, σ_e^2 to the phenotypic variance and n to the number of replicates [39], the variance being set to the sum of the squared deviations from the mean.

Estimation of selection and genetic drift intensities

We developed a method for estimating the parameters of a multi-allelic Wright-Fisher model with selection and genetic drift for a haploid population. The parameters and state variables of the model and the observed variables are summarized in Table 1.

Table 1. Main notations for the observations and the model.

	Designation (unit) [reference value]
Observed variables	
$\boldsymbol{x}^p(t_d) = (x_1^p(t_d), \ldots, x_{n_{var}}^p(t_d))$	Variant sequence counts in virus population p at sampling time t_d (*seq* [a])
$\boldsymbol{f}^p(t_d) = (f_1^p(t_d), \ldots, f_{n_{var}}^p(t_d))$	Variant frequencies in virus population p at sampling time t_d (no unit)
State variables	
$\boldsymbol{\lambda}^p(t) = (\lambda_1^p(t), \ldots, \lambda_{n_{var}}^p(t))$	Theoretical variant frequencies in virus population p at time t of a Wright-Fisher model (no unit)
Parameters of interest	
$\boldsymbol{r} = (r_1, \ldots, r_{n_{var}})$	Variant relative intrinsic rates of increase ($generation^{-1}$) [a]
$\boldsymbol{\eta}_e = (\eta_e^{IO}, \eta_e^{S_1}, \eta_e^{S_2}, \eta_e^{S_3})$	Successive virus effective population sizes (*individuals*) [b]
Fixed parameters	
$\boldsymbol{\lambda}^{inoc}$	Vector of variant frequencies in the virus inoculum (no unit)
$\boldsymbol{T}$	Vector of measurement dates (*day*) [(0, 6, 10, 14, 20, 27, 34)]
Additional notations	
$\boldsymbol{N_e} = (N_e(1), \ldots, N_e(34))$	Vector of virus effective population sizes (piecewise constant function of η_e)
$N_e^h(t_d)$	Harmonic mean of virus effective population sizes at sampling time t_d
$\sigma[f_i^\bullet(t_d)]$	Standard deviation of the frequencies of virus variant i at sampling time t_d over the virus populations p
$\boldsymbol{\lambda}^{det}(t) = (\lambda_1^{det}(t), \ldots, \lambda_{n_{var}}^{det}(t))$	Vector of variant frequencies at time t for an infinite size Wright-Fisher model

[a] The abbreviation "seq" is the number of sequences representing the virus population or a given variant in this population.

[b] The mean intrinsic rate of increase $\bar{r}$ of all virus variants is one.

[c] With the full model $\mathfrak{M}_4$, η_e^{IO} in the inoculated organ for $t \in [1, 6]$, η_e^{S1} at the onset of systemic infection for $t \in [7, 10]$, η_e^{S2} for $t \in [11, 14]$ and η_e^{S3} for $t \in [15, 34]$.

https://doi.org/10.1371/journal.ppat.1006702.t001

Notation: Observed variables, state variables and parameters of interest. $\boldsymbol{\lambda}^{inoc}$ denotes the vector of the observed variant frequencies in the parental virus population, i.e. the inoculum used to inoculate all host plants in the experiment. Thereafter, the measurement date vector $\boldsymbol{T} = (0, 6, 10, 14, 20, 27, 34)$ is indexed by t_d ($d = 0, \ldots, 6$). In particular, $t_0 = 0$ is the inoculation date and t_d ($d = 1, \ldots, 6$) are the sampling dates. The time, indexed by $t = 1, \ldots, 34$, is the number of days post-inoculation and indicates the viral generation, as the generation time was assumed to be one day [40]. For a given plant genotype, at each sampling date t_d, a sample of $n_{inf}(t_d)$ infected plants was observed, for which we measured the vectors $\boldsymbol{x}^p(t_d) = (x_1^p(t_d), \ldots, x_{n_{var}}^p(t_d))$, with $x_i^p(t_d)$ the number of sequences obtained for virus variant i ($1 \leq i \leq n_{var}$) in population p ($1 \leq p \leq n_{inf}(t_d)$). Thereafter, replacing an index in a notation by $\bullet$ is equivalent to summing over the corresponding index set. The total number of sequences obtained from virus population p at time t_d is thus $x_\bullet^p(t_d)$. Finally, $\boldsymbol{f}^p(t_d) = (f_1^p(t_d), \ldots, f_{n_{var}}^p(t_d))$, with $f_i^p(t_d)$ the observed frequency of virus variant i in population p at sampling date t_d. It is calculated as $\boldsymbol{x}^p(t_d)/x_\bullet^p(t_d)$.

The state variable of interest is $\boldsymbol{\lambda}^p(t) = (\lambda_1^p(t), \ldots, \lambda_{n_{var}}^p(t))$, with $\lambda_i^p(t)$ the frequency of virus variant i ($1 \leq i \leq n_{var}$) in virus population p at any date $t = 1, \ldots, 34$. Virus variant dynamics are represented by a Wright-Fisher model (see below) to infer $\boldsymbol{\theta} = (\boldsymbol{r}, \boldsymbol{\eta}_e)$, the vector of parameters describing the underlying evolutionary forces. $\boldsymbol{r} = (r_1, \ldots, r_{n_{var}})$ is the vector of the intrinsic rates of increase r_i of each virus variant i. We assumed that the mean intrinsic rate of increase $\bar{r}$ over all variants was one, as we were interested only in the relative intrinsic rates of increase. The selection coefficient of a variant i is usually computed as $s_i = r_i - 1$. The vector parameter $\boldsymbol{\eta}_e$ defines a piecewise function describing effective population sizes $N_e(t)$. We determined the temporal variation of effective population sizes, using four models with $\boldsymbol{\eta}_e$ having one to four parameters. With the more general model $\mathfrak{M}_4$, $\boldsymbol{\eta}_e = (\eta_e^{IO}, \eta_e^{S_1}, \eta_e^{S_2}, \eta_e^{S_3})$. η_e^{IO} is the effective population size of the viral population in the inoculated organ; this stage lasts $t_1 = 6$ days in our experimental design. $\eta_e^{S_1}$ is the effective population size during the onset of systemic infection; this stage lasts $t_2 - t_1 = 4$ days. $\eta_e^{S_2}$ is the effective population size during the next $t_3 - t_2 = 4$ days and $\eta_e^{S_3}$ the effective population size later on, during the last $t_6 - t_3 = 20$ days of survey. Accordingly, we define $N_e(t)$ as follows:

$$N_e(t) = \begin{cases} \eta_e^{IO} & t \in [1, \ldots, t_1] \\ \eta_e^{S_1} & t \in [t_1 + 1, \ldots, t_2] \\ \eta_e^{S_2} & t \in [t_2 + 1, \ldots, t_3] \\ \eta_e^{S_3} & t \in [t_3 + 1, \ldots, t_6] \end{cases} \quad (1)$$

With model $\mathfrak{M}_3$, $\boldsymbol{\eta}_e = (\eta_e^{IO}, \eta_e^{S_1}, \eta_e^{S_2})$. $N_e(t)$ has three parameters: (i) $N_e(t) = \eta_e^{IO}$ when $t \in [1, 6]$, (ii) $N_e(t) = \eta_e^{S_1}$ when $t \in [7, 14]$ and (iii) $N_e(t) = \eta_e^{S_2}$ when $t \in [15, 34]$. With model $\mathfrak{M}_2$, $\boldsymbol{\eta}_e = (\eta_e^{IO}, \eta_e^{S})$. $N_e(t)$ has two parameters: (i) $N_e(t) = \eta_e^{IO}$ when $t \in [1, 6]$ and (ii) $N_e(t) = \eta_e^{S}$ when $t \in [7, 34]$. Finally, with model $\mathfrak{M}_1$, $\boldsymbol{\eta}_e = (\eta_e)$: the effective population size for the virus remains constant throughout the experiment ($N_e(t) = \eta_e$ for $t \in [1, 34]$). The effective population size at any sampling date of interest t_d is given, approximately, by the harmonic mean of the effective sizes of the successive generations $N_e^h(t_d) = \left(\frac{1}{t_d}\sum_{j=1}^{t_d}\frac{1}{N_e(j)}\right)^{-1}$.

The multi-allelic Wright-Fisher model with selection and genetic drift. The Wright-Fisher model occupies a central position in population genetics [41]. It assumes an ideal population: a randomly mating haploid population of finite size reproducing in discrete non-

overlapping generations, with no structure. By definition, N_e is the size of an ideal population (i.e. obeying previous assumptions) that would display the same degree of randomness in variant frequencies as the real population studied [2]. As for any model-based approach, the use of this concept requires the actual population being not too far from an ideal Wright-Fisher model with suitable parameters [10]. Formally, the Wright-Fisher model is very similar to the quasispecies model describing the evolution of DNA (or RNA) sequences in finite populations [42]. In practice, the Wright-Fisher model has been used to infer the evolutionary history of viruses (e.g. [11]) and it can also be used to describe the stochastic dynamics of the frequency of the n_{var} virus variants considered here. Let $\boldsymbol{z}(t)$ be the vector of the number of each virus variant i ($1 \leq i \leq n_{var}$) in generation t and, with previous notations, let $\boldsymbol{\lambda}(t) = \dfrac{\boldsymbol{z}(t)}{N_e(t)}$ be the corresponding vector of variant frequencies. The dynamics of $\boldsymbol{z}(t)$ are shaped by random genetic drift and selection. Let $\boldsymbol{pr}(t) = (pr_1(t), \ldots, pr_{n_{var}}(t))$ be the vector of the probabilities of sampling each virus variant from generation t to generation $t+1$. For $t \geq 1$, the distribution of $\boldsymbol{z}(t+1)$ conditionally on $\boldsymbol{z}(t)$ follows a multinomial distribution [41]:

$$\begin{cases} \boldsymbol{z}(t+1)|\boldsymbol{z}(t) \sim \text{Mult}(\text{size} = N_e(t), \text{prob} = \boldsymbol{pr}(t)) & (2) \\ pr_i(t) = \dfrac{r_i \lambda_i(t)}{\sum_{j=1}^{n_{var}} r_j \lambda_j(t)} \text{ with } \lambda_i(t) = \dfrac{z_i(t)}{N_e(t)} \text{ and } \boldsymbol{\lambda}(0) = \boldsymbol{\lambda}^{inoc} & (3) \end{cases}$$

In the model, selection reweights the different genotypes according to their constant fitness. Fitness does not depend on the composition of the population (*i.e.* there is no frequency-dependent selection). As population size tends to ∞ (*i.e.* genetic drift becomes negligible), the stochastic process of variant frequencies described by eq (2) converges on deterministic recursion describing $\lambda_i^{det}(t)$ which approximates the mean frequency of variant i at generation t. For $t \geq 1$:

$$\lambda_i^{det}(t+1)|\boldsymbol{r} = \frac{r_i \lambda_i^{det}(t)}{\sum_{j=1}^{n_{var}} r_j \lambda_j^{det}(t)} \quad \text{with} \quad \boldsymbol{\lambda}^{det}(0) = \boldsymbol{\lambda}^{inoc} \qquad (4)$$

Parameter estimation. We propose an approach combining a first step relying on maximum-likelihood followed by a step relying on ABC, to estimate the parameters of interest $\boldsymbol{\theta} = (\boldsymbol{r}, \boldsymbol{\eta}_e)$. The first step estimates the vector of the relative intrinsic rates of increase of each virus variant $\boldsymbol{r}$ by maximum-likelihood methods. Let the vector $\boldsymbol{x}^{\bullet}(t_d) = (x_1^{\bullet}(t_d), \ldots, x_{n_{var}}^{\bullet}(t_d))$ be the total number of sequences of virus variant i obtained in the $n_{inf}(t_d)$ infected plants (of a given plant genotype) at sampling date t_d. This step assumes that $\boldsymbol{x}^{\bullet}(t_d) \sim Mult\left(\text{size} = \sum_{i=1}^{n_{var}} x_i^{\bullet}(t_d), \text{prob} = \boldsymbol{\lambda}^{det}(t_d)|\boldsymbol{r}\right)$. Let $\boldsymbol{x}$ denote the vector of all total sequence counts, at all sampling time-points, constituting one dataset. As the samples are independent between sampling dates, the likelihood function is:

$$l(\boldsymbol{x}|\boldsymbol{r}) = \prod_{d=1}^{6} dM\left(\text{size} = \sum_{i=1}^{n_{var}} x_i^{\bullet}(t_d), \text{prob} = \boldsymbol{\lambda}^{det}(t_d)|\boldsymbol{r}\right),$$

dM being the probability density function (pdf) of the multinomial distribution. Under these hypotheses, $\boldsymbol{r}$ can be inferred by minimizing $-log(l(\boldsymbol{x}|\boldsymbol{r}))$, assuming that the mean intrinsic rate of increase $\bar{r}$ of all variants is one. The estimate of $\boldsymbol{r}$, denoted $\hat{\boldsymbol{r}}$, was obtained straightforwardly, using the 'nlminb' optimization routine implemented in R software version 3.0.2 [38].

The second step estimates the vector of effective population sizes $\boldsymbol{\eta}_e$ of a given model $\mathfrak{M}_j$ ($j = 1, \cdots, 4$) with ABC, conditionally to $\hat{\boldsymbol{r}}$. All ABC algorithms involve the simulation of a large number of possible datasets by sampling the parameters of interest (here $\boldsymbol{\eta}_e$) from prior probability distributions $\pi(\boldsymbol{\eta}_e)$. We used independent log-uniform priors on [10, 2500] for each parameter of $\boldsymbol{\eta}_e$. For a given $\boldsymbol{\eta}_e^{sim}$ sampled in $\pi(\boldsymbol{\eta}_e)$, a dataset was simulated as follows. The first step was to simulate 48 (= 6 sampling dates * 8 plants/date) independent dynamics of evolution of virus variant frequencies lasting 34 days (max($\boldsymbol{T}$)) with the Wright-Fisher model (eqs 1, 2 and 3) parameterized by $(\hat{\boldsymbol{r}}, \boldsymbol{\eta}_e^{sim})$. Let $\boldsymbol{\lambda}_{sim}^{l}(t)$ ($l = 1, \ldots, 48$; $t = 1, \ldots, 34$) be this set of simulated dynamics. The second step was to simulate the experimental design. Eight plants were analyzed with HTS at each sampling date, $x_{sim}^{tot,l}$ being the total number of sequences obtained for plant l. Variant counts $\boldsymbol{x}_{sim}^{l}(t_d)$ for HTS are sampled from multinomial distributions. A single sample was obtained for each dynamic l: at 6 dpi, $\boldsymbol{x}_{sim}^{l}(6) \sim \text{Mult}(\text{size} = x_{sim}^{tot,l}, \text{prob} = \boldsymbol{\lambda}_{sim}^{l}(6))$ with $l \in [1, 8]$; at 10 dpi, $\boldsymbol{x}_{sim}^{l}(10)$ is obtained similarly with $l \in [9, 16]$, and so on, until 34 dpi, with $l \in [41, 48]$. The observed frequencies are $\boldsymbol{f}_{sim}^{l}(t_d) = \dfrac{\boldsymbol{x}_{sim}^{l}(t_d)}{x_{sim}^{tot,l}}$. The last step was to calculate the vector of summary statistics from the simulated dataset, $\boldsymbol{S}_{sim} = (S_{sim}^{t_1}, \ldots, S_{sim}^{t_6})$. A single summary statistic was calculated for each sampling date. This statistic is the inverse of the mean of the standard deviation of the variant frequencies at sampling date t_d. Formally, $S_{sim}^{t_d} = \left(\dfrac{1}{n_{var}} \sum_{i=1}^{n_{var}} \sigma[f_{sim,i}^{\bullet}(t_d)] \right)^{-1}$ where $\sigma[f_{sim,i}^{\bullet}(t_d)]$ is the standard deviation (over the infected hosts at sampling date t_d) of $f_{sim,i}^{l}(t_d)$, the observed frequency of variant i. In practice, estimation was performed with the adaptive ABC algorithm of Lenormand *et al.* [43] implemented in the R package EasyABC with tuning parameters $nb_{simul} = 5000$, $p_{accmin} = 0.04$ and $\alpha = 0.5$. Models $\mathfrak{M}_1$, $\mathfrak{M}_2$, $\mathfrak{M}_3$ and $\mathfrak{M}_4$ (embedding $N_e(t)$ functions with one to four parameters) were compared, using the multinomial logistic regression method implemented in the ABC package (function postpr with 2.10^5, 2.5×10^5, 7.5×10^5 and 1.5×10^6 simulated summary statistics under models $\mathfrak{M}_1$, $\mathfrak{M}_2$, $\mathfrak{M}_3$ and $\mathfrak{M}_4$, respectively, and tuning parameter tol = 5×10^{-4}). The estimation code will be made available upon request.

Numerical simulations

Before using the estimation method on the datasets corresponding to the biological experiment, we performed several batches of simulations to assess its ability to infer effective population sizes and selection coefficients accurately (see S2 Text for details). Briefly, in experiment 1, we first simulated the changes in frequencies of five virus variants under 750 selection and genetic drift regimes with a Wright-Fisher model for haploid individuals. The simulations were designed to fit the experimental setup of our datasets (48 independent host plants regularly analyzed at 6 sampling dates). For each of the 750 datasets obtained, the true parameters $\boldsymbol{\theta}_{true}$ were known and could be compared with the estimated parameters $\hat{\boldsymbol{\theta}}$. In experiment 2, we assessed the sensitivity of the estimation method to the presence of a sixth undetected virus variant. This sixth variant was selectively neutral (its selection coefficient is zero), present in the inoculum at a frequency of 3% and still present at the last sampling date (34 dpi) in all plants analyzed, at frequencies ranging from 1% to 6%. It affected the dynamics of the five variants of interest in all plants but was not detected, so the variant frequencies measured by HTS (and used to estimate $\hat{\boldsymbol{\theta}}$) are noisy with respect to their true values. In all, 350 simulated datasets were analyzed in this second test.

Results

In this section, we will (i) describe the virus dynamics observed in the biological experiment with 15 plant genotypes, (ii) validate the method for estimating selection and genetic drift with numerical simulations and (iii) describe the estimates obtained in the biological experiment.

Virus variant dynamics in the 15 plant genotypes

The frequencies of the five virus variants were assessed in completely isolated populations during the course of infection, in 15 different plant genotypes. For each of these 15 pepper genotypes, 48 plants were inoculated with the same equimolar mixture of the five variants, and the frequencies of the virus variants were determined in eight plants at each of six sampling dates, from 6 to 34 days post-inoculation (Fig 2). In a few cases, no viruses were detected in plant samples (lacking bars in Fig 2). These negative samples may reflect the presence of an extreme bottleneck at inoculation, leading to virus population extinction, or a long time lag to systemic infection of the plant (for measurements from 10 to 34 dpi), resulting in the sampling of leaves not yet infected (e.g. DH line 2321). Negative samples were most frequent for the first two dates on which systemically infected leaves were analyzed, *i.e.* at 10 and 14 dpi, probably indicating a time lag to systemic infection in some DH lines. Negative samples were observed in only four DH lines (e.g. DH lines 219 and 2321). No infection was observed in a mean of 3.5 (resp. 2.0) plant samples 10 (resp. 14) dpi for the four DH lines concerned.

The virus populations present in all infected plants and in the common inoculum were analyzed by HTS, to determine the frequencies of the five PVY variants. Inoculum analysis confirmed that all variants were present in roughly equimolar proportions, with 22.6% of variant G, 17.5% of N, 20.6% of K, 17.1% of GK and 22.2% of KN.

The raw data for variant frequency dynamics provided considerably different patterns between the 15 pepper genotypes (Fig 2, S2 and S3 Figs). Variant frequencies were similar between virus populations sampled on the same date in some plant genotypes, consistent with weak genetic drift (e.g. DH lines 240 and 2430, Fig 2A and 2B), whereas they differed in other plant genotypes (e.g. DH lines 2321 and 219, Fig 2D and 2E). Furthermore, the heterogeneity of variant frequencies between the eight plants analyzed fluctuated between dates, probably due to changes in effective population size during the course of infection (e.g. DH line 2344, Fig 2C). The four pepper genotypes for which some samples were virus-negative were also characterized by the highest heterogeneity in variant frequencies, consistent with an extreme bottleneck at inoculation and/or during systemic movement of the virus (see DH lines 2321, 219, 2256 and 2400 in Fig 2D and 2E, S2D and S3I Figs). Selection regimes also differed between lines. In some DH lines, all variants remained present at all dates (e.g. DH line 240, Fig 2A), whereas one variant (e.g. DH line 219, Fig 2E), or up to two variants (e.g. DH lines 2430, 2344, 2321, Fig 2B–2D) became extinct in others.

Validation of the estimation method with numerical simulations

Before its application to the experimental dataset, we validated the estimation method proposed by numerical simulations of a Wright-Fisher model with selection and genetic drift for haploid individuals.

Range of selection and genetic drift intensities explored. The 750 datasets generated in experiment 1 corresponded to very different selection and genetic drift regimes (S4 Fig). Randomly sampled effective population sizes in the inoculated organ and at the onset of systemic infection were combined independently, to simulate highly diverse dynamics of effective population size (see S2 Text). This led to the exploration of a large range of harmonic means of

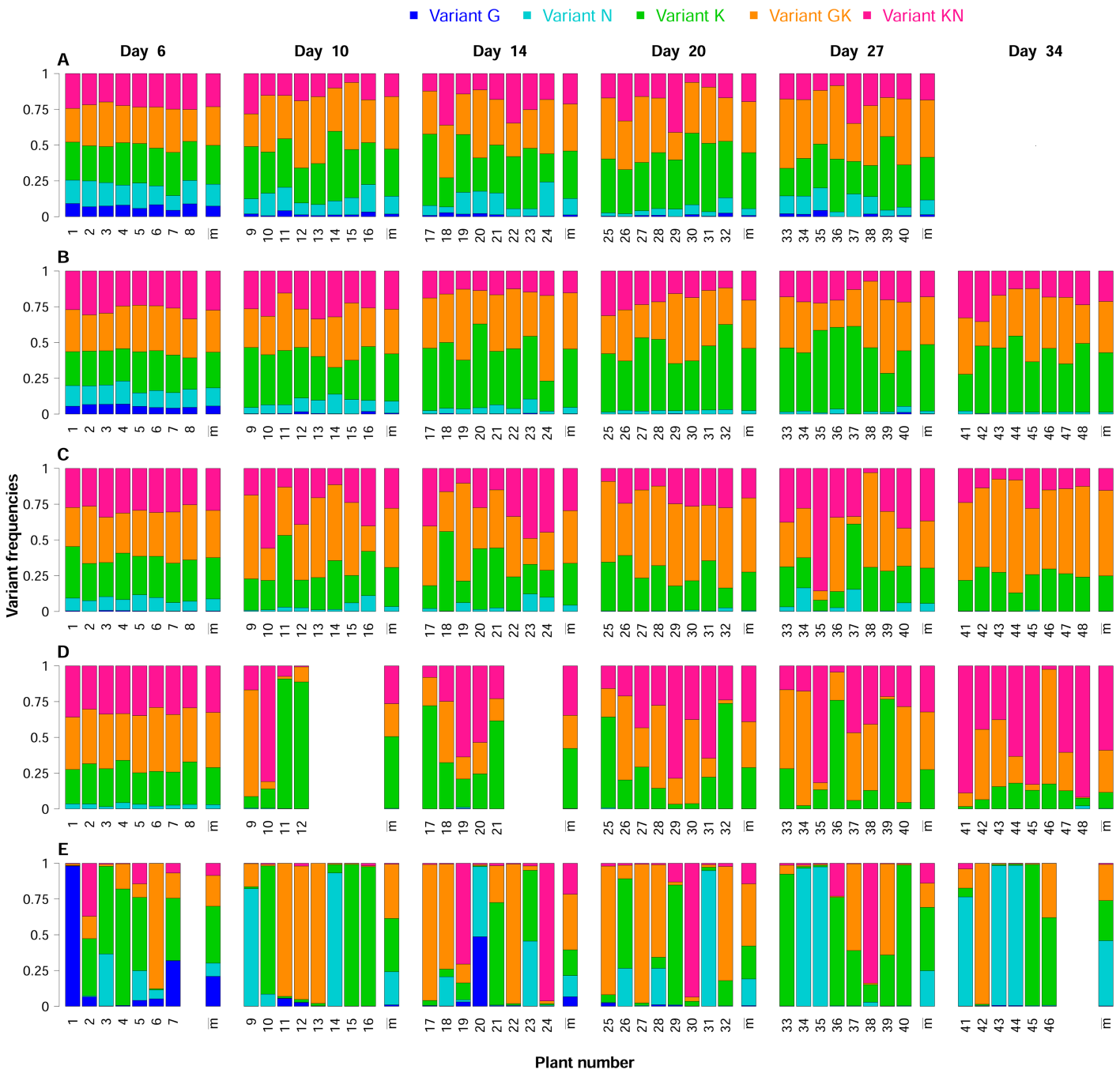

Fig 2. Five contrasting datasets obtained in the biological experiment. Each line of bar plots represents the dynamics of virus variants in a single DH line over time: (A) 240, (B) 2430, (C) 2344, (D) 2321 and (E) 219. We inoculated 48 plants per DH line, and we sampled eight plants, which were subsequently removed from the experiment, at each of the six sampling dates (6, 10, 14, 20, 27 and 34 days post-inoculation). Within each bar plot, the frequencies of the five variants (see top of the figure for the color code) in each infected plant sample are represented by single bars (labeled from 1 to 48). The missing bars correspond to plant samples for which no viruses were detected. The last bar indicates the mean viral composition in the infected plants. Each individual bar plot corresponds to a single sampling date, indicated at the top of each column of barplots. The five DH lines displayed contrasting virus variant dynamics, consistent with contrasting patterns of selection and genetic drift. We could not sample plants of DH line 240 (A) 34 days post-inoculation, because severe necrosis symptoms invading the stem led to the death of all plants at this sampling date.

https://doi.org/10.1371/journal.ppat.1006702.g002

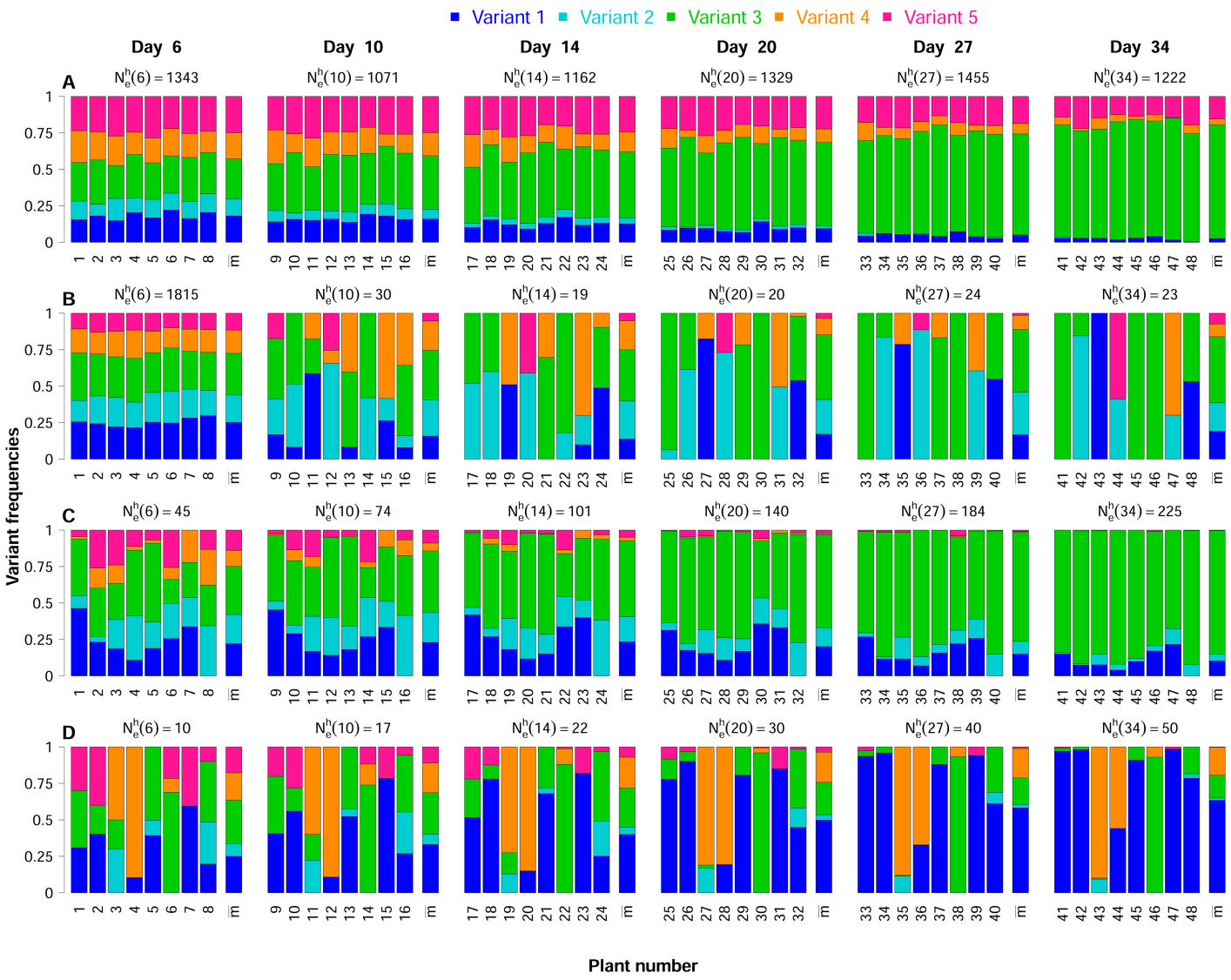

Fig 3. Contrasting datasets obtained in numerical experiment 1. For each dataset (series A to D), the composition of eight populations was observed at six sampling dates, from 6 to 34 days post-inoculation, in independently sampled hosts. Within each plot, each bar represents the composition of the population in one plant at one date, and the last bar shows the mean frequencies over these populations. The color code at the top is used to distinguish the five variants. The harmonic mean of effective population size is indicated in the main title of each plot. The parameter values used for the simulations are: series (A) $\boldsymbol{r}$ = (0.971, 0.92, 1.09, 0.992, 1.027), $N_e^{I^O} = 1343$, $N_e^{S_1} = 822$; series (B) $\boldsymbol{r}$ = (1.05, 1.005, 1.077, 0.963, 0.904), $N_e^{I^O} = 1815$, $N_e^{S_1} = 12$; series (C) $\boldsymbol{r}$ = (1.045, 1.031, 1.12, 0.879, 0.924), $N_e^{I^O} = 45$, $N_e^{S_1} = 1473$; series (D) $\boldsymbol{r}$ = (1.105, 0.943, 0.999, 1.041, 0.912), $N_e^{I^O} = 10$, $N_e^{S_1} = 1025$. Note that $N_e^{S_1}$ is used for the iterative computation of a sequence of effective population sizes varying each five generations during the systemic infection stage.

https://doi.org/10.1371/journal.ppat.1006702.g003

effective population size $N_e^h(t_d)$, ranging from 10 to 1996 (5% quantile = 22, mean = 230, 95% quantile = 873). Relative fitness values (r_i) ranged from 0.75 to 1.27, independently of effective population size (S4 Fig). They reflect a mean absolute selection coefficient $|s|$ of 0.08 (5% quantile = 0.007, 95% quantile = 0.18). As a result, highly diverse simulated datasets were obtained. Overall, the patterns encountered in the experimental datasets (Fig 2) were similar to some of those obtained for the simulated datasets. The simulated datasets also included a number of datasets with more extreme patterns of selection and genetic drift regimes. We illustrate the differences in the genetic drift regimes observed in Fig 3. For a given dataset, variant frequencies could be roughly similar (Fig 3A) or very different (Fig 3D) between populations, at all

Table 2. Performance of the estimators of the harmonic mean of effective population sizes $N_e^h(t_d)$ and variant fitness *r* obtained with the two numerical experiments.

Experiment[a]	Parameter[b]	R^2	Intercept	Slope	Accuracy of 90% CI	Mean bias [95% CI]
Experiment 1	$\hat{r}_i$ (all variants)	0.93	0.02	0.98	91%	10^{-4} [-0.05;0.05]
Experiment 1	$\hat{N}_e^h$ (all dates)	0.86	28	0.91	90%	0.18 [-0.42; 1.48]
Experiment 2	$\hat{r}_i$ (all variants)	0.92	0.06	0.94	89%	8.10^{-4} [-0.07;0.07]
Experiment 2	$\hat{N}_e^h$ (all dates)	0.85	20	0.87	87%	0.03 [-0.53; 1.17]

[a] Experiment 1: 750 simulated datasets with 5 variants under a wide range of selection and genetic drift regimes. Experiment 2: 350 simulated datasets with an additional and undetected sixth variant.

[b] For each parameter, the determination coefficient R^2, the slope and the intercept of the best linear model fit between predicted and true values are given, together with the percentage of true parameter values included in the 90% confidence interval and the mean relative bias and its 95% confidence interval.

https://doi.org/10.1371/journal.ppat.1006702.t002

sampling dates. Moreover, the independent sampling of effective population sizes in the inoculated organ and during systemic infection generated genetic drift regimes that varied over time. For example, we observed strong similarities between populations at the first sampling date, but greater heterogeneity at subsequent dates (Fig 3B) and the opposite pattern (Fig 3C).

Parameter estimation accuracy. Effective population sizes $N_e^h(t_d)$ and intrinsic rates of increase of each variant r_i were inferred for each of the 750 datasets simulated in experiment 1 (with 5 virus variants) and the 350 datasets simulated in experiment 2 (with the 5 virus variants and an additional undetected sixth variant) using the more general model $\mathfrak{M}_4$. True parameters (*i.e.* known parameter values used in the simulations) and estimated parameter values were compared, to assess estimation accuracy.

In numerical experiment 1, HTS analysis provided samples of the true frequencies of the virus variants in the simulated Wright-Fisher populations. The estimates of the intrinsic rates of increase $\hat{r}_i$ were very accurate, with an R^2 of the best-fit line of 0.93, a slope close to 1 (0.98) and an intercept of 0.02 (Table 2, Fig 4A). The estimates of the harmonic mean of the effective population size $\hat{N}_e^h(t_d)$ were also accurate, with a best-fit line close to the first bisector (Table 2, Fig 4B) ($R^2 = 0.86$), despite a slight trend towards overestimation (slope = 0.91, intercept = 28). In both cases, mean relative bias was small and its 95% confidence interval included zero. The 90% confidence intervals of all estimated parameters were highly accurate. They included the true parameter values in nearly 90% (resp. 91%) of the cases for $\hat{N}_e^h(t_d)$ (resp. $\hat{r}_i$) (Table 2).

In numerical experiment 2, we assessed the sensitivity of the estimation method to the presence of a sixth undetected virus variant (see S2 Text). This additional variant was neutral and initially present in the inoculum at a frequency of 3%. It affected virus population dynamics in all 48 host plants of the dataset, because we retained only Wright-Fisher simulations in which the frequency of this sixth variant ranged from 0.01 to 0.06 at 34 dpi. In the 350 simulated datasets, the mean frequency of the sixth variant at all sampling dates and in all plants was 0.07 (5% quantile = 0.01, median = 0.04, 95% quantile = 0.24). However, this variant was considered to be undetected by the HTS method. Thus, HTS analysis provided noisy estimates of the true frequencies of the five virus variants of interest: the mean relative difference between their true frequencies in the simulated population and their measured frequencies was 0.08 (5% quantile = 0.01, median = 0.05, 95% quantile = 0.29). Moreover, inference was performed assuming, as in numerical experiment 1, that the inoculum was an equimolar mixture of the five variants of interest. Despite this detection bias, the estimates of both $N_e^h(t_d)$ and r_i remained consistent (Table 2). The systematic presence of an undetected virus variant at a mean frequency of 0.07 only slightly affected the performance of the estimators (mean relative bias confidence

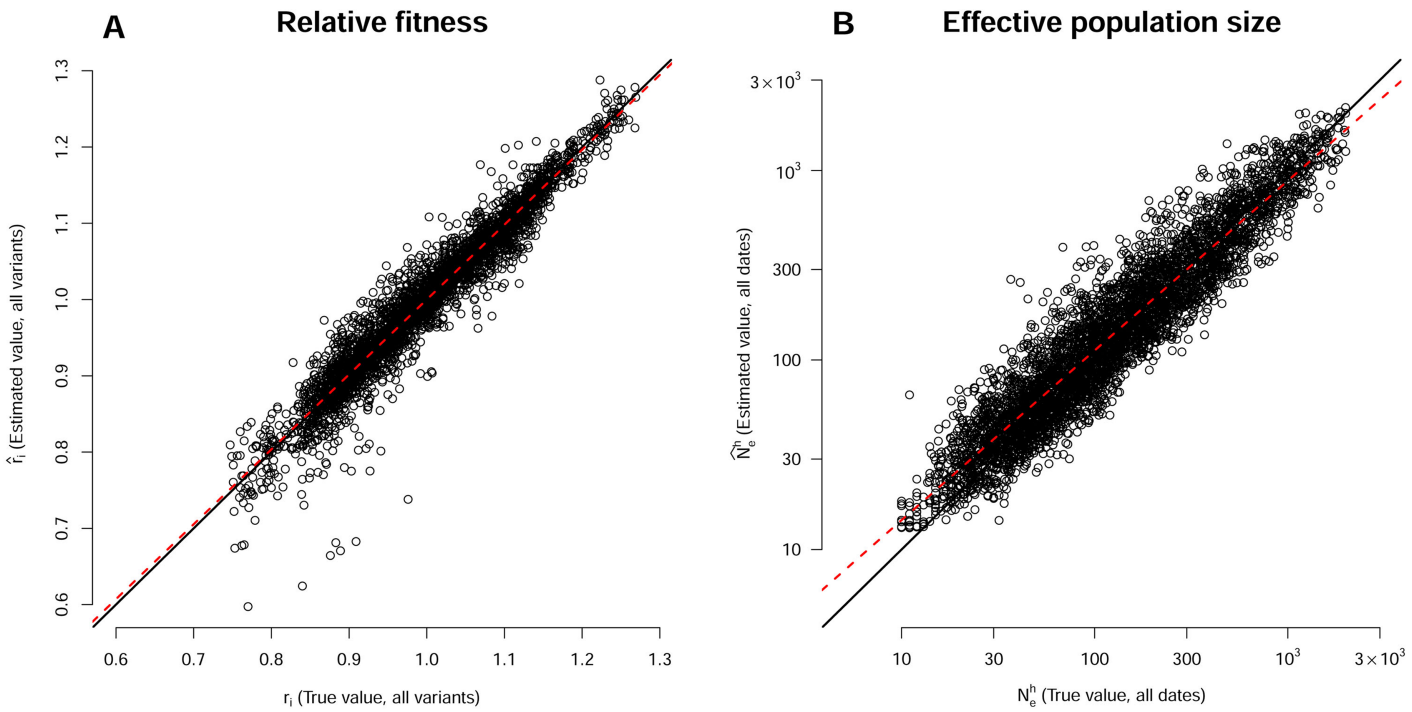

Fig 4. Inferences for variant fitness *r* and for the harmonic mean of effective population size $N_e^h(t_d)$, for the 750 datasets simulated with five virus variants. (A) Correlation between true r_i (x-axis) and estimated $\hat{r}_i$ (y-axis) (all variants considered together). (B) Correlation between true N_e^h (x-axis) and estimated $\hat{N}_e^h$ (y-axis) (all sampling dates considered together, logarithmic scale). In both panels, the black line is the first bisector and the red dashed line is the best-fitting linear model. In panel A, the 9 points with $\hat{r}_i < 0.7$ correspond to datasets in which a highly counterselected variant was observed in only a few plants (5, on average, of the 48 plants) due to an initial low effective population size.

https://doi.org/10.1371/journal.ppat.1006702.g004

intervals systematically included zero, the R^2 of the best-fit lines remained unchanged, 90% of confidence intervals remained highly precise).

In a nutshell, from these numerical simulations with known parameter values, we can conclude that the proposed inference method provides accurate estimates of the intrinsic rates of increase r_i of each variant i, and, thus, of their selection coefficient, together with the dynamics of effective population size $N_e^h(t_d)$ during the time course of the experiment.

Estimation of effective population sizes and variant fitness in the 15 plant genotypes

We estimated the $N_e(t)$ and r_i of the PVY populations in each DH line with a Wright-Fisher model including selection and genetic drift. By contrast to the numerical experiments, the evolutionary parameters underlying the true dynamics of virus populations in their hosts were unknown. The Wright-Fisher model fitted the data very satisfactorily (Fig 5). The best-fit line between observed and fitted mean variant frequencies (averaged over all virus populations and sampling times) was very close to the first bisector (Fig 5A; slope = 0.92, intercept = 0.01, R^2 = 0.92). This was also the case for the variability of variant frequencies between virus populations at each sampling date t_d (Fig 5B; slope = 0.92, intercept = -0.09, R^2 = 0.84). A Wright-Fisher model including selection and genetic drift accurately described the mean evolutionary dynamics of a virus population and the variability of these dynamics between hosts. Due to an identifiability issue (we observed the relative frequencies of variants rather than variant densities), we had to fix the number of generations per day γ. We set this number to 1, a value close

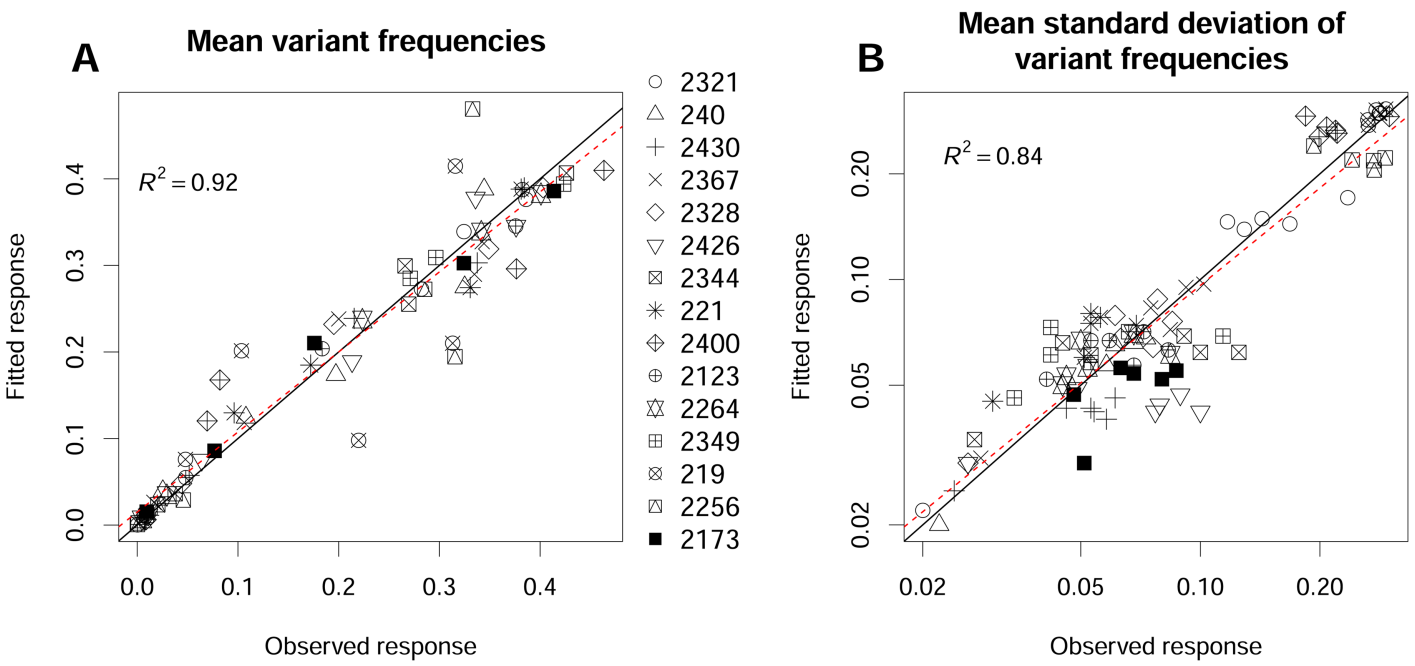

Fig 5. Goodness-of-fit of the Wright-Fisher model $\mathfrak{M}_4$ with the data of the biological experiments. (A) Correlation between the observed mean frequencies of the five virus variants (averaged over all virus populations and sampling times ($mean[f_i^\bullet(\bullet)]$) and their fitted values ($n = 75$). (B) Correlation between the logarithm of the observed mean (averaged over the variants) standard deviation of variant frequencies (between virus populations) at each sampling date t_d ($1/n_{var} \sum_{i=1}^{n_{var}} \sigma[f_i^\bullet(t_d)]$) and their fitted values ($n = 87$). In both panels, the black line is the first bisector and the red dashed line is the best-fitting linear model.

https://doi.org/10.1371/journal.ppat.1006702.g005

to that reported by Khelifa *et al.* [40]. Different γ values would change r_i and $N_e(t)$ estimates to $r_i^{1/\gamma}$ and $\gamma N_e(t)$, but would have no effect on their ranking.

Relative fitness values (r_i) ranged from 0.43 to 1.25 (corresponding to $|s|$: 5% quantile = 0.004, mean = 0.12, 95% quantile = 0.27) and were associated with narrow 90% confidence intervals (S3 Table). The fitness ranks of the PVY variants were very similar in most DH lines (Fig 6A and 6C). Variant G was the weakest in all DH lines, followed by variant N in 13 DH lines. Variant GK was the fittest variant in 13 DH lines, with variant K the fittest variant in the remaining two lines (DH lines 2256 and 2430). Overall, variants K and GK were the two fittest variants in 12 DH lines; variants GK and KN were the two fittest in DH lines 2349 and 2321, and variants N and GK the two fittest in DH line 219. The fitness difference between the weakest and the fittest variants ranged from 0.14 for DH line 219 to 0.81 for DH line 2349.

We further estimated the dynamics of effective population size over the time course of the experiment, as modeled by a piecewise function $N_e(t)$, using a model selection procedure. Four models with one to four parameters were considered. The most general model $\mathfrak{M}_4$ distinguished four successive effective population sizes (one in the inoculated organ and three during systemic infection). $\mathfrak{M}_4$ was the model best supported by the data for five DH lines (2173, 2321, 2328, 2344 and 2367). Model $\mathfrak{M}_3$ distinguished three successive effective population sizes (one in the inoculated organ and two during systemic infection). It was best supported by the data for five DH lines (219, 221, 2256, 240 and 2430). Model $\mathfrak{M}_2$, which distinguished two successive effective population sizes (one in the inoculated organ and one during systemic infection), was selected for a single DH line (2426). Finally, with $\mathfrak{M}_1$, the effective population size of the virus population remained constant. This model was selected in the four remaining DH lines (2123, 2264, 2349 and 2400). The corresponding posterior probabilities of each

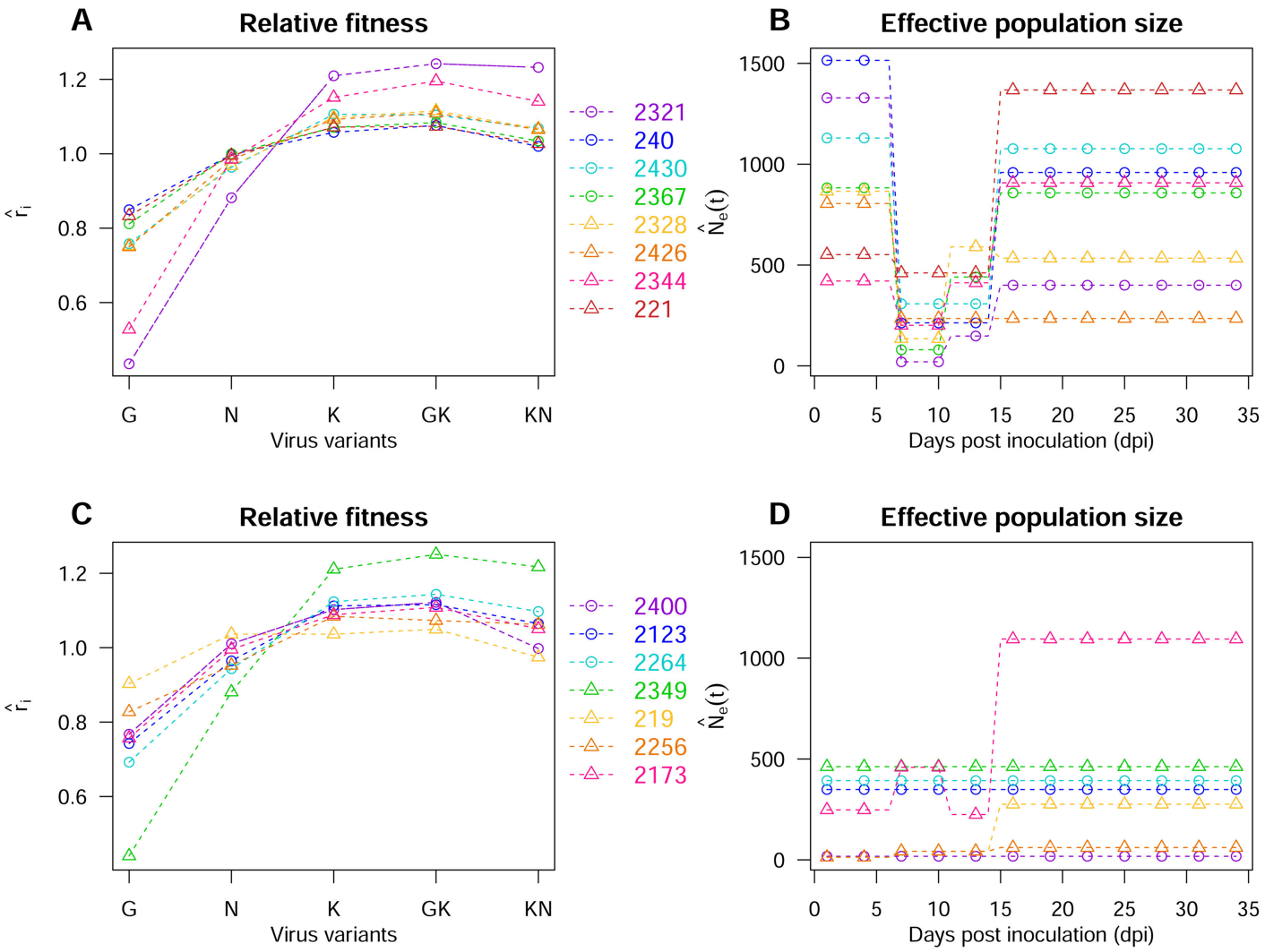

Fig 6. Fitness of virus variants and effective population size estimates for the 15 plant genotypes. (A) Estimates of intrinsic rates of increase $\hat{r}_i$ for each variant *i* for the DH lines 2321, 240, 2430, 2367, 2328, 2426, 2344 and 221. (B) Estimates of effective population size $\hat{N}_e(t)$ during the time course of the experiment for the DH lines listed in (A) and the model best supported by the data. (C) As for (A) for DH lines 2400, 2123, 2264, 2349, 219, 2256 and 2173. (D) As for (B) for the DH lines listed in (C).

https://doi.org/10.1371/journal.ppat.1006702.g006

model are shown in S4 Table, together with effective population size estimates and 90% credibility intervals.

At the first sampling date, considerable variability was observed (Fig 6B and 6D), with effective population sizes ranging from 13 for DH lines 219 and 2256 to 1515 for DH line 240. This was not surprising, given that we chose the DH lines on the basis of the density of primary infection foci in inoculated organs [34] (S1 Fig). A much narrower range of effective population sizes, from 18 to 462, was observed across all plant genotypes at 10 dpi, the first date on which systemic infection was observed. From 6 to 10 dpi, effective population sizes decreased in eight DH lines (Fig 6B), remained approximately constant in six DH lines (Fig 6D) and increased slightly in a single plant genotype (DH line 2173, Fig 6D). Later on, from 10 to 34 dpi, effective population size increased in eight DH lines (mostly DH lines

displaying a bottleneck from 6 to 10 dpi, Fig 6B) and remained approximately constant in the others (mostly in DH lines with lower, *i.e.* < 500, effective population sizes in the inoculated organ, Fig 6D).

Heritability of the intensities of selection and genetic drift exerted by plants on virus populations

By creating two dataset replicates of 24 randomly chosen plants for each DH line, we estimated the heritability of two plant traits corresponding to the evolutionary forces exerted by the plant on virus populations: selection and genetic drift. These forces were estimated by (i) intrinsic rates of increase in viral variants and (ii) effective population sizes for PVY. With 24 plants in each dataset, we used the function $N_e(t)$ of model $\mathfrak{M}_2$ with two parameters. In this approach, we used the contrasting behavior of PVY populations, which were fixed and identical at the time of inoculation in all plants, on different pepper genotypes to characterize the phenotype of each host. Very high heritability estimates were obtained for the intrinsic rates of increase (mean heritability over the five variant estimates: $h^2 = 0.94$). Somewhat lower, but nevertheless substantial heritability estimates were obtained for effective population size in the inoculated organ (mean heritability, $h^2 = 0.64$) and for effective population size during systemic infection (mean heritability, $h^2 = 0.63$). The details of the calculation are provided in S3 Text.

Discussion

Advances in sequencing technologies are revolutionizing the study of microbial evolution [13]. To our knowledge, this study is, for example, the first to suggest such strong variability in the selection and genetic drift regimes experienced by plant viruses in closely related host genotypes (Fig 2). This new type of data paves the way for the estimation of population genetics parameters influencing the fate of pathogen variants of special interest in medicine and agriculture (e.g. variants resistant to pesticides and drugs [44] or, as in this study, variants adapted to host resistance genes). However, estimation methods encompassing the whole range of variation of these parameters are still lacking.

A method for estimating genetic drift and selection from microbial experimental evolution

We present here a method for the estimation of selection and genetic drift in a haploid and asexual organism, as modeled by a Wright-Fisher process. As for any model-based approach, the population of interest must not be too far from an ideal Wright-Fisher population with suitable parameters [10]. The estimation method did not require neutral markers. It was validated for small effective population sizes ($N_e \ll 100$) and a wide range of both positive and negative selection coefficients (weak ($|s| \simeq 0.01$) or strong ($|s| \simeq 0.15$) selection), using simulated datasets. Recent reviews [23, 32] have highlighted the small number of methods available for the inference of selection and genetic drift over the whole range of variation, particularly in the case of small effective population sizes ($N_e \ll 1000$) and strong selection coefficients ($|s| \simeq 0.1$). Indeed, these conditions do not fulfill the hypotheses underlying most approximations of the Wright-Fisher model. The classical approximation, with a standard diffusion process, requires both selection and genetic drift to be weak [23]. Approximations based on Gaussian diffusion require the stochastic effects of genetic drift to decrease more rapidly than the effects of selection [23]. The work of Foll *et al.* [11, 32] constituted a major step forward, but their method requires a large proportion of the genetic markers studied to be neutral. This assumption is not valid for many pathogens with small genomes,

such as viruses. For example, only 22.7% of 66 randomly chosen mutations in the genome of *Tobacco etch virus* (TEV, genus *Potyvirus*), a plant RNA virus, were found to be consistent with neutrality [45]. As the statistical power to detect departure from neutrality is limited, the true proportion of neutral mutations is probably much lower. Similar results have been obtained for bacteria (e.g. [46]).

The estimation method proposed does not require neutral markers, an appealing feature for studying pathogens with small genomes. Lacerda and Seoighe [47] recently developed another method that does not require neutral markers. Their method provided satisfactory estimates of both N_e and s (estimated at a single locus) for a relatively small effective population size of 1000 individuals and values of s up to 0.5. They did not test the performance of their method for $N_e \ll 1000$. By comparison, the method developed here was effective for much lower N_e values, in the range of a few tens of individuals, and for inferring the time course of N_e over a few tens of generations. However, although the range of selection coefficients s included cases of strong selection ($|s| \simeq 0.1$, as defined by Malaspinas [23]), none of the simulation experiments included values as high as 0.5. It may be possible to infer such high selection coefficients with the estimation method proposed, provided that the first generations are sampled more densely, typically every day after inoculation in our set-up. Lacerda and Seoighe [47], for example, used samples taken at each generation, for 20 generations. This makes it possible to record the trajectories of variant frequencies before variant loss or fixation.

The use of the proposed estimation method requires observation of the evolution of isolated populations derived from the same parental population, each population being sampled only once. This design is particularly suitable for studying within-host microbial evolution when several genetically-identical hosts (48 plants for each pepper genotype in our case study) can easily be included in the experiment. With this experimental design, we observed a set of variant frequencies at several time points, in independent hosts. This set contained footprints of selection and genetic drift. In the method developed, selection is evaluated from the mean trajectories of variant frequencies. Genetic drift is evaluated at several time points, by assessing differences in variant frequencies between the replicated populations during the time-course of the experiment. Even for populations with small effective sizes, for which genetic drift and selection have confounding effects on the fate of variants (Fig 2), a moderate number of replicates contains sufficient information to disentangle the two mechanisms. Here, we estimated four selection coefficients and four effective population sizes (*i.e.* 8 parameters) with 48 samples (6 sampling dates × 8 replicates).

The proposed estimation method could be improved further. It explicitly accounts for the technical sampling noise resulting from the assessment of variant frequencies from finite counts of virus sequences. However, HTS also introduces sequencing errors, albeit at a low rate of about 1 substitution per 400 bases for MiSeq technology [48], which were not explicitly accounted for in our framework. Several models have been proposed for separating true genetic variation from technical artifacts [48], and these models could be integrated into the method through a hierarchical Bayesian modeling framework [49], for example. Finally, the method could be extended to take mutation and recombination into account, particularly for experiments over longer periods, in which new variants might appear and displace those currently most abundant. In our short-term experiment, we have already observed *de novo* substitutions in a few plants (removed plant samples, see S1 Text). The inclusion of recombination is not relevant for our case study, as the nucleotide positions differentiating the variants are located only a few codons apart. Recombination can thus be ignored in this study [50], particularly given the small number of generations considered [32].

Plant genotypes modulate genetic drift and selection within virus populations

On the host side, our experiment involved 15 DH lines of pepper, all carrying the major resistance gene *pvr2*^{3}, but differing in terms of their genetic backgrounds [12]. These DH lines were derived from the F_1 hybrid between two pepper lines, Perennial and Yolo Wonder. Consequently, on average, any pair of DH lines have 50 percent of alleles in common for markers differentiating between Perennial and Yolo Wonder. This is the first study, to our knowledge, to show such a high level of diversity in selection and genetic drift regimes experienced by virus populations from the same viral inoculum in closely related host genotypes (Fig 2, S2 and S3 Figs). On the pathogen side, we used five virus variants: the G and N variants displayed weaker adaptation to *pvr2*^{3} than the K, GK and KN variants. The ranking of the selection coefficients of the five variants was mostly identical in the 15 plant genotypes. We were therefore unable to identify any host genotype, among those tested, able to counterselect against the virus variants best adapted to *pvr2*^{3}. This may be due to (i) the strong selective effect exerted by the major-effect resistance gene *pvr2*^{3}, which is present in all the DH lines studied here and probably exceeds the additional selective effect of the plant genetic background and/or (ii) the close genetic relatedness of the DH lines analyzed. Other genetic resources for pepper should be explored, to identify genotypes capable of counterselecting against the K, GK and KN variants, which were the fittest in our study. The best candidates for this would be pepper genotypes carrying *pvr2* resistance alleles other than *pvr2*^{3}, with a different specificity in the face of PVY diversity [51], or pepper genotypes devoid of resistance alleles at the *pvr2* locus, as shown by Quenouille *et al.* [12]. Combinations of plant genotypes exerting opposite selective pressures on pathogen populations are particularly interesting for the sustainable management of plant resistance at landscape level, and can be implemented in cultivar rotations, mixtures or mosaics [52]. However, in our study, the difference in fitness between the weakest and fittest variants differed between host genotypes. The dynamics of selection for the fittest variants were under plant genetic control and could therefore be modulated by the choice of plant genotypes grown. For example, growing the pepper DH lines with the smallest differential selection between the five PVY variants would be particularly useful for delaying PVY adaptation in *pvr2*^{3}-carrying plants, in which a two-step mutational trajectory may be required [12]. Indeed, the G and N variants are most likely to appear initially, because they require transitions, whereas the K variant requires a transversion, and transitions are more frequent than transversions [53]. However, an additional substitution, in a second step, is required to confer a sufficient level of fitness for the emergence of GK and KN variants. These mutational trajectories were observed in PVY adaptation to the Perennial pepper genotype, the resistant parent of all the DH lines studied here [12].

We also inferred the time course of the genetic drift experienced by the viruses in the 15 host environments during the experiment. Genetic drift intensities were highly variable with time and between plant genotypes, revealing an unprecedented level of variability between closely related host genotypes. Our estimates of $N_e(t)$ ranged from 18 to 462 just after the colonization of apical leaves at 10 dpi, and from 13 to 1515 in the inoculated leaves four days previously (at 6 dpi). Eight of the 15 DH lines displayed a high N_e in the inoculated leaves at 6 dpi (from 421 to 1515), a decrease at 10 dpi (N_e (10 dpi) values of 1.5 to 83.5% of the value at 6 dpi) and a subsequent increase (Fig 6B). This pattern suggests a founder effect, in which a new PVY population in apical leaves is set up by a few members of the original population in the inoculated leaf. In the remaining seven DH lines, the N_e of the inoculated leaves at 6 dpi was much lower (from 13 to 462), and N_e values often remained low in the apical leaves (Fig 6D). However, an increase in N_e was observed in DH lines 219 and 2173, after 14 dpi. This result sheds

new light on the importance of the within-host bottlenecks experienced by virus populations, as discussed in a recent article by Zwart *et al.* [54], who reported that the N_e of TEV in the first systemically infected leaf of tobacco plants was determined largely by inoculum viral load. They then hypothesized that genetic drift occurred mostly during the inoculation process. Previous estimations of N_e for viruses did not focus on N_e dynamics at the whole-plant level as in this study. Instead, they considered the multiplicity of infection (MOI) during cell-to-cell movement or N_e during the colonization of apical leaves (for a comprehensive review, see Gutiérrez *et al.* [4]). Direct comparisons with these studies are, therefore, not appropriate. Gutiérrez *et al.* [55] recently showed that *Turnip mosaic virus* (genus *Potyvirus*) infections are characterized by a very low MOI ($\simeq 1$) when cells are infected with virus particles moving in the plant vasculature, and a much higher MOI ($\simeq 30$) during subsequent cell-to-cell movement in the mesophyll. The general picture that emerges when we consider both these MOI patterns and plant growth dynamics is consistent with our observations. Indeed, the lowest N_e values were observed at 10 dpi, corresponding to the onset of systemic infection, when plants were small and consisted essentially of a few infected leaves. N_e tends often to increase with time, because (i) increasing numbers of leaves are infected and behave as virus sources as the plant grows and (ii) leaf areas increase, probably increasing the relative proportion of cell-to-cell, as opposed to long-distance, virus movement.

One of the key results of this study is the finding that the effective population size of PVY is a heritable plant trait. The high heritability estimated for N_e (partially due to the use of a DH progeny of pepper genotypes) indicates that plant resistance could potentially be improved through breeding programs. Indeed, our findings pave the way for the breeding of plant cultivars exposing viruses to greater genetic drift. This would provide a twofold benefit against viruses. First, in asexual populations, genetic drift favors the accumulation of deleterious mutations, decreasing viral fitness (Muller's ratchet) [56]. Second, genetic drift decreases the fixation probability of beneficial mutations, such as those responsible for overcoming plant resistance genes [57]. Breeding for greater genetic drift in virus populations would thus constitute a novel approach to increasing the durability of resistance to plant viruses in agricultural landscapes [52, 58, 59]. Another key result is the finding that the Wright-Fisher model accurately captures the major processes driving the within-host dynamics of a set of virus variants (Fig 5), despite being much simpler than the underlying mechanisms involved in the infection of highly structured hosts. Over longer periods, mutation and recombination increase in importance and this can easily be encompassed in the Wright-Fisher model [60]. This model can thus serve as a valuable cornerstone for linking the within- and between-host scales of disease dynamics and studying, for example, how breeding for greater genetic drift can delay the emergence of a new pathogen variant.

Supporting information

S1 Text. Sequence analyses to detect PVY mutations.
(PDF)

S2 Text. Numerical experiments.
(PDF)

S3 Text. Heritability of the intensity of selection and genetic drift exerted by plants on virus populations.
(PDF)

S1 Fig. Resistance-breakdown (RB) frequency, viral accumulation and mean number of primary infection foci for the 15 DH lines studied. Pepper genotypes are represented as

points, with their nomenclature (DH line number) given above each point. We estimated the mean number of primary infection foci for the 15 DH lines with the *Potato virus Y* (PVY, genus *Potyvirus*) variant K, carrying a green fluorescent marker (green fluorescent protein, GFP) [34]. The resistance-breakdown (RB) frequency and the relative viral accumulation were estimated by Quenouille *et al.* [12]. The RB frequency corresponds to the percentage of infected plants when inoculated with an avirulent variant regarding the allele of resistance $pvr2^3$, carried by all DH lines. The relative viral accumulation, or relative viral concentration, was measured by double antibody sandwich enzyme-linked immunosorbent assay (DAS-E-LISA).
(PDF)

S2 Fig. Five datasets obtained by high-throughput sequencing in the biological experiment. Each line of bar plots represents the dynamics of virus variants in a single DH line over time: (A) 221, (B) 2123, (C) 2173, (D) 2256 and (E) 2264. Within each bar plot, the frequencies of the five variants (see top of the figure for the color code) in each infected plant sample are represented by single bars (labeled from 1 to 48). The missing bars correspond to plant samples for which no viruses were detected. The last bar indicates the mean viral composition in the infected plants. Each individual bar plot corresponds to a single sampling date, indicated at the top of each column of barplots.
(PDF)

S3 Fig. Five datasets obtained by high-throughput sequencing in the biological experiment. Each line of bar plots represents the dynamics of virus variants in a single DH line over time: (F) 2328, (G) 2349, (H) 2367, (I) 2400 and (J) 2426. Within each bar plot, the frequencies of the five variants (see top of the figure for the color code) in each infected plant sample are represented by single bars (labeled from 1 to 48). The missing bars correspond to plant samples for which no viruses were detected. The last bar indicates the mean viral composition in the infected plants. Each individual bar plot corresponds to a single sampling date, indicated at the top of each column of barplots.
(PDF)

S4 Fig. Variability of the selection and genetic drift regimes obtained among the simulated datasets in numerical experiment 1. In the diagonal, each histogram represents the distribution of input parameters r_1 (intrinsic rate of increase of variant 1), N_e^{IO} (effective population size in the inoculated organ) and $N_e^{S_1}$ (effective population size at the onset of the systemic infection) used to simulate the 750 datasets. Off-diagonal scatter plots are two by two combinations of parameters.
(PDF)

S1 Table. Tag sequences used to distinguish each plant sample after pooling and MiSeq Illumina high-throughput sequencing. The forward (Fwd.) primer sequence was the same for all amplifications and was bound to the sequence tag, just after it. Its binding site corresponds to positions 5971 to 5990 of PVY isolate SON41p (accession number AJ439544). The binding site of the reverse (Rev.) primer sequence corresponds to positions 6095 to 6114 of PVY isolate SON41p. RT-PCR amplifications were done according to the following profile: 1h at 42˚C, 10 min at 95˚C, 35 times the following sequence (45s at 95˚C, 30s at 50˚C and 20s at 72˚C) and finally 10 min at 72˚C.
(PDF)

S2 Table. Number of sequences and composition of the virus populations in each sample of the biological experiment. In all, 708 samples were analyzed: 15 doubled-haploid (DH)

lines of pepper × 6 sampling dates (dpi: days post-inoculation) × 8 plants per date, except for virus-negative samples, and 4 samples for the initial inoculum. Columns indicate (i) the name of each DH line, (ii) the sampling date in dpi, (iii) the number of the sequence tag used (see S1 Table), (iv) the plant number (as in Fig 2, S2 and S3 Figs), (v) the infection status of each sample (0: not infected / 1: infected), (vi) the total number of cleaned sequences of variants G, N, K, GK and KN assigned to each sample after filtering, (vii-xi) the number of sequences for each inoculated viral variant (G, N, K, GK and KN), and (xii-xiv) the number of sequences of the three additional variants SON41p, GN and GKN, based on the three codon positions of interest in the VPg cistron.
(XLSX)

S3 Table. Estimations of the relative intrinsic rates of increase of the virus variants for the 15 plant genotypes. Virus variants are indexed as follows: (i) r_1 virus variant G, (ii) r_2 virus variant N, (iii) r_3 virus variant K, (iv) r_4 virus variant GK, (v) r_5 virus variant KN. The 90% confidence intervals are calculated as $\hat{r}_i \pm 1.645.\hat{\sigma}_i$.
(PDF)

S4 Table. Model selection and estimations of the effective population sizes for the 15 plant genotypes. The posterior probabilities of the four models considered for the piecewise function describing the temporal variation of the effective population sizes during the time course of the experiment (models $\mathfrak{M}_1$, $\mathfrak{M}_2$, $\mathfrak{M}_3$ and $\mathfrak{M}_4$) are first indicated. The bold value corresponds to the model that is best supported by the data. The next columns indicate the estimation of the effective population sizes of the model selected and the extent of the 90% credibility intervals.
(PDF)

Acknowledgments

We would like to thank Fabio Zanini (Max Planck Institute) and Nicolas Parisey (INRA Rennes) for their invaluable help with FFPopSim and Rcpp softwares, respectively, used in an earlier version of this manuscript. We also would like to thank Grégory Girardot and Baptiste Lederer for their precious help during experiments. The simulations were carried out with the Avakas (Bordeaux University) and Migale (INRA Jouy en Josas) computer clusters.

Author Contributions

Conceptualization: Elsa Rousseau, Benoît Moury, Ludovic Mailleret, Rachid Senoussi, Alain Palloix, Frédéric Grognard, Frédéric Fabre.

Data curation: Elsa Rousseau, Benoît Moury, Vincent Simon, Sophie Valière.

Formal analysis: Elsa Rousseau, Benoît Moury, Ludovic Mailleret, Frédéric Grognard, Frédéric Fabre.

Funding acquisition: Benoît Moury, Ludovic Mailleret, Alain Palloix, Frédéric Grognard, Frédéric Fabre.

Investigation: Elsa Rousseau, Benoît Moury, Alain Palloix, Vincent Simon, Sophie Valière, Frédéric Fabre.

Methodology: Elsa Rousseau, Benoît Moury, Ludovic Mailleret, Rachid Senoussi, Alain Palloix, Frédéric Grognard, Frédéric Fabre.

Supervision: Benoît Moury, Ludovic Mailleret, Frédéric Grognard, Frédéric Fabre.

Visualization: Elsa Rousseau, Frédéric Fabre.

Writing – original draft: Elsa Rousseau, Benoît Moury, Ludovic Mailleret, Rachid Senoussi, Frédéric Grognard, Frédéric Fabre.

Writing – review & editing: Elsa Rousseau, Benoît Moury, Frédéric Fabre.

References

1. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 2009; 10:195–205. https://doi.org/10.1038/nrg2526 PMID: 19204717
2. Wright S. Evolution in Mendelian populations. Genetics. 1931; 16(2):97–159. PMID: 17246615
3. Vucetich JA, Waite TA, Nunney L. Fluctuating population size and the ratio of effective to census population size. Evolution. 1997; 51(6):2017–2021. https://doi.org/10.1111/j.1558-5646.1997.tb05123.x PMID: 28565105
4. Gutiérrez S, Michalakis Y, Blanc S. Virus population bottlenecks during within-host progression and host-to-host transmission. Current Opinion in Virology. 2012; 2:546–555. https://doi.org/10.1016/j.coviro.2012.08.001 PMID: 22921636
5. Waples RS, Antao T, Luikart G. Effects of overlapping generations on linkage disequilibrium estimates of effective population size. Genetics. 2014; 197:769–780. https://doi.org/10.1534/genetics.114.164822 PMID: 24717176
6. Wright S. Statistical genetics in relation to evolution. In: Hermann, editor. Exposés de Biométrie et de Statistique Biologique. Paris; 1939.
7. Kimura M, Crow JF. The measurment of effective population number. Evolution. 1963; 17(3):279–288.
8. Caballero A. Developments in the prediction of effective population size. Heredity. 1994; 73:657–679. https://doi.org/10.1038/hdy.1994.174 PMID: 7814264
9. Motro U, Thomson G. On heterozygosity and the effective size of populations subject to size changes. Evolution. 1982; 36(5):1059–1066. https://doi.org/10.2307/2408083 PMID: 28567820
10. Rouzine IM, Rodrigo A, Coffin JM. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application in virology. Microbiol Mol Biol Rev. 2001; 65:151–185. https://doi.org/10.1128/MMBR.65.1.151-185.2001 PMID: 11238990
11. Foll M, Poh YP, Renzette N, Ferrer-Admettla A, Bank C, Shim H, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 2014; 10(2):e1004185. https://doi.org/10.1371/journal.pgen.1004185 PMID: 24586206
12. Quenouille J, Montarry J, Palloix A, Moury B. Farther, slower, stronger: how the plant genetic background protects a major resistance gene from breakdown. Molecular Plant Pathology. 2013; 14:109–118. https://doi.org/10.1111/j.1364-3703.2012.00834.x PMID: 23046402
13. Brockhurst MA, Colegrave N, Rozen DE. Next-generation sequencing as a tool to study microbial evolution. Molecular Ecology. 2011; 20:972–980. https://doi.org/10.1111/j.1365-294X.2010.04835.x PMID: 20874764
14. Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci. 2004; 101(22):8396–8401. https://doi.org/10.1073/pnas.0400146101 PMID: 15159545
15. Elena SF, Fraile A, García-Arenal F. Evolution and Emergence of Plant Viruses. Adv Virus Res. 2014; 88:161–191.
16. Nei M, Tajima F. Genetic drift and estimation of effective population size. Genetics. 1981; 98:625–640. PMID: 17249104
17. Waples RS. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics. 1989; 121:379–391. PMID: 2731727
18. Williamson EG, Slatkin M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics. 1999; 152:755–761. PMID: 10353915
19. Anderson EC, Williamson EG, Thompson EA. Monte Carlo evaluation of the likelihood for N_e from temporal spaced samples. Genetics. 2000; 156:2109–2118. PMID: 11102399
20. Berthier P, Beaumont MA, Cornuet JM, Luikart G. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. Genetics. 2002; 160:741–751. PMID: 11861575

21. Vitalis R, Gautier M, Dawson KJ, Beaumont MA. Detecting and measuring selection from gene frequency data. Genetics. 2014; 196:799–817. https://doi.org/10.1534/genetics.113.152991 PMID: 24361938

22. Illingworth CJR, Parts L, Schiffels S, Liti G, Mustonen V. Quantifying selection acting on a complex trait using allele frequency time series data. Molecular Biology and Evolution. 2012; 29(4):1187–1197. Available from: http://mbe.oxfordjournals.org/content/29/4/1187.abstract. PMID: 22114362

23. Malaspinas AS. Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. Molecular Ecology. 2016; 25(1):24–41. Available from: http://dx.doi.org/10.1111/mec.13492. PMID: 26613371

24. Bollback JP, York TL, Nielsen R. Estimation of $2N_es$ from temporal allele frequency data. Genetics. 2008; 179:497–502. https://doi.org/10.1534/genetics.107.085019 PMID: 18493066

25. Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-serial data. Genetics. 2012; 192:599–607. https://doi.org/10.1534/genetics.112.140939 PMID: 22851647

26. Mathieson I, McVean G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics. 2013; 193:973–984. https://doi.org/10.1534/genetics.112.147611 PMID: 23307902

27. Steinrücken M, Bhaskar A, Song YS. A novel spectral method for inferring general diploid selection from time series genetic data. Ann Appl Stat. 2014; 8(4):2203–2222. Available from: http://dx.doi.org/10.1214/14-AOAS764. PMID: 25598858

28. French RC, Stenger DC. Evolution of Wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. Annu Rev Phytopathol. 2003; 41:199–214. https://doi.org/10.1146/annurev.phyto.41.052002.095559 PMID: 12730393

29. García-Arenal F, Fraile A, Malpica JM. Variation and evolution of plant virus populations. Int Microbiol. 2003; 6:225–232. https://doi.org/10.1007/s10123-003-0142-z PMID: 13680390

30. Sacristán S, Malpica JM, Fraile A, García-Arenal F. Estimation of population bottlenecks during systemic movement of *Tobacco mosaic virus* in tobacco plants. J Virol. 2003; 77(18):9906–9911. https://doi.org/10.1128/JVI.77.18.9906-9911.2003 PMID: 12941900

31. Elena SF, Bedhomme S, Carrasco P, Cuevas JM, de la Iglesia F, Lafforgue G, et al. The evolutionary genetics of emerging plant RNA Viruses. Mol Plant Microbe In. 2011; 24(3):287–293. https://doi.org/10.1094/MPMI-09-10-0214 PMID: 21294624

32. Foll M, Shim H, Jensen JD. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular Ecology Resources. 2014; 15 (1): 87–98. https://doi.org/10.1111/1755-0998.12280 PMID: 24834845

33. Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genet. 2011; 7(3):e1001336. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1001336. PMID: 21437274

34. Tamisier L, Rousseau E, Baraillé S, Nemouchi G, Szadkowski M, Mailleret L, Grognard F, Fabre F, Moury B, Palloix A. Quantitative trait loci in pepper control the effective population size of two RNA viruses at inoculation. Journal of General Virology. 2017; 98:1923–1931. Available from: 10.1099/jgv.0.000835 PMID: 28691663

35. Fabre F, Montarry J, Coville J, Senoussi R, Simon V, Moury B. Modelling the evolutionary dynamics of viruses within their hosts: a case study using high-throughput sequencing. PLoS Pathog. 2012; 8:1–9. https://doi.org/10.1371/journal.ppat.1002654

36. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology. 2007; 8(7):R143 Available from: 10.1186/gb-2007-8-7-r143 PMID: 17659080

37. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27(21):2957–2963. https://doi.org/10.1093/bioinformatics/btr507 PMID: 21903629

38. R Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2013. Available from: http://www.R-project.org/.

39. Gallais A. Théorie de la sélection en amélioration des plantes. Paris: Masson; 1990.

40. Khelifa M, Massé D, Blanc S, Drucker M. Evaluation of the minimal time of *Cauliflower mosaic virus* in different hosts. Virology. 2010; 396:238–245. PMID: 19913268

41. Ewens WJ. Mathematical population genetics 1—Theoretical introduction. Antman S, Marsden J, Sirovich L, Wiggins S, editors. Springer-Verlag; 2004.

42. Musso F. On the relation between the Eigen model and the asexual Wright–Fisher model. Bulletin of Mathematical Biology. 2012; 74 (1):103–115. https://doi.org/10.1007/s11538-011-9666-0 PMID: 21656088

43. Lenormand M, Jabot F, Deffuant G. Adaptive approximate Bayesian computation for complex models. Computational Statistics. 2013; 28:2777–2796 doi: 10.1007/s00180-013-0428-3

44. Consortium, R. E. X. Heterogeneity of selection and the evolution of resistance. Trends in Ecology & Evolution. 2013; 2:110–118. Available from: http://www.sciencedirect.com/science/article/pii/S0169534712002352.

45. Carrasco P, de la Iglesia F, Elena SF. Distribution of fitness and virulence effects caused by single-nucleotide substitutions in *Tobacco etch virus*. J Virol. 2007; 81(23):12979–12984. https://doi.org/10.1128/JVI.00524-07 PMID: 17898073

46. Kassen R, Bataillon T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nature Genetics. 2006; 38:484–488. 4. Available from::// 000236340500024. https://doi.org/10.1038/ng1751 PMID: 16550173

47. Lacerda M, Seoighe C. Population genetics inference for longitudinally-sampled mutants under strong selection. Genetics. 2014; 198(3):1237–1250. Available from: http://genetics.org/content/198/3/1237. PMID: 25213172

48. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Briefings in Bioinformatics. 2016; 17(1):154–179. https://doi.org/10.1093/bib/bbv029 PMID: 26026159

49. Clark JS. Why environmental scientists are becoming Bayesians. Ecology Letters. 2005; 8:2–14. https://doi.org/10.1111/j.1461-0248.2004.00702.x

50. Rhodes TD, Nikolaitchik O, Chen J, Powell D, Hu WS. Genetic recombination of Human immunodeficiency virus Type 1 in one round of viral replication: Effects of genetic distance, target cells, accessory genes, and lack of high negative interference in crossover events. J Virol. 2005; 79(3):1666–1677. https://doi.org/10.1128/JVI.79.3.1666-1677.2005 PMID: 15650192

51. Moury B, Janzac B, Ruellan Y, Simon V, Khalifa MB, Fakhfakh H, Fabre F, Palloix A. Interaction Patterns between *Potato Virus Y* and eIF4E-Mediated Recessive Resistance in the *Solanaceae*. J Virol. 2014; 88(17):9799–9807. Available from: 10.1128/JVI.00930-14. PMID: 24942572

52. Djidjou-Demasse R, Moury B, Fabre F. Mosaics often outperform pyramids: insights from a model comparing strategies for the deployment of plant resistance genes against viruses in agricultural landscapes. New Phytologist. 2017; 216:239–253. Available from: 10.1111/nph.14701 PMID: 28776688

53. Ayme V, Souche S, Caranta C, Jacquemond M, Chadoeuf J, Palloix A, Moury B. Different mutations in the genome-linked protein VPg of *Potato virus Y* confer virulence on the *pvr2*^{3} resistance in pepper. Mol Plant Microbe Interact. 2006; 19:557–563. https://doi.org/10.1094/MPMI-19-0557 PMID: 16673943

54. Zwart MP, Daròs JA, Elena SF. One is enough: *in vivo* effective population size is dose-dependent for a plant RNA virus. PLoS Pathog. 2011; 7:1–12. https://doi.org/10.1371/journal.ppat.1002122

55. Gutiérrez S, Pirolles E, Yvon M, Baecker V, Michalakis Y, Blanc S. The multiplicity of cellular infection changes depending on the route of cell infection in a plant virus. J Virol. 2015; 89(18):9665–9675. Available from: http://jvi.asm.org/content/89/18/9665. PMID: 26178988

56. de la Iglesia F, Elena SF. Fitness declines in *Tobacco etch virus* upon serial bottleneck transfer. J Virol. 2007; 81(10):4941–4947. https://doi.org/10.1128/JVI.02528-06 PMID: 17344305

57. Patwa Z, Wahl LM. The fixation probability of beneficial mutations. J R Soc Interface. 2008; 5:1279–1289. https://doi.org/10.1098/rsif.2008.0248 PMID: 18664425

58. Fabre F, Rousseau E, Mailleret L, Moury B. Durable strategies to deploy plant resistance in agricultural landscapes. New Phytologist. 2012; 193:1064–1075. https://doi.org/10.1111/j.1469-8137.2011.04019x PMID: 22260272

59. Fabre F, Rousseau E, Mailleret L, Moury B. Epidemiological and evolutionary management of plant resistance: optimizing the deployment of cultivar mixtures in time and space in agricultural landscapes. Evolutionary Applications. 2015; 8(10):919–932. Available from: http://dx.doi.org/10.1111/eva.12304. PMID: 26640518

60. Zanini F, Neher RA. FFPopSim: an efficient forward simulation package for the evolution of large populations. Bioinformatics. 2012; 28 (24):3332–3333. https://doi.org/10.1093/bioinformatics/bts633 PMID: 23097421