Journal of General Virology (2012), **93**, 1831–1840 DOI 10.1099/vir.0.042622-0

Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus

H. E. Simmons,[1,2] J. P. Dunham,[3] J. C. Stack,[1] B. J. A. Dickins,[4] I. Pagán,[1,5] E. C. Holmes[1,6] and A. G. Stephenson[1]

Correspondence
H. E. Simmons
hsimmons@iastate.edu

[1]Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

[2]Seed Science Center, Iowa State University, Ames, IA 50011, USA

[3]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90033, USA

[4]The Huck Institutes for the Life Sciences and Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA

[5]Centro de Biotecnología y Genómica de Plantas (UPM-INIA), Campus de Montegancedo, Universidad Politécnica de Madrid, 28223, Pozuelo de Alarcón (Madrid), Spain

[6]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

The genetic diversity present in populations of RNA viruses is likely to be strongly modulated by aspects of their life history, including mode of transmission. However, how transmission mode shapes patterns of intra- and inter-host genetic diversity, particularly when acting in combination with *de novo* mutation, population bottlenecks and the selection of advantageous mutations, is poorly understood. To address these issues, this study performed ultradeep sequencing of zucchini yellow mosaic virus in a wild gourd, *Cucurbita pepo* ssp. *texana*, under two infection conditions: aphid vectored and mechanically inoculated, achieving a mean coverage of approximately 10 000×. It was shown that mutations persisted during inter-host transmission events in both the aphid vectored and mechanically inoculated populations, suggesting that the vector-imposed transmission bottleneck is not as extreme as previously supposed. Similarly, mutations were found to persist within individual hosts, arguing against strong systemic bottlenecks. Strikingly, mutations were seen to go to fixation in the aphid-vectored plants, suggestive of a major fitness advantage, but remained at low frequency in the mechanically inoculated plants. Overall, this study highlights the utility of ultradeep sequencing in providing high-resolution data capable of revealing the nature of virus evolution, particularly as the full spectrum of genetic diversity within a population may not be uncovered without sequence coverage of at least 2500-fold.

Received 1 March 2012
Accepted 13 May 2012

INTRODUCTION

Understanding the factors that generate and maintain genetic diversity is the central goal of evolutionary genetics. RNA viruses are ideally suited for the study of the determinants of genetic variation because of their extremely high mutation rates, itself due to the lack of error correction associated with replication by an RNA-dependent RNA polymerase, and their rapid replication (Duffy *et al.*, 2008). This capacity to generate genetic diversity is central to the ability of RNA viruses to breakdown host resistance mechanisms (Acosta-Leal *et al.*, 2010; Feuer *et al.*, 1999; Lech *et al.*, 1996), to adapt to new niches (Roossinck, 1997), including new hosts (Jerzak *et al.*, 2008), and for changes in virulence (Acosta-Leal *et al.*, 2011).

For any RNA virus, the extent and structure of the genetic variation that occurs within individual hosts is due to a combination of *de novo* mutation, genetic diversity generated through mixed infection, natural selection and stochastic processes such as genetic drift and the population bottlenecks that occur both within and among hosts. However, the roles played by these differing processes in

The GenBank/EMBL/DDBJ accession numbers for the consensus nucleotide sequences determined in this study are JN192405–JN192428 and JQ716413.

A supplementary table is available with the online version of this paper.

shaping intra-host genetic variation are uncertain. For example, given the extremely large population sizes that plant RNA viruses can achieve [e.g. tobacco mosaic virus (TMV) has been documented to reach levels of 10^{11}–10^{12} virions per infected leaf; García-Arenal *et al.*, 2001], it might be expected that selection would act efficiently within hosts. However, several studies indicate that the effective population size (N_e) of RNA viruses in nature is several orders of magnitude lower than the census population number (García-Arenal *et al.*, 2001, 2003; Hughes, 2009), and the duration of infection in a single host may be of insufficient length to enable natural selection to fix beneficial mutations. As such, stochastic processes may be more important determinants of genetic diversity than selection at the intra-host level.

Population bottlenecks may be particularly important in plant RNA viruses. Such bottlenecks are thought to occur during both inter-host vector transmission and systemic movement within an individual plant. For example, it has been estimated that an average of three virions transmit from mechanically infected squash plants to healthy plants via aphids (Ali *et al.*, 2006), and even lower numbers have been observed in cucumber mosaic virus infection (Betancourt *et al.*, 2008). Likewise, very low numbers of virions were involved in the transmission of potato virus Y (PVY) from an artificial medium (Moury *et al.*, 2007). Similarly, drastic population bottlenecks have been reported during systemic movement. Estimates of the founding population in a new leaf after systemic movement during TMV infection ranged between two and 20 virions (Sacristán *et al.*, 2003), and only four virions of wheat streak mosaic virus were involved in the invasion of new tillers of wheat (French & Stenger, 2003). Population bottlenecks have also been observed at a cellular level. For example, for TMV, six viral genomes have been shown to infect a cell, which decreases to one to two virions as the virus moves systemically (González-Jara *et al.*, 2009). Similarly, the cell-to-cell bottleneck for soil-borne wheat mosaic virus may be approximately six virions for the initial movement from the infected cell and five virions in subsequent movements (Miyashita & Kishino, 2010). Although these studies suggest that population bottlenecks will have major effects on plant virus evolution, to date there has been no analysis of the impact of viral population bottlenecks on intra- and inter-host genetic diversity. Ultradeep genome sequencing is an excellent tool to address this issue, particularly as very high coverage levels facilitates the detection of mutations present at very low frequencies.

To gain a fuller understanding of the extent of intra-host genetic diversity in plant RNA viruses and the processes that have generated this variation, we used ultradeep sequencing to analyse the extent of genetic variation, and particularly the effect of population bottlenecks, in zucchini yellow mosaic virus (ZYMV) infecting its natural host, *Cucurbita pepo* ssp. *texana* (a wild gourd). ZYMV infects wild and agronomically important members of the plant family Cucurbitaceae (squash, melon and cucumber), causing symptoms that include yellowing and stunting of the plant, as well as severe leaf and fruit deformities (Desbiez & Lecoq, 1997). This emerging RNA virus attained a worldwide distribution within two decades of its description (Lisa *et al.*, 1981), and the importance of ZYMV as a crop pathogen is underscored by the fact that it has been shown to reduce agricultural yields by up to 94 % (Blua & Perring, 1989). ZYMV has a single-stranded, positive-sense RNA genome of approximately 9.6 kb. A single ORF encodes a large polyprotein precursor that is processed into ten putative proteins by three virally encoded proteases (P1, HC-Pro and Nla) (Gal-On, 2007), with an additional ORF (PIPO) embedded in the P3-coding region (Chung *et al.*, 2008).

Transmission of ZYMV occurs primarily via aphids in a non-persistent manner (Pfosser & Baumann, 2002; Urcuqui-Inchima *et al.*, 2001), and 26 aphid species have been shown to be capable of transmitting ZYMV (Katis *et al.*, 2006). An interaction between two conserved regions of the HC-Pro – KITC/KLSC (which interacts with the aphid stylet) and PTK (which interacts with the conserved DAG region in the coat protein (CP) – results in viral transmission (Urcuqui-Inchima *et al.*, 2001). This has been termed the 'helper strategy', as HC-Pro acts as a bridge between the CP and the aphid stylet, and differs from the 'capsid strategy' whereby the CP interacts directly with the aphid mouthparts (Pirone & Blanc, 1996). In addition, vertical transmission via seed has been shown to occur in *C. pepo* at low rates (1.6 %; Simmons *et al.*, 2011a).

To determine the extent and structure of genetic diversity in intra-host populations of ZYMV, and particularly how this diversity is likely to be shaped by population bottlenecks, we undertook ultradeep sequencing of ZYMV populations infecting *C. pepo* ssp. *texana* under two modes of horizontal transmission: aphid vectored and mechanically inoculated (i.e. without aphids). From the aphid-vectored experiment, we produced epidemiological-scale data, from which we could determine the extent of the bottleneck imposed by the aphid during inter-host transmission, as well as intra-host genetic variation over the course of infection. As a new leaf sample was collected at each time point, we were able to determine not only the mutational spectrum maintained within individual plants over time but also how intra-host viral genetic diversity is affected by bottlenecks during systemic movement. ZYMV was also inoculated mechanically across eight generations in a serial passaging experiment carried out in a greenhouse. Comparison of these data with those from the field study allowed us to analyse the pattern of viral evolution with and without the aphid-imposed bottleneck.

RESULTS

Genome coverage

Twenty-five samples were successfully sequenced: the inoculant, 16 aphid-vectored and eight mechanically

inoculated samples. The field plants were named as follows: the first letter and number, for example F8, designates the plant coordinates within the field grid, whilst the number in parentheses denotes the order in which samples were collected from an individual plant. The proportion of the genome that was sequenced ranged from 98.1 to 99.0 % (mean 98.8 %). After filtering, coverage ranged from 2132 to 14 544 reads per individual sample with a mean coverage of 10 051-fold. Given the high levels of coverage attained, we used these data as a baseline to run simulations in which we resampled the Illumina reads (excluding the inoculant) using a 1 % cut-off (to control for methodological errors) to determine the coverage level at which all variants in the population were revealed. This analysis suggested that at very low levels of coverage (75-fold or less) variants tended to be oversampled, leading to an overestimate of the number of mutations in the population. In contrast, coverage levels from 100- to 1000-fold led to an overall underestimation of the mutational spectrum. Saturation, defined as the ability to sample all variants in that population, was reached at approximately 2500-fold coverage. As we averaged 10 051-fold coverage, it is likely that we successfully uncovered the majority of the variants in our populations.

To determine further the power of our Illumina coverage to detect low-frequency variants, we performed a bootstrap resampling analysis using the minor variants found in the CP gene. This region was chosen as we had previously cloned and Sanger sequenced the CP of these samples (Simmons *et al.*, 2011b). Six CP mutations were uncovered in the current study (excluding the inoculant), none of which was detected in the previous study. Four of these mutations were sampled only once and ranged from 3 to 4.8 % in frequency: nt 8547 (3 %), 8631 (4.8 %), 8971 and 9009 (3.7 %). The other two were found in more than one sample with frequencies averaging 11.7 % (nt 8715) and 3.7 % (nt 8971). Accordingly, we found the level of coverage needed to detect a least one read for each variant frequency to be: 3 %, approx. 150-fold; 3.7 %, approx. 125-fold; 4.8 %, approx. 100-fold and 11.7 %, approx. 50-fold (Fig. 1). Hence, attaining sufficient coverage is extremely important for detecting low-frequency variants and for obtaining an accurate characterization of genetic diversity in viral populations.

Frequency and pattern of nucleotide variants

A total of 136 variants (i.e. polymorphic mutations at a frequency >1 %), and ranging in frequency from 1.1 to 49.8 %, were found across the dataset as a whole: 105 were found in a single sample (although not all within the same plant) and 31 were found in at least two samples. Of these 31 mutations, 30 were found in more than one plant, suggesting that they were spread between hosts, whilst the remaining mutation was found at different time points within the same plant. Within these two groups of variants, 49/105 and 11/31 were non-synonymous mutations. In addition, 74/105 and 11/31 were unique to the field samples; 31/105 were unique to the greenhouse samples; and 20/31 were shared between both experimental conditions.

Strikingly, among the 136 variants detected, six were present in every time point within a plant, or in all eight of the greenhouse samples. Hence, these mutations were maintained during the course of infection and through any

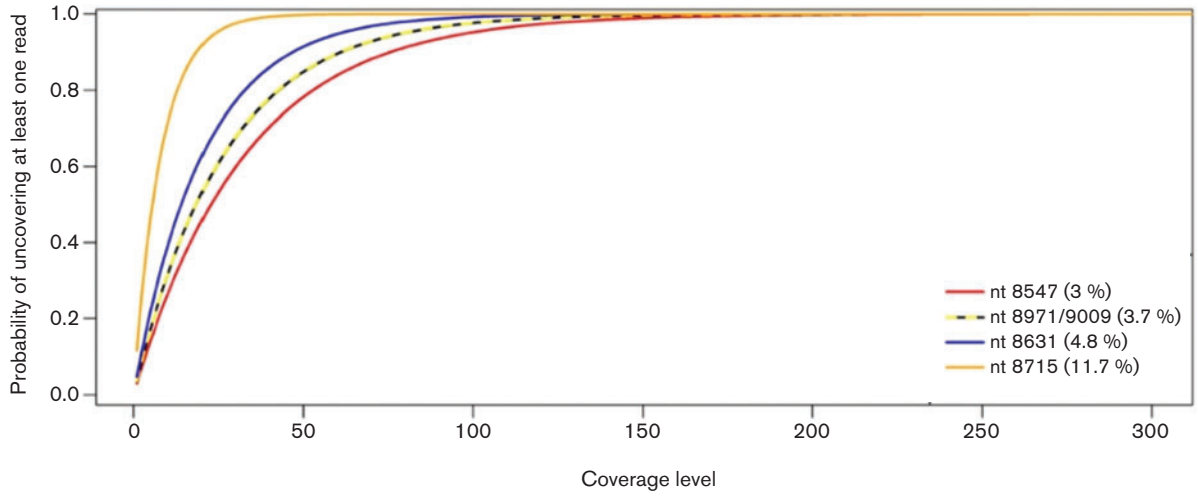

Fig. 1. Effect of coverage in the probability (estimated assuming a binomial distribution) of detecting the ZYMV CP variants uncovered in this study. Each colour represents a different mutation, labelled with their nucleotide position in the genome and variant frequency in parentheses. Six CP mutations were uncovered in the current study. Four of these mutations were sampled only once and ranged from 3 to 4.8 % in variant frequency: nt 8547 (3 %), 8631 (4.8 %), 8971 and 9009 (3.7 %). The other two were found in more than one sample with variant frequencies averaging 11.7 % (nt 8715) and 3.7 % (nt 8971).

intra- and inter-host bottlenecks that occurred. These included: two mutations in F8 (nt 2205 and 7688), four mutations in F7 (nt 7317, 7688, 7821 and 8971), three mutations in G7 (nt 6294, 7317 and 7688), two mutations in E8 (nt 2205 and 7688) and one mutation in G6 (nt 7688). In addition, six mutations were present in at least one time point in every single field plant (nt 1254, 2205, 4626, 7317, 7688 and 7821), such that they were maintained during inter-host transmission and again through any population bottlenecks. All but one of these mutations (nt 1254) were also found in at least one greenhouse sample. In the greenhouse samples, one mutation was shared across serial passages (nt 7688).

The mean number of mutations between our samples and the reference strain NC_003224.1 (a Taiwanese isolate) was 464 (5.78 %), which is comparable with previous studies using consensus sequences (Simmons *et al.*, 2008). We also compared the variants found here with the other 24 full-length ZYMV genomes published in GenBank. Of the 105 mutations observed in a single sample, 48 were present in the GenBank sequences, as were 17 of the 31 polymorphic variants (Table S1, available in JGV Online), including all six mutations that were present at least once in every plant. This suggests that these variants may exist as polymorphic sites in natural populations.

Changes in variant frequency

Of the 31 variants present in more than one sample, we found two cases (nt 2205 and 7317) in which the originally 'minor' variant (defined as initially <50 % frequency; in these cases, 30.6 and 7.9 %, respectively) approached fixation in later samples (both frequencies reached 98 %; Fig. 2a, b). These mutations were present at frequencies of 2.67 and 0.3 %, respectively, in the inoculant. In addition, these fixation events occurred rapidly, taking only 59 days in both cases. Interestingly, these same two nucleotide

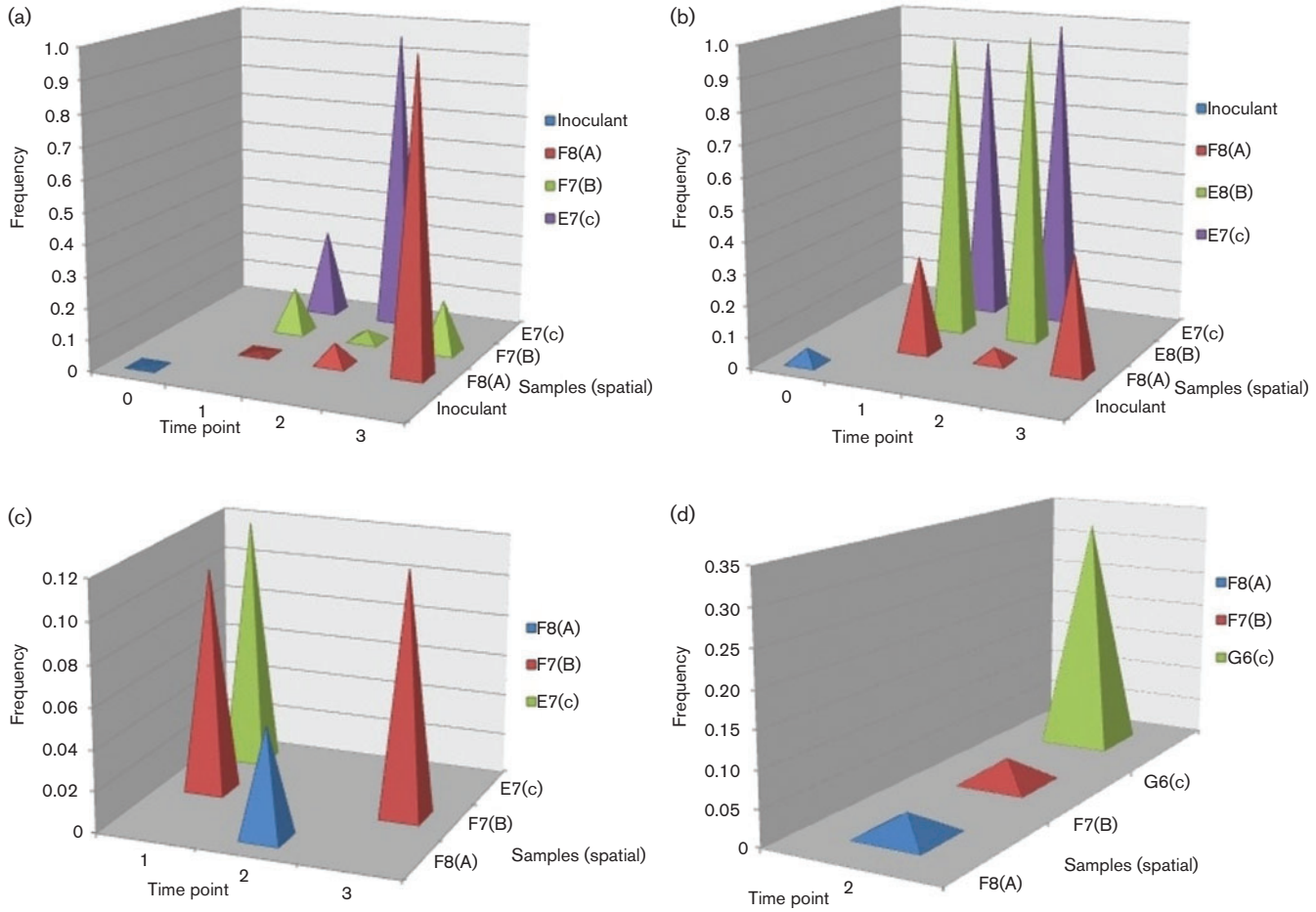

Fig. 2. Variation in variant frequency over time and space of ZYMV variants in the field experiment. The three-dimensional graphics show changes in variant frequency, with each colour representing a different plant. The *y*-axis shows how the variant frequency increases over the course of the infection. The *x*-axis shows variation over time within each plant or intra-host variation (moving from left to right). The *z*-axis shows variation over space, or between each plant (moving from the front to the back of the graph). These four graphs depict fluctuations in variant frequencies at nt 2205 (a), 7317 (b), 4626 (c) and 1254 (d).

positions are present as polymorphic sites in the 25 ZYMV genome sequences in GenBank (nt 2205 in 11/25 and nt 7317 in 7/25), suggesting that they may confer a selective advantage in some host genotypes (or host species) or under some environmental conditions. This idea was supported by the fact that these changes in variant frequencies appeared to be affected by environmental conditions. For instance, the minor variant at nt 7317 reached fixation in the field but after an initial decrease remained constant through transmission events in the greenhouse, where environmental conditions are relatively constant; in the first greenhouse sample, the variant frequency was 10 % but dropped to 1.7 % by the last host (Fig. 3). Two additional cases where the minor variant increased as the virus spread in the field from the first infected plant were also observed (nt 1254 and 4626), although they did not approach fixation (Fig. 2c, d). At nt 1254, the frequency in the first field plant was 3.1 % and subsequently increased to 33 %. Similarly, at nt 4626, the frequency increased from 5.4 to 12 %. Position 1254 lies within HC-Pro, which is believed to be involved in suppression of RNA silencing; indeed, this nucleotide (nt 185 of HC-Pro) lies within a region thought to inhibit the methyltransferase activity of Hua Enhancer 1 (Jamous *et al.*, 2011).

Spatial distribution of mutations along the genome

Interestingly, mutations identified in field-grown plants were clustered in the genome ($P<1\times10^{-4}$). In contrast, there was no significant spatial clustering of mutations in the greenhouse samples ($P\approx1$) (Fig. 4). Using a χ^2 goodness-of-fit test (R Development Core Team, 2011), we determined that the number of mutations per gene was greater than would be expected by chance in only two regions: Nlb in the field samples and HC-Pro in the greenhouse samples. We also found one region in the greenhouse samples (CI) in which the number of mutations was less than would be expected by chance alone, although these results were strongly dependent on the level of coverage attained. Despite the relatively high number of mutations observed, those genomic regions previously suggested to constitute conserved domains in ZYMV were also conserved in our analysis, indicating that mutations in these regions are strongly deleterious. For instance, all of the regions known to be necessary for aphid transmission – the PTK and KLSC regions in HC-Pro (Huet *et al.*, 1994) and the DAG region in CP (Atreya *et al.*, 1990) – were conserved in our samples.

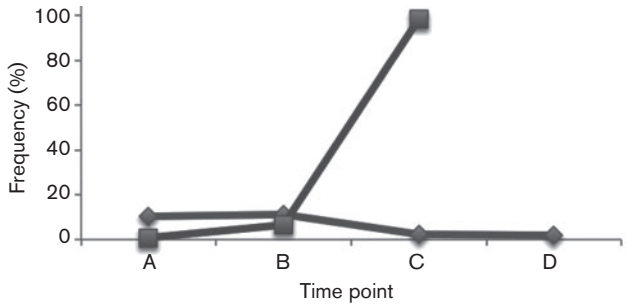

Fig. 3. Fluctuations in variant frequency (nt 7317) in aphid-transmitted versus mechanically transmitted plants. The variant frequency in the first infected field plant (■; F8) increased from >1 % to 98.7 % over the course of infection. In contrast, in the greenhouse plants (◆), the frequency in the first plant was 10 % and dropped to 1.7 % by the last host.

DISCUSSION

Although population bottlenecks are expected to be strong both within and between hosts, nearly one-quarter of the variants detected within our viral populations were found in more than one sample, either within the same or a different plant, and some at relatively high frequencies. As such, the population bottlenecks that shape the evolution of plant RNA viruses may not be as large as previously suggested, although this will clearly vary in a virus-specific manner. Of equal importance was the observation that three of the initially 'minor' variants rapidly went to fixation in the aphid-vectored plants, but remained at low frequency in the mechanically inoculated plants, suggesting that they are selectively advantageous under field conditions. The dramatic increase in frequency for some of these variants in the aphid-vectored plants (e.g. <1 % in the inoculant to 98 % by the end of the season at nt 7317) was observed in more than one plant. This argues strongly for natural selection and against genetic drift as the main mechanism generating the differences between the variant frequencies in the greenhouse and field populations, as the latter process is expected to result in fixation events over much longer timescales; indeed, the mean time for fixation of a neutral mutation in a haploid population is $N_e\times$ generation time, which will generally equate to timescales measured in years, whereas the change in frequency recorded here occurred over a time period of only 2 months.

Also of interest in this context was the observation that regions known to be involved in aphid transmission were conserved in all of the samples analysed. Hence, the natural selection we observed is unlikely to be directly linked to transmission events. However, the natural selection we observed in the form of variant frequency changes may be indirectly linked to transmission events through host–virus or host–vector rather than vector–virus interactions. Specifically, it is believed that compositional differences in saliva among aphid species may result in differential viral transmission (Pirone & Perry, 2002). There is also evidence that the virus may manipulate host factors to increase the plant's attractiveness to potential vectors by modulating colour changes associated with infection (Ajayi & Dewar, 1983) and olfactory cues in the form of volatile

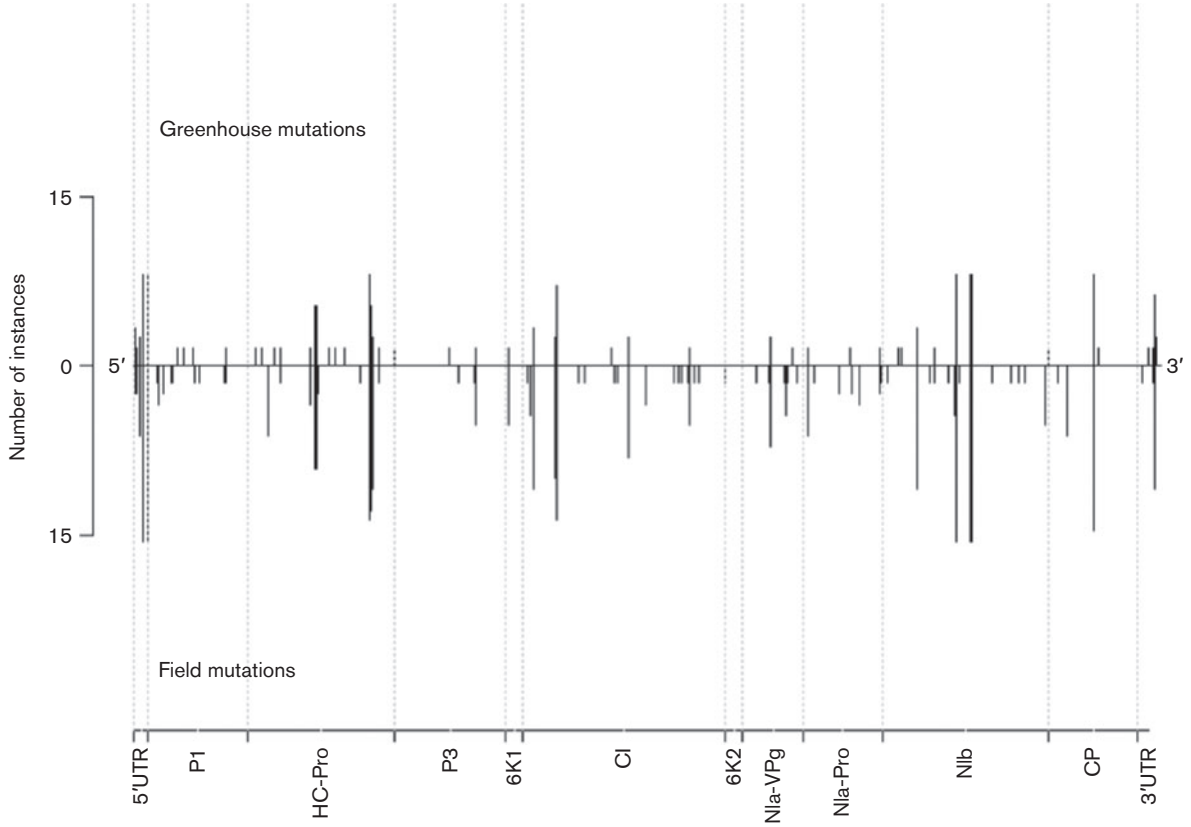

Fig. 4. Distribution of mutations across the ZYMV genome under field (below) and greenhouse (above) conditions. The length of the lines indicates the relative number of samples with that particular mutation, and the schematic of the viral genome indicates in which gene each mutation occurred.

compounds (Mauck *et al.*, 2010; Medina-Ortega *et al.*, 2009; Ngumbi *et al.*, 2007), as well as altering the mechanisms involved in virus acquisition. In addition, host factors may be involved in optimizing vector transmission. For example, in cauliflower mosaic virus, virus inclusion bodies have been shown to control aphid-mediated transmission (Espinoza *et al.*, 1991; Khelifa *et al.*, 2007). Although little is known about the specific mechanisms underlying these processes, it is possible that the differences in variant frequencies observed here may be due to the absence of the aphid vector in the greenhouse experiment. This possibility notwithstanding, the effect of other environmental differences between the field and the greenhouse experiments on variant frequencies should also be investigated. For example, the greenhouse environment is relatively stress free, as the plants are watered regularly, maintained within a narrow range of temperatures, have ample room and light, and are sprayed regularly with insecticide to prevent herbivory. In addition, the greenhouse plants were exposed only to the mechanically introduced virus population. This is in direct contrast to the field plants, which were subjected to the vagaries of nature and experienced a variety of biotic and abiotic stresses, such as drought, herbivory and multiple viral populations introduced by aphids migrating into the field, as well as competition.

As the transmission events undertaken in the greenhouse represent a release from the aphid-imposed population bottleneck, and the inoculum dose was large (half a leaf, which ensured inoculation at saturation), we might have expected the amount of genetic diversity being transmitted between greenhouse plants to be significantly greater than in the field. It was therefore surprising that our results indicated that greater genetic diversity was transmitted in the field experiment. Indeed, a mean of only 0.5–3.2 PVY virions were transmitted per aphid in *in vitro* experimental systems (Moury *et al.*, 2007), with similar numbers reported *in vivo* (Betancourt *et al.*, 2008). However, these estimations do not consider the huge number of aphids that may be involved in transmitting the virus, which could potentially overwhelm the population bottlenecks induced by single transmission events. Indeed, experiments using suction traps found that, although aphid population size tends to fluctuate both in terms of year and location, very high population numbers can be achieved (Katis *et al.*, 2006), with up to 40 000 aphids being counted in one location in 1 year (range 2179–41 851). Similarly, studies have revealed up to four alatae (winged) and 400 apterous

(non-winged) aphids per leaf per time point on *C. pepo* (Hooks *et al.*, 1998). As the incidence of ZYMV appears to be correlated with total aphid numbers (Basky *et al.*, 2001), the effect of aphid population size on the effective population size of viral populations in individual plants clearly needs to be examined in more detail.

It is also possible that helper-dependent transmission, such as occurs with ZYMV, may be less prone to severe bottlenecks than transmission where the virions interact directly with the aphid stylet via the CP (i.e. the capsid strategy). Specifically, HC-Pro and the virion do not have to be acquired simultaneously. As long as the HC-Pro is capable of interacting with the aphid stylet, it can assist in the transmission of virions acquired from other parts of the host or even from different hosts, thus ameliorating the effect of the population bottleneck. This is in direct contrast to viruses that interact directly with the vector (Pirone & Blanc, 1996). Thus, it is possible that multiple aphids transmitting the virus between hosts as well as the fact that ZYMV is vector transmitted via HC-Pro maintained high levels of genetic diversity in our field experiment.

The genetic resolution we have achieved in this study is clearly a reflection of our ultradeep sequencing strategy. A previous study using some of the same samples, for which cloning and Sanger sequencing of the CP region was undertaken, revealed that no mutations were transmitted between or within plants (Simmons *et al.*, 2011b), in marked contrast to the results obtained here. Our simulations revealed that, to reach saturation and detect all variants in the population (assuming a 1 % cut-off), a coverage level of approximately 2500-fold was needed in order to sample all of the variants present in our populations. Similarly, we determined that the probability of detecting a variant that comprises approximately 12 % of the population at least once requires approximately 50-fold coverage, and to detect a variant present at 3 % frequency at least once requires a minimum coverage of 150-fold. Given that in our previous study we averaged 35 clones per sample, it is not surprising that we were unable to uncover these mutations.

More than three-quarters of the mutations observed in this study were observed in a single sample only (105/136). Thus, although there was some transmission of variants both inter- and intra-host, the majority of the mutations generated were not transmitted either inter- or intra-host. Whether this is the result of population bottlenecks restricting viral genetic diversity, purifying selection acting on the viral population or some combination of both still needs to be determined. However, the majority of single-nucleotide substitutions in RNA viruses are likely to be deleterious (Sanjuán *et al.*, 2004). Hence, given that we previously detected the mean d_N/d_S ratios (ratio of the number of non-synonymous to synonymous substitutions per site) among these populations to be approximately 0.6 (the CP region only) (Simmons *et al.*, 2011b), it is probable that many of the mutations that occurred in only one sample were also deleterious and were subsequently purged from the population.

Our study also revealed that some viruses are clearly transmitted both within and among hosts, despite the presence of population bottlenecks. Although stochastic processes undoubtedly play a role in structuring viral populations, these processes alone may be insufficient to negate the action of natural selection. This point was dramatically highlighted by the fact that we uncovered minor variants that later approached fixation, strongly suggesting that they were selectively advantageous. Overall, these results attest to a complex pattern of changing genetic diversity in an emerging RNA virus, and will contribute to a more complete understanding of the dynamics of evolutionary change with implications for the management of emerging viral diseases.

METHODS

Field experiment. The field experiment was conducted using *C. pepo* ssp. *texana* at the Pennsylvania State University Agriculture Experiment Station. One 0.4 ha field with 180 plants (each approx. 6 m apart) was laid out as a grid labelled A–L and 1–15. In 2007, the plant situated in the middle of the field, F8, was inoculated mechanically with ZYMV that was isolated by us during the previous field season (the inoculant). When this plant exhibited viral symptoms, a leaf was collected. As neighbouring plants became infected, a leaf sample was collected weekly from each plant from the onset of visible symptoms until host death (approx. 9 weeks). The presence of ZYMV in the leaf samples was detected immunologically using double-antibody sandwich ELISA (Agdia) and confirmed by RT-PCR. All samples were stored at −80 °C. Although samples were collected from all of the infected plants, a subset of samples that were spatially related to F8 was selected so that a total of 16 samples representing six individual plants were used for sequencing. This subset comprised one plant that was sampled at four time points: F8 (24 July, 8, 13 and 28 August); two plants sampled at three time points: F7 (30 August, 13 and 20 September) and G7 (30 August, 6 and 20 September); and three plants sampled at two time points: E7 and E8 (13 and 20 September), and G6 (20 and 30 September).

Greenhouse experiment. A *C. pepo* plant was inoculated mechanically in a greenhouse in January 2008 with a ZYMV sample taken from the first diseased plant from the 2007 season [F8(1)]. Inoculum was prepared from infected plant tissue diluted in a phosphate buffer (0.1 M Na_2H/KH_2PO_4) at a 1 : 3 (v/v) ratio. Carborundum powder (500 g) was then rubbed on the surface of the leaf and the inoculum subsequently applied to the leaf surface with a pestle. When the plants displayed disease symptoms and exhibited at least an additional eight leaves of growth from the inoculation site (typically at 4–5 weeks), half of the fifth leaf was used to inoculate another plant. This process was repeated up to the eighth generation. The other half of the leaf was stored at −80 °C and subsequently used for sequencing.

RNA isolation and RT-PCR. RNA was isolated from frozen leaf samples using an RNeasy Plant Mini kit (Qiagen). First-strand cDNA was synthesized from the extracted RNA using five genome-specific primers, which were designed based on the reference strain (GenBank accession no. NC_003224.1), following the protocol provided by the supplier using a Superscript III First-strand Synthesis kit (Invitrogen). The target cDNA was then amplified directly via PCR using Phusion High-Fidelity PCR Master Mix (Finnzyme) following the manufacturer's protocols. The following PCR conditions were used: 98 °C for 1 min, followed by 20 cycles of 98 °C for 10 s, 58 °C for 20 s and 72 °C for 1 min 20 s, with a 5 min final extension at 72 °C. The five

primers were designed with 560, 19, 141 and 151 bp overlaps between amplicons across the genome. The primer sequences were: ZYMC_F1, 5′-AGAAATCAACGAACAAGCAGACGA-3′ (nt 27–50 of the reference strain), and ZYMC_R1, 5′-GCAACATCCATCAACGAAGGC-3′ (nt 2199–2219); ZYMC_F2, 5′-GGGGGAAAGAGGGTATCATT-3′ (nt 1689–1708), and ZYMC_R2, 5′-CCAAGGGGCGTGTAGGTT-3′ (nt 3956–3973); ZYMC_F3, 5′-TGAACCTACACGCCCCTTG-3′ (nt 3956–3974), and ZYMC_R3, 5′-TGCCCTTGCCCATAAAATA-3′ (nt 6070–6088); ZYMC_F4, 5′-GACGAAAGCACCCATACAGACATA-3′ (nt 5947–5970), and ZYMC_R4, 5′-TGACCGACCCACCAATCCT-3′ (nt 7808–7826); and ZYMV_F5-2, 5′-GGTGGTTGGGATAGATT-GATGAG-3′ (nt 7675–7697) and ZYMV_R5-2, 5′-TCCGACAGGAC-TACGGCATT-3′ (nt 9515–9534). Amplicons from these primers covered 99 % of the viral genome with lengths of 2192, 2314, 2134, 1879 and 1859 bp, respectively. The five PCR products per viral sample were pooled and gel extracted using a Zymoclean Gel Recovery kit (Zymo Research) and quantified using a Qubit fluorometer (Invitrogen).

Illumina library construction. Once quantified, samples (300 ng) were sheared using Next dsDNA Fragmentase (New England Biolabs) following the manufacturer's recommendations. The reaction was terminated by adding 5 µl cold 0.5 M EDTA and cleaned with a DNA Clean & Concentrator-5 kit (Zymo Research). The fragmented samples were used for library construction following the protocol of Mortazavi *et al.* (2008) starting with blunt-end repair. The following exceptions were made: each cleaning step was conducted using the DNA Clean & Concentrator kit, and blunt-end repair and ligation reactions were conducted using reagents from NEB. Samples were amplified and indexes were incorporated following standard indexing protocols with the following PCR cycles: 98 °C for 1 min, 18 cycles of 98 °C for 10 s, 65 °C for 30 s and 72 °C for 30 s, and a 5 min extension at 72 °C. Samples were PCR purified, quantified using a Qubit fluorometer (Invitrogen) and diluted to 10 nM for Illumina sequencing. Sequencing was performed at the University of Southern California on an Illumina GAIIx with multiplexing (12 samples per lane for the first two lanes and eight on the last lane) for a total of three lanes on the same flow cell. The inoculant was sequenced on a separate run at the University of Southern California on the same machine.

Consensus sequences. We pooled the reads from all samples (approx. 62 million reads) and used Velvet to assemble a *de novo* consensus sequence (Zerbino & Birney, 2008; www.ebi.ac.uk/~zerbino/velvet/). As Velvet is unable to build a consensus from such a large pool of reads, we randomly sampled 24 subsets (1.5 % each of the total) of reads. Apart from length, there were no discrepancies between the resulting 24 contigs, which we subsequently collapsed to form a consensus sequence of the samples. To confirm the consensus sequence, we used Velvet to create *de novo* consensus sequences for each sample individually, which did not differ from the consensus apart from polymorphic sites.

Read accuracy and identification of variant sites. We used a standard workflow for the identification of variant sites using Galaxy (Blankenberg *et al.*, 2010; Goecks *et al.*, 2010; Goto *et al.*, 2011; http://main.g2.bx.psu.edu/heteroplasmy). We altered the workflow by increasing the maximum edit distance to seven. The reads were mapped to the *de novo* consensus sequence using a Burrows–Wheeler alignment mapper (Li & Durbin, 2009) and subsequently transformed and filtered using Galaxy tools. Given that both strand bias and low-quality scores have been shown to increase the number of miscalls obtained whilst mapping Illumina reads (Minoche *et al.*, 2011; Nakamura *et al.*, 2011), we controlled for both of these. Strand bias was accounted for such that any variation found at a site was validated in both strands in order to be considered a true variant. To control for mapping quality, we excluded any sites that had a Phred-scaled quality score of <30 compared with the Illumina supplied control (ΦX 174); according to Illumina, with this quality score the inferred base call accuracy is 99.9 %. Finally, to account for methodological errors introduced as a result of the experimental procedures, we conservatively excluded: (i) any mutations that were present at a frequency of <1 %, and (ii) any sites where the coverage was <500-fold.

To validate the polymorphisms uncovered by the Galaxy pipeline, we used an alternative method, the Genome Analysis Toolkit (McKenna *et al.*, 2010; www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit), to analyse one sample [F8(1)] and were able to recover the same polymorphisms.

Mutation analysis. The consensus sequence for each sample was aligned manually to the ZYMV reference strain (GenBank accession no. NC_003224.1) using Se-Al, version 2.0a11 (kindly provided by Andrew Rambaut, University of Edinburgh, UK). Counts of the number of mutations in each sample were undertaken manually.

Given that we determined the frequency of 'minor' variants (i.e. those <50 % in the population), we then used a binomial distribution (R Development Core Team, 2011) to determine the probability of uncovering that variant at increasing coverage levels (number of reads). In addition, we resampled our Illumina data at progressively lower levels of coverage to determine how lower coverage levels could bias the discovery of true minor variants. We ran a simulation (R Development Core Team, 2011) in which we resampled our Illumina data at each base position in the genome. As we had excluded any variants that occurred at <1 % frequency, we calculated the minimum threshold as the 99th percentile of a binomial distribution. Not only did this analysis indicate the coverage level at which all variants would be uncovered, it also revealed how, at low levels of coverage, the discovery of true minor variants tends to be biased.

Spatial distribution of mutations. We used a bootstrap method to infer whether mutations were clustered spatially across the genome compared with a null model, which assumed random mutation placement. Bootstrap distributions and null distributions were calculated for the index of dispersion statistic and then compared using a Mann–Whitney U test (R Development Core Team, 2011).

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Doctorial Dissertation Grant Program grant no. 1010881, and the Biotechnology Risk Assessment Program grant no. 2009-33120-20093 from the USDA National Institute of Food and Agriculture. We thank Tony Omeis for assistance and use of the Biology Greenhouse, and R. Oberheim and his staff for use of the Horticulture Farm at the Penn State Agriculture Experiment Station at Rock Springs, PA, USA.

REFERENCES

Acosta-Leal, R., Bryan, B. K. & Rush, C. M. (2010). Host effect on the genetic diversification of *Beet necrotic yellow vein virus* single-plant populations. *Phytopathology* **100**, 1204–1212.

Acosta-Leal, R., Duffy, S., Xiong, Z., Hammond, R. W. & Elena, S. F. (2011). Advances in plant virus evolution: translating evolutionary insights into better disease management. *Phytopathology* **101**, 1136–1148.

Ajayi, O. & Dewar, A. M. (1983). The effect of barley yellow dwarf virus on field populations of the cereal aphids, *Sitobion avenae* and *Metopolophium dirhodum*. *Ann Appl Biol* **103**, 1–11.

Ali, A., Li, H., Schneider, W. L., Sherman, D. J., Gray, S., Smith, D. & Roossinck, M. J. (2006). Analysis of genetic bottlenecks during horizontal transmission of *Cucumber mosaic virus*. *J Virol* **80**, 8345–8350.

Atreya, C. D., Raccah, B. & Pirone, T. P. (1990). A point mutation in the coat protein abolishes aphid transmissibility of a potyvirus. *Virology* **178**, 161–165.

Basky, Z., Perring, T. & Tobias, I. (2001). Spread of zucchini yellow mosaic potyvirus in squash in Hungary. *J Appl Entomol* **125**, 271–275.

Betancourt, M., Fereres, A., Fraile, A. & García-Arenal, F. (2008). Estimation of the effective number of founders that initiate an infection after aphid transmission of a multipartite plant virus. *J Virol* **82**, 12416–12421.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. & Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Unit 19.10**, 1–21.

Blua, M. & Perring, T. (1989). Effect of zucchini yellow mosaic virus on development and yield of cantaloupe (*Cucumis melo*). *Plant Dis* **73**, 317–320.

Chung, B. Y., Miller, W. A., Atkins, J. F. & Firth, A. E. (2008). An overlapping essential gene in the *Potyviridae*. *Proc Natl Acad Sci U S A* **105**, 5897–5902.

Desbiez, C. & Lecoq, H. (1997). Zucchini yellow mosaic virus. *Plant Pathol* **46**, 809–829.

Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267–276.

Espinoza, A. M., Medina, V., Hull, R. & Markham, P. G. (1991). Cauliflower mosaic virus gene II product forms distinct inclusion bodies in infected plant cells. *Virology* **185**, 337–344.

Feuer, R., Boone, J. D., Netski, D., Morzunov, S. P. & St Jeor, S. C. (1999). Temporal and spatial analysis of Sin Nombre virus quasispecies in naturally infected rodents. *J Virol* **73**, 9544–9554.

French, R. & Stenger, D. C. (2003). Evolution of wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. *Annu Rev Phytopathol* **41**, 199–214.

Gal-On, A. (2007). *Zucchini yellow mosaic virus*: insect transmission and pathogenicity – the tails of two proteins. *Mol Plant Pathol* **8**, 139–150.

García-Arenal, F., Fraile, A. & Malpica, J. M. (2001). Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol* **39**, 157–186.

García-Arenal, F., Fraile, A. & Malpica, J. M. (2003). Variation and evolution of plant virus populations. *Int Microbiol* **6**, 225–232.

Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team, T. & Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86.

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

Goto, H., Dickins, B., Afgan, E., Paul, I. M., Taylor, J., Makova, K. D. & Nekrutenko, A. (2011). Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* **12**, R59.

Hooks, C. R. R., Valenzuela, H. R. & Defrank, J. (1998). Incidence of pests and arthropod natural enemies in zucchini grown with living mulches. *Agric Ecosyst Environ* **69**, 217–231.

Huet, H., Gal-On, A., Meir, E., Lecoq, H. & Raccah, B. (1994). Mutations in the helper component protease gene of zucchini yellow mosaic virus affect its ability to mediate aphid transmissibility. *J Gen Virol* **75**, 1407–1414.

Hughes, A. L. (2009). Small effective population sizes and rare nonsynonymous variants in potyviruses. *Virology* **393**, 127–134.

Jamous, R. M., Boonrod, K., Fuellgrabe, M. W., Ali-Shtayeh, M. S., Krczal, G. & Wassenegger, M. (2011). The helper component-proteinase of the *Zucchini yellow mosaic virus* inhibits the Hua Enhancer 1 methyltransferase activity *in vitro*. *J Gen Virol* **92**, 2222–2226.

Jerzak, G. V. S., Brown, I., Shi, P.-Y., Kramer, L. D. & Ebel, G. D. (2008). Genetic diversity and purifying selection in West Nile virus populations are maintained during host switching. *Virology* **374**, 256–260.

Katis, N. I., Tsitsipis, J. A., Lykouressis, D. P., Papapanayotou, A., Margaritopoulos, J. T., Kokinis, G. M., Perdikis, D. C. & Manoussopoulos, I. N. (2006). Transmission of *Zucchini yellow mosaic virus* by colonizing and non-colonizing aphids in Greece and new aphid vectors of the virus. *J Phytopathol* **154**, 293–302.

Khelifa, M., Journou, S., Krishnan, K., Gargani, D., Espérandieu, P., Blanc, S. & Drucker, M. (2007). Electron-lucent inclusion bodies are structures specialized for aphid transmission of cauliflower mosaic virus. *J Gen Virol* **88**, 2872–2880.

Lech, W. J., Wang, G., Yang, Y. L., Chee, Y., Dorman, K., McCrae, D., Lazzeroni, L. C., Erickson, J. W., Sinsheimer, J. S. & Kaplan, A. H. (1996). In vivo sequence diversity of the protease of human immunodeficiency virus type 1: presence of protease inhibitor-resistant variants in untreated subjects. *J Virol* **70**, 2038–2043.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Lisa, V., Boccardo, G., D'Agostino, G., Dellavalle, G. & D'Aquilio, M. (1981). Characterization of a potyvirus that causes zucchini yellow mosaic. *Phytopathology* **71**, 667–672.

Mauck, K. E., De Moraes, C. M. & Mescher, M. C. (2010). Deceptive chemical signals induced by a plant virus attract insect vectors to inferior hosts. *Proc Natl Acad Sci U S A* **107**, 3600–3605.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. & other authors (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303.

Medina-Ortega, K. J., Bosque-Pérez, N. A., Ngumbi, E., Jiménez-Martínez, E. S. & Eigenbrode, S. D. (2009). *Rhopalosiphum padi* (Hemiptera: Aphididae) responses to volatile cues from barley yellow dwarf virus-infected wheat. *Environ Entomol* **38**, 836–845.

Minoche, A. E., Dohm, J. C. & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12**, R112.

Miyashita, S. & Kishino, H. (2010). Estimation of the size of genetic bottlenecks in cell-to-cell movement of *Soil-borne wheat mosaic virus* and the possible role of the bottlenecks in speeding up selection of variations in *trans*-acting genes or elements. *J Virol* **84**, 1828–1837.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628.

Moury, B., Fabre, F. & Senoussi, R. (2007). Estimation of the number of virus particles transmitted by an insect vector. *Proc Natl Acad Sci U S A* **104**, 17891–17896.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A. & other authors (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**, e90.

Ngumbi, E., Eigenbrode, S. D., Bosque-Pérez, N. A., Ding, H. & Rodriguez, A. (2007). *Myzus persicae* is arrested more by blends than by individual compounds elevated in headspace of PLRV-infected potato. *J Chem Ecol* **33**, 1733–1747.

Pfosser, M. F. & Baumann, H. (2002). Phylogeny and geographical differentiation of zucchini yellow mosaic virus isolates (*Potyviridae*) based on molecular analysis of the coat protein and part of the cytoplasmic inclusion protein genes. *Arch Virol* **147**, 1599–1609.

Pirone, T. P. & Blanc, S. (1996). Helper-dependent vector transmission of plant viruses. *Annu Rev Phytopathol* **34**, 227–247.

Pirone, T. P. & Perry, K. L. (2002). Aphids: non-persistent transmission. *Adv Bot Res* **36**, 1–19.

R Development Core Team (2011). R: a language and environment for statistical computing, http://www.R-project.org. R Foundation for Statistical Computing, Vienna, Austria.

Roossinck, M. J. (1997). Mechanisms of plant virus evolution. *Annu Rev Phytopathol* **35**, 191–209.

Sacristán, S., Malpica, J. M., Fraile, A. & García-Arenal, F. (2003). Estimation of population bottlenecks during systemic movement of tobacco mosaic virus in tobacco plants. *J Virol* **77**, 9906–9911.

Sanjuán, R., Moya, A. & Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A* **101**, 8396–8401.

Simmons, H. E., Holmes, E. C. & Stephenson, A. G. (2008). Rapid evolutionary dynamics of zucchini yellow mosaic virus. *J Gen Virol* **89**, 1081–1085.

Simmons, H. E., Holmes, E. C., Gildow, F. E., Bothe-Goralczyk, M. A. & Stephenson, A. G. (2011a). Experimental verification of seed transmission in *Zucchini yellow mosaic virus*. *Plant Dis* **95**, 751–754.

Simmons, H. E., Holmes, E. C. & Stephenson, A. G. (2011b). Rapid turnover of intra-host genetic diversity in Zucchini yellow mosaic virus. *Virus Res* **155**, 389–396.

Urcuqui-Inchima, S., Haenni, A. L. & Bernardi, F. (2001). Potyvirus proteins: a wealth of functions. *Virus Res* **74**, 157–175.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829.