Virology 484 (2015) 346–353

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Mapping genetic determinants of viral traits with F_{ST} and quantitative trait locus (QTL) approaches

Juliette Doumayrou [a,b], Gaël Thébaud [c], Florence Vuillaume [a], Michel Peterschmitt [a], Cica Urbino [a,*]

[a] CIRAD, UMR BGPI, F-34398 Montpellier, France
[b] Department of Plant Pathology, 351 Bessey Hall, Iowa State University, Ames, IA 50011, USA
[c] INRA, UMR 385 BGPI, F-34398 Montpellier, France

ARTICLE INFO

Article history:
Received 13 March 2015
Returned to author for revisions
16 April 2015
Accepted 16 June 2015
Available online 15 July 2015

Keywords:
QTL
F_{ST}
Viral accumulation
Geminivirus
Artificial recombinant genome

ABSTRACT

The genetic determinism of viral traits can generally be dissected using either forward or reverse genetics because the clonal reproduction of viruses does not require the use of approaches based on laboratory crosses. Nevertheless, we hypothesized that recombinant viruses could be analyzed as sexually reproducing organisms, using either a quantitative trait loci (QTL) approach or a locus-by-locus fixation index (F_{ST}). Locus-by-locus F_{ST} analysis, and four different regressions and interval mapping algorithms of QTL analysis were applied to a phenotypic and genotypic dataset previously obtained from 47 artificial recombinant genomes generated between two begomovirus species. Both approaches assigned the determinant of within-host accumulation—previously identified using standard virology approaches—to a region including the 5' end of the replication-associated protein (*Rep*) gene and the upstream intergenic region. This study provides a proof of principle that QTL and population genetics tools can be extended to characterize the genetic determinants of viral traits.

Introduction

Assigning functions to viral genes is achieved either with biochemical approaches or with the following classic genetic approaches: forward genetics in which genomes of viral strains and mutants exhibiting contrasting phenotypes are compared, and reverse genetics in which phenotypes induced by mutagenesis or homologous recombination of a particular gene are compared. Although these approaches have provided valuable information for many viruses (Dawson, 1978; Saenz et al., 2001), they may be limited by the absence of prior knowledge of suspected regions or for virus traits determined by several loci or genes with possible epistasis. An illustration of such largely unknown multi-loci genetic determinism in viruses was recently provided with a genome-wide association approach and a variety of permutation tests which revealed that intra-genome interaction networks are broadly preserved by selection (Martin et al., 2011).

The genetic determinism of complex traits has been investigated for eukaryotes using various molecular approaches. Complex genetic determinism of a trait is generally identified with molecular markers associated at different degrees with the variation of the trait; the loci involved in such a determinism were named quantitative trait loci (QTL) (Beavis et al., 1991; Lynch and Walsh, 1998; Paterson et al., 1991; Stuber et al., 1987). Typically, QTL analysis is based on the measure of a phenotypic trait of the parents and their recombinant progeny and provides a direct correlation between a genomic region and a phenotypic trait. Alternatively, the measure of the fixation index (F_{ST}), commonly used in population genetics to assess how populations differ genetically within and between populations (Wright, 1949; Holsinger and Weir, 2009), may also be used for this purpose through computing one F_{ST} value per locus. It is based on the variance of allele frequency between populations, and is usually estimated for population with diploid genomes. However, as described in *Material and Methods*, it is possible to implement F_{ST} analysis to virus genomes which are mostly haploid.

Unfortunately, as QTL approaches need recombinant progenies derived from sexual crossings, they have so far not been used for viruses because of their clonal multiplication. However, as viruses of some genera or families are highly recombinogenic, it can be foreseen that such approaches can be extended to determine the complex

* Corresponding author.
E-mail addresses: Juliette.doumayrou@gmail.com (J. Doumayrou), gael.thebaud@supagro.inra.fr (G. Thébaud), florence.vuillaume@gmail.com (F. Vuillaume), michel.peterschmitt@cirad.fr (M. Peterschmitt), cica.urbino@cirad.fr (C. Urbino).

http://dx.doi.org/10.1016/j.virol.2015.06.019
0042-6822/© 2015 Elsevier Inc. All rights reserved.

genetic determinisms of some traits. For these viruses, recombinant genomes could be straightforwardly generated *in vivo* by co-infecting two parental clones to a host, but natural recombination breakpoints are non-random due to intrinsic molecular properties of the recombination mechanism and to within-plant selection. Alternatively, the required library of full-length hybrid genomes can be generated *in vitro* (Martin and Rybicki, 2002; Stratford and Covey, 1989) or by shuffling (Vuillaume et al., 2011). Due to the small size of most viral genomes, single nucleotide polymorphisms (SNPs) can be used as the most accurate molecular marker for genotyping the parental and the recombinant viral genomes at all the nucleotide positions that discriminate the parental genomes. The different recombinant genomes can be screened for infectivity and various traits following their inoculation to a host. The data collected from the progeny should fulfill the following conditions required by the QTL and F_{ST} approaches: (i) the parental genomes are differentiated using discriminating loci as the single-nucleotide polymorphisms (SNP) and the SNP of each recombinant can be assigned to one of the two parents and (ii) the genetic distance between the SNPs is known. Especially for QTL analysis, the two parents can be distinguished by a phenotypic trait which segregates in the recombinant progeny while, for F_{ST} analysis, the phenotypic trait has a multinomial distribution to compare the pseudo-populations.

In this study, we demonstrate for the first time that molecular methods developed and used in eukaryotes (*i.e.* locus-by-locus F_{ST} and QTL mapping) for analyzing genetic determinism can be used in virology. We took advantage of the first virus library of full-length recombinant genomes ever generated *ex vivo* with DNA shuffling (Vuillaume et al., 2011). Although this study has shown that random recombination rarely has lethal and/or large deleterious effects according to infectivity and within-host viral DNA accumulation, the genetic determinism of these traits had not been analyzed. The two shuffled parental viruses are from the genus *Begomovirus* (family *Geminiviridae*). They have single-stranded DNA genomes of 2.8 kb with six genes including *Rep*/C1encoding for the Rep protein, the only viral protein necessary for replicating the viral genome (Gutierrez, 2000; Hanley-Bowdoin et al., 1999; Jeske et al., 2001; Elmer et al., 1988). The C4 gene, embedded in C1, encodes for a multifunctional protein involved in virus movement (Jupin et al., 1994; Bisaro, 2006) and gene silencing (Vanitharani et al., 2004), functions which could impact directly or indirectly on intra-host viral accumulation. The intergenic region (IR) contains the origin of replication and typical sequence motifs involved in replication. The Vuillaume et al. (2011) data sets were found to fulfill the conditions required by the QTL and/or F_{ST} approaches.

We show here with F_{ST} index and QTL mapping, that early infectivity and viral accumulation of begomovirus recombinant genomes within tomato plants is determined by the loci located at the 5' end of *Rep*/C1 gene and in the 5' end of the IR. The positions of these loci were consistent with the genomic regions previously reported to be important for the initiation of infection and for viral replication and accumulation, which validates the use of the F_{ST} index and QTL analysis to characterize the genetic determinism of viral traits.

Results

In a previous study, 47 randomly chosen recombinants were generated between *Tomato yellow leaf curl virus* (TYLCV-Mld) and *Tomato leaf curl Mayotte virus* (ToLCKMV-[Dem]) designated Tyx and Tox respectively. The parental genomes could be differentiated by 534 loci (SNPs) disseminated along the genome and all the recombinants were assignable to one of the two parental genomes at each locus (see Additional file 1). The within-host accumulation of Tyx in tomato at 22 days post inoculation (dpi) was significantly higher than that of Tox genome (Fig. 3 in (Vuillaume et al., 2011)) and the distribution of the effect of recombination on virus accumulation was bimodal, with each mode centered on one parental phenotype (Fig. 4 in (Vuillaume et al., 2011)). Likewise, the percentage of PCR-positive plants at 15 dpi, a proxy of the speed at which viral DNA accumulates at the onset of the infection, was significantly higher for Tyx than for Tox (Fig. 2A in (Vuillaume et al., 2011)). Therefore, the data sets fulfilled the conditions required for the QTL and/or F_{ST} approaches.

Identification of viral genomic determinants involved in within-host accumulation by an F_{ST} approach

In comparison with the parental genomes, three accumulation groups were distinguished within the recombinant genomes: 10 recombinants accumulated significantly less than Tyx and not significantly differently from Tox (named Tox-type), 11 accumulated significantly more than Tox and not significantly differently

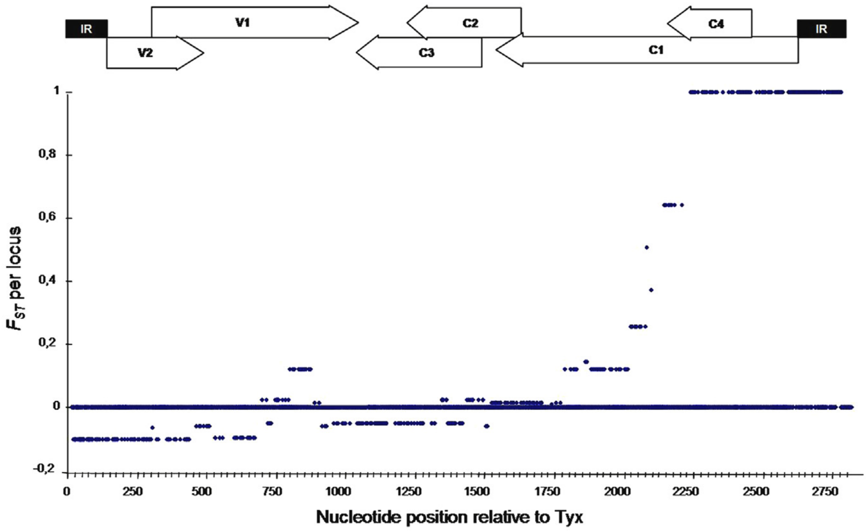

Fig. 1. Locus-by-locus F_{ST} between two groups of *begomovirus* recombinant genomes. The two groups differ in their within-host viral DNA accumulation, which is either similar to the Tyx or Tox parent. Each point corresponds to the F_{ST} per locus. A viral genome linearized at the virion strand origin of replication is presented with the open reading frames (horizontal arrows) and the intergenic region (IR) at the top of the graph.

from Tyx (named Tyx-type) and 26 were not significantly different from any parent (intermediate between Tyx and Tox) (Fig. 3 in (Vuillaume et al., 2011)). The locus-by-locus F_{ST} approach was used to detect the potential genomic regions that determine the difference in accumulation of the genomes belonging to the Tyx-type or Tox-type groups. The F_{ST} calculated at each of the 534 genomic Tyx/Tox discriminating loci revealed that 166 had a value of 1 (*i.e.* 100% difference between the two groups, locus-by-locus AMOVA, $P<0.0001$). They were all located between positions 2223 and 2761 (positions determined from the alignment of parental and recombinant genomes; the origin of replication was taken as position 1 (Fig. 1). The 11 recombinant genomes belonging to the Tyx-type accumulation group (SH-9, -9A, -24, -29, -31, -36, -39, -46, -96, -98 and -109) exhibited the Tyx allele at these 166 loci (Additional file 2). On the contrary, the 10 recombinant genomes belonging to the Tox-type accumulation group (SH-7, -15, -26, -40, -48, -72, -74, -82, -93 and -103) exhibited the Tox allele at these 166 loci (Additional file 2). The 166 loci were located within the 384 5′-end nucleotides of the C1 gene (88 loci) and the 149 nucleotides of the intergenic region located upstream C1 (78 loci). It is noteworthy that the C4 gene, embedded in the C1 gene, includes 41 of the 166 discriminating loci.

In order to validate the identified region on an independent set of genomes, the 26 recombinant genomes belonging to the intermediate accumulation group were separated in two genetic sub-groups according to the Tyx or Tox origin of their alleles at the 166 loci identified above. Six recombinant genomes (SH-1, -2, -3, -55, -75, and -97) exhibited 100% Tyx alleles at the 166 loci, whereas 18 recombinant genomes (SH-5, -5A, -10, -13, -32, -34, -41, -56, -59, -63, -71, -88, -90, -95, -100, -101, -104 and -108) exhibited 100% Tox alleles (Additional file 2); as recombinant genomes SH-64 and SH-54 exhibited mixed allele types, they were subsequently excluded from F_{ST} analysis. A hierarchical ANOVA was applied to the accumulation data of the two genetic sub-groups to test the biological effect of the 166-loci region on viral accumulation. The biological effect of this 166-loci region was confirmed because the viral accumulation of the recombinants with 100% Tyx alleles in this region was significantly higher than that of the recombinants with 100% Tox alleles (Tyx-type *vs.* Tox-type group; hierarchical ANOVA; mean ± se -0.67 ± 0.048 *vs.* -0.90 ± 0.034, respectively; $df=1$, $F=15.74$, $P<0.0001$; Fig. 2). Analysis of the mean accumulation of the two recombinant genomes which exhibited recombination breakpoints in this 166-loci region showed that SH-64, which has a Tyx sequence only between positions 2223 and 2237, would be classified between SH-104 and SH-95 of the Tox group (mean ± se -1.05 ± 0.42), while SH-54 which has a Tyx sequence only between positions 2702 and 2761 would be classified between SH-1 and SH-97 of the Tyx group (-0.51 ± 0.22). This result suggests that the determinant region could be shorter than the one including the 166 loci and, more specifically, that the major determinant of virus accumulation might be located between positions 2702 and 2761.

As some of the nucleotide differences at the 166 discriminating loci corresponded to non-synonymous mutations in the C1 and C4 gene, an F_{ST} was estimated for the amino acid sequences of C1 and C4 protein of the two groups of accumulation (Fig. 3); the intermediate recombinant genomes were included in this analysis except SH-64 and SH-54 which exhibited recombination breaks in this region. Consistently with the F_{ST} calculated with the polymorphic nucleotide loci, an F_{ST} of 1 was only detected at the amino acid loci included in the region of the 166 polymorphic nucleotide loci.

Identification of viral genomic determinants involved in within-host accumulation and early infectivity by QTL approaches

The data set previously generated with the 47 recombinant genomes (Vuillaume et al., 2011) was submitted to QTL analysis to detect the loci involved in within-host accumulation and early infectivity. The quantitative traits were the relative viral DNA accumulation at 22 dpi previously determined by qPCR and the percentage of PCR-positive plants at 15 dpi. For virus accumulation, the analyses were performed in parallel with the medians and the averages of the logarithm of the calibrated Normalized Relative Quantity (logNRQ) calculated for each recombinant genome (Vuillaume et al., 2011). The 534 Tyx/Tox-discriminating loci distributed throughout the genomes were targeted. A LOD score (logarithm of odds score) was calculated at each locus and plotted all along the viral genome (Fig. 4 and Additional file 3); the LOD score is related to the probability that a QTL is present at the tested locus. Four different analytical approaches of QTL analysis were used: simple interval mapping (SIM), single-trait MLE (CIM-MLE), and single-trait multiple IM (MIM) consider the genetic distance between the markers, and single marker regression (SMR) which does not. According to the threshold significance level performed by permutation test for SMR, SIM and MLE methods (Additional file 4), only the markers located between positions 2066 and 2761 can explain the difference of accumulation between the low and high virus accumulation groups of recombinant genomes (Fig. 4, Table 1, Additional file 3 and 5). Some loci explained 19–25% of the difference of accumulation while others explained more than 60% of it. The MIM method detected only 8 loci explaining the difference. Three of these (loci 2721, 2723 and 2725) explained 77.3% of the viral accumulation (see R^2 in Table 1 and Additional file 5), were also detected with both the averages and the medians of log NRQ and were included in the region containing the loci for which an F_{ST} of 1 was detected between low and high viral accumulation groups of recombinant genomes (positions 2223–2761).

The distribution of the effect of recombination on early infectivity did not show a bimodal distribution (Additional file 6A), in spite of a significant positive correlation between early infectivity and viral accumulation at 22 dpi ($n=47$, Pearson correlation coefficient $r=0.58$, $P<0.0001$; Additional file 6B). Therefore only QTL methods could be applied to this data set. According to the threshold significance level obtained by permutation tests ($\text{LOD}=2.953$ at $P=0.01$, based on 5000 permutations), only the markers located between positions 2223 and 2761 were detected to be involved in early infectivity with the SMR QTL method (Additional file 6C). The same result was obtained when the SIM method was used (data not shown). The different positions explained between 35.8% and 41.6% of the variance of early infectivity (Additional file 6D). Thus, the same loci explain the early infectivity and the DNA accumulation at 22 dpi.

Discussion

F_{ST} and QTL approaches identify the same viral region involved in within-host accumulation

F_{ST} and QTL analysis, a qualitative and quantitative approach respectively, were used to identify genetic determinants implicated in within-host viral DNA accumulation of tomato begomoviruses. The viral accumulation data set obtained by Vuillaume et al. (2011) for two parental and 47 recombinant genomes fulfilled all the conditions for the implementation of these methods. Both approaches revealed that DNA markers located between nucleotides 2223 and 2761 of the viral genomes were involved in within-host accumulation (Figs. 1 and 4, Table 1, Additional files 3 and 5). Early infectivity was showed to be positively correlated with viral accumulation; using the SMR QTL method, the same genomic region was detected to be a genetic determinant of this trait (Additional file 6C). This region encompasses the 5' end of the

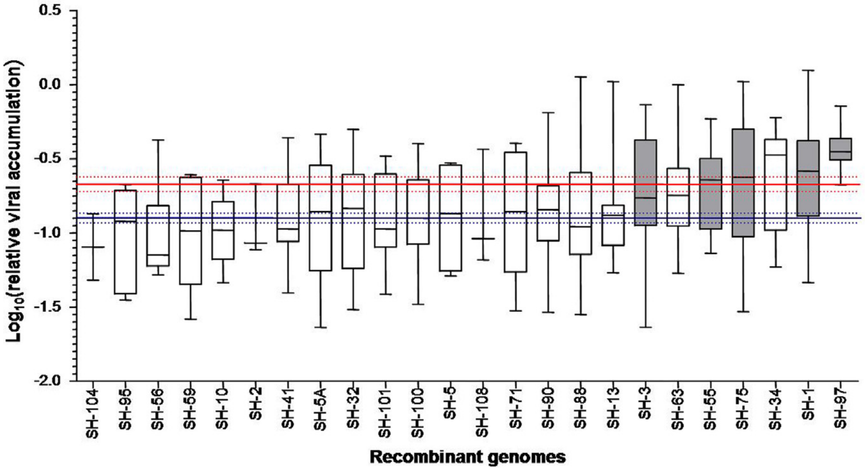

Fig. 2. Relative within-host viral accumulation of 26 recombinant genomes belonging to the intermediate accumulation group. Accumulation was assessed in tomato plants 22 days post-inoculation. Each box represents the quartile range (25–75%) and the median accumulation for each isolate. Recombinant genomes were distinguished according to the Tyx origin (gray boxes) or Tox origin (white boxes) of the fragment located between positions 2222–2761 of the parental genomes alignment. Red and blue lines correspond to the mean viral accumulation for Tyx and Tox groups, respectively. Dashed lines correspond to the standard deviation of each group. The recombinant genomes are ordered from left to right by increasing mean accumulation.

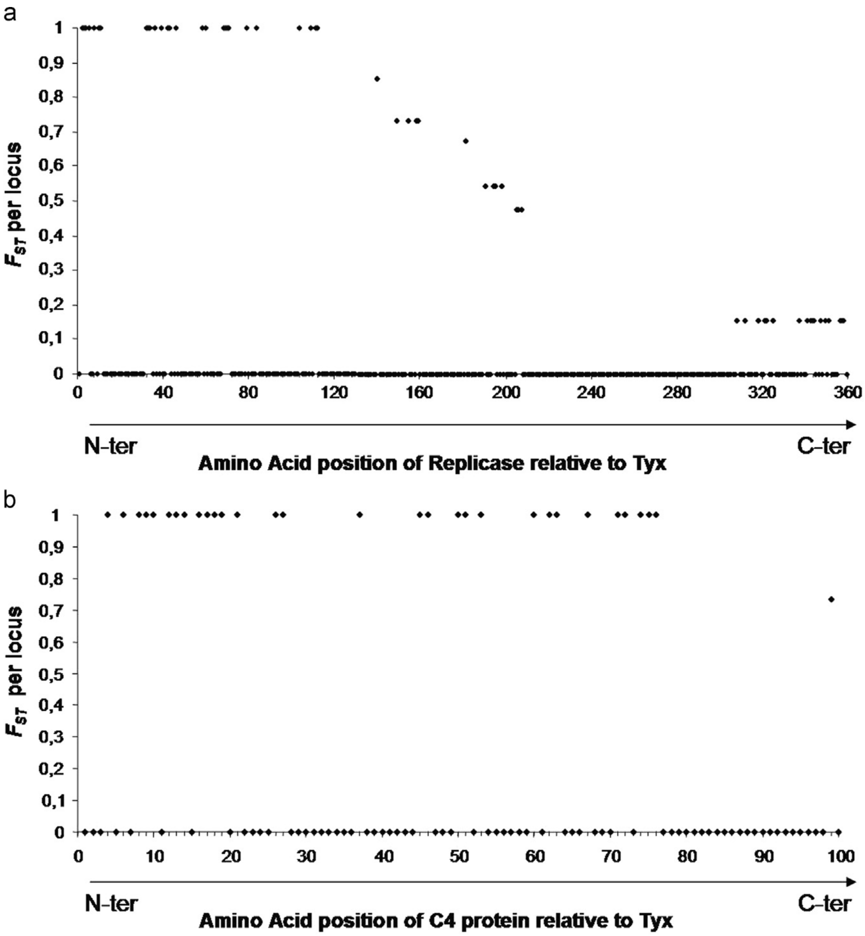

Fig. 3. Locus-by-locus F_{ST} between two groups of *begomovirus* recombinant genomes for the Rep and C4 proteins. The two groups differ in their within-host viral DNA accumulation, which is either similar to the Tyx or Tox parent. Each point corresponds to the F_{ST} per locus.

IR located upstream of the C1 gene and the 5' end of the C1 gene which overlaps the two thirds of the C4 gene. This genomic region was previously reported to determine the viral replication of begomoviruses (Hanley-Bowdoin et al., 1999).

The QTL detection methods were developed many years ago (Tanksley, 1993). Most studies dealing with QTLs have focused on plant species with agronomical interest to develop genomic selection programs to improve crop yield and disease resistance

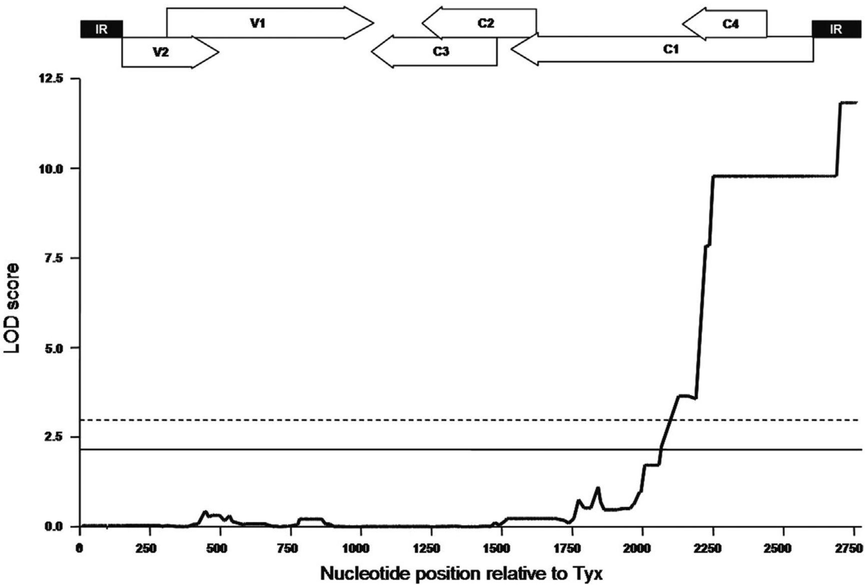

Fig. 4. LOD score distribution on *begomovirus* genome for the viral accumulation using the single-marker regression method. Black and dashed lines correspond to significance thresholds based on 5000 permutations at $P=0.05$ (solid line) and $P=0.01$ (dashed line), respectively. A viral genome linearized at the virion strand origin of replication is presented with the open reading frames (horizontal arrows) and the intergenic region (IR) at the top of the graph.

Table 1
Genomic regions involved in the viral accumulation of two begomoviruses using four different QTL methods.

Methods	Genomic region position in the alignment of TYX and TOX genomes	R^2
Single marker regression (SMR)	2127–2189 **2223–2761**	30.1% **53.6–68.6%**
Simple interval mapping (SIM)	2094–2221 **2223–2557**	26.8–52.8% **53.6–68.6%**
Single-trait MLE	2087–2221 **2223–2759**	28.6–54.0% **53.6–68.6%**
Single-trait multiple IM (MIM)	**2721–2725**	**77.3%**

The R^2 values correspond to the proportion of phenotypic variation explained by the QTL marker with a significance thresholds based on 5000 permutations at $P=0.01$. Bold numbers correspond to genetic regions which overlap with the loci for which an F_{ST} index of 1 was detected between recombinants genomes of low and high viral accumulation groups (positions 2223–2761). Position 1 corresponds to the origin of replication.

(Price et al., 2006) and on humans for identification of genes underlying diseases (*e.g.* Hubner et al. (2005)), but some researchers have also used this method to deal with questions about evolutionary biology and ecology (Erickson et al., 2004; Mauricio, 2001; Mitchell-Olds, 1995). In our study, we have demonstrated that QTL approaches could also be extended to viruses to identify genetic determinants of viral traits like within-host accumulation and early infectivity. In the literature, the theory and computer simulations argue for a large population size, at least several hundred, to correctly identify QTLs (Beavis, 1998; Beavis et al., 1994). In practical, proper QTL analyses were performed with 200 and 300 plant progenies (*e.g.* Melchinger et al. (1998); Wang et al. (1994)). In this study, QTLs of viral accumulation and infectivity were detected with only 47 artificial recombinant genomes of begomoviruses at a 500 bp resolution that plant breeders would envy. Similarly, we show that locus-by-locus F_{ST} analysis can be extended to a virus model and can help for the identification of genetic determinants of viral traits such as within-host accumulation.

The same genetic determinants were detected by F_{ST} and three QTL methods when R^2 values (*i.e.* proportion of phenotypic variation explained by the QTL marker) above 50% were considered (Figs. 1 and 4, Table 1, Additional files 3a, 3b and 5). The consistency between the two approaches indicates that the use of base pair distances between loci instead of the generally used cM distance (*e.g.* distances in Qgene; (Nelson, 1997)), did not hamper the relevance of the analyses. Such a result was not unexpected as cM unit is positively correlated to base pair. The broad 2223–2761 region which was inferred from both approaches to be associated with the measured traits might be potentially shorter. Indeed, the inclusion in the data set of two recombinant genomes (SH-54 and SH-64) which exhibit a mixed origin in this region has obviously pointed towards a shorter region beyond the 2700 nt position. This is consistent with the results obtained with the single-trait multiple interval mapping analysis (MIM, Table 1, Additional files 3c and 4d) and the higher R^2 values for nucleotide positions after nt 2700 with the 3 other QTL analyses (Table 1 and Additional file 5). The capacity of MIM model to use multiple marker intervals simultaneously to construct multiple putative QTL (Kao et al., 1999) explains the restriction of loci between 2721 and 2725. The fact that we identified the same genomic region already known to determine the replication of geminivirus genome strongly supports the interest of F_{ST} and QTL approaches in the identification of genomic regions influencing quantitative traits.

The DNA accumulation of the recombinant genomes exhibited a bimodal distribution which was supposed to be conducive for identifying the underlying genetic determinants through F_{ST} and QTL methods (See Fig. 4B in Vuillaume et al. (2011)). Although early infectivity and viral accumulation traits were positively correlated to each other (Additional file 6C), the data set for early infectivity did not exhibit a clear bimodal distribution, suggesting that the genetic determinism of this trait could be more complex. However, the QTL method allowed identifying genetic determinants of this trait (Additional file 6C), confirming the robustness of these approaches and their potential to identify traits with a more complex genetic determinism.

The QTL region encompasses the viral Rep,C4 and the 5′ end of the intergenic region

Contribution of individual QTLs (*i.e.* each polymorphic locus) to total phenotypic variation for viral accumulation ranged from 19.7% to 68.6% (with single-marker regression method; Table 1 and Additional file 5) in the last third of the TYLCV-Mld × ToLCKMV-[Dem] recombinant genomes (positions 1861–2761). The loci detected to be associated to virus accumulation by both F_{ST} and QTL approaches were located in two overlapping genes coding for *Rep*/C1 and C4 proteins and the 5' end of the IR and contribute from 53.6% to 68.6% variation in within-host accumulation (significant LOD score, Figs. 1, 3 and 4, Table 1 and Additional file 5). It was not unexpected to detect major SNP loci with R^2 above 50% in this region because *Rep* gene and IR were both reported to be major determinants of *Begomovirus* replication cycle (see review Gutierrez (2000)) and conserved functional motifs had been identified in both of them (Additional file 7). Two different DNA elements in the IR are functional targets of peptide motifs of the Rep: (i) a tandemly repeated motif located at variable distances from the conserved hairpin (iteron), which is bound specifically by the iteron-related domain (IRD) of its cognate Rep and functions as a major recognition element of the replication origin in *begomoviruses* and *curtoviruses* (Argüello-Astorga et al., 1994; Choi and Stenger, 1995; Fontes et al., 1994) and (ii) the nonameric motif 5'-TAATATT↓AC-3', invariably located at the loop of a conserved “hairpin” element, where Rep introduces a site-specific nick (↓) to initiate virus replication *via* a rolling circle mechanism (Jupin et al., 1995; Laufs et al., 1995). TYLCV-Mld and ToLCKMV-[Dem] differ in the sequences of the iterons, the IRD (Additional file 7) and RCR2, RCR3 and GRS motifs in Rep. However, the 3D protein structures of this region for TYLCV-Mld and ToLCKMV-[Dem] predicted by i-TASSER protein prediction (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) are typically that of the Rep of *begomovirus* (N-terminal domain of the *Tomato yellow leaf curl Sardinia virus* Rep, PDB database ID 1L5I, DOI:10.2210/pdb1l5i/pdb), principally the location of the three motifs ((Campos-Olivas et al., 2002; Nash et al., 2011); Additional file 7b). According to the close correlation reported between the sequence of IRD and that of cognate iterons in geminiviruses (Argüello-Astorga and Ruiz-Medrano, 2001), it is predicted that TYLCV/ToLCV recombinant genomes which would present an heterologous association of these sequences would not replicate (Jupin et al., 1995). All the recombinant genomes tested in this study have iteron, IRD and all the motifs involved in viral replication coming from the same parental genome (either Tyx or Tox), allowing the specific function of Rep. Therefore, although our results suggest that the entire region is responsible of the level of within-host viral accumulation, it could be one or some of the motifs present in the 128N-terminal amino acids of Rep or the IRD which determine the viral accumulation trait (Additional file 7b). In addition to the iterons and the nonameric motif, the 5' side of the IR contains the promoter region of all the complementary-sense genes *Rep*/C1 gene, Transcriptional activator protein *TrAP*/C2 gene and replication enhancer gene *REn*/C3, which indirectly influence the viral accumulation. The generation of other recombinant genomes harboring recombination events within the QTL region detected here to be important for virus accumulation may permit to determine the relative importance of reported IR and Rep motifs and possibly detect new ones.

Our results suggest that the C-terminal amino acids of Rep did not significantly explain viral accumulation (Fig. 3a). This is consistent with the current knowledge that this region which confers oligomerisation activity, interaction with transcriptional activator protein (*TrAP*) and ATP hydrolysis (Hanley-Bowdoin et al., 1999; Desbiez et al., 1995) does not affect viral accumulation. Since ORF C4 is embedded in the Rep/C1 ORF, our results cannot conclude on its specific involvement in within-host accumulation (Figs. 1 and 3b). However, as the C4 multifunctional protein is involved in symptom determination (Vanitharani et al., 2004; Rigden et al., 1994), virus movement (Jupin et al., 1994; Bisaro, 2006) and gene silencing (Vanitharani et al., 2004), it may indirectly impact viral accumulation.

Geminiviruses need to interact efficiently with their host from which they rely for replication and transcription because their small genome does not encode any polymerase. Thus, it cannot be excluded that the QTL profile for a particular trait is potentially host dependant. However, the genomic region involved in viral accumulation in tomato plants was the same in this study and in previous studies conducted with different *begomoviruses* and different host plants (Orozco et al., 1998; Orozco et al., 1997).

Extension of the QTL and F_{ST} approaches to viruses

The approaches reported here rely on a library of recombinants. When viruses are prone to recombination as *Begomovirus* (Martin et al., 2011; Urbino et al., 2013), *Caulimovirus* (Froissart et al., 2005) or *Coronavirus* (Baric et al., 1990), a library can be generated in a host co-infected with representatives of two parental species or strains. As selection and genetic drift may reduce the diversity of recombinant genomes, it may be recommended to isolate recombinant genomes as soon as possible after co-infection; for begomoviruses it was reported that 50% of the genomes were recombinant from 120 days post inoculation (Martin et al., 2011; Urbino et al., 2013; García-Andrés et al., 2007). An early isolation of the recombinant genomes may also limit the frequency of random mutations. Indeed, although QTL approaches are powerful, they cannot be used if recombinant genomes present mutations which cannot be assigned to any of the parental genome. For viral genera which are not prone to recombination, representative parental genomes may be shuffled *in vitro* as reported for begomoviruses (Gutierrez, 2000). Alternatively, genome-wide association studies may be used on natural viral populations from which numerous and diverse genomes may be isolated, sequenced and associated to contrasted phenotypes.

In addition to viral accumulation and early infectivity, other quantitative traits such as vector transmission efficiency or virulence may be studied. Transmission efficiency might be very informative because it is expected to be determined by at least two genes: the coat protein gene, which is the determinant of transmission (Hohnle et al., 2001; Noris et al., 1998), and the Rep gene, which is the determinant of viral accumulation (which in turn determines virus availability) (Lapidot et al., 2001).

Conclusions

Using a library of recombinant viral genomes previously generated *in vitro* between two begomoviruses and the data set of their within-host accumulation, a genomic region determining this trait was identified with locus-by-locus F_{ST} and QTL mapping. The relevance of this approach was confirmed by the fact that both F_{ST} and QTL mapping pointed to the same genomic region and, most importantly, that common virological approaches previously identified this region as being involved in viral replication. This study provides evidence that it is possible to extend QTL and population genetics tools to virology for the identification of genetic determinants governing phenotypic traits. Using a second data set related to early infectivity, we validated the robustness of these methods by detecting a significant QTL region. Moreover, the F_{ST} method can be efficiently extended to qualitative viral data for the mapping of genetic determinants. The major limitation will be the size of the recombinant library, the preparation of infectious clones and the phenotyping throughput. Depending on these constraints, viruses may be more or less amenable to these new approaches.

Materials and methods

Within-host viral accumulation data set from parental and recombinant viral genomes

The parental begomoviruses, *Tomato yellow leaf curl virus* (TYLCV-Mld, accession no. AJ865337 here referred to as Tyx) and *Tomato leaf curl Mayotte virus* (ToLCKMV-[Dem] accession no. AJ865341, here referred to as Tox) (Delatte et al., 2005), were both used previously to generate a series of recombinant genomes *in vitro* by DNA shuffling (European Patent 1104457, US Patent 6951719; Proteus, Nîmes, France). Multiple genome alignment using Clustal ω version 2.0 ((Larkin et al., 2007); http://www.ebi.ac.uk/Tools/clustalw2/) and whole genomes of recombinant clones are presented in Additional file 8. Agroinfectious clones were obtained for these 47 distinct recombinant clones and were tested for their within-host viral accumulation in tomato plants at 22 dpi. The logarithm of the calibrated Normalized Relative Quantity (logNRQ) reflecting the within-host accumulation of the virus were obtained from real-time PCR assays. The average and median of the log-transformed data presented in (Vuillaume et al., 2011) were used in this study. In addition, the efficiency of the 47 clones to reach a detectable viral DNA accumulation after inoculation was measured by the percentage of PCR-positive plants at 15 dpi which is here referred to as early infectivity (13).

F_{ST} analysis

Genetic differentiation between the genomes of the Tyx-type and Tox-type accumulation groups was assessed with the fixation index F_{ST} (Weir and Cockerham, 1984) using the analysis of molecular variance (AMOVA) procedure implemented in the Arlequin 3.0 software (Excoffier et al., 2005). A locus-by-locus AMOVA was used to detect significant differentiation in allele frequencies among the two groups. Analyses were performed on the whole recombinant genomes previously aligned by the program Clustalω2. The software was not designed for haploid organisms like viruses. We thus used the classical approach for such organisms: each sequence was coded as doubled haploid (*i.e.* $A=1\ 1$, $T=2\ 2$, $G=3\ 3$, $C=4\ 4$ and deletion$=5\ 5$), and the heterozygosity of a group of recombinants was calculated at each polymorphic locus by multiplying the frequency of both parental alleles supposing a theoretical Hardy-Weinberg equilibrium. F_{ST} and F-statistics were estimated for each locus of the alignment separately. A locus with an F_{ST} equal to 0 shows no genetic difference between the two groups of recombinants while F_{ST} equal to 1 shows a 100% difference between the two accumulation groups. A locus with an F_{ST} different from 0 was considered as a polymorphic locus in this study. A negative F_{ST} reflects the fact that more variance exists within than across the two groups of recombinant genomes (Excoffier et al., 2005). The same procedure was performed on the amino acid sequence of Rep and C4 proteins. The coding of amino acids is presented in Additional file 9.

QTL analysis

We used four different analytical approaches implemented in Qgene 3.06 (Nelson, 1997): single marker regression (SMR), simple interval mapping (SIM), Single-trait MLE (CIM-MLE), and single-trait multiple IM (MIM) methods. Recombinant sequences were also coded as doubled haploids: for each of the 534 polymorphic loci, the alleles were coded as either Tyx or Tox depending on the genome they derived from (Additional file 10). The allelic frequency per locus was of 30–62% for Tyx alleles and 38–70% for Tox alleles (Additional file 10). Thus, the genomic data set met the condition of relatively balanced frequency of each polymorphic site, and all of them were conserved in the compared recombinants. The genetic linkage distances used by QTL software are expressed in centimorgan (cM). However, as viruses have small genomes that enable obtaining full genome sequences for many individuals, physical distances were directly expressed in base pairs (bp) units. QTL that were significant in at least three approaches with $P<0.01$ were reported as a putative QTL for the within-host viral accumulation trait. R^2 represents the proportion of phenotypic variation explained by the QTL marker (Nagelkerke, 1991).

For early infectivity (percentage of PCR-positive plants at 15 dpi), a generalized linear model (GLM) was used with a binomial distribution and Firth's bias-adjusted estimates. The LOD score from the single marker regression QTL mapping method was calculated at each locus and plotted all along the viral genome. Only the markers with a LOD score above the threshold significance level obtained by permutation tests were considered significant.

Statistical analyses

A generalized linear model (GLM) with a binomial distribution and Firth's bias-adjusted estimates was used to calculate the early infectivity of each recombinant clone. A hierarchical ANOVA was used to compare the viral accumulation of the two genetic subgroups. Both analyses were performed with JMP 10 (SAS Institute Inc., Cary, North Carolina).

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

JD, GT and MP conceived the analysis. JD and CU designed the analysis. JD, CU, MP, and GT wrote the manuscript. JD analyzed the data. FV acquired the biological data. All authors have read and approved the final version of the manuscript.

Acknowledgments

Nathalie Boissot is acknowledged for valuable comments and suggestions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2015.06.019.

References

Argüello-Astorga, G.R., Ruiz-Medrano, R., 2001. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid–base pairs by a comparative approach. Arch. Virol. 146 (8), 1465–1485.

Argüello-Astorga, G.R., Guevara-González, R.G., Herrera-Estrella, L.R., Rivera-Bustamante, R.F., 1994. Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. Virology 203 (1), 90–100.

Baric, R.S., Fu, K., Schaad, M.C., Stohlman, S.A., 1990. Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. Virology 177 (2), 646–656.

Beavis, W.D., 1998. QTL analysis: Power, precision, and accuracy. In: Paterson, A.H. (Ed.), Molecular dissection of complex traits. CRC Press, Boca Raton, FL, pp. 145–161.

Beavis, W.D., Grant, D., Albertsen, M., Fincher, R., 1991. Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theor. Appl. Genet. 83 (2), 141–145.

Beavis, W.D., Smith, O.S., Grant, D., Fincher, R., 1994. Identification of quantitative trait loci using a small sample of topcrossed and F4 progeny from maize. Crop Sci. 4, 882–896.

Bisaro, D.M., 2006. Silencing suppression by geminivirus proteins. Virology 344 (1), 158–168.

Campos-Olivas, R., Louis, J., Clérot, D., Gronenborn, B., Gronenborn, A., 2002. Letter to the Editor: 1H, 13C, and 15N assignment of the N-terminal, catalytic domain of the replication initiation protein from the geminivirus TYLCV. J. Biomol. NMR, 24; , pp. 73–74.

Choi, I.R., Stenger, D.C., 1995. Strain-specific determinants of Beet curly top geminivirus DNA replication. Virology 206 (2), 904–912.

Dawson, W.O., 1978. Isolation and mapping of replication-deficient, temperature-sensitive mutants of cowpea chlorotic mottle virus. Virology 90 (1), 112–118.

Delatte, H., Martin, D.P., Naze, F., Goldbach, R., Reynaud, B., Peterschmitt, M., et al., 2005. South West Indian Ocean islands tomato *begomovirus* populations represent a new major monopartite begomovirus group. J. Gen. Virol. 86 (5), 1533–1542.

Desbiez, C., David, C., Mettouchi, A., Laufs, J., Gronenborn, B., 1995. Rep protein of tomato yellow leaf curl geminivirus has an ATPase activity required for viral DNA replication. Proc. Natl. Acad. Sci. USA 92 (12), 5640–5644.

Elmer, S.J., Brand, L., Sunter, G., Gardiner, W.E., Bisaro, D.M., Rogers, G., 1988. Genetic analysis of *Tomato golden mosaic virus*. II. The product of AL1 coding sequence is required for replication. Nucleic Acids Res. 16, 7043–7060.

Erickson, D.L., Fenster, C.B., Stenøien, H.K., Price, D., 2004. Quantitative trait locus analyses and the study of evolutionary process. Mol. Ecol. 13 (9), 2505–2522.

Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform Online 1, 47–50.

Fontes, E.P., Eagle, P.A., Sipe, P.S., Luckow, V.A., Hanley-Bowdoin, L., 1994. Interaction between a geminivirus replication protein and origin DNA is essential for viral replication. J. Biol. Chem. 269 (11), 8459–8465.

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

García-Andrés, S., Tomás, D.M., Sánchez-Campos, S., Navas-Castillo, J., Moriones, E., 2007. Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. Virology 365 (1), 210–219.

Gutierrez, C., 2000. DNA replication and cell cycle in plants: learning from geminiviruses. EMBO J. 19 (5), 792–799.

Hanley-Bowdoin, L., Settlage, S.B., Orozco, B.M., Nagar, S., Robertson, D., 1999. Geminiviruses: models for plant DNA replication, transcription, and cell cycle regulation. Crit. Rev. Plant Sci. 18 (1), 71–106.

Hohnle, M., Hofer, P., Bedford, I.D., Briddon, R.W., Markham, P.G., Frischmuth, T., 2001. Exchange of three amino acids in the coat protein results in efficient whitefly transmission of a nontransmissible Abutilon mosaic virus isolate. Virology 290 (1), 164–171.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat. Rev. Genet. 10 (9), 639–650.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., et al., 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. 37 (3), 243–253.

Jeske, H., Lütgemeier, M., Preiß, W., 2001. DNA forms indicate rolling circle and recombination-dependent replication o*f Abutilon mosaic virus*. EMBO J. 20 (21), 6158–6167.

Jupin, I., De Kouchkovsky, F., Jouanneau, F., Gronenborn, B., 1994. Movement of *Tomato yellow leaf curl geminivirus* (TYLCV): involvement of the protein encoded by ORF C4. Virology 204 (1), 82–90.

Jupin, I., Hericourt, F., Benz, B., Gronenborn, B., 1995. DNA replication specificity of TYLCV geminivirus is mediated by the amino-terminal 116 amino acids of the Rep protein. FEBS Lett. 362 (2), 116–120.

Kao, C.-H., Zeng, Z.-B., Teasdale, R.D., 1999. Multiple interval mapping for quantitative trait loci. Genetics 152 (3), 1203–1216.

Lapidot, M., Friedmann, M., Pilowsky, M., BenJoseph, R., Cohen, S., 2001. Effect of host plant resistance to tomato yellow leaf curl virus (TYLCV) on virus acquisition and transmission by its whitefly vector. Phytopathology 91 (12), 1209–1213.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al., 2007. Clustal w and clustal x version 2.0. Bioinformatics 23 (21), 2947–2948.

Laufs, J., Jupin, I., David, C., Schumacher, S., Heyraud-Nitschke, F., Gronenborn, B., 1995. Geminivirus replication: genetic and biochemical characterization of Rep protein function, a review. Biochimie 77 (10), 765–773.

Lynch, M., Walsh, B., 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, MA.

Martin, D.P., Rybicki, E.P., 2002. Investigation of maize streak virus pathogenicity determinants using chimaeric genomes. Virology 300 (2), 180–188.

Martin, D.P., Lefeuvre, P., Varsani, A., Hoareau, M., Semegni, J.Y., Dijoux, B., et al., 2011. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog. 7 (9), e1002203.

Mauricio, R., 2001. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat. Rev. Genet. 2 (5), 370–381.

Melchinger, A.E., Utz, H.F., Schön, C.C., 1998. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149 (1), 383–403.

Mitchell-Olds, T., 1995. The molecular basis of quantitative genetic variation in natural populations. Trends Ecol. Evol. 10 (8), 324–328.

Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika 78 (3), 691–692.

Nash, T.E., Dallas, M.B., Reyes, M.I., Buhrman, G.K., Ascencio-Ibañez, J.T., Hanley-Bowdoin, L., 2011. Functional analysis of a novel motif conserved across geminivirus Rep proteins. J. Virol. 85 (3), 1182–1192.

Nelson, J., 1997. QGENE: software for marker-based genomic analysis and breeding. Mol. Breed. 3 (3), 239–245.

Noris, E., Vaira, A.M., Caciagli, P., Masenga, V., Gronenborn, B., Accotto, G.P., 1998. Amino acids in the capsid protein of tomato yellow leaf curl virus that are crucial for systemic infection, particle formation, and insect transmission. J. Virol. 72 (12), 10050–10057.

Orozco, B.M., Miller, A.B., Settlage, S.B., Hanley-Bowdoin, L., 1997. Functional domains of a Geminivirus replication protein. J. Biol. Chem. 272 (15), 9840–9846.

Orozco, B.M., Gladfelter, H.J., Sharon, B.S., Eagle, P.A., Gentry, R.N., Hanley-Bowdoin, L., 1998. Multiple cis elements contribute to geminivirus origin function. Virology 242, 346–356.

Paterson, A.H., Tanksley, S.D., Sorrells, M.E., 1991. DNA markers in plant improvement. In: Sparks, D.L. (Ed.), Advances in Agronomy, 46. Academic Press, New York, pp. 39–90.

Price, R.N., Uhlemann, A.-C., van Vugt, M., Brockman, A., Hutagalung, R., Nair, S., et al., 2006. Molecular and pharmacological determinants of the therapeutic response to artemether-lumefantrine in multidrug-resistant *Plasmodium falciparum* malaria. Clin. Infect. Dis. 42 (11), 1570–1577.

Rigden, J.E., Krake, L.R., Rezaian, M.A., Dry, I.B., 1994. ORF C4 of tomato leaf curl geminivirus is a determinant of symptom severity. Virology 204 (2), 847–850.

Saenz, P., Quiot, L., Quiot, J.B., Candresse, T., Garcia, J.A., 2001. Pathogenicity determinants in the complex virus population of a *Plum pox virus* isolate. Mol. Plant Microbe Interact. 14 (8), 1032.

Stratford, R., Covey, S.N., 1989. Segregation of *Cauliflower mosaic virus* symptom genetic determinants. Virology 172 (2), 451–459.

Stuber, C.W., Edwards, M.D., Wendel, J.F., 1987. Molecular marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop. Sci. 27 (4), 639–648.

Tanksley, S.D., 1993. Mapping polygenes. Annu. Rev. Genet. 27, 205–233.

Urbino, C., Gutiérrez, S., Antolik, A., Bouazza, N., Doumayrou, J., Granier, M., et al., 2013. Within-host dynamics of the emergence of *Tomato yellow leaf curl virus* recombinants. PLoS One 8 (3), e58375.

Vanitharani, R., Chellappan, P., Pita, J.S., Fauquet, C.M., 2004. Differential roles of AC2 and AC4 of cassava geminiviruses in mediating synergism and suppression of posttranscriptional gene silencing. J. Virol. 78 (17), 9487–9498.

Vuillaume, F., Thébaud, G., Urbino, C., Forfert, N., Granier, M., Froissart, R., et al., 2011. Distribution of the phenotypic effects of random homologous recombination between two virus species. PLoS Pathog. 7 (5), e1002028.

Wang, G.L., Mackill, D.J., Bonman, J.M., McCouch, S.R., Champoux, M.C., Nelson, R.J., 1994. RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. Genetics 136 (4), 1421–1434.

Weir, B.S., Cockerham, C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38 (6), 1358–1370.

Wright, S., 1949. The genetical structure of populations. Ann. Eugen. 15 (1), 323–354.