Genetic diversity and potential vectors and reservoirs of *Cucurbit aphid-borne yellows virus* in southeastern Spain

Mona Kassem, Miguel Juarez, Pedro Gómez, Carmen M. Mengual, Raquel N. Sempere, Santiago F. Elena, Aranzazu Moreno, Alberto Fereres, and Miguel A. Aranda

First, third, fourth, fifth, and ninth authors: Centro de Edafología y Biología Aplicada del Segura (CEBAS), Consejo Superior de Investigaciones Científicas (CSIC), apdo. correos 164, 30100 Espinardo, Murcia, Spain; third author: Current address: Centre for Ecology and Conservation, School of Biosciences, University of Exeter, Cornwall Campus, Penryn TR10 9EZ, UK; second author: Escuela Politécnica Superior de Orihuela, Universidad Miguel Hernández de Elche, Ctra. de Beniel km 3.2, 03312 Orihuela, Alicante, Spain; sixth author: Instituto de Biología Molecular y Celular de Plantas (IBMCP), CSIC-Universidad Politécnica de Valencia, Ingeniero Fausto Elio s/n, 46022 Valencia, Spain, and The Santa Fe Institute; Santa Fe, 1399 Hyde Park Road, NM 87501, USA; seventh and eighth authors: Instituto de Ciencias Agrarias (ICA), CSIC, Serrano 115, 28006, Madrid, Spain.

Correspondence: Miguel A. Aranda
Address: CEBAS-CSIC. Apdo. correos 164. 30100, Espinardo, Murcia, Spain.
Phone: +34-968396355. Fax: +34-968396213
E-mail address: m.aranda@cebas.csic.es

Keywords: *Cucumis melo*, *Cucurbita pepo*, gene flow, luteovirus, melon, Polerovirus, resistance, selection, squash, virus emergence.

Number of words in the main text of the paper: 6136
Number of words in the Summary: 221
Number of figures: 5
Supplementary material: 2

SUMMARY

The genetic variability of a *Cucurbit aphid-borne yellows virus* (CABYV) (genus *Polerovirus,* family *Luteoviridae*) population was evaluated by determining the nucleotide sequences of two genomic regions of CABYV isolates collected in open field melon and squash crops during three consecutive years in Murcia (southeastern Spain). A phylogenetic analysis showed the existence of two major clades. The sequences did not cluster according to host, year or locality of collection, and nucleotide similarities among isolates were 97-100% and 94-97% within and between clades, respectively. The ratio of non-synonymous to synonymous nucleotide substitutions reflected that all open reading frames have been under purifying selection. Estimates of the population's genetic diversity were of the same magnitude as those previously reported for other plant virus populations sampled at larger spatial and temporal scales, suggesting either the presence of CABYV in the surveyed area long before it was first described, multiple introductions or a particularly rapid diversification. We also determined the full-length sequences of three isolates, identifying the occurrence and location of recombination events along the CABYV genome. Furthermore, our field surveys indicated that *Aphis gossypii* was the major vector species of CABYV and the most abundant aphid species colonizing melon fields in the Murcia (Spain) region. Our surveys also suggested the importance of the weed species *Ecballium elaterium* as an alternative host and potential virus reservoir.

INTRODUCTION

Cucurbit aphid-borne yellows virus (CABYV) is a member of the genus *Polerovirus* within the family *Luteoviridae*. CABYV was first described in 1992 in France (26), but it was later detected infecting cucurbits in many other countries in the world (2,23,26,28,54). In addition to cucurbits, CABYV can infect other crop species, including lettuce (*Lactuca sativa*) and fodder beet (*Beta vulgaris*), as well as common weeds of the species *Capsella bursa-pastoris*, *Lamium amplexicaule*, *Senecio vulgaris*, *Ecballium elaterium*, *Bryonia dioica*, *Papaver rhoeas*, *Montia perfoliata* and *Crambre abyssinica*, which are thought to be virus reservoirs. CABYV is a phloem-limited virus that is transmitted in a persistent, non-propagative mode by aphids, including *Aphis gossypii* and *Myzus persicae* (26). Typical symptoms of CABYV infection include yellowing and thickening of basal and older leaves. In melon and cucumber, a high percentage of flower abortion was detected with no effects on fruit quality (26).

CABYV particles are isometric, of about 25 nm in diameter, and encapsidate the CABYV genome, which consists of a single-stranded positive sense RNA molecule of 5.7 kb in length (10). The CABYV genome contains at least six ORFs organized into two clusters (33). The first three ORFs (1, 2 and 3) are expressed from the genomic RNA, whereas the other three (4, 5 and 6) are expressed from a subgenomic RNA (sgRNA1). Two additionally predicted ORFs (7 and 8) might be expressed from a second sgRNA (sgRNA2) (1). The protein encoded by ORF 1 (P1) is a strong silencing suppressor and an enhancer of pathogenicity (42,43). The proteins, putatively encoded by ORF 2 (P2) and ORF 3 (P3), have regions of amino acid sequence similarity with serine proteases and genome-linked viral proteins (VPgs) of other poleroviruses, as well as amino acid motifs typical of RNA-dependent RNA polymerases (RdRps) (15), respectively. By analogy with other poleroviruses, it is likely that CABYV ORF 3 is translated as a result of a ribosomal frameshift in the C-terminal portion of ORF 2 to

generate a P2–P3 fusion protein (15). ORFs 4, 5 and 6 encode proteins P4 (the coat protein, CP), which is involved in viral transmission, particle packaging and viral accumulation within the plant, P5 (the movement protein, MP) and P6. P6 is expressed only as a read-through protein (P4-P6), and is needed for aphid transmission of the virus (5,7,14,56).

Based on phylogenetic analysis, CABYV isolates have been classified into the Asian and Mediterranean groups (46). Manarri-Hattab *et al.* (34) further distinguished Tunisian, Italian and French isolates from Spanish and Chinese isolates, and suggested that gene flow could have had a fundamental role in this geographically-based distribution. Among poleroviruses, very low genetic diversity was described for *Potato leafroll virus* (PLRV) (16) and *Cotton leafroll dwarf virus* (CLRDV) (47), while *Barley yellow dwarf virus-PAV* (BYDV-PAV) (3,31,52) and viruses belonging to the beet polerovirus complex (18) are known to be more variable. Recombination may have had a fundamental role in maintaining luteovirus diversity, as appears to be the case for CABYV, where the continuous generation of recombinant strains has lead to great variability in the Asian group (24,46,53). Recombination within members of the family *Luteoviridae* has been extensively analyzed, showing that recombination break-points are distributed along various regions of the virus genomes, although the main putative recombination hotspot seems to lie between the end of ORF 3 and the beginning of ORF 4 (24,38). Pagán and Holmes (38) have shown that luteovirus genes appear to be subjected to purifying selection, except for the *MP* gene that seems to be evolving neutrally in the majority of cases. Torres *et al.* (50) have shown that the *CP* gene of luteoviruses appears to be under heterogeneous selective forces; furthermore, highly conserved sites under purifying selection were identified mainly on the CP surface,

107 perhaps associated to constraints imposed by the conservation of receptors in the aphid
108 and the host plant (50).

109 CABYV has been shown to be the prevalent virus in cucurbit crops in
110 southeastern Spain (23). With this study, our goal was to understand the epidemiology
111 and population genetics of CABYV. We focused on an ecologically confined area,
112 Campo de Cartagena (Murcia, Spain), where cucurbit cultivation has high economic
113 importance. Thus, we examined the genetic variability of a CABYV population sampled
114 in this area during 3 consecutive years by sequencing two genomic regions of CABYV
115 isolates; the full-length sequence of three of these isolates was also determined to
116 identify occurrence and location of recombination events along the CABYV genome. In
117 addition, we studied the dynamics of aphid populations and sampled aphids landing on
118 melon crops (vector activity) in this same geographical area to determine the proportion
119 of viruliferous aphids and also which species were potentially acting as major vectors of
120 CABYV (vector propensity) (20). Finally, we also sampled weeds and surrounding
121 crops and determined the presence of CABYV to identify potential alternative hosts and
122 virus reservoirs.

123

124 **MATERIALS AND METHODS**

125

126 **CABYV surveys, virus detection and isolates**. Surveys were performed in open field
127 melon (*Cucumis melo*) and squash (*Cucurbita pepo*) crops of the Murcia (Spain)
128 province during 2003, 2004 and 2005 as described by Kassem *et al.* (23). During these
129 three years, more than 1000 samples were obtained from symptomatic plants from 48
130 randomly distributed fields in a geographical area of approximately 300 km^2 (Fig. 1A).
131 All samples were analyzed to detect the presence of CABYV by dot-blot hybridization.

Approximately 80% of the total of samples were positive for CABYV. CABYV isolates analyzed in this study were chosen randomly from this collection. The same geographical area was surveyed during 2009 to identify alternative hosts and potential virus reservoirs. In this case, 239 samples were collected from 28 weed and 6 alternative crop species (see below).

Partial sequencing of CABYV isolates. Complementary DNAs (cDNAs) to two regions of the CABYV genome (Fig. 1B) were synthesized by RT-PCR. The primer pairs CE-124 (5′-CTCCTTCCGATATTGGCTCG-3′) and CE-123 (5′-CCCATTCTGCGCCGC AGTGG-3′), and CE-9 (5′-GAATACGGTCGCGGCTAGAAATC-3′) and CE-10 (5′-C TATTTCGGGTTCTGGACCTGGC-3′) (23) were designed based on the nucleotide sequence of isolate CABYV-N (GenBank accession no. NC003688) to amplify fragments from nucleotide 2,316 to 2,979 of ORF 3 and nucleotide 3,507 to 4,104 of ORF 4, respectively (Fig. 1B). Note that ORF 4 overlaps ORF 5 for most of its sequence (nucleotides 3,534 to 4,104). Reverse transcription and PCR were carried out using Expand Reverse Transcriptase (Roche, Mannheim, Germany) and Expand High Fidelity PCR System (Roche), respectively, following the recommendations of the manufacturer. RT-PCR products were purified from agarose gels after electrophoresis by using Geneclean turbo columns (MP Biomedicals, Illkrich, France) and then sequenced (Secugen, Madrid, Spain) using the same primers.

Sequences of ORF 4 from worldwide isolates were retrieved from GeneBank (http://www.ncbi.nlm.nih.gov/). Accession numbers of the corresponding sequences are listed in the corresponding figures or tables.

Phylogenetic and diversity analysis. Nucleic and amino acid sequences were aligned using ClustalX and BioEdit software (www.mbio.ncsu.edu/bioedit). To calculate the

157 percentages of nucleotide and amino acid identity and similarity, GeneDoc Editor v
158 2.7.000 software was used. Further sequence and cluster analyses were performed with
159 the MEGA4 package. Phylogenetic trees were constructed by using the minimum-
160 evolution (ME) method, using distance matrices based on the Kimura 2-parameter
161 (K2P) method. A bootstrap value for each internal node of the tree was calculated by
162 using 1000 random pseudoreplicates. Maximum likelihood and maximum parsimony
163 trees were constructed using programs from the phylogeny inference package PHYLIP.
164 The mean genetic distance for each gene was calculated using the K2P distance as
165 implemented in the MEGA4 package. The ratio of nonsynonymous (d_N) to synonymous
166 (d_S) nucleotide substitutions ($\omega = d_N/d_S$) for each ORF was calculated using the Pamilo-
167 Bianchi-Li method. The MEGA4 package was also used to perform Tajima´s *D*
168 statistical tests and to compute the total nucleotide diversity π. Nucleotide diversity was
169 estimated using the K2P distance estimator, and was expressed as the average number
170 of nucleotide substitutions per site in each pair of sequences. To assess selection
171 pressures, a Maximum Likelihood (ML) approach was used. In this case, non-
172 synonymous and synonymous differences that correlated with phylogenetic
173 relationships were estimated using the HyPhy package (http://www.hyphy.org). The ω
174 ratio was calculated by the MG94 model as implemented in HyPhy.

175 A Bayesian Markov Chain Monte Carlo (MCMC) approach, implemented in the
176 Bayesian Tip-associated Significance testing (BaTS) software, was employed to explore
177 the overall structure of the sequence dataset in relation with host adaptation. This
178 analysis was based on the trees that were constructed by using the BEAST program
179 version 1.4.8 (http:// beast.bio.ed.ac.uk), removing 10% of burn-ins and using 1,000
180 replicates. From these trees, the significance of the Parsimony Score (PS) and
181 Association Index (AI) statistics was computed to assess the strength of host clustering

182 in the phylogeny. The PATH-O-GEN software
183 (http://tree.bio.ed.ac.uk/software/pathogen/) was employed to explore the potential
184 correlation between genetic structure and year of sampling. ML trees were randomly
185 generated using the best substitution models HKY+G+I (for ORFs 3 and 4) and GTR+G
186 (for ORF 5). A Mantel test was used to examine the correlation between geographical
187 and genetic distances. Geographical distances were determined from the geographical
188 coordinates of each sample by using the Google Earth software version 5
189 (http://www.google.com/earth/index.html). Mantel's test was performed with the ZT
190 software using 50,000 nonparametric permutations for significance testing.

191 **Complete sequencing of CABYV isolates**. To determine the full-length genome
192 sequence of CABYV isolates, five overlapping cDNA fragments covering the entire
193 length of the genome were amplified from three randomly chosen isolates. The five
194 primer pairs used were: CE-56 (5′-ACAAAAGATACGAGCGGGTGATG-3′)/CE-92
195 (5′-GCATACCCGGCATGCGATCCGTC-3′) (nt 1 to 823); CE-93 (5′-
196 GGCGACTTGCCAATTTCCCT-3′)/CE-95 (5′-GAACTGGGGCCACCCGAAGC-3′)
197 (nt 646 to 2,016); CE-94 (5′- GCCATGAAGAAACCGAGGTCGCG-3′)/ CE-96 (5′-
198 GGCGTCGGCGGTTTCTTCG-3′) (nt 1,762 to 3,649); CE-9/CE-10 (nt 3,507 to 4,104)
199 (23) and CE-97 (5′-GGACCCCCACTGCAAGCTTAGC-3′)/CE-57 (5′-
200 ACACCGAAACGCCAGGGGGAATC-3′) (nt 3,885 to 5,669). All cDNAs were
201 sequenced in both directions using the adequate primers. The assembly of the CABYV
202 genomic sequences was carried out using BioEdit.

203 **Recombination and other sequence analyses**. Putative recombinant genomes were
204 evaluated using the RDP3 package (http://darwin.uvigo.es/rdp/rdp.html). RDP3
205 implements six recombination detection programs: RDP , GENECONV, Maximum
206 Chi-Square (MaxChi), Chimera, BootScan and SisterScan (SiScan). The default

207 detection thresholds were used. Events detected by more than two methods were
208 retained.

209 RNA secondary structure predictions were performed using the RNAalifold
210 program included in the Vienna RNA package v1.8.2 (http://rna.tbi.univie.ac.at). The
211 transmembrane regions were predicted by using the TMHMM, available at
212 http://www.cbs.dtu.dk/services, and TMPred, available at http://www.ch.embnet.org,
213 programs. Protein motifs and domains were identified by using motif libraries included
214 in the Prosite, PRINTS, ProDom and Pfam, available at http://motif.genome.jp,
215 databases.

216 **Aphid population dynamics and estimation of CABYV-viruliferous aphids landing**
217 **in melon fields.** Aphids were sampled on six randomly distributed melon fields in the
218 geographical area were the virus surveys were carried out. The sampling procedure was
219 designed to estimate the number of aphids alighting on the crop (vector activity) and the
220 number of aphids colonizing the crop (population density) at different time intervals
221 during the melon growth cycle. Aphids alighting on the selected experimental plots
222 were monitored weekly with a horizontal mosaic-green tile trap (19) during the spring-
223 summer melon-growing season of two consecutive years, 2010 and 2011. Aphids from
224 trap samples were separated in the laboratory using a stereomicroscope (Nikon Systems
225 Europe, Amstelveen, The Netherlands). All aphid specimens were preserved in 70%
226 ethanol and they were prepared and identified following the protocol described in Pérez
227 *et al*. (41). Additionally, aphid populations colonizing the crop (alate adults, apterae
228 adults and nymphs) were monitored weekly in the six selected melon fields during
229 summer, 2010. Four rows were selected in the sampled plots and 5 plants were chosen
230 within each of the rows for a complete total of 20 plants sampled each time. All leaves
231 of each sampled plant were visually inspected and the numbers and morph of aphids of

each species were recorded.

A sample of the aphid species caught in the field from the horizontal mosaic-green tile traps in 2011 was used to estimate the number of CABYV-viruliferous aphids visiting melon fields. Individual aphids were analyzed by quantitative real-time RT-PCR according to Moreno *et al*. (35).

CABYV transmission from *E. elaterium* to melon. To confirm the capacity of *E. elaterium* to act as a virus reservoir and infective source for the transmission of CABYV to melon crops, a total of eight *E. elaterium* plants were agroinoculated with CABYV (7). Twelve days post-inoculation plants were visually inspected and tested (see above) to detect CABYV. Transmission assays were carried out using *E. elaterium* infected plants, placed inside insect-proof boxes, as a virus source. Three to eight two-week old melon seedlings were introduced inside the boxes together with one CABYV-source *E. elaterium* plant. Then, approximately 100 virus-free *A. gossypii* adults were released into the boxes, which were then maintained for more than two weeks in an insect-proof glasshouse. Virus-free *A. gossypii* colonies were reared as described in Moreno *et al*. (35). Test melon seedlings were visually inspected for CABYV symptoms appearance, and presence of CABYV (see above) was determined 12 days after aphid release.

RESULTS

Phylogenetic relationships among CABYV isolates. Surveys were carried out in field plots used for commercial intensive production during the 2003, 2004 and 2005 growing seasons (23) (Fig. 1A). We sequenced two genomic fragments, one in ORF 3 and another in the overlapping ORFs 4 and 5 (Fig. 1B), for 50 isolates (Table 1),

257 covering ca. 21% of the CABYV genome. To analyze the phylogenetic relationships
258 among CABYV isolates, a matrix of genetic distances among isolates was computed
259 after the sequences were aligned. Each genomic region was used independently to infer
260 phylogenetic trees by various methods, including minimum evolution, maximum
261 parsimony and maximum likelihood; topologies inferred by the different methods were
262 all similar. For both regions, two major clades were identified, dividing the population
263 into groups Ia and Ib for ORF3 (Fig. 2A) and IIa and IIb for ORFs 4/5 (Fig. 2B).
264 Nucleotide similarities among isolates of each group ranged between 97% and 100%,
265 whereas nucleotide similarities among isolates from different groups ranged between
266 94% and 97%. These phylogenies did not reveal any obvious association of host (i.e.,
267 melon or squash) or year of collection with phylogenetic grouping (Fig. 2). Indeed, the
268 statistical methods implemented in BaTS and Path-O-Gen found no significant
269 association between host species and phylogenetic groups, or divergence time and
270 genetic diversity, respectively (data not shown). To analyze whether the geographic
271 origin of isolates could possibly be associated with genetic grouping, we looked for
272 potential correlations between geographic and genetic distances among pairs of isolates;
273 however, no significant correlation was detected (Mantel test: $P > 0.1$). We then tested
274 whether the inclusion of worldwide CABYV isolates would influence the resulting tree
275 topology. This was done only for the region covering ORFs 4/5, since databases
276 contained a significant number of sequences for this region, but not for other CABYV
277 genomic regions. In the corresponding deduced minimum evolution tree, Asian and
278 Mediterranean isolates clustered in different groups (Fig. 3). The Spanish isolates were
279 in the two clades containing Mediterranean isolates, one with Spanish but also with
280 isolates from other Mediterranean regions (Fig. 3, group IIa), and the other with Spanish
281 isolates only (Fig. 3, group IIb). Therefore, the CABYV genetic structure suggested by

282 phylogenetic analyses appeared to be related to geography only when the sampling scale
283 was large enough, but not at a local-scale.
284 **Genetic diversity and selective pressures in non-overlapping and overlapping**
285 **CABYV genes.** Values of nucleotide diversity were estimated independently for each
286 of the ORFs analyzed, ranging from 0.00 to 0.05 (Table 2). A pairwise comparison
287 among means showed that nucleotide diversity estimated for ORF 3 was significantly
288 higher than for ORF 4 (t = 10.897 $P < 0.0001$) and ORF 5 (t = 10.263 $P < 0.0001$), and
289 that there were no differences between ORF 4 and 5 (t = 0.122 $P < 0.906$), indicating
290 that variability was not evenly distributed throughout the viral genome. Furthermore, we
291 also estimated the relative rates of change at non-synonymous (d_N) and synonymous
292 positions (d_S) for each ORF by the use of $\omega = d_N/d_S$, in order to assay the direction and
293 strength of the selection pressure acting on coding regions within CABYV populations.
294 We found a low ratio ($\omega < 1$) for the three ORFs (Table 2), indicating that purifying
295 selection was restricting variability for all three ORFs. Moreover, the ω ratio of ORF 3
296 was approximately of one order of magnitude smaller than for the other two ORFs,
297 suggestive of purifying selection being stronger for this ORF than for the other two.
298 After further evaluating the selective constraints operating in each region, codons under
299 selection were detected using the three methods implemented in HyPhy: SLAC, FEL
300 and REL. The best substitution models were K81 for ORF 3 and HKY85 for ORF 4 and
301 ORF 5. We found no significant evidence of positive selection on codons ORF 3 and 4,
302 and only one codon (position 145) was found to be significantly under positive selection
303 in ORF 5 (Fig. 4). In contrast, several codons were found to be under negative selection,
304 but only six codons (positions 356, 409, 417, 429, 432 and 445) on ORF 3, five codons
305 (positions 16, 64, 109 and 175) on ORF 4, and two codons (positions 12 and 99) on
306 ORF 5 were detected by all three methods (Fig. 4). Alignment of CABYV amino acid

307 sequences around the only codon subjected to positive selection (Supplementary Fig.
308 S1) showed that proline (P) co-existed with glutamine (Q) in this position in 2003,
309 becoming prevalent in the following years.

310 In order to further analyze whether host adaptation could have contributed to
311 genetic differentiation of the CABYV population, nucleotide diversity values were
312 estimated within and between CABYV subpopulations, considering a subpopulation as
313 the group of isolates that were originally collected from melon or from squash.
314 Between-subpopulations diversity values were of the same order of magnitude than
315 within-subpopulation diversity values, and the $\omega < 1$ in all cases (Table 3). The
316 Tajima's *D* test was applied for each ORF, showing that *D* values were not significantly
317 different than zero ($P > 0.1$), indicating that the hypotheses of selective neutrality
318 cannot be rejected and further suggesting that there is no differentiation of the
319 population according to the host species from which the isolates were sampled.

320 **Complete genome sequencing of three CABYV isolates: Recombination events in**
321 **CABYV genomes.** Recombination has been shown to significantly contribute to
322 luteovirus diversity. An analysis of potential recombination events within the two
323 regions sequenced of the isolates characterized here did not reveal any significant result
324 (data not shown). However, when comparing phylogenies for the two regions
325 sequenced, a phylogenetic incongruence was detected for some isolates. For instance,
326 isolate Sq/2005/2.2 was included in group Ia for the ORF 3 tree but in group IIb for the
327 ORFs 4/5 tree (Fig. 2). To analyze this aspect in more detail and along the whole
328 CABYV genome, we took three isolates at random and determined their full sequences.

329 Genomic RNAs of isolates Sq/2003/7.2 (GenBank accession no. JF939812) and
330 Sq/2004/1.9 (GenBank accession no. JF939814) were both 5,672 nucleotides long,
331 whereas genomic RNA of isolate Sq/2005/9.2 (GenBank accession no. JF939813)

332 contained a 3-nucleotide insertion resulting in 5,675 nucleotides in length. The genomic
333 structure of the three isolates did not differ from that described for the other two,
334 completely sequenced CABYV isolates, one from France (CABYV-N) (15) and the
335 other from China (CABYV-C) (53). Similarly, all motifs previously identified in the
336 CABYV putative proteins were also identified in these three Spanish isolates, including
337 the short motif LPLLI in P1 (39), the chymotrypsin-related serine protease and VPg
338 motifs in P2 (45,51), three transmembrane regions in the amino-terminus of the P2 (32),
339 the GDD motif in P2-P3 (33), the arginine-rich RNA binding motif and epitopes 5 and
340 10 in P3 (CP) (6), and a proline-rich motif in RT (15). Isolates Sq/2003/7.2 and
341 Sq/2004/1.9 shared 98% of their nucleotide sequence, whereas isolate Sq/2005/9.2
342 showed only a 91% nucleotide sequence identity with the other two isolates (Table 4).
343 Nucleotide and amino acid identities among Sq/2003/7.2 and other poleroviruses were
344 computed for the whole genome and independently for the six ORFs integrating the
345 CABYV genome, showing that the less conserved is ORF 1 and the most conserved are
346 ORFs 4 and 6; comparing different polerovirus species, the least conserved are ORFs 1
347 and 6 and the most conserved are ORFs 3 and 4 (Table 4).

348 Recombination events were identified using the RDP3 program, as it implements
349 several recombination-detection methods. The sequences of CABYV-N and CABYV-C
350 (15,53) were also included in the analysis. A total of 29 potential recombination events
351 were detected by at least one of the methods, but only three events (Fig. 5A) were
352 accepted based on a triple criteria: detection by more than one method, consensus score
353 in RDP3 higher than 40 and $P < 001$. In CABYV-C, the region between nucleotides
354 4,469 and 5,669, which forms part of ORF 6, may have originated by recombination of
355 Sq/2004/1.9 (major parent) and an unknown minor parent, with the event detected by
356 methods RDP ($P = 9.627 \times 10^{-26}$) and BootScan ($P = 1.602 \times 10^{-32}$). In Sq/2005/9.2, the

region between nucleotides 1,296 and 1,617, which forms part of ORF 2, may have been the result of recombination of the major parent CABYV-N with an unknown minor parent, this event being detected by RDP ($P = 5.76 \times 10^{-3}$), BootScan ($P = 2.26 \times 10^{-2}$) and Chimaera ($P = 2.46 \times 10^{-4}$) methods. Also in Sq/2005/9.2, the region between nucleotides 4,785 and 5,669, which includes parts of ORF 6 and the 3´-non coding region, is closely related to Sq/2004/1.9, whereas for the rest of the genome, Sq/2005/9.2 is distantly related to Sq/2004/1.9; this event was detected by five methods: RDP ($P = 3.74 \times 10^{-19}$), GENCONV ($P = 9.85 \times 10^{-17}$), BootScan ($P = 4.18 \times 10^{-22}$), MaxChi ($P = 5.96 \times 10^{-13}$), Chimaera ($P = 1.09 \times 10^{-13}$) and SisCan ($P = 3.19 \times 10^{-16}$). For this latter case, assignment of parental and daughter sequences was illustrated by comparing the minimum evolution trees for the coding regions upstream and downstream of the putative recombination point (Fig. 5B). This recombination point is in close proximity to the putative transcription origin of the subgenomic RNA 2, a genomic location that has been proposed to be a recombination hotspot for poleroviruses (18). Interestingly, a conserved secondary RNA structure can be predicted in this region, with the recombinant isolate Sq/2005/9.2 having an insertion of three nucleotides in a loop of this structure (Fig. 5C), suggesting a potential role of this structure in the recombination mechanism.

Aphid population dynamics and identification of CABYV-viruliferous aphid species. *Aphis gosypii* was the most abundant aphid species caught on the green tile traps located in the open melon fields sampled both in 2010 and 2011, with 41.56 and 28.57% of the total cumulative aphid captures, respectively. *Aphis spiraecola* was the second most abundant aphid species that landed on melon fields with 12.18% of the total aphid catches in 2010 and 10.20% in 2011. Other abundant species that landed on melon were *Acyrthosiphon pisum* (7.47%) in 2010, *Aploneura lentisci* (10.20%) in 2011

382 and *Hyalopterus pruni*, found in the green tile traps both in 2010 and 2011 (2.92% and
383 12.24%, respectively). Other less abundant species found in the 2010 and 2011 surveys
384 were *A. fabae*, *Brachycaudus helichrysi*, *Macrosiphum euphorbiae* and *Uroleucon* sp.
385 Only a few individuals belonging to two aphid species contained CABYV when aphids
386 caught in the green tile traps in 2011 were analyzed by real-time RT-qPCR. The main
387 aphid species positive for CABYV was again *A. gossypii.* The percentage of potentially
388 CABYV-viruliferous *A. gossypii* was 69.23 % (9 positive aphids/13 analyzed aphids).
389 The only one individual of the species *M. euphorbiae* analyzed was CABYV-positive.
390 In addition to the sampling using green tile traps, the aphid species colonizing
391 melon plants were also estimated in surveys carried out during 2010. In agreement with
392 the above data, the most abundant species was *A. gossypii,* with its population peak
393 occurring around mid-June, coinciding with the middle of the melon-growing season
394 (data not shown).
395 **Alternative CABYV hosts and potential reservoirs.** Twenty-eight weed and six
396 alternative crop species, belonging to 20 different families, were sampled during 2009
397 to identify alternative hosts and potential CABYV reservoirs. In total, 239 samples were
398 tested for CABYV infection. Out of these, 19 were infected, and belonged to the species
399 *Abutilion theophrasti*, *Malva parviflora*, *Chenopodium murale*, *Ecballium elaterium*,
400 *Senecio vulgaris*, *Sinapis arvensis* and *Sonchus oleraceus* (Supplementary Table S2).
401 Interestingly, more than 40% of the *E. elaterium* (family *Cucurbitaceae*) samples were
402 infected by CABYV, and frequent yellowing symptoms were observed in *E. elaterium*
403 plants close to commercial cucurbit fields. Given the potential importance of plants of
404 this species as alternative and/or reservoir hosts, transmission experiments from *E.*
405 *elaterium* to melon plants were performed. *E. elaterium* plants were infected with
406 CABYV through agroinoculation (7); three out of eight plants were infected and

showed yellowing symptoms 12 days after inoculation. CABYV-infected *E. elaterium* plants were placed together with 32 healthy melon seedlings inside insect-proof boxes. Adult *A. gossypii* aphids were then released inside the boxes. Symptoms typically induced by CABYV appeared in 6 melon plants two weeks after the release of the aphids. CABYV infection of melon plants was confirmed by molecular hybridization, indicating that *A. gossypii* can transmit CABYV from *E. elaterium* to melon at least under our experimental conditions.

DISCUSSION

In this report, we analyzed the genetic structure of a CABYV population sampled within a limited geographical area and during a relatively short period of time. We found that in spite of these restrictions, the virus population appeared to be genetically quite diverse. In comparison with two cases of virus populations sampled in southeastern Spain using similar spatial and temporal scales, the values of total nucleotide diversity estimated for *Cucumber vein yellowing virus* (CVYV) and *Cucurbit yellow stunting disorder virus* (CYSDV) Spanish populations were approximately 20 times lower than those estimated for CABYV (21,30) alone. Indeed, CABYV nucleotide diversity values were of the same order of magnitude than those found, for instance, for *Watermelon mosaic virus* (WMV) populations sampled during a longer period of time and at a relatively much larger geographical scale (36). This is remarkable, because CABYV was detected for the first time in 1992 in France (26) and described in cucurbit crops of the Murcia region rather recently (22). Thus, the diversity observed in the CABYV population does not agree with a recent introduction in the area under study, such as in the case of CVYV mentioned above; it rather fits with an older

presence of CABYV and/or with multiple introduction events. Alternatively, CABYV might have been introduced recently in Murcia and then it may have evolved rapidly, a hypothesis that is congruent with the high substitution rates reported by Pagán & Holmes (38) for the *CP* of CABYV (> 1×10^{-2} substitution/site/year). Further studies, expanding the spatial and temporal scales of sampling, are required to clarify this issue.

As is evident after inspection of some branches of the phylogenetic trees deduced in this work, mutation accumulation has been a source of CABYV diversification. However, recombination events may have also played a role. Recombination events were not detected within ORF 3 or ORFs 4/5, in agreement with data deduced by Pagán & Holmes (38). Potential recombination events in the non-coding region between these two genes have been reported for many polero- and luteo-viruses, including CABYV (24,38,46). Indeed, the topological incongruence of ORF 3 and ORFs 4/5 trees for some isolates in this study suggests that this is a likely possibility. However, this hypothesis has to be taken with caution, because ORF 3 and ORFs 4/5 sequences come from independent PCRs from field samples, and mixed infections have been reported in samples of the same collection (23). To expand the genomic information available for CABYV, and to explore if recombination may have taken place in the CABYV population, sequencing of three complete genomes was carried out. Thus, recombination events were detected, although the putative recombination breakpoints did not correspond to the above-mentioned hotspot. Interestingly, the putative recombination point described by Gibbs & Cooper (13) at the 3´-terminus of the RT-domain of three poleroviruses (including CABYV) coincides with a trinucleotide insertion in the recombinant isolate Sq/2005/9.2, which could be the result of a recombination event that occurred at this point. This hypothesis is strongly supported by the close presence of the sgRNA2 initiation site and the detection of a

457 conserved secondary RNA structure in this region, two features which have been
458 described in many studies associated with the occurrence of RNA recombination
459 (18,25).

460 Regardless of how genetic diversity was generated, it is evident that purifying
461 selection influenced the genetic structure of the CABYV population, though with
462 different intensities for different genomic regions. ORF 3 showed the lowest ω value
463 indicating that strong purifying selection is acting on it, and fits with the identification
464 of RdRp as a functionally important protein of RNA viruses, it being the least prone to
465 fix changes (52). The highest ω value was observed for ORF 5, which codes for the MP,
466 in agreement with results obtained for the *MP* genes of two other luteoviruses, BYDV
467 and *Cereal yellow dwarf virus* (CYDV) (17,38). Here, it is important to remark on the
468 overlapping nature of the *CP* and *MP* genes, where any point mutation would have an
469 impact in both genes simultaneously, possibly constraining the functional evolution of
470 such genes (55). ORF 3 had several codons under negative selection also, in agreement
471 with the presence of conserved motifs in viral RdRps (15,33). Negative selection was
472 also detected acting on the arginine-rich N-terminal domain (R domain) and the shell
473 central domain (S domain) (44) of the CP (ORF 4). Indeed, most residues under
474 purifying selection were identified in close proximity to epitopes 5 and 10 of the S
475 domain (6,49), in analogy with results obtained by Torres *et al*. (50). Regions located on
476 the surface of the CP protein are presumably critical for virus assembly and stability,
477 plant systemic infection and virus transmission (27), by means of interactions with viral
478 and non-viral factors (4,9). In contrast, we only detected one positively selected amino
479 acid mutation, located in the *MP* gene, for which no specific function has been
480 identified yet for CABYV. Further studies for the identification of the biological

481 functions associated with this particular change could help us understand the process of
482 selection on the *MP* gene.

483 Our study suggests that the CABYV population is genetically structured; this
484 could be explained by constraints intrinsic to the viral genome (37). However,
485 ecological factors that influence long- and short-range transmission, persistence in the
486 field, etc. could play an important role in the structure of the plant virus population
487 (reviewed in 12). Here, we analyzed the potential association of genetic structure with
488 host (melon or squash), location of virus isolation and sampling season, but we did not
489 identify any specific relationship. This is in contrast with the identification of CABYV
490 genetic groups associated to broad geographical regions (i.e., Mediterranean and Asian
491 clades in tree of Supplementary Fig. 1). Therefore, barriers to CABYV gene flux do not
492 appear to exist in the area where the CABYV isolates come from, although they seem to
493 play a role in CABYV diversity at a global scale. This raises the question of what
494 factors are responsible for the observed genetic structure of our local CABYV
495 population. To try to answer this question we studied two additional aspects of CABYV
496 epidemiology in the surveyed region; alternative hosts and vector transmission. The
497 alternative host with the highest potential impact on CABYV epidemiology is *E.*
498 *elaterium*. As shown in this study, *E. elaterium* plants grown in proximity of melon and
499 squash crops are frequently infected with CABYV. *E. elaterium* persists from crop
500 season to crop season (M. Juárez, *unpublished data*), thus could be a virus reservoir of
501 significant importance (48). Interestingly, diversity among *E. elaterium* populations in
502 the Iberian Peninsula has been reported (40). On the other hand, the main aphid species
503 caught on the green tile traps and found colonizing melon crops in the Murcia region
504 was *A. gossypii*. This aphid species is known to be an important vector of CABYV,
505 requiring prolonged feeding times and direct contact with vascular tissues, spreading

506 CABYV mainly by colonizing vectors (11). By means of real-time quantitative RT-
507 PCR it was possible to demonstrate that *A. gosypii* was not only the most abundant
508 aphid species in the fields but also the aphid species with the highest number of aphids
509 carrying CABYV. Other CABYV-viruliferous aphid species found in melon crops in
510 Murcia during this study was *M. euphorbiae*. Although *M. euphorbiae* has been
511 previously described as a potential CABYV vector (7), the low density of this non-
512 colonizing aphid species suggests a very limited role on CABYV spread in this region.
513 Taken together, this information indicates that *A.gossypii* is the main aphid species
514 involved in the spread of CABYV in melon crops in the Murcia region. Significantly,
515 morphological differences were observed between individuals of different *A. gossypii*
516 colonies in Murcian melon crops (M. Juárez and M.A. Aranda, *unpublished data*);
517 therefore, it is tempting to speculate that genetic differences among aphids that colonize
518 CABYV hosts may exist and these might be associated with CABYV transmission
519 specificities (8,29) and, thus, with CABYV diversity.

520 To summarize, our results support the following conclusions: First, the Spanish
521 CABYV population is diverse and recombination might have had an important role in
522 diversity generation and maintenance. The main evolutionary force identified was
523 purifying selection. Second, the population appears to be genetically structured, with at
524 least two genetic groups that coexist in the field. This structure did not seem to be
525 associated to crop species, date or locality of sampling and, therefore, other factors must
526 be responsible for this observation. Among ecological factors that could potentially
527 condition the genetic structure of CABYV populations, are vector transmission and
528 alternative host adaptation. In this regard, it would be particularly interesting to study
529 the potential adaptation of CABYV to *A. gossypii*, its main vector in the surveyed area,
530 and to *E. elaterium*, its main alternative host in the same area.

ACKNOWLEDGEMENTS

This work was supported by grants 11934/PI/09 (Fundación Séneca de la Región de Murcia, Spain) and AGL2009-07552/AGR (Ministerio de Ciencia e Innovación, Spain). M. Kassem was recipient of a fellowship from the Agencia Española de Cooperación Internacional-Ministerio de Asuntos Exteriores (AECI-MAE). We thank Verónica Truniger for critically reading the manuscript, and Mari Carmen Montesinos and Blanca Gosalvez for their excellent technical assistance. M. Fon (mariogfon@gmail.com) checked the English.

REFERENCES

1. Ashoub, A., Rohde, W., and Prüfer, D. 1998. In plant transcription of a second subgenomic RNA increases the complexity of the subgroup 2 luteovirus genome. Nucleic Acids Res. 26:420-426.
2. Bananej, K., Vahdat, A., Predajna, L., and Glasa, M. 2009. Molecular characterization of geographically different *Cucurbit aphid-borne yellows virus* isolates. Acta Virol. 53:61-64.
3. Bencharki, B., Mutterer, J., El Yamani, M., Ziegler-Graff, V., Zaoui, D., and Jonard, G. 1999. Severity of infection of Moroccan barley yellow dwarf virus PAV isolates correlates with variability in their coat protein sequences. Ann. Appl. Biol. 134:89-99.
4. Bencharki B., Boissinot S., Revollon S., Ziegler-Graff V., Erdinger M., Wiss L., and Dinant S. Renard D., Beuve M., Lemaitre-Guillier C., and Brault V. 2010. Phloem Protein Partners of Cucurbit aphid borne yellows virus: Possible Involvement of Phloem Proteins in Virus Transmission by Aphids. Mol. Plant Microbe-Interact. 23:799–810.
5. Brault, V., van den Heuvel, J. F. J. M., Verbeek, M., Ziegler-Graff, V., Reutenauer, A., Herrbach, E., Garaud, J.C., Guilley, H., Richards, K., and Jonard, G. 1995. Aphid transmission of beet western yellows luteovirus requires the minor capsid read-through protein P74. EMBO J. 14:650-659.
6. Brault, V., Bergdoll, M., Mutterer, J., Prasad, V., Pfeffer, S., Erdinger, M., Richards, K. E., and Ziegler-Graff, V. 2003. Effects of Point Mutations in the Major Capsid Protein of Beet Western Yellows Virus on Capsid Formation, Virus Accumulation, and Aphid Transmission. J. Virol. 77:3247–3256.
7. Brault, V., Périgon, S., Reinbold, C., Erdinger, M., Scheidecker, D. Herrbach,

E., Richards, K., and Zielgler-Graff, V. 2005. The Polerovirus minor capsid protein determines vector specificity and intestinal tropism in the aphid. J. Virol. 79:9685-9693.

8. Burrows, M. E., Caillaud, M. C., Smith, D. M., Benson, E. C., Gildow, F. E., and Gray, S. M. 2006. The genetic regulation of polerovirus and luteovirus transmission in the aphid *Schizaphis graminum*. Phytopathology 96:828-837.
9. Chare E. R., and Holmes E. C. 2004. Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. J. Gen. Virol. 85:3149–3157.
10. D'Arcy, C. J., and Domier L. L. 2005. Luteoviridae. In Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses. Edited by M.A. Mayo, J. Maniloff, U. Desselberger, L.A. Ball and Claude M. Fauquet. Academic Press. New York., NY.
11. Fereres, A., and Moreno, A. 2009. Behavioural aspects influencing plant virus transmission by homopteran insects. Virus Res. 141:158-168.
12. García-Arenal, F., Fraile, A., and Malpica, J. M. 2001. Variability and genetic structure of plant virus population. Annu. Rev. Phytopathol. 39:157-186.
13. Gibbs, M., and Cooper, J. 1995. A recombinational event in the history of luteoviruses probably induced by base-pairing between the genomes of two distinct viruses. Virology 206:1129-1132.
14. Gildow, F. E., Reavy, B., Mayo, M. A., Duncan, G. H., Woodford, J. A. T., Lamb, J. W. and Hay, R. T. 2000. Aphid acquisition and cellular transport of potato leafroll virus-like particles lacking P5 readthrough protein. Phytopathology 90;1153-1161..

590 15. Guilley, H., Wipf-Scheibel, C., Richards, K., Lecoq, H., and Jonard, G. 1994.
591 Nucleotide sequence of cucurbit aphid-borne yellows luteovirus. Virology
592 202:1012-1017.
593 16. Guyader, S., and Ducray, D. G. 2002. Sequence analysis of Potato leafroll virus
594 isolates reveals genetic stability, major evolutionary events and differential
595 selection pressure between overlapping reading frame products. J. Gen. Virol.
596 83:1799-1807.
597 17. Hall, G. 2006. Selective constraint and genetic differentiation in geographically
598 distant barley yellow dwarf virus populations. J. Gen. Virol. 87:3067-3075.
599 18. Hauser, S., Stevens, M., Beuve, M., Lemaire, O. 2002. Biological properties and
600 molecular characterization of Beet chlorosis virus (BChV). Archives of
601 Virology 147:745-762.
602 19. Irwin, M. E. 1980. Sampling aphids in soybean fields. In Sampling Methods in
603 Soybean Entomology. Edited by M. Kogan and D. C. Herzog, Springer-Verlag.
604 New York., NY
605 20. Irwin, M. E., and Ruesink, W. G. 1986. Vector intensity: A product of
606 propensity and activity. In Plant virus epidemics: Monitoring, modelling and
607 predictions outbreaks. 1st edition ed, vol. Chapter 2. Edited by G. D., McLean,
608 R. G. Garrett and W. G. Ruesink, Sydney, Australia
609 21. Janssen, D., Velasco, L., Martín, G., Segundo, E., and Cuadrado, I. M. 2007.
610 Low genetic diversity among *Cucumber vein yellowing virus* isolates from
611 Spain. Virus Genes 34:367-371.
612 22. Juárez, M., Truniger, V., and Aranda, M. A. 2004. First report of *Cucurbit*
613 *aphid-borne yellows virus* in Spain. Plant Dis. 88:907.

614 23. Kassem, M. A., Sempere, R. N., Juárez, M., Aranda, M. A., and Truniger, V.
615 2007. Cucurbit aphid-borne yellows virus is prevalent in field-grown cucurbit
616 crops of southeastern Spain. Plant Dis. 91:232-238.
617 24. Knierim, D., Deng, T. C., Tsai, W. S., Green, S. K., and Kenyon, L. 2010.
618 Molecular identification of three distinct Polerovirus species and a recombinant
619 Cucurbit aphid-borne yellows virus strain infecting cucurbit crops in Taiwan.
620 Plant Pathol. 59:991-1002.
621 25. Lai, M. M. 1992. RNA recombination in animal and plant viruses. Microbiol.
622 Rev. 56:61-79.
623 26. Lecoq, H., Bourdin, D., Wipf-Scheibel, C., Bon, M., Lot, H., Lemaire, O.,
624 Herbach, E. 1992. A new yellowing disease of cucurbits caused by a luteovirus,
625 cucurbit aphid-borne yellows virus. Plant Pathol. 41:749-761.
626 27. Lee, L., Kaplan, I. B., Ripoll, D. R., Liang, D., Palukaitis, P., and Gray S.M.
627 2005. A Surface Loop of the *Potato Leafroll Virus* Coat Protein Is Involved in
628 Virion Assembly, Systemic Movement, and Aphid Transmission. J. Virol.
629 79:1207-1214.
630 28. Lemaire, O., Gubler, W. D., Valencia, J., Lecoq, H., and Falk, B. W. 1993. First
631 report of cucurbit aphid-borne yellows luteovirus in the Unites States. Plant Dis.
632 77:1169.
633 29. Lucio-Zavaleta, E., Smith, D. M., and Gray, S. M. 2001. Variation in
634 transmission efficiency among Barley yellow dwarf virus-RMV isolates and
635 clones of the normally inefficient aphid vector, Rhopalosiphum padi.
636 Phytopathology 91:792-796.
637 30. Marco, C. F., and Aranda M. A. 2005. Genetic diversity of a natural population
638 of *Cucurbit yellow stunting disorder virus*. J. Gen. Virol. 86:815-822.

639 31. Mastari, J., and Lapierre, H. 1999. Sequences of the 3' halves of the genome of
640 barley yellow dwarf virus-PAV CpA isolates that vary in symptom severity. Eur.
641 J. Plant Pathol. 105:801-811.
642 32. Mayo M. A., Robinson D. J., Jolly C. A., and Hyman L. J. 1989. Nucleotide
643 Sequence of Potato Leafroll Luteovirus RNA. Gen. Virol. 70:1037-1051.
644 33. Mayo, M. A., and Ziegler-Graff, V. 1996. Molecular biology of luteoviruses.
645 Adv. Virus Res. 46:413-460.
646 34. Mnari-Hattab, M., Gauthier, N., and Zouba, A. 2009. Biological and molecular
647 characterization of the Cucurbit aphid-borne yellows virus affecting cucurbits in
648 Tunisia. Plant Dis. 93:1065-1072.
649 35. Moreno, A., Garzo, E., Fernandez-Mata G., Kassem, M., Aranda, M. A., and
650 Fereres, A. 2011. Aphids secrete watery saliva into plant tissues from the onset
651 of stylet penetration. Entomol. Exp. Appl. 139:145-153.
652 36. Moreno, I. M., Malpica, J. M., Díaz-Pendón, J. A., Moriones, E., Fraile, A., and
653 García-Arenal, F. 2004. Variability and genetic structure of the population of
654 watermelon mosaic virus infecting melon in Spain. Virology 318:451- 460.
655 37. Moury B., and Simon V. 2011. dN/dS-Based Methods Detect Positive Selection
656 Linked to Trade-Offs between Different Fitness Traits in the Coat Protein of
657 Potato virus Y. Mol. Biol. Evol. 28:2707–2717.
658 38. Pagán, I., and Holmes, E. C. 2010. Long-term evolution of the Luteoviridae:
659 time-scale and mode of virus speciation. J. Virol. 84:6177–6187.
660 39. Pazhouhandeh, M., Dieterle, M., Marrocco, K., Lechner, E., Berry, B., Brault,
661 V., Hemmer, O., Kretsch, T., Richards, K. E., Genschik P., and Ziegler Graff V.
662 2006. F-boxlike domain in the Polerovirus protein P0 is required for silencing
663 suppressor function. Proc. Natl. Acad. Sci. U S A 103:1994-1999.

664 40. Perez Chiscano, J. L. 1985. Distribución geográfica de *Ecballium elaterium* (L.)
665 Richard (Cucurbitaceae) en la Península Ibérica e Islas Baleares. Stvdia
666 Botanica 4: 57
667 41. Perez, P., Collar, J. L., Avilla, C., Duque, M., and Fereres, A. (1995). Estimation
668 of vector propensity of potato virus Y (PVY) in open-field pepper crops of
669 central Spain. J. Econ. Entomol. 88:986-991.
670 42. Pfeffer, S., Dunoyer, P., Heim, F., Richards, K. E., Jonard, G., and Ziegler-
671 Graff, V. 2002. P0 of *Beet western yellow virus* is a suppressor of
672 posttranscriptional gene silencing. J. Virol. 76:6815-6824.
673 43. Prüfer, D., Kawchuk, L. M., and Rohde, W. 2006. Polerovirus ORF0 genes
674 induce a host-specific response resembling viral infection. Can. J. Plant Pathol.
675 28:302-309.
676 44. Rao, A. L. 2006. Genome Packaging by Spherical Plant RNA Viruses. Annu.
677 Rev. Phytopathol. 44:61-87.
678 45. Sadowy, E., Juszczuk, M., David, C., Gronenborn, B., and Hulanicka, M. D.,
679 2001. Mutational analysis of the proteinase functions of *Potato leafroll virus*. J.
680 Gen. Virol. 82:1517- 1527.
681 46. Shang, Q., Xiang, H., Han, C., Li, D., and Yua, J. 2009. Distribution and
682 molecular diversity of three cucurbit-infecting poleroviruses in China. Virus
683 Res. 145:341-346.
684 47. Silva, T. F., Corrêa, R. L., Castilho, Y., Silvie, P., Bélot J-L., and Vaslin M. F.
685 S. 2008. Widespread distribution and a new recombinant species of Brazilian
686 virus associated with cotton blue disease. Virol. J. 5:123-135.

687 48. Srinivasan, R., and Alvarez, J. M. 2008. Hairy nightshade as a potential Potato
688 leafroll virus (Luteoviridae: Polerovirus) inoculum source in Pacific Northwest
689 potato ecosystems. Phytopathology 98:985-991.
690 49. Torrance, L. 1992. Analysis of epitopes on potato leafroll virus capsid protein.
691 Virology 191:485-489.
692 50. Torres, M. W., Corrêa, R. L., and Schrago, C. G. 2005. Analysis of differential
693 selective forces acting on the coat protein (P3) of the plant virus family
694 Luteoviridae. Genet. Mol. Res. 4:790-802.
695 51. van der Wilk, F., Verbeek, M., Dullemans, A. M., and van den Heuvel J. F. J.
696 M. 1997. The genome-linked protein of potato leafroll virus is located
697 downstream of the putative protease domain of the ORF1 product. Virology
698 234:300-303.
699 52. Wu, B., Blanchard-Letort, A., Liu, Y., Zhou, G., Wang, X., and Elena, S. F.
700 2011. Dynamics of Molecular Evolution and Phylogeography of *Barley yellow*
701 *dwarf virus*- PAV. PLoS ONE 6 (2): e16896
702 53. Xiang H. Y., Shang Q. X., Han C. G., Li D. W., and Yu J. L. 2008. Complete
703 sequence analysis reveals two distinct poleroviruses infecting cucurbits in
704 China. Arch. Virol. 153:1155-1160.
705 54. Yardimci, N., and Özgönen, H. 2007. First report of *Cucurbit aphid-borne*
706 *yellows virus* in Turkey. Australasian Plant Disease Notes 2:59.
707 55. Zhang, D., Chen, J., Deng, L., Mao, Q., Zheng, J., Wu, J., Zeng, C., and Li, Y.
708 2010. Evolutionary selection associated with the multi-function of overlapping
709 genes in the hepatitis B virus. Infect. Genet. Evol 10:84-88.
710 56. Ziegler-Graff, V., Brault, V., Mutterer, J. D., Simionis, M. T., Herrbach, E.,
711 Guilley, H., Richards, K. E., and Jonard, G. 1996. The coat protein of Beet

712 Western Yellows Luteovirus is essential for systemic infection but the viral gene
713 products P29 and P19 are dispensable for systemic infection and aphid
714 transmission. Mol. Plant-Microbe Interact. 9:501-510.

715
716

TABLES

TABLE 1. CABYV isolates analyzed in this study, grouped by host and year of collection

	Year			
Host	2003	2004	2005	Total
Melon	12	11	11	34
Squash	7	7	6	20
Total	19	18	17	54

TABLE 2. Nucleotide diversities[a] for different regions of the CABYV genome

	π	d_N	d_S	d_N/d_S
ORF 3	0.0232	0.00288 ± 0.00110	0.04591 ± 0.00918	0.063
ORF 4	0.0152	0.00933 ± 0.00273	0.02233 ± 0.00595	0.417
ORF 5	0.0155	0.01105 ± 0.00274	0.02097 ± 0.00665	0.526

[a] Nucleotide diversity is defined here as the mean number of nucleotide substitutions per site (π). Nucleotide diversities were computed separately for synonymous (*dS*) and non-synonymous (*dNS*) positions using the PBL method. Standard error of means was calculated by using the bootstrap method implemented in the Mega4 program.

748

749

750

751 **TABLE 3.** Within- and between-subpopulation nucleotide diversities[a] in the CABYV
752 ORFs 3, 4 and 5

ORF3		Squash	Melon
Squash	d_{NS}	0.0037 ± 0.0015	0.0031 ± 0.0013
	d_S	0.0648 ± 0.0118	0.0499 ± 0.0095
	d_{NS}/d_S	0.0571	0.0621
Melon	d_{NS}		0.0023 ± 0.0009
	d_S		0.0340 ± 0.0075
	d_{NS}/d_S		0.0676

753

ORF4		Squash	Melon
Squash	d_{NS}	0.0119 ± 0.0034	0.0102 ± 0.0032
	d_S	0.0273 ± 0.0074	0.0234 ± 0.0066
	d_{NS}/d_S	0.4359	0.4359
Melon	d_{NS}		0.0072 ± 0.0023
	d_S		0.0190 ± 0.0052
	d_{NS}/d_S		0.3789

754

ORF4		Squash	Melon
Squash	d_{NS}	0.0136 ± 0.0034	0.0102 ± 0.0032
	d_S	0.0270 ± 0.0081	0.0234 ± 0.0066
	d_{NS}/d_S	0.5037	0.4373
Melon	d_{NS}		0.0096 ± 0.0023
	d_S		0.0160 ± 0.0056
	d_{NS}/d_S		0.6000

755

756 [a] Nucleotide diversity was computed separately for nonsynonymous (*dN*) and for
757 synonymous (*dS*) sites by using the PBL method. Between subpopulations diversity
758 values correspond to net nucleotide substitutions. Sequences were grouped according to
759 the host (squash or melon) from which CABYV was originally isolated.
760

761

762
763
764
765

766 **TABLE 4.** Nucleotide and amino acid sequence comparisons of Sq/2003/7.2 ORFs
767 *versus* the corresponding ORFs of other CABYV isolates and other poleroviruses

Virus	Nucleotide Identity [a]	Amino acid identity (similarity) [a]					
		ORF 1	ORF 2	ORF 3	ORF 4	ORF 5	ORF 6
Sq/2004/1.9	98	98 (98)	98 (98)	99 (99)	97 (97)	98 (98)	98 (99)
Sq/2005/9.2	91	77 (86)	82 (90)	92 (96)	95 (99)	90 (95)	97 (98)
CABYV-N	95	97 (97)	96 (98)	96 (97)	98 (99)	94 (97)	95 (97)
CABYV-C	88	80 (89)	86 (93)	91 (95)	94 (95)	91 (93)	90 (95)

768

Virus[b]	Nucleotide identity [a]	Amino acid identity (similarity) [a]					
		ORF1	ORF2	ORF3	ORF4	ORF5	ORF6
MABYV	69	69 (82)	61 (74)	74 (82)	81 (89)	66 (76)	51 (63)
BWYV	60	48 (67)	55 (69)	72 (82)	63 (72)	48 (56)	33 (49)
BMYV	56	41 (61)	45 (60)	65 (77)	62 (72)	47 (55)	33 (49)
TuYV	52	23 (44)	30 (47)	57 (68)	64 (73)	47 (56)	34 (49)
BChV	51	18 (33)	26 (43)	57 (69)	64 (73)	46 (55)	32 (49)
CpCSV	48	18 (30)	25 (42)	55 (67)	67 (75)	56 (65)	39 (52)
TVDV	45	31 (47)	27 (45)	56 (66)	57 (69)	37 (48)	17 (30)
WYDV	45	16 (35)	26 (43)	54 (67)	60 (71)	43 (53)	18 (30)

769
770 [a] The percentages of identity (similarity) between aligned sequences were calculated as the number of
771 residues identical (similar) shared by compared sequences multiplied by 100 and divided by the number
772 of residues of the shorter sequence excluding insertions or deletions.
773 b*Melon aphid-borne yellows virus* (MABYV; GenBank accession no. NC_010809), *Beet western yellows*
774 *virus* (BWYV; NC_004756), *Wheat yellow dwarf virus-GPV* (WYDV;-GVP NC_012931), *Tobacco vein*
775 *distorting virus* (TVDV; NC010732), *Chickpea chlorotic stunt virus* (CpCSV; NC_008249)*,Turnip*
776 *yellows virus* (TuYV; NC_003743), *Beet mild yellowing virus* (BMYV; NC_003491), *Beet chlorosis*
777 *virus* (BChV; NC_002766).
778

779 **FIGURE LEGENDS**

780

781 **FIGURE 1.** Sampling area and regions sequenced in the CABYV genome. (A) Map of
782 the sampling area were CABYV isolates were sampled. (B) Schematic representation of
783 the CABYV genome; the two regions sequenced are indicated by R3, a fragment of
784 ORF 3 located between nucleotides 2,316 and 2,979, and R4/5, corresponding to the
785 ORF 4 and its overlapping part of ORF5, located between nucleotides 3,507 and 4,104.
786 **FIGURE 2.** Phylogenetic relationships of 50 Spanish CABYV isolates inferred from
787 sequences from ORF 3 (A) and ORF 4/5 (B). Minimum evolution phylogenetic trees
788 were inferred using the Mega 4 program. Bootstrap values (1,000 pseudoreplicates)
789 above 50% are shown. The sequence of a French isolate was introduced as an outgroup
790 and appears underlined in each tree. Isolate Sq/2005/2.2 (marked with an arrow)
791 belongs to group Ia for the ORF 3 but to group IIb for the ORFs 4/5. Isolates fully
792 sequenced are marked with a diamond.
793 **FIGURE 3.** Phylogenetic tree constructed based on sequences of the ORF 4/5 of 107
794 CABYV isolates from all over the world, including the 50 Spanish isolates
795 characterized here (gray shading) and using the Minimum Evolution method. Bootstrap
796 values (1,000 pseudoreplicates) above 50% are shown. Ch refers to China, Sp to Spain,
797 Fr to France, Sl to Slovakia, It to Italy, Tu to Tunisia, Ir to Iran and Ta to Taiwan.
798 **FIGURE 4**. Genetic diversity of CABYV isolates in synonymous (d_S) and
799 nonsynonymous (d_N) positions at each codon of the ORFs 3 (A), 4 (B) and 5 (C). The
800 differences between the estimated diversity values (d_N - d_S) are represented in the
801 vertical axis and the codon positions are represented in the horizontal axis. Bold black
802 lines marked with arrows represent sites under selection detected by the three methods
803 implemented in HyPhy: SLAC, FEL and REL.

804 **FIGURE 5.** Recombination in CABYV genomes. (A) Recombination hypotheses
805 generated by more than one algorithm of the RDP3 program with a consensus score
806 above 40 and a *P*-value < 1%. On top there is a schematic representation of the CABYV
807 genome with the ORFs indicated as dark gray boxes. Long dashed boxes represent
808 CABYV genomes (isolate code above the box); internal pale light gray segments
809 indicate recombinant regions and short black boxes represent the minor parents. The
810 codes of the major (first) and minor (second) parents are indicated. For example,
811 recombinant isolate CABYV-C (recombinant region indicated by the internal gray
812 segment) has Sq/2004/1.9 as a major parent and an unknown minor parent (short black
813 box). (B) Minimum evolution trees for the coding regions upstream and downstream of
814 the putative recombination point in ORF 6 of Sq/2005/9.2. Isolates with incongruent
815 phylogenetic relationships are marked with arrows. (C) RNA secondary structure of the
816 region of the putative transcription origin of the subgenomic RNA2 predicted by the
817 RNAalifold program. The putative recombination breakpoint is marked with an arrow
818 and coincides with a trinucleotide insertion in the Sq/2005/9.2 genome.
819
820
821 **SUPLEMENTARY MATERIAL**
822
823 **SUPPLEMENTARY FIGURE S1.** Alignment of CABYV amino acid sequences
824 around the codon under positive selection in ORF 5. Shading represents year of
825 sampling. A rectangular box represents the amino acid putatively under positive
826 selection.
827 **SUPPLEMENTARY TABLE S2.** Weed and crop species surveyed during 2009 to
828 identify alternative hosts and potential CABYV reservoirs. Sampled plants grew in the

proximity of melon and squash fields in the geographical area were CABYV surveys were carried out (see Fig. 1).

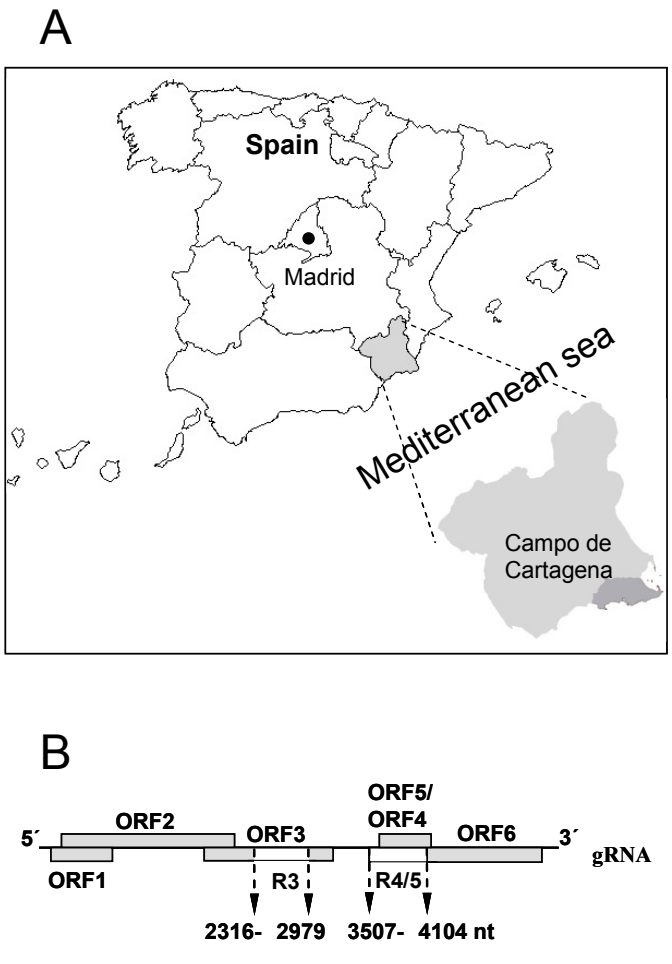


Figure 1, Kassem, Phytopathology

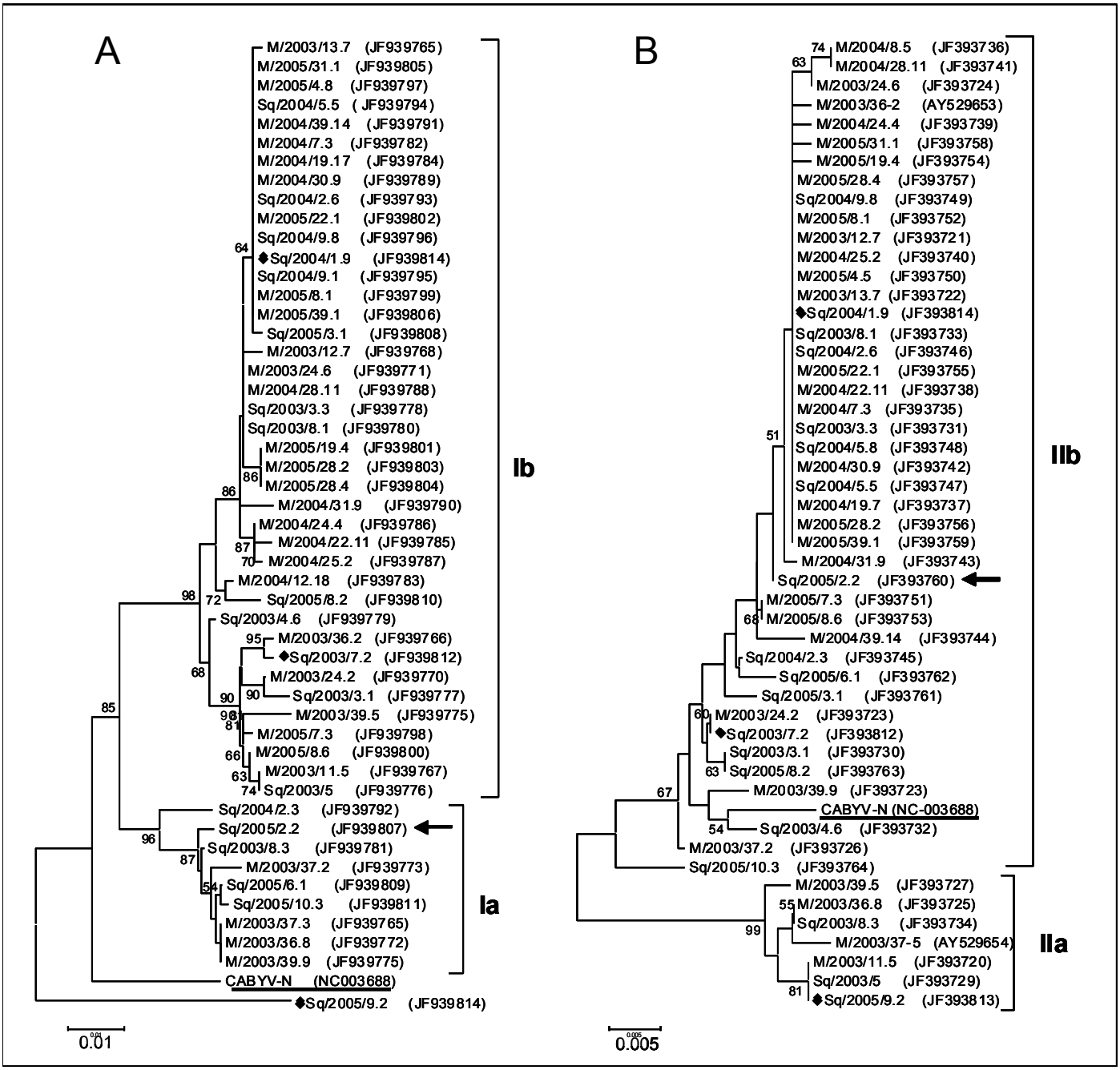


Figure 2, Kassem, Phytopathology

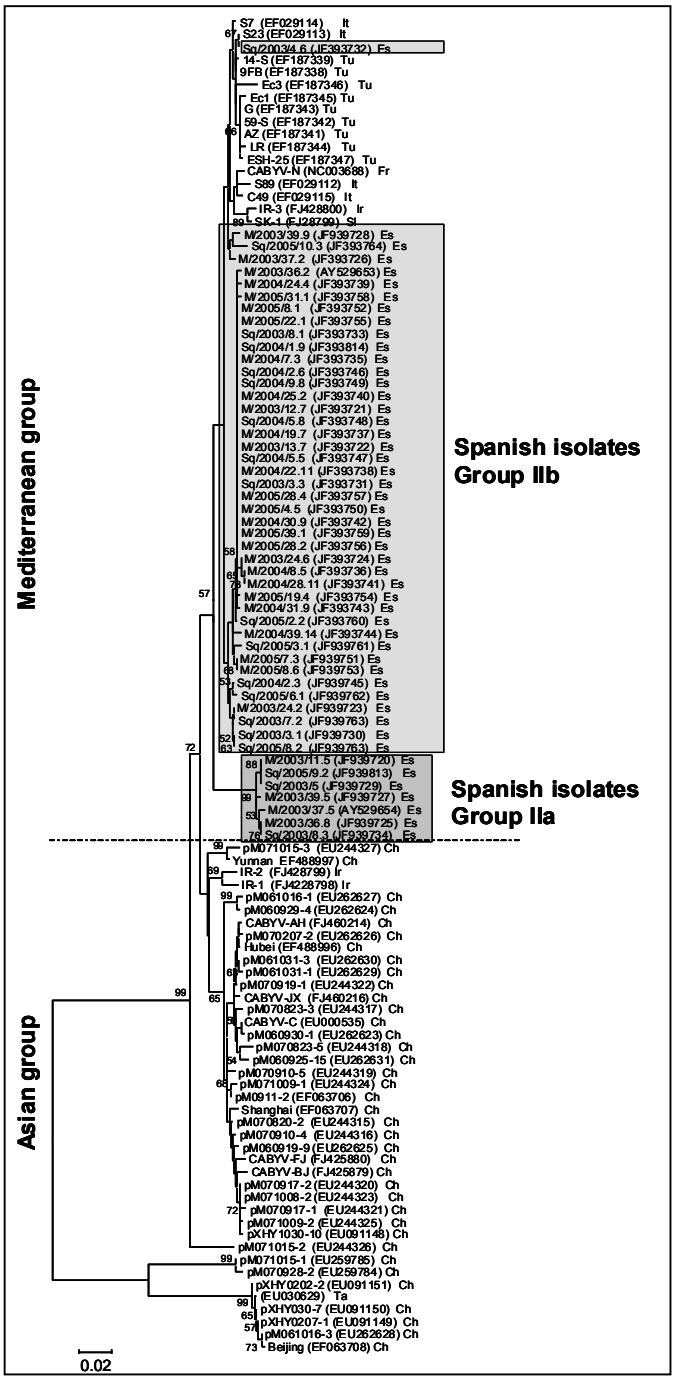


Figure 3, Kassem, Phytopathology

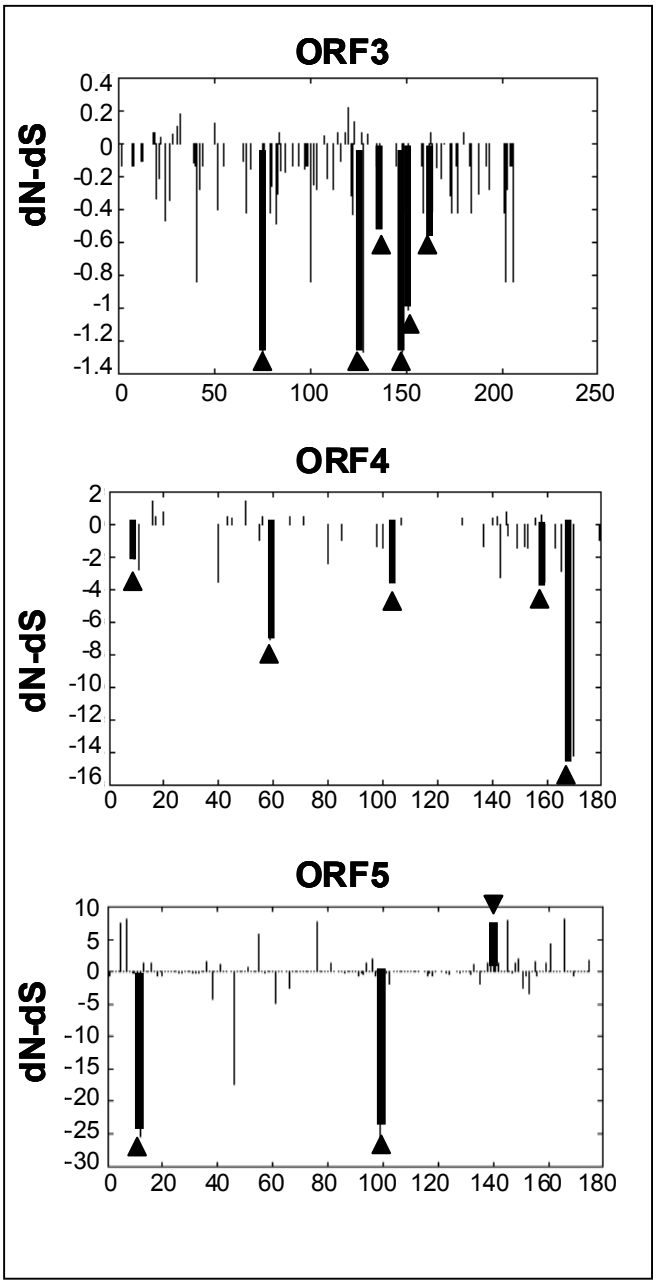


Figure 4, kassem, Phytopathology

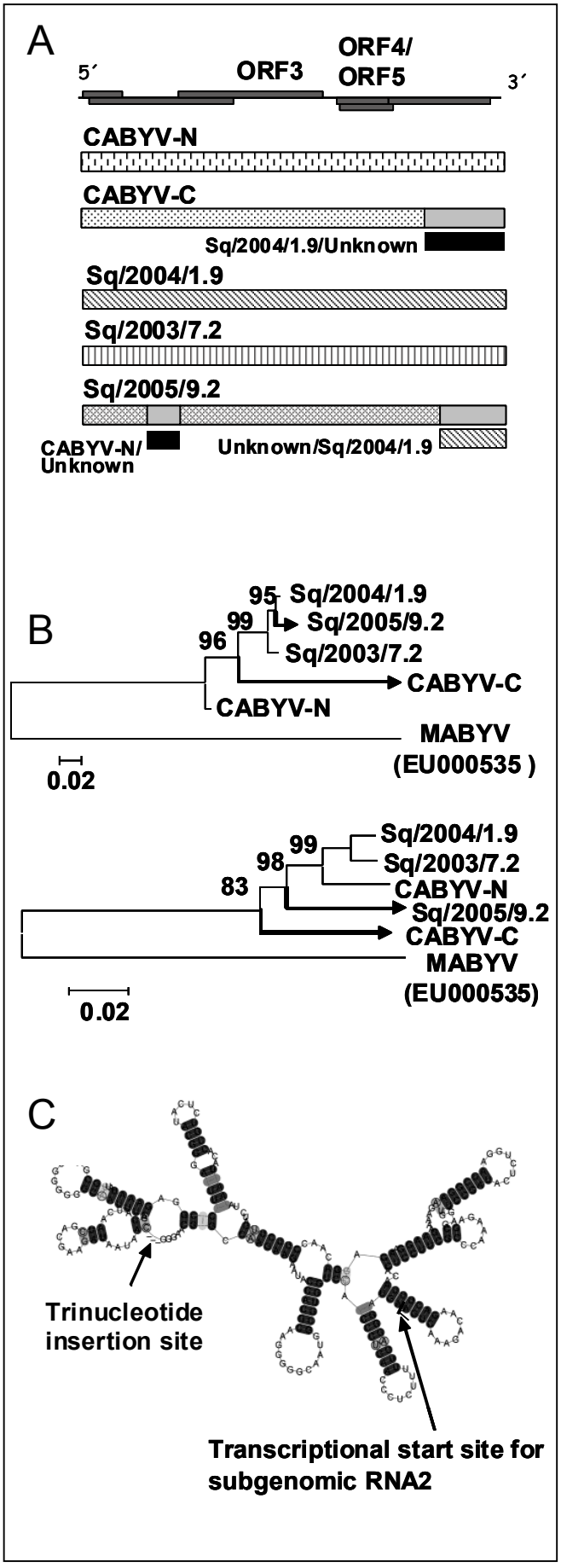
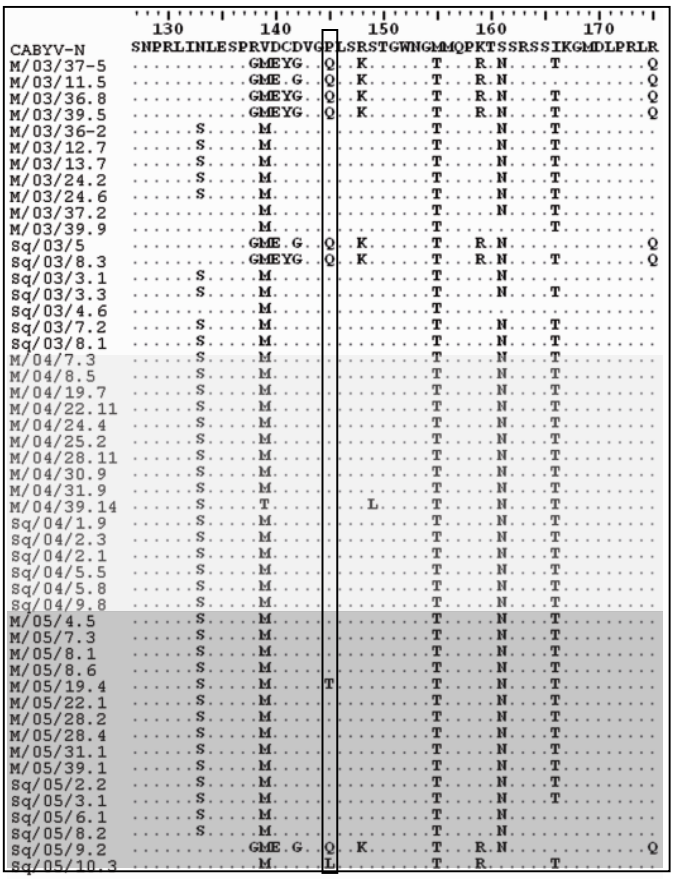


Figure 5, kassem, Phytopathology



Supplementary figure S1, kassem, Phytopathology

Supplementary table S2: Weed and crop species surveyed during 2009 to identify alternative hosts and potential CABYV reservoirs. Sampled plants grew in the proximity of melon and squash fields, in the geographical area were CABYV surveys were carried out (see Fig. 1).

Weeds		
Species	**Family**	**Infected/tested samples**
Abutilon theophrasti Medicus	*Malvaceae*	2/10
Lavatera cretica L.	*Malvaceae*	0/4
Malva neglecta Wallr.	*Malvaceae*	0/3
Malva parviflora L.	*Malvaceae*	1/8
Amaranthus blitoides S. Wats.	*Amaranthaceae*	0/6
*Amaranthus retroflexu*s L.	*Amaranthaceae*	0/8
Cardaria draba (L.)Desv.	*Cruciferae (Brassicaceae)*	0/7
Chenopodium album L.	*Chenopodiaceae*	0/4
Chrysanthemum coronarium L.	*Compositae (Asteraceae)*	0/5
Chenopodium murale L.	*Chenopodiaceae*	5/8
*Convolvulus arvensi*s L	*Convolvulaceae*	0/6
Ecballium elaterium (L.) A. Rich.	*Cucurbitaceae*	7/17
Echium plantagineum L.	*Boraginaceae*	0/8
Erodium chium (L.) Willd.	*Geraniaceae*	0/7
Galium aparine L. Subesp. *aparine*	*Rubiaceae*	0/9
Inula viscosa (L.) Ait.	*Compositae (Asteraceae)*	0/6
Lamium amplexicaule L.	*Labiatae (Lamiaceae)*	0/5
Moricandia arvensis (L.) DC.	*Cruciferae (Brassicaceae)*	0/7
Portulaca oleracea L.	*Portulacaeae*	0/9
Senecio vulgaris L.	*Compositae (Asteraceae)*	1/7
Sinapis arvensis L.	*Cruciferae (Brassicaceae)*	2/10
Datura stramonium L.	*Solanaceae*	0/4
Solanum nigrum L.	*Solanaceae*	0/8
Sonchus oleraceus L.	*Compositae (Asteraceae)*	1/6
Conyza bonariensis (L.) Cronq.	*Compositae (Asteraceae)*	0/6
Cyperus rotundus L.	*Cyperaceae*	0/5
Urtica urens L.	*Urticaceae*	0/6
Setaria viridis(L.) P. Beauv.	*Poaceae*	0/8
Crops		
Species	**Family**	**Infected/tested samples**
Beta vulgaris L. *var. cruenta* Alef.	*Amaranthaceae*	0/6
Solanun tuberosum L.	*Solanaceae*	0/10

Lactuca sativa L.	*Asteraceae*	0/7
Vicia faba L.	*Fabaceae*	0/8
Apium graveolens L.	*Umbelliferae (Apiaceae)*	0/6
Phaseolus vulgare L.	*Leguminosae*	0/5