Sixty years from the first disease description, a novel badnavirus associated with chestnut mosaic disease

Armelle Marais, Sergio Murolo, Chantal Faure, Yoann Brans, Clement Larue, François Maclot, Sébastien Massart, Michela Chiumenti, Angelantonio Minafra, Gianfranco Romanazzi, et al.

Sixty years from the first disease description, a novel badnavirus associated with chestnut mosaic disease

Armelle Marais*†[1], Sergio Murolo†[2], Chantal Faure[1], Yoann Brans[3], Clément Larue[4,5], François Maclot[6], Sébastien Massart[6], Michela Chiumenti[7], Angelantonio Minafra[7], Gianfranco Romanazzi[2], Marie Lefebvre[1], Teresa Barreneche[1], Cécile Robin[4], Rémy J. Petit[4], Thierry Candresse[1]

[1] Univ. Bordeaux, INRAE, UMR BFP, Villenave d'Ornon, France

[2] Department Agricultural, Food and Environmental Sciences, UNIVPM, Ancona, Italy

[3] Laboratoire de Virologie et de Biologie moléculaire, CTIFL, Prigonrieux, France

[4] Univ. Bordeaux, INRAE, UMR Biogeco, Cestas, France

[5] INVENIO, Maison Jeannette, Douville, France

[6] Plant Pathology Laboratory, TERRA-Gembloux Agro-BioTech, University of Liège, Belgium

[7] CNR Institute for Sustainable Plant Protection —Bari, Italy

* Corresponding author: A. Marais; Email: armelle.marais-colombel@inrae.fr

† AM and SM contributed equally to this work

Keywords: High throughput sequencing, *Castanea* sp., etiology, viral disease, *Badnavirus*, *Caulimoviridae*

The nucleotide sequences reported here have been deposited in GenBank under the accession numbers MT261366, MT269853, MT270664-MT270682, and MT339503-MT339590.

ABSTRACT

Although the chestnut mosaic disease (ChMD) was described several decades ago, its etiology is still not elucidated. Here, using classical approaches in combination with high throughput sequencing (HTS) techniques, we identify a novel *Badnavirus* that is a strong etiological candidate for ChMD. Two disease sources from Italy and France were submitted to HTS-based viral indexing. Total RNAs were extracted, ribodepleted and sequenced on an Illumina NextSeq500 (2x150 or 2x 75 nt). In each source, we identified a single contig of about 7.2 kilobases that corresponds to a complete circular viral genome and shares homologies with various badnaviruses. The genomes of the two isolates have an average nucleotide identity of 90.5% with a typical badnaviral genome organization comprising three open reading frames. Phylogenetic analyses and sequence comparisons show that this virus is a novel species for which we propose the name *Chestnut mosaic virus* (ChMV). Using a newly developed molecular detection test, we systematically detected the virus in symptomatic graft-inoculated indicator plants (chestnut and American oak), as well in chestnut trees presenting typical ChMD symptoms in the field (100% and 87% in France and Italy surveys, respectively). Datamining of publicly available chestnut SRA transcriptomic data allowed the reconstruction of two additional complete ChMV genomes from two *Castanea mollissima* sources from the USA, as well as ChMV detection in *C. dentata* from the USA. Preliminary epidemiological studies, performed in France and in Central Eastern Italy, showed that ChMV has a high incidence in some commercial orchards, with a low within-orchard genetic diversity.

European chestnut (*Castanea sativa* Mill.) has a long-standing tradition of cultivation in many European countries. It is an important species, both economically as a source of timber and fruit and ecologically through the multiple ecosystemic services it provides. In Europe, chestnut covers about 2.5 million hectares, mainly concentrated in France, Italy, Spain, Portugal, Switzerland, the Balkan regions, and Southern England (Conedera et al. 2016). Chestnut (*Castanea* spp.) can be heavily affected by various pathogens. The most detrimental are caused by fungal-like organisms (Oomycetes) and fungi such as *Phytophthora cambivora Petri and P. cinnamomic* Rands., the agents of ink disease, or *Cryphonectria parasitica,* the causal agent of chestnut blight, which all provoke disorders that can lead to tree mortality (Prospero et al. 2012; Rigling and Prospero, 2018). In Italy, Gualaccini (1958) described a chestnut disease associated with viral symptoms (mosaic, shoots with asymmetric leaf blade deformation) which was again reported in Campania during the 80s, (Ragozzino and Lahoz 1986), and in the Marche region (central eastern Italy) in 2000 (Antonaroli and Perna; 2000). In France, the disease was first identified circa 1987 on cultivars of *C. sativa* x *C. crenata* hybrids from commercial orchards located in the south-west of the country. Desvignes (1999b) made a more detailed description of the symptoms, which are characterized by necrotic lesions in the bark and wood that turns into cankers, chlorotic lesions and yellow stripes on leaf veins and partial limb atrophy, and called this disease Chestnut Mosaic Disease (ChMD). This disease can heavily affect the production of both young and secular trees (Antonaroli and Perna, 2000). It has also been reported in Japan and Hungary (Shimada, 1962; Horvath et al. 1975). Even if its etiology has long remained unknown, researchers hypothesized that the causal agent of ChMD could be a virus, introduced in Europe between 1940 and 1960 when a number of *C. crenata* cultivars were imported from Japan for breeding purposes. Investigations in France and Italy established that the causal agent can be eliminated by thermotherapy, is aphid-transmissible, and is graft-transmissible to *Castanea* and *Quercus* species in which it may elicit symptoms (Desvignes

and Lecocq, 1995; Desvignes, 1999b; Vettraino et al. 2005). The susceptibility to the ChMD agent of *Castanea* species/cultivars has been evaluated in several studies (Desvignes, 1992; 1999b; Desvignes and Lecoq, 1995). Three categories of cultivars could thus be defined from tolerant to moderately and fully susceptible. Graft incompatibility was also observed when cultivars of different susceptibilities are assembled by grafting. Most of the *C. sativa* cultivars and hybrids are tolerant to ChMD, although some well-known French hybrids like 'Maraval' (Ca 74) are fully susceptible, and used for indexing purposes to detect the ChMD agent in tolerant cultivars (Desvignes and Lecoq, 1995).

In the last decade, a number of studies have highlighted the potential of non-targeted molecular diagnostics based on high-throughput sequencing (HTS) to elucidate the etiology of viral plant diseases and to provide viral sequence data from which rapid diagnostic molecular assays can be developed (Martin et al. 2016; Villamor et al. 2019). Since 2009, HTS combined with bioinformatics have been used for the discovery, characterization, and *de novo* assembly of the genome of known and novel plant viruses and viroids (Rott et al. 2017; Kreuze et al. 2009). This has accelerated the application of HTS technologies in the field diagnostic of diseases (Massart et al. 2014), and in quarantine regulations (Martin et al. 2016; Massart et al. 2017).

Badnaviruses are plant pararetroviruses belonging to the family *Caulimoviridae* that have emerged as serious pathogens causing severe yield losses in a wide range of economically important crops all over the world (Bhat et al. 2016). The genome of badnaviruses is composed of a non-covalently closed, circular double-stranded DNA (ranging from 7.2 to 9.2 kbp) and is encapsidated in bacilliform virions. This genome typically harbors three open reading frames (ORFs) encoding, respectively, a protein of unknown function, the virion-associated protein (VAP), and a polyprotein containing functional and structural domains [movement protein (MP), coat protein (CP), aspartic protease (AP), reverse-transcriptase (RT) and RNase H)] (Hohn and Rothnie, 2013; Bhat et al. 2016).

Badnaviruses can also be present as integrated sequences in some host plant genomes (endogenous badnaviruses) (Staginnus et al. 2009; Bhat et al. 2016). The contribution of these integrated sequences to host and virus evolution is still poorly understood (Geering et al. 2014).

Given the very limited knowledge on the etiology of ChMD, and based on previously published studies (Desvignes, 1992; 1999a, 1999b; Desvignes and Lecoq, 1995; Desvignes and Cornaggia, 1996), we investigated the hypothesis that a virus might be involved in this disease. Combining HTS-based viral indexing and classical approaches, we report here the complete genome sequence of a novel badnavirus species for which the name *Chestnut mosaic virus* (ChMV) is proposed. We further show that there is a strict correlation between the presence of the virus and the appearance of typical ChMD symptoms in various graft-inoculated indicator plants. Preliminary epidemiological studies carried out in Italy and in France reveal that the virus can have high incidence in some orchards, and, as expected, can be associated with symptomatic or asymptomatic infections.

MATERIALS AND METHODS

Plant samples and virus isolates. Virus isolates included in this study are listed in Supplementary Table S1. Isolate LC1224H is originated from a red oak (*Quercus rubra*) artificially inoculated in 1992 with a chestnut mosaic source from a hybrid *Castanea sativa* x *Castanea crenata* included in a French breeding program. Leaves of grafted oaks displayed typical symptoms characterized by chlorotic mottle, yellow veins, and mosaic (Desvignes and Lecoq 1995) (Figure 1A). Isolate FRlc1224A was derived from the same source, and is the result of a back-inoculation by grafting of LC1224H to the natural chestnut hybrid Maraval (Ca 74; *C. crenata* x *C. sativa*) indicator (Desvignes et al. 1992). Isolate LC1224F originated from a Maraval indicator inoculated by aphid transmission from an initial ChMD source in a *C. crenata* x *C. sativa* French hybrid (Desvignes and Cornaggia, 1996). The LCA552 and LCA584 isolates were collected from *C. sativa* trees in France in 2009 and

2018, while the T32018 disease source was isolated from a French hybrid *C. crenata* x *C. sativa* in 2018. All of these isolates have been held and propagated on 'Maraval' indicator plants at the CTIFL virology laboratory (Lanxade, France).

In the framework of a survey carried out in Italian chestnut orchards to monitor chestnut blight (Acquasanta Terme (AP), locality Umito, Italy) (Murolo et al. 2018), typical leaf symptoms of ChMD were recorded in 2016. Six symptomatic plants were collected, pooled (10 -15 symptomatic shoots) and included in the HTS analysis (ITumito39 source).

In order to evaluate the incidence of ChMV, chestnut trees from INRAE chestnut biological resource center (https://www6.bordeaux-aquitaine.inrae.fr/biogeco/Ressources) were sampled. This orchard is located on the Villenave d'Ornon INRAE center (France) with trees distributed in three plots (A, E, or Port, Table S1). A total of 43 *C. sativa,* 14 *C. mollissima,* six *C. crenata* and 32 hybrid chestnut trees were sampled, corresponding to a total of 38 symptomatic trees with typical ChMD symptoms, 47 asymptomatic trees, and 10 trees with dubious or atypical symptoms. In addition, in the Central eastern Italy Marche region, leaves from 60 symptomatic and from 10 asymptomatic grafted *C. sativa* cv. Marrone trees of different ages were collected in a commercial chestnut orchard (Plot I, Table S1).

Isolates FRlc1224A and ITumito39 were used for the HTS analysis, whereas all other samples were included either in the incidence analysis or in the causal relationship analysis (Table S1).

Total RNA extraction and RNA-Seq analysis. Symptomatic leaves from a 'Maraval' indicator (FRlc1224A) were collected and used to extract total RNAs according to the protocol described by Reid et al. (2006). For the Italian material, total RNAs were extracted from symptomatic leaves according to the protocol described by Gambino et al. (2008). Total RNAs were then submitted to a DNAse treatment following the manufacturer's recommendations (Fisher Scientific, Illkirch, France). Ribosomal RNAs were removed using a RiboMinus Plant Kit for RNA-Seq (Invitrogen, Fisher

Scientific, Illkirch, France) before cDNA library synthesis with the Illumina TruSeq Stranded RNA library Prep kit (Illumina, Inc., San Diego, CA) and sequenced on an Illumina NextSeq500 (2x150 nt or 2x75 nt) in a multiplexed format (GIGA-Genomics facility, Université de Liège, Belgium).

Bioinformatic analysis. Primary quality analyses were performed using Geneious Prime 2019.2.1 Software (https://www.geneious.com). *De novo* assemblies of quality filtered reads were performed using Velvet (Zerbino and Birney, 2008), Geneious R 11 (https://www.geneious.com), and Spades (Bankevich et al. 2012), or using the CLC genomics workbench 8.0 http://www.clcbio.com). Contigs were annotated by BlastN and BlastX comparisons with nucleotide and non-redundant protein GenBank databases, respectively. Blast results were screened using e-value thresholds of 10^{-6} and 10^{-4} for BlastN and BlastX, respectively. Publicly available chestnut RNA-Seq transcriptomic data were retrieved from the NCBI Sequence Read Archive (SRA) and downloaded reads were mapped against the sequence of the FRIc1224A isolate using CLC Genomics Workbench 11.0. When needed, *de novo* assembly and contig annotations were also performed as described above.

Total DNA extraction and PCR confirmation of genome completeness and circularity. In order to verify both the completeness of the assembled genome sequences and genome circularity, pairs of specific outward-facing primers were designed for each isolate (Ch-Bad-6976F/ Ch-Bad-252R for the isolate FRIc1224A and Bad-Ch-6481F/Bad-Ch-325R for the isolate ITumito39, Table S2). Leaf tissues (0.5 g) were pulverized in liquid nitrogen and total DNA was extracted in CTAB buffer (2% cetyl trimethylammonium bromide, 100 mM Tris-HCl, 1.4 M NaCl, 20 mM EDTA), adding 3% polyvinyl pyrrolidone 40, and 0.5% sodium metabisulfite (Doyle and Doyle, 1990). Finally, the DNA pellets were resuspended in 50 µl sterile water. PCRs were performed in a 50 µl reaction volume containing 10 mM Tris-HCl (pH 8.5), 2 mM $MgCl_2$, 50 mM KCl, 0.2 mM dNTPs, forward and reverse primers at 1µM each, and 1.25U of Dream Taq (ThermoFisher) using 50 ng of the template. After an

initial denaturation step at 95°C for 4 min, 40 or 35 cycles, respectively, were set at 94˚C for 30 sec, 60°C (Ch-Bad-6976F/ Ch-Bad-252R) or 55˚C (Bad-Ch-6481F/325R) for 30 sec, and 72˚C for 90 sec, followed by a final extension step of 10 min at 72˚C**.** PCR amplification products were sequenced on both strands (GATC, Eurofins, Ebersberg, Germany)**.**

ChMV molecular detection and variant analysis by PCR. For the molecular detection of ChMV, two sets of primers were designed in conserved regions of ORF3 designed using the sequences of isolates FRlc1224A and ITumito39. One primer pair (Ch-Bad-1466F/Ch-Bad-1800R, Table S2) allows the amplification of a genomic region (335 nt) in the MP domain (Figure 2), whereas the second pair (Ch-Bad-5860F/Ch-Bad-6109R, Table S2) amplifies a 232-nt fragment in the RH domain (Figure 2). An aliquot of 25 ng of total DNA was used for the PCR assays in a 50 µl volume containing 10 mM Tris-HCl (pH 8.5), 2 mM $MgCl_2$, 50 mM KCl, 0.2 mM dNTPs, forward and reverse primers at 1 µM each, and either 1.25U of DreamTaq or 1U of GoTaq. After an initial denaturation step at 95°C for 4 min, 35 cycles were set at 94˚C for 30 sec, 56˚C for 30 sec, and 72˚C for 90 sec, followed by a final extension step of 10 min at 72˚C**.** Amplicons were analyzed by electrophoresis on 1.5% agarose gel and were directly sequenced on both strands (GATC).

Possible phytoplasma infection was evaluated using primer pair P1/P7 (Deng and Hiruki. 1991; Smart et al. 1996) and, in nested PCR, primers R16F2n/R2 (Gundersen and Lee 1996).

Sequence and phylogenetic analyses. The full-length genomes were analyzed by ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) to identify putative ORFs in the viral genome. Deduced amino acid (aa) sequences were analyzed for conserved protein domains gathered in Conserved Domains Database (CDD) (http://www.ncbi.nlm.nih.gov/structure/cdd.shtml) and theoretical molecular weights were calculated using ExPASy (http://web.expasy.org/compute_pi/). Multiple alignments of nucleotide (nt) or amino acid (aa) sequences were performed using the ClustalW program (Thompson et al. 1994) implemented in MEGA version 7.0 (Kumar et al. 2016).

Genetic distances (p-distances using strict nt or aa identity) were calculated using MEGA 7.0. Phylogenetic trees were reconstructed using the neighbor-joining method implemented in MEGA 7.0 and robustness of nodes was assessed from 1,000 bootstrap resamplings.

RESULTS

Determination of the complete genome sequence of a novel badnavirus from two chestnut disease sources. Two ChMD sources were included in the HTS analysis. The French source (FRlc1224A) showed typical ChMD symptoms, with leaf deformation, yellow veins and chlorotic diffuse mottling (Figure 1B) and the Italian source (ITumito39) is a mixture of six plants showing intensive vein banding and leaf blade deformation (Figure 1C). High-throughput sequencing of ribodepleted RNAs extracted from the sources FRlc1224A and ITumito39 yielded a total of 10,737,052 reads and 4,135,330 reads, respectively. *De novo* assembly and Blast annotation allowed for the identification of a single long contig with significant homology with badnaviruses. These contigs were respectively 7,264 and 7,214 bp long and showed short terminal redundancies, consistent with the structure of the long RNA transcript involved in the replication of badnaviruses (Teycheney et al. 2020) and suggesting they represented the full coverage of a circular badnaviral genome. A total of 39,657 reads were integrated in the FRlc1224A contig, representing 0.37% of total reads, with a mean coverage depth of 795X, whereas 611 reads (0.015% of total reads) were integrated in the ITumito39 contig, with a mean coverage depth of 14.4X. The circularity and completion of the DNA genome sequence of each isolate were validated by a PCR on purified DNA extracted from the host plants and using specific outward-facing primers designed from the contig sequences. The respective 436 nt- and 1,007 nt fragments were amplified and sequenced, confirming DNA genome completeness and circularity (data not shown). The assembled sequences

have been deposited in GenBank under accession numbers MT269853 and MT261366, respectively. No other plant virus was detected in the two datasets during the Blast annotation of contigs.

Genome organization of chestnut mosaic virus and determination of its phylogenetic relationships. The badnaviral genomes characterized independently from the French and Italian ChMD sources are respectively 7,160 bp and 7,161 bp long, within the range of badnavirus genome sizes (Teycheney et al. 2020). The genomic organization is the same for both isolates, comprising three open reading frames (ORFs) encoded on the positive strand (Figure 2), and is typical for badnaviruses (Teycheney et al. 2020). The ORF1 (nt 245-751, numbering according to the isolate FRlc1224A sequence) encodes a protein of 169 aa (19.8 kDa), the ORF2 (nt 751-1161) encodes a 137-aa protein (15 kDa), and the third ORF (nt 1,163-6,721) encodes a polyprotein of 1,853 aa (211.7 kDa) with five conserved protein domains (Figure 2): a viral movement protein (MP, cl03100), a zinc-binding motif (ZnF, pfam00098), a retroviral aspartyl protease domain (RVP, pfam00077), a reverse transcriptase domain (RT, cd01647) and a ribonuclease H domain (RH, cl14782). The two "Cys" motives (C-X_2-C-X_4-H-X_4-C, and C-X_2-C-X_{11}-C-X_2-C-X_4-C-X_2-C) usually found in the coat protein of badnaviruses (Bath et al. 2016) were also detected in the ORF3-deduced protein, between aa positions 777-790 and 902-928.

Both isolates are closely related, with an overall 90.5% nt identity. The three indels observed between the two sequences are located in the intergenic region, the isolate ITumito39 ended up being one nucleotide longer. The three ORFs have the same sizes, are strictly colinear and the encoded proteins share respectively 95.2% (ORF1), 95.5% (ORF2) and 94.8% (ORF3) aa identity.

To characterize the phylogenetic relationships and taxonomic position of the chestnut badnavirus, a phylogenetic tree was reconstructed using an alignment of full genome nucleotide sequences of *Badnavirus* genus members, with the rice tungro bacilliform virus used as an outgroup (Figure 3). Both isolates cluster in group 3 defined by Wang et al. (2014), together with gooseberry

vein banding virus (GVBV), rubus yellow net virus (RYNV), grapevine vein clearing virus (GVCV), birch leafroll-associated virus (BLRaV), wisteria badnavirus 1 (WBV1), and pagoda yellow mosaic-associated virus (PYMaV) (Figure 3). Nevertheless, they are clearly distant from all of these species, defining a novel branch, supported by a 99% bootstrap value (Figure 3). Tree topology was similar when using an alignment of representative badnaviral ORF3 protein sequences (Figure S1). To confirm these analyses, pairwise comparisons of genome sequences showed that the isolate FRlc1224A has only weak identity levels with representative members of the genus *Badnavirus*, comprised between 42.1% nt identity (sugarcane bacilliform IM virus; 42.5% for the isolate ITumito39) and 50.9% (WBV1; 50.8% for the isolate ITumito39). The same tendency is observed when considering the genome proteins. The ORF1-encoded protein shows only weak homology with the corresponding proteins of WBV1 (27.8% aa identity) and PYMAV (26.1% aa identity), and the ORF2-encoded protein shares only 33.1% aa identity with the corresponding protein of the most closely related virus, WBV1. The polyprotein encoded by ORF3 shares 49.5% aa identity with the corresponding protein of the closest relative, PYMAV. Using the ORF3 region (RT and RH domain) used for taxonomical discrimination in the family *Caulimoviridae* (Teycheney et al. 2020), the FRlc1224A isolate shows between 64% (with GVBV) and 68.4% (with BLRaV) nt identity (Table 1), which is below the 80% nt identity value used as the species demarcation threshold in the family. Therefore, this virus represents a novel species in the family *Caulimoviridae*, for which we propose the name *Chestnut mosaic virus* (ChMV). In the same taxonomically informative region, the isolates FRlc1224A and ITumito39 share 91.9% nt identity (97.8% aa identity), indicating that they belong to the same viral species (Table 1).

Identification of ChMV in publicly available chestnut HTS data. The datamining of chestnut HTS data from various chestnut sources publicly available at GenBank [EST sequences, whole genome assembly, RNA-Seq and Genotyping-by-sequencing (GBS) reads available as Sequence Read

Archives (SRA)] allowed the identification of ChMV in several of those datasets (Table S3). In particular, two complete genomes were obtained from datasets involving *C. mollissima* cv. Vanuxem in the USA, one from the whole genome assembly (JRKL01079565) and the other by *de novo* assembly of RNA-Seq data (SRX4015368) with 99.2% and 97.4% nt identity, respectively, with the FRIc1224A isolate over the whole genome (Figure 4). In addition, partial ChMV genome assemblies of >3 kbp could be obtained from a range of other datasets generated in the USA or in China from *C. mollissima* (Table S3), all of which showed significant relatedness with the FRIc1224A sequence as shown by a phylogenetic tree reconstructed using nucleotide alignments of concatenated ChMV sequences retrieved from the various datasets (Figure 4). In addition, partial ChMV genomes could also be reconstructed from two datasets obtained from *C. dentata* in the USA. Interestingly, one of these two *C. dentata* isolate sequences shows closest relationship with ITumito39 sequence (Figure 4) with only 89.2% nt identity with the isolate FRIc1224A as compared to 93.9% nt identity with ITumito39. The second isolate of *C. dentata* appears to be equally related to the FRIc1224A and ITumito39 isolates, with 90.9% and 90.6% nt identity, respectively.

Incidence and genetic variability of ChMV in France and Italy. The incidence and genetic variability of ChMV were investigated by analyzing two genomic regions of ORF3, one 335-nt long located in the MP domain amplified using primer pair Ch-Bad-1466F/Ch-Bad-1800R and the other 232-nt long in the RNase H domain and amplified with primer pair Ch-Bad-5860F/Ch-Bad-6109R (Table S2, and Figure 2). The two primer pairs were designed to be able to detect both isolates FRIc1224A and ITumito39. In Italy, a total of 70 *C. sativa* cv Marrone samples were collected in the same location, while in France, 95 chestnut accessions belonging to three different *Castanea* species or hybrids were sampled in three plots (A, Port, E). Both symptomatic and asymptomatic samples were collected, as well as some samples with atypical or dubious symptoms. Globally, ChMV was frequent in the surveyed plots, with 57/70 (81.5%) infected *C. sativa* samples in Italy and 65/95 trees

(68%) in France (Table 2). In the Italian orchard, half of the asymptomatic trees were found to be infected by ChMV, compared to 87% of the symptomatic ones (Table 2). None of the analyzed samples were found positive using a phytoplasma-specific PCR assay. In the French collection, the virus was detected in 100% (38/38) of the trees showing typical ChMD symptoms, and in 49% (23/47) of the asymptomatic trees, including two trees that were symptomless but showed strong symptoms on rootstock off-shoots (Figure S2). ChMV was also detected in four out of the 10 trees showing atypical/doubtful symptoms.

The genetic variability of ChMV was evaluated by analyzing the sequences of the two PCR amplicons generated for the incidence survey. Considering the relative homogeneity of the origin of the Italian samples, the number of samples included in this analysis was limited to 13 (four from asymptomatic trees, and nine symptomatic ones) (Table S1). In total, the final dataset consisted in 53 isolates for which the sequence of the two genomic regions were available (49 from the incidence survey and four from independent ChMD sources held in collection at CTIFL, see below). As illustrated by the unrooted neighbor-joining tree reconstructed from the alignments of RT-RnaseH domain nucleotide sequences (Figure S3A), ChMV diversity is structured into two clusters, defined by the geographical origin of the samples (Italy and France). The sequences determined from the four independent French disease sources (FRlc1224A, T30218, LCA552, LCA584) belong to the same «French» cluster. Overall, the level of genetic diversity is very low in this genomic region, with an average pairwise nt divergence (diversity) of 2.2% +/- 0.5%. This value is even lower when considering the intra-group diversity, as 0.2% +/- 0.1% within the French cluster and 0.1% +/- 0.1% within the Italian ones. In contrast, the inter-group diversity reaches 6.3% +/- 1.5%, confirming the existence of two geographical clusters. The same trends are observed when analyzing the genomic region located in the MP domain (Figure S3B). The same geographical clustering could be observed, with the exceptions of three French isolates that seem to be more closely related to the Italian

cluster (Figure S3B). Another French isolate, 20971-E remains isolated and does not fit in either group. The average nt divergence in this region is slightly higher than in the RT-RNaseH region, (5.8% +/- 0.7%), and the inter-group diversity reaches a value of 13.4% +/- 1.8%, as compared to the 6.3% value for the other region.

DISCUSSION

Since the seminal work of Desvignes and collaborators in the 1990s, it has been acknowledged that the agent responsible for ChMD is most likely a thermosensitive, graft-transmissible virus that can be transmitted experimentally and probably naturally by the aphid *Myzocallis castanicola* Desvignes, 1992; 1999a; 1999b; Desvignes and Lecoq, 1995; Desvignes and Cornaggia, 1996). Depending on the chestnut genotype, this infection can either be asymptomatic or can result in the expression of severe and conspicuous ChMD symptoms. In chestnut orchards in the Marche region (Italy), both young and mature plants were shown to be affected, significantly decreasing chestnut production. Symptoms have also been observed in some *Quercus* species following experimental graft inoculation. To date, however, the causal agent remains to be identified.

Here, by using HTS-based viral indexing, we were able to identify and characterize, in two independent ChMD sources, two isolates of the same novel virus. Phylogenetic and sequence analyses showed that this virus belongs to the genus *Badnavirus*, in the family *Caulimoviridae*, and could be considered as a new species, for which the name *Chestnut mosaic virus* (ChMV) is proposed. Interestingly, this new virus clusters with a group of badnaviruses that includes RYNV, GVBaV, and GVCV.

There is unambiguous evidence that ChMV as reported here is an episomal virus. It was detected in graft-inoculated indicators, and not in non-inoculated control plants of the same variety, demonstrating its graft-transmissibility, a property of episomal viruses. This line of evidence is further reinforced by the detection of ChMV in symptomatic, graft-inoculated indicator *Quercus*

plants and, again, not in the corresponding control plants. In parallel, the HTS detection of ChMV from DNAse-treated RNAs, the failure to detect ChMV in a range of the surveyed chestnut trees and the sequence diversity identified in ChMV all rule out a scenario in which an endogenous ChMV genome, integrated in the chestnut genome could be responsible for the HTS and PCR results reported here. There was in fact no indication of ChMV in the chestnut genome assembly (JRKL01079565) since no integration borders could be identified and a single contig, representing a complete unintegrated viral genome transcript, was identified. Integration of ChMV as an endogenous viral element (EVE, Bhat et al. 2016) therefore does not appear to be a general genomic feature of chestnut.

According to the simplified hierarchical approach proposed by Fox (2020) for assessing causal relationships in plant virology, ChMV appears as a good candidate, if not as the causative agent of ChMD. There are several arguments and experimental evidence supporting this idea. Following HTS analyses, ChMV was the sole virus detected in the French source FRlc1224A, coming from a ChMD source initially involving a *C. sativa* x *C. crenata* hybrid. It was also the sole virus detected in the Italian ChMD source analyzed by HTS. Using molecular detection tests developed in this work, the virus was consistently found in other symptomatic accessions derived from the same diseased source (LC1224H, a *Q. rubra* artificially inoculated and LC1224F, an indicator plant inoculated by aphid transmission; Figure 5). In addition, three other independent chestnut sources shown by biological indexing on the 'Maraval' indicator to be affected by ChMD were found to be infected by ChMV (LCA552, LCA584 and T32018 in Figure 5). There is therefore a correlation between the appearance of ChMD symptoms and the presence of ChMV in the graft-inoculated indicators, supporting the hypothesis of a causal relationship between ChMV infection and ChMD. In total, five independent ChMD sources collected between 1990 and 2018 in two countries (Italy and France)

were ChMV positive, satisfying the Bradford-Hill's experimental and consistency criteria (Bradford Hill, 1965; Fox, 2020).

Preliminary studies indicate that ChMV is highly prevalent in the analyzed orchards in France and Italy, confirming the earlier results of Desvignes (1999a). In parallel, the identification of ChMV sequences in publicly available HTS data provides a strong indication of the presence of ChMV in *C. mollissima* in the USA and in China and in *C. dentata* in the USA. In the surveys, ChMV was not systematically associated with symptomatic infections, although its frequency was systematically higher in symptomatic plants. This result was expected since previous grafting experiments had demonstrated that not all chestnut varieties/species are susceptible to ChMD and develop symptomatic infections (Desvignes and Lecoq, 1995; Desvignes, 1992; 1999b). Biological indexing on the susceptible 'Maraval' indicator has in particular identified latent ChMV infections in many symptomless *C. sativa* varieties or *C. sativa x C crenata* hybrids (Desvignes and Lecoq, 1995; Desvignes, 1992; 1999b). On the other hand, all surveyed symptomatic plants in France were found to harbor the virus, while it was detected in 52/60 (87%) of tested symptomatic Italian trees. The failure to detect ChMV in eight symptomatic Italian trees might reflect sequence variability and an incomplete inclusiveness of the PCR primers used or low or uneven virus accumulation. Indeed, using biological indexing, Desvignes et al. have previously found an uneven distribution of the ChMD agent in infected trees leading to a failure to detect it in parts of some infected trees (Desvignes and Lecoq, 1995; Desvignes et al. 1999b).

Taken together, and even though Koch's postulates were not fully verified, the experiments reported here make a very strong case for a role of ChMV as the causal agent of the chestnut mosaic disease. The low ChMV diversity observed in France and Italy are consistent with the scenario of its recent introduction in Europe (Desvignes and Lecoq, 1995), while the genetic separation of the Italian and French clusters is suggestive of separate introduction events. These results and the

associated development of molecular tools for the detection of ChMV will help speed up the selection of virus-free mother plants and mitigate the virus spread in new chestnut orchards and layerings. However, many questions remain regarding the variability of symptom intensity in relationship to cultivar susceptibility, ChMV-induced graft incompatibility, the impact of pedoclimatic conditions and of synergic and competitive interferences with other chestnut pathogens, and silvicultural management.

ACKNOWLEDGMENTS

The authors thank the INRAE Vine Experimental Unit of Bordeaux (UEVB) and X. Capdevielle for taking care of chestnut orchards. We would like to thank Ascenzio Santini (Azienda Agricola Laga Nord, Umito, Aquasanta Terme, Italy) for the essential support during the phytosanitary monitoring in chestnut orchards. Thanks to D. Cornaggia and P. Gentit for the information provided on the past work of CTIFL, for the characterization of the symptoms of the disease on biological indicators and the conservation of the isolates. Thanks to Eden Darnige for proofreading the manuscript. Part of this work was based upon COST Action FA1407 (DIVAS), supported by COST (European Cooperation in Science and Technology).

LITERATURE CITED

Antonaroli, R., and Perna, M.R. 2000. Una fitopatia ad eziologia ancora incerta: il giallume dl castagno in Emilia Romagna e nelle Marche. Sherwood 6:43-46.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 9:455-477.

Bhat, A. I., Hohn, T., and Selvarajan, R. 2016. Badnaviruses: the current global scenario. Viruses 8:177; doi:10.3390/v8060177.

Bradford Hill, A. 1965. The environment and disease: association or causation? Proc. R. Soc. Med. 58:295–300.

Conedera, M., Tinner, W., Krebs, P., de Rigo, D., and Caudullo, G. 2016. *Castanea sativa* in Europe: distribution, habitat, usage and threats. Pages 78-79. in: European Atlas of Forest Tree Species. J. San-Miguel-Ayanz, D. de Rigo, G. Caudullo, T. Houston Durrant, A. Mauri eds.

Deng, S. and Hiruki, C. 1991. Amplification of 16S rRNA genes from culturable and non-culturable mollicutes. J. Microbiol. Meth. 14:53-61.

Desvignes, J. C. 1992. Characterization of the chestnut mosaic. Acta Hortic. 309:353-358.

Desvignes, J. C., and Lecocq, G. 1995. New knowledges on the chestnut mosaic virus disease. Acta Hortic. 386:578-584.

Desvignes, J. C., and Cornaggia, D. 1996. Mosaïque du chataignier: Transmission par le puceron *Myzocallis castanicola*. Phytoma 481:39-41.

Desvignes, J. C. 1999a. Mosaïque du chataignier. Pages 179-181. in: Maladies à virus des arbres fruitiers. Ctifl Ed. France

Desvignes, J. C. 1999b. Sweet chestnut incompatibility and mosaics caused by the chestnut mosaic virus (ChMV). Acta Hortic. 494:451-454.

Doyle, J. J., and Doyle, J. L. 1990. Isolation of plant DNA from fresh tissue. Focus 12:13–15.

Fox, A. 2020. Reconsidering causal association in plant virology. Plant Pathol. doi: 10.1111/ppa.13199

Gambino, G., Perrone, I., and Gribaudo, I. 2008. A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem. Anal. 19:520-525.

426 Geering, A. D. W., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., McTaggart, A. R.,
427 Scalabrin, S., Vezzulli, S., Wing, R. A., Quesneville, H. and Teycheney, P. Y 2014. Endogenous
428 florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat.
429 Comm. 5:5269
430 Gualaccini, G. 1958. Una virosi nuova del castagno. Boll. Staz Patol. Veg. Roma, 16:67-75.
431 Gundersen, D. E. and Lee, I. M. 1996. Ultrasensitive detection of phytoplasmas by nested-PCR assays
432 using two universal primer pairs. Phytopathol. Mediterr. 35:144-151.
433 Hohn, T., and Rothnie, H. 2013. Plant pararetroviruses: replication and expression. Curr. Opin. Virol.
434 3:621-628.
435 Horvath, J., Ecke, I., Gal, T., and Dezcery, M. 1975. Demonstration of virus-like particles in sweet
436 chestnut and oak with leaf deformations in Hungary. Z. Pflkrankh. Pfsdrntz. 82:498-502.
437 Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., and Barker, I. 2009. Complete viral
438 genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic
439 method for diagnosis, discovery and sequencing of viruses. Virology 388:1-7.
440 Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA 7: Molecular Evolutionary Genetics Analysis
441 version 7 for bigger datasets. Mol. Biol. Evol. 33:1870-1874.
442 Martin, R. R., Constable, F., and Tzanetakis, I. E. 2016. Quarantine regulations and the impact of
443 modern detection methods. Annu. Rev. Phytopathol. 54:189-205.
444 Massart, S., Olmos, A., Jijakli, H., and Candresse, T. 2014. Current impact and future directions of
445 high throughput sequencing in plant virus diagnostics. Virus Res. 188:90-96.
446 Massart, S., Candresse, T. Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J. S., Rumbou, A.,
447 Saldarelli, P., Škorić, D., Vainio, E. J., Valkonen, J. P. T., Vanderschuren, H., Varveri, C., and Wetzel,
448 T. 2017. A framework for the evaluation of biosecurity, commercial, regulatory, and scientific

449 impacts of plant viruses and viroids identified by NGS technologies. Front. Microbiol. 8:45. doi:
450 10.3389/fmicb.2017.00045
451 Murolo, S., De Miccolis Angelini, R. M., Faretra F., and Romanazzi, G. 2018. Phenotypic and
452 molecular investigations on hypovirulent *Cryphonectria parasitica* in Italy. Plant Dis. 102:540-
453 545.
454 Prospero, S., Vannini, A., and Vettraino, A. M. 2012. Phytophthora on *Castanea sativa* Mill. (sweet
455 chestnut). JKI Data Sheets 6, DOI 10.5073/jkidspdd.2012.006
456 Ragozzino, A., and Lahoz, E. 1986. Una malattia virus-simile del castagno in provincia di Avellino.
457 Giornate di studio sul castagno. Soc. Orti. Italiana 307-311.
458 Reid, K. E, Olsson, N., Schlosser, J., Peng, F., and Lund, S. T. 2006. An optimized grapevine RNA
459 isolation procedure and statistical determination of reference genes for real-time RT-PCR during
460 berry development. BMC Plant Biol. 6:26-37.
461 Rigling, D., and Prospero, S. 2018. *Cryphonectria parasitica*, the causal agent of chestnut blight:
462 invasion history, population biology and disease control. Mol. Plant Pathol. 19:7-20.
463 Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C.,
464 Bhagwat, B., and Rast, H. 2017. Application of Next Generation Sequencing for diagnostic testing
465 of tree fruit viruses and viroids. Plant Dis. 101:14891499.
466 Shimada, S. 1962. Chestnut Yellows. Pl. Prot. Tokyo, 16:253-254.
467 Smart, C. D., Schneider, B., Blomquist, C. L., Guerra, L. J., Harrison, N. A., Ahrens, U., Lorenz, K. H.,
468 Seemuller, E., and Kirkpatrick, B. C. 1996. Phytoplasma-specific PCR primers based on sequences
469 of 16s-23s rRNA spacer region. Appl. Environ. Microbiol. 62:2988-2993.
470 Staginnus, C., Iskra-Caruana, M. L., Lockhart, B., Hohn, T., and Richert-Pöggeler, K. R. 2009.
471 Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. Arch. Virol.
472 154:1189-1193.

473 Teycheney, P. Y., Geering, A. D. W., Dasgupta, I., Hull, R., Kreuze, J.F., Lockhart, B., Muller, E.,
474 Olszewski, N., Pappu, H., Pooggin, M. M., Richert-Pöggeler, K. R., Schoelz, J. E.,Seal, S., Stavolone,
475 L., Umber, M. and ICTV Report Consortium. 2020. ICTV Virus Taxonomy Profile: *Caulimoviridae*.
476 J. Gen. Virol. DOI 10.1099/jgv.0.001497
477 Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTALW: Improving the sensitivity of
478 progressive multiple sequence alignment through sequence weighting, position-specific gap,
479 penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.
480 Vettraino, M., Vannini, A., Flamini, L., Lagnese, R., Pizzichini, L., Talevi, S., and Fulbright, D. W. 2005.
481 A new transmissible symptomology on sweet chestnut in Italy. Acta Hortic. 693:547-550.
482 Villamor, D. E. V., Ho, T., Al Rwahnih, M., Martin, R. R., and Tzanetakis, I. E. 2019. High Throughput
483 Sequencing for plant virus detection and discovery. Phytopathology 109:716-725.
484 Wang, Y., Cheng, X., Wu, X., Wang, A., and Wu, X. 2014. Characterization of complete genome and
485 small RNA profile of pagoda yellow mosaic virus, a novel badnavirus in China. Virus Res. 188:103-
486 108.
487 Zerbino, D. R., and Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de
488 Bruijn graphs. Genome Res. 18:821-829.

489 **TABLE 1.** Percentage of identity between the ORF3 region encoding the Reverse Transcriptase -
490 Rnase H[a] of chestnut mosaic virus (ChMV) isolate FRlc1224A and the corresponding genomic regions
491 of the isolate ITumito39 and of the most closely related members of the genus *Badnavirus*

Virus[b]	Nucleotide identity (%)	Amino acid identity (%)
ChMV ITumito39	91.9%	97.8%
RYNV	65.1%	71.6%
GVBV	64%	68.9%
GVCV	66.8%	72.6%
BLRaV	68.4 %	72.6%
WBV1	68.2%	71.6%
PYMaV	67.7%	71.9%

492 [a] This region is the one typical used for taxonomic discrimination in the family *Caulimoviridae*
493 (Teycheney et al. 2020)

494 [b] Acronyms used: RYNV, rubus yellow net virus; GVBV, gooseberry vein banding virus; GVCV,
495 grapevine vein-clearing virus; BLRaV, birch leafroll-associated virus; WBV1, wisteria badnavirus 1;
496 PYMaV, pagoda yellow mosaic-associated virus.

497

498

499 **TABLE 2.** Number and percentage of chestnut mosaic virus-infected plants regarding the plot, the
500 *Castanea* species sampled, and the symptomatology

Origin of the sampled plants	Infected/total plants (%)	Infected/asymptomatic plants (%)	Infected/symptomatic plants (%)	Infected/plants with atypical symptoms (%)
France (overall)	65/95 (68.4%)	23/47 (48.9%)	38/38 (100%)	4/10 (40%)
Plot A	13/23 (56.5%)	6/15 (40%)	6/6 (100%)	1/2 (50%)
Plot E	48/66 (72.7%)	15/28(53.6%)	30/30 (100%)	3/8 (37.5%)
Plot Port	4/6 (66.6%)	2/4 (50%)	2/2 (100%)	na[b]
C. crenata	5/6 (83%)	1/1 (100%)	4/4 (100%)	0/1
C. sativa	30/43 (70%)	11/20 (55%)	18/18 (100%)	1/5 (20%)
C. mollissima	13/14 (93%)	1/2 (50%)	10/10 (100%)	2/2 (100%)
Hybrid[a]	17/32 (53.1%)	10/24 (41.6%)	6/6 (100%)	1/2 (50%)
Italy *Castanea sativa*	57/70 (81.5%)	5/10 (50%)	52/60 (87%)	na[b]

501 [a] Interspecific hybrids between *C. crenata, C. mollisima* and *C. sativa*

502 [b] not applicable

503

504 **CAPTIONS FOR FIGURES**

505 **Fig. 1. Symptoms of chestnut mosaic disease on various hosts.** (A) Isolate LC1224H: Red oak
506 (*Quercus rubra*) graft-inoculated with a diseased source; (B) Isolate FRlc1224A: 'Maraval' Ca 74
507 graft-inoculated with LC1224H; (C) Isolate ITumito39: symptomatic leaves from cv Marrone grafted
508 onto *Castanea sativa*; (D) non-inoculated *Q. rubra*; (E) non-inoculated 'Maraval' Ca 74; (F)
509 Asymptomatic leaves from cv Marrone grafted onto *C. sativa*.

Fig. 2: Schematic representation of the genomic organization of the chestnut mosaic virus. The tRNA binding site is indicated and defines the position 1 on the genome. The three open reading frames (ORFs) are shown as grey arrows, as well as their position in parentheses. Five conserved motifs are identified in the ORF3 polyprotein: MP, Viral movement protein (pfam01107); ZnF, Zinc finger (pfam00098); RVP, Retroviral aspartyl protease (pfam00077); RT, Reverse transcriptase (cd01647); RH, Ribonuclease H (cd09274)

Fig. 3. Phylogenetic tree reconstructed using the complete genome sequences of badnavirus members. Virus names as well as GenBank accession numbers are indicated. The tree was reconstructed using the neighbor-joining method, and randomized bootstrapping was performed to evaluate the statistical significance of branches (1,000 replicates). Bootstrap values above 70% are shown. The scale bar represents 5% nucleotide divergence between sequences. The groups as defined by Wang et al. (2014) are indicated. Chestnut mosaic virus isolates determined in this work are indicated by black triangles. Rice tungro bacilliform virus was used as outgroup.

Fig. 4. Unrooted neighbor-joining phylogenetic tree reconstructed from the alignment of concatenated nucleotide sequences related to chestnut mosaic virus detected by datamining of publicly available transcriptomic chestnut data. Randomized bootstrapping was performed to evaluate the statistical significance of branches (1,000 replicates). Bootstrap values above 70% are shown. The scale bar represents 10% nucleotide divergence between sequences.

Fig. 5. Detection of chestnut mosaic virus in various samples by PCR using primers pairs Ch-Bad1466F/1800R (A) and Ch-Bad5860F/6109R (B). Lane 1: LC1224F; Lane 2: LC1224H; Lane 3: FRlc1224A; Lane 4: T32018; Lane 5: LCA552; Lane 6: LCA584; Lane 7: 'Maraval' Ca 74 non-inoculated plant; Lane 8: *Quercus rubra* non-inoculated plant; Lane 9: no template; L: molecular weight marker. Horizontal bars on the left of the figure indicate the size of the amplification products. The isolates are listed in Table S1.

e-EXTRA FIGURE CAPTIONS AND TABLE TITLES

Supplementary Fig. S1. Phylogenetic tree reconstructed using the ORF3-deduced amino acid sequences of badnavirus members. Virus names as well as GenBank accession numbers are indicated. The tree was reconstructed using the neighbor-joining method, and randomized bootstrapping was performed to evaluate the statistical significance of branches (1,000 replicates). Bootstrap values above 70% are shown. The scale bar represents 10% amino acid divergence between sequences. Chestnut mosaic virus isolates determined in this work are indicated by black triangles. Rice tungro bacilliform virus was used as outgroup.

Supplementary Fig. S2. Leaves of the sample 20893. (A) rootstock leaves (*Castanea sativa*); (B) cultivar leaves

Supplemental Fig. S3. Unrooted neighbor-joining phylogenetic trees reconstructed from the alignment of nucleotide sequences of the PCR fragments targeted partial RT-RnaseH domain (A) and partial MP domain (B) obtained from a range of chestnut mosaic virus isolates (listed in Table S1). Randomized bootstrapping was performed to evaluate the statistical significance of branches (1,000 replicates). Bootstrap values above 70% are shown. The scale bar represents 5% (A) or 10% (B) nucleotide divergence between sequences.

Supplemental Table S1. List of chestnut samples used in the present study together with relevant ChMV accession numbers

Supplemental Table S2. Primers used for genome circularity confirmation and for molecular detection of chestnut mosaic virus by PCR

565 **Supplemental Table S3. Datamining of publicly available chestnut HTS data for chestnut mosaic**
566 **virus sequences**

A

D

B

E

C

F

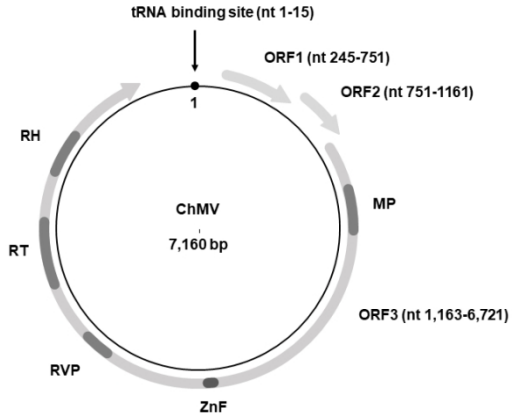


Fig. 2: Schematic representation of the genomic organization of the chestnut mosaic virus. The tRNA binding site is indicated and defines the position 1 on the genome. The three open reading frames (ORFs) are shown as grey arrows, as well as their position in parentheses. Five conserved motifs are identified in the ORF3 polyprotein: MP, Viral movement protein (pfam01107); ZnF, Zinc finger (pfam00098); RVP, Retroviral aspartyl protease (pfam00077); RT, Reverse transcriptase (cd01647); RH, Ribonuclease H (cd09274)

338x190mm (96 x 96 DPI)

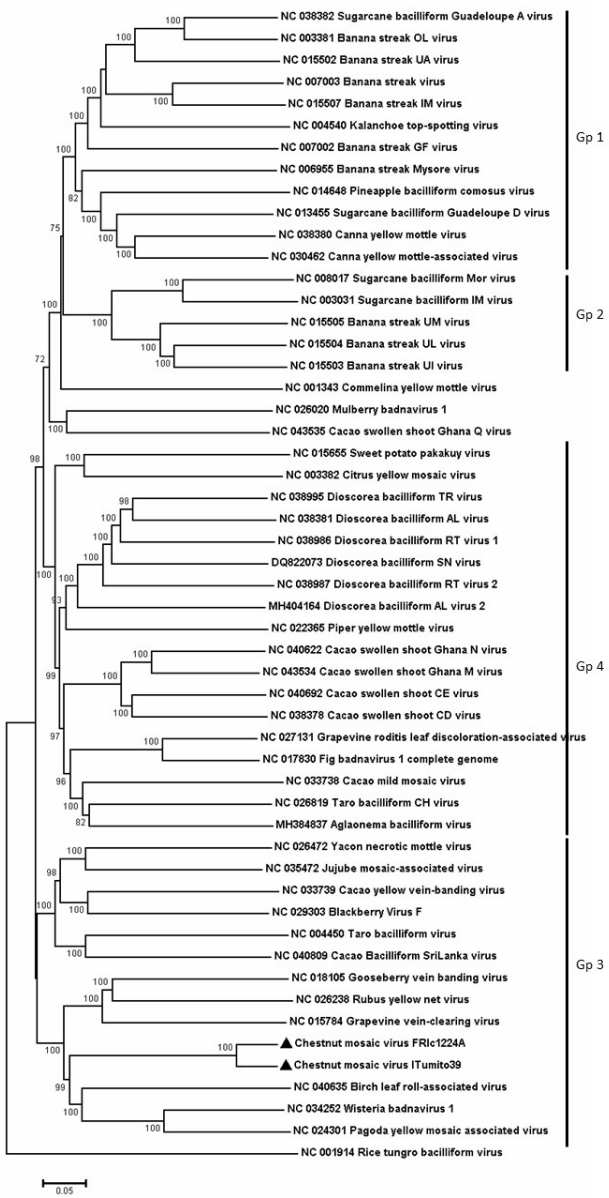
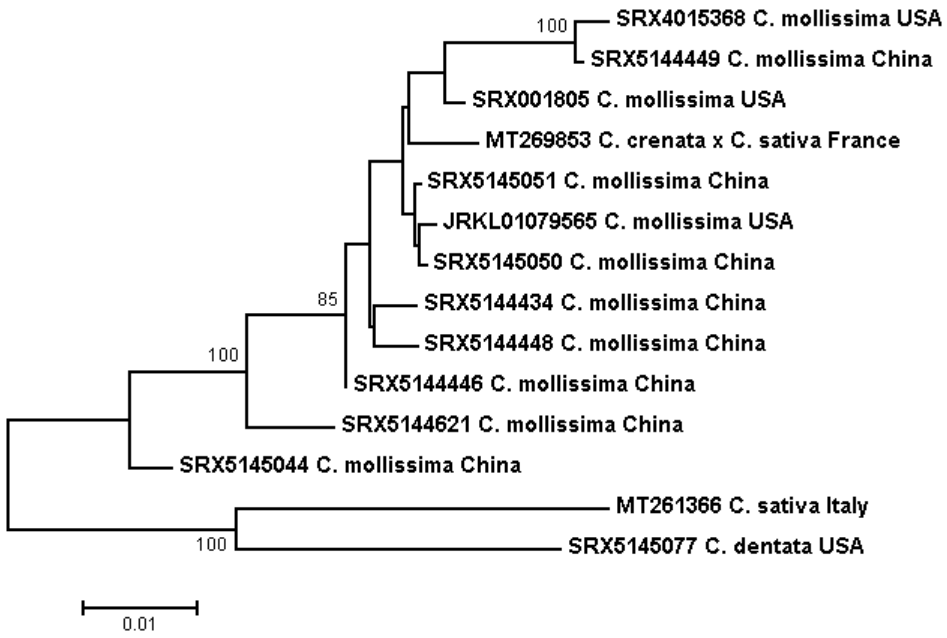


Fig. 3. Phylogenetic tree reconstructed using the complete genome sequences of badnavirus members. Virus names as well as GenBank accession numbers are indicated. The tree was reconstructed using the neighbor-joining method, and randomized bootstrapping was performed to evaluate the statistical significance of branches (1,000 replicates). Bootstrap values above 70% are shown. The scale bar represents 5% nucleotide divergence between sequences. The groups as defined by Wang et al. (2014) are indicated. Chestnut mosaic virus isolates determined in this work are indicated by black triangles. Rice tungro bacilliform virus was used as outgroup.

190x338mm (96 x 96 DPI)


100
SRX4015368 C. mollissima USA
SRX5144449 C. mollissima China
SRX001805 C. mollissima USA
MT269853 C. crenata x C. sativa France
SRX5145051 C. mollissima China
JRKL01079565 C. mollissima USA
SRX5145050 C. mollissima China
85
SRX5144434 C. mollissima China
SRX5144448 C. mollissima China
100
SRX5144446 C. mollissima China
SRX5144621 C. mollissima China
SRX5145044 C. mollissima China
MT261366 C. sativa Italy
100
SRX5145077 C. dentata USA
0.01

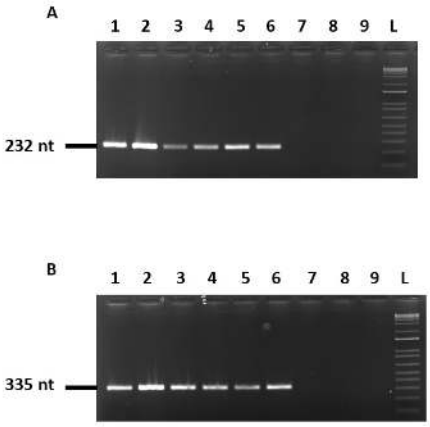


Fig. 5. Detection of chestnut mosaic virus in various samples by PCR using primers pairs Ch-Bad1466F/1800R (A) and Ch-Bad5860F/6109R (B). Lane 1: LC1224F; Lane 2: LC1224H; Lane 3: FRlc1224A; Lane 4: T32018; Lane 5: LCA552; Lane 6: LCA584; Lane 7: 'Maraval' Ca 74 non-inoculated plant; Lane 8: Quercus rubra non-inoculated plant; Lane 9: no template; L: molecular weight marker. Horizontal bars on the left of the figure indicate the size of the amplification products. The isolates are listed in Table S1.

338x190mm (96 x 96 DPI)

Fig. S1

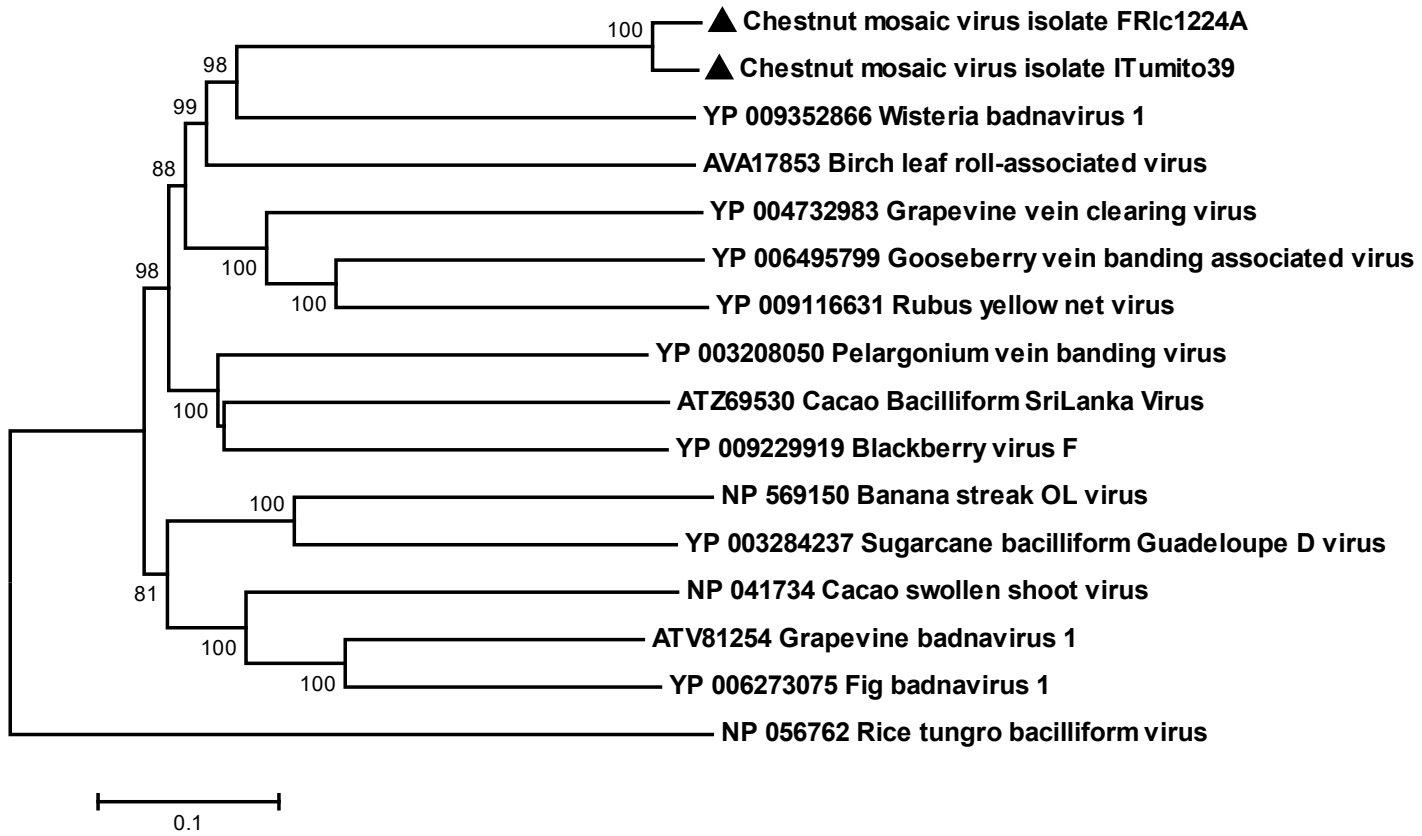

Chestnut mosaic virus isolate FRIc1224A
Chestnut mosaic virus isolate ITumito39
YP 009352866 Wisteria badnavirus 1
AVA17853 Birch leaf roll-associated virus
YP 004732983 Grapevine vein clearing virus
YP 006495799 Gooseberry vein banding associated virus
YP 009116631 Rubus yellow net virus
YP 003208050 Pelargonium vein banding virus
ATZ69530 Cacao Bacilliform SriLanka Virus
YP 009229919 Blackberry virus F
NP 569150 Banana streak OL virus
YP 003284237 Sugarcane bacilliform Guadeloupe D virus
NP 041734 Cacao swollen shoot virus
ATV81254 Grapevine badnavirus 1
YP 006273075 Fig badnavirus 1
NP 056762 Rice tungro bacilliform virus
100
98
99
88
98
100
100
100
100
81
100
100
0.1

Fig. S2

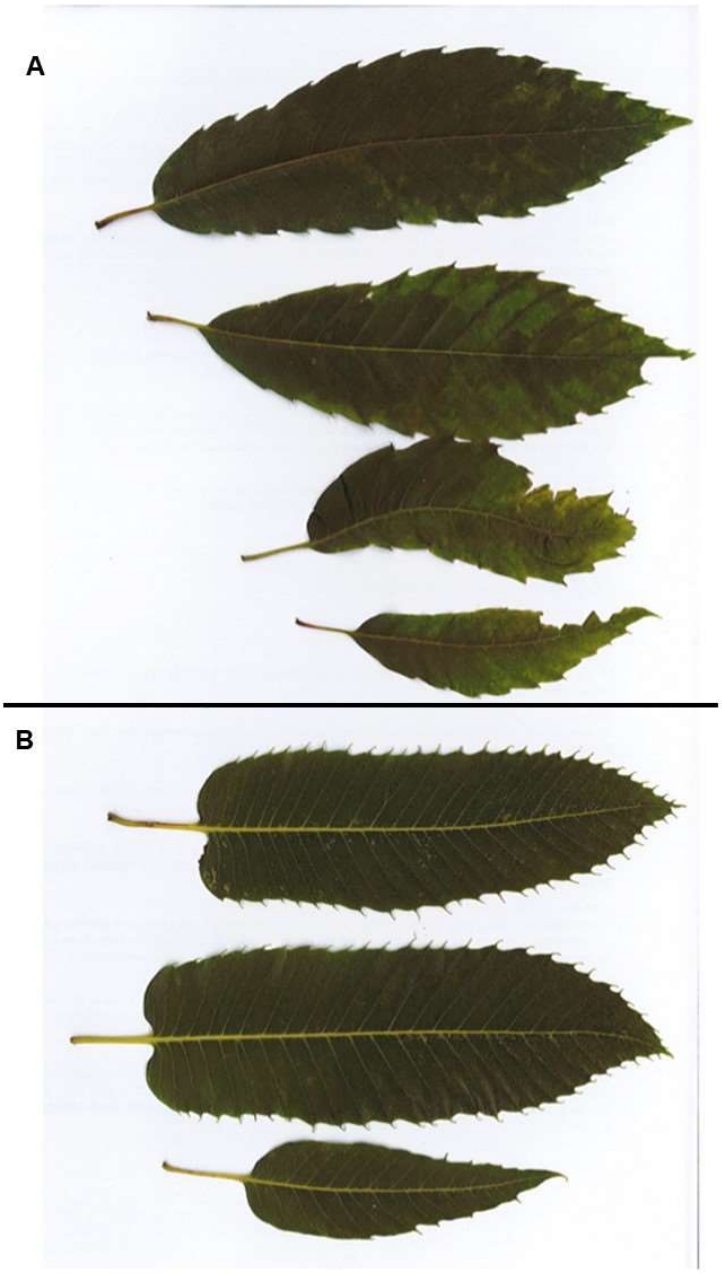


Fig. S3A

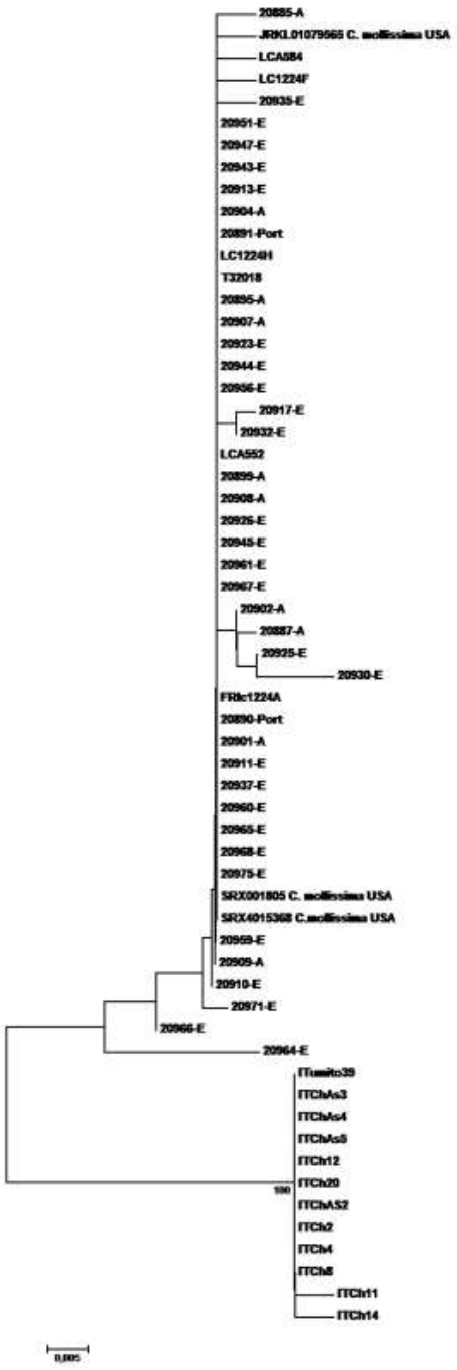
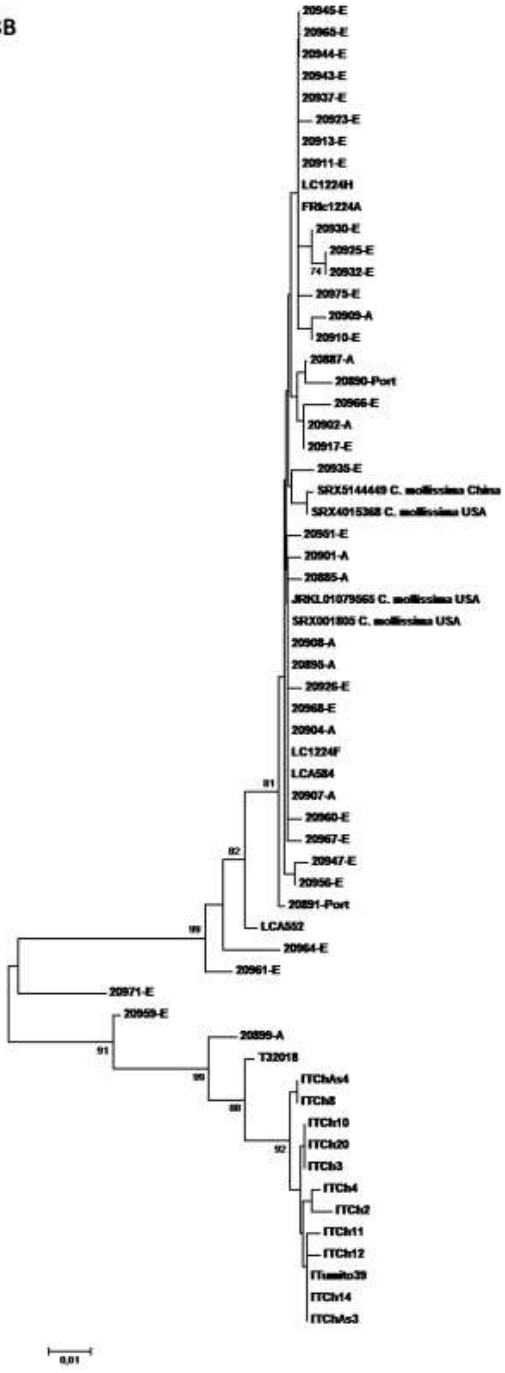

20885-A
JRKL01079566 C. scalfiensis USA
LCA584
LC1224F
20935-E
20951-E
20947-E
20943-E
20913-E
20904-A
20891-Port
LC1224H
T32018
20895-A
20907-A
20923-E
20944-E
20956-E
20917-E
20932-E
LCA552
20899-A
20908-A
20926-E
20945-E
20961-E
20967-E
20902-A
20887-A
20925-E
20930-E
FRtc1224A
20890-Port
20901-A
20911-E
20937-E
20960-E
20965-E
20968-E
20975-E
SRX001805 C. scalfiensis USA
SRX4015368 C.scalfiensis USA
20959-E
20909-A
20910-E
20971-E
20966-E
20964-E
ITtumito39
ITChAs3
ITChAs4
ITChAs5
ITCh12
ITCh20
ITChAS2
ITCh2
ITCh4
ITCh8
ITCh11
ITCh14
100
0.005

Fig. S3B


20945-E
20965-E
20944-E
20943-E
20937-E
20923-E
20913-E
20911-E
LC1224H
FRdc1224A
20930-E
20925-E
74
20932-E
20975-E
20909-A
20910-E
20887-A
20890-Port
20966-E
20902-A
20917-E
20935-E
SRX5144449 C. mollissima China
SRX4015368 C. mollissima USA
20961-E
20901-A
20885-A
JRKL01079565 C. mollissima USA
SRX001805 C. mollissima USA
20908-A
20896-A
20926-E
20968-E
20904-A
LC1224F
LCA584
81
20907-A
20960-E
20967-E
82
20947-E
20956-E
20891-Port
99
LCA552
20964-E
20971-E
20959-E
91
20899-A
T32018
99
ITChAs4
ITCh8
88
ITCh10
ITCh20
92
ITCh3
ITCh4
ITCh2
ITCh11
ITCh12
ITkunito39
ITCh14
ITChAs3
0,01

Supplementary Table S1. List of chestnut samples used in the present study together with relevant ChMV accession numbers

Isolate ID [a]	Chestnut Accession ID	Chestnut Accession name	Species [b]	Rootstock	Country	Plot [c]	Accession numbers [d]	Use [e]
20885	Ca04x Ca03	-	*C. crenata*	Not graft	France	A	MT339547; MT339503	Incidence
20887	Ca 715	'Merle'	*C. sativa*	G1. Ca15	France	A	MT339548; MT339504	Incidence
20890	Ca 663	'Trigueira'	*C. sativa*	unknown	France	Port	MT339549; MT339505	Incidence
20891	Ca 664	'Longal Special'	*C. sativa*	Ca 74	France	Port	MT339550; MT339506	Incidence
20895	Ca 43	'Vignols'	*C. sativa x C. crenata*	Ca 116	France	A	MT339551; MT339507	Incidence
20899	Ca 599	'Ibuki'	*C. crenata*	G1.Ca 02	France	A	MT339552; MT339508	Incidence
20901	Ca 598	'Rihei'	*C. crenata x C. mollissima*	G1. Ca102	France	A	MT339553; MT339509	Incidence
20902	Ca 75	'Fertil'	*C. mollissima*	Ca 07	France	A	MT339554; MT339510	Incidence
20904	Tree-A68Ks [i]	-	*C. mollissima*	Not graft	France	A	MT339555; MT339511	Incidence
20907	Ca 118	'Marlhac'	*C. sativa x C. crenata*	Not graft	France	A	MT339556; MT339512	Incidence
20908	Ca 124	'Maridonne'	*C. sativa x C. crenata*	G1.Ca116	France	A	MT339557; MT339513	Incidence
20909	Ca 125	'Bouche de Bétizac'	*C. sativa x C. crenata*	Ca 74	France	A	MT339558; MT339514	Incidence
20910	Ca 860	-	hybrid SC	unknown	France	E	MT339559; MT339515	Incidence
20911	Ca 860	-	hybrid SC	unknown	France	E	MT339560; MT339516	Incidence
20913	Ca 844 [i]	-	*C. mollissima*	Not graft	France	E	MT339561; MT339517	Incidence
20917	Ca 837 [h]	-	*C. mollissima*	Not graft	France	E	MT339562; MT339518	Incidence
20923	Ca 846 [h]	-	*C. mollissima*	Not graft	France	E	MT339563; MT339519	Incidence
20925	Ca 741	'Dauphine'	*C. sativa*	Ca 07	France	E	MT339564; MT339520	Incidence
20926	Ca 665	'Longal'	*C. sativa*	Ca 07	France	E	MT339565; MT339521	Incidence
20930	Ca 860	-	hybrid SC	unknown	France	E	MT339566; MT339522	Incidence
20932	Ca 564	'Ipharra 16'	*C. crenata*	Ca 07	France	E	MT339567; MT339523	Incidence
20935	Ca 576	'Sardonne'	*C. sativa*	Ca 07	France	E	MT339568; MT339524	Incidence
20937	Ca 138	'Marron de Redon'	*C. sativa*	Ca 07	France	E	MT339569; MT339525	Incidence
20943	unknown	-	*C. sativa*	Ca 74	France	E	MT339570; MT339526	Incidence
20944	Ca 520	'Montagne'	*C. sativa*	Ca 07	France	E	MT339571; MT339527	Incidence
20945	Ca 106	'Marron Comballe'	*C. sativa*	Ca 07	France	E	MT339572; MT339528	Incidence

20947	Ca 111	'Marron de Lyon'	*C. sativa*	Ca 07	France	E	MT339573; MT339529	Incidence
20951	Ca 126	-	*C. sativa x C. crenata*	Ca 07	France	E	MT339574; MT339530	Incidence
20956	Ca 105	'Sardonne'	*C. sativa*	Ca 07	France	E	MT339575; MT339531	Incidence
20959	Ca 03	-	*C. crenata*	Ca 74	France	E	MT339576; MT339532	Incidence
20960	Ca 127	-	*(C. crenata x C. sativa) x C. mollisima*	Ca 07	France	E	MT339577; MT339533	Incidence
20961	Ca 127	-	*(C. crenata x C. sativa) x C. mollisima*	Ca 07	France	E	MT339578; MT339534	Incidence
20964	Ca 112	'Bournette'	*C. crenata x C. sativa*	Ca 07	France	E	MT339579; MT339535	Incidence
20965	Ca 501	-	*C. sativa*	Ca 07	France	E	MT339580; MT339536	Incidence
20966	Ca 501	-	*C. sativa*	Ca 07	France	E	MT339581; MT339537	Incidence
20967	Ca 665	'Longal'	*C. sativa*	Ca 07	France	E	MT339582; MT339538	Incidence
20968	Ca 48	'Précoce Migoule'	*(C. crenata x C. sativa)*	Ca 07	France	E	MT339583; MT339539	Incidence
20971	Ca 151	'Bouche Rouge'	*C. sativa*	Ca 07	France	E	MT339584; MT339540	Incidence
20975	unknown	-	*C. sativa x C. crenata*	Ca 07	France	E	MT339585; MT339541	Incidence
20889	unknown	-	*C. sativa*	unknown	France	Port	na	Incidence
20893	Ca 106	'Marron Comballe'	*C. sativa*	unknown	France	Port	na	Incidence
20898	Ca 105	'Sardonne'	*C. sativa*	G1. Ca394	France	A	na	Incidence
20900	Ca 501	-	*C. sativa*	G1.Ca486	France	A	na	Incidence
20903	Ca744	'Quing Zha'	*C. mollissima*	G1.Moll	France	A	na	Incidence
20912	Ca 845 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20914	Ca 843 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20915	Ca 842 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20918	Ca 838 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20919	Ca 839 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20920	Ca 840 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20921	Ca 841 [h]	-	*C. mollissima*	Not graft	France	E	na	Incidence
20922	Ca 74	'Maraval'	*C. crenata x C. sativa*	Not graft	France	E	na	Incidence
20924	Ca 825	'Précoce de Besse'	*C. sativa*	Ca 07	France	E	na	Incidence
20940	Ca 393	'Marron de Chevanceaux'	C. sativa	Ca 07	France	E	na	Incidence
20941	Ca 460	-	(*C. crenata x C. sativa*) x (*C. crenata x C. sativa*)	Ca 74	France	E	na	Incidence
29942	Ca 511	'Marrone di Chiusa Pesio'	*C. sativa*	Ca 74	France	E	na	Incidence
20948	Ca 116	-	*C. sativa x C. crenata*	Ca 74	France	E	na	Incidence
20949	Ca 135	'Précoce des Vans'	*C. sativa*	Ca 74	France	E	na	Incidence
20952	Ca 521	-	*C. crenata*	Ca 07	France	E	na	Incidence

20955	Ca 399	'Comballe'	*C. sativa*	Ca 07	France	E	na	Incidence
20962	Ca 639	-	*C. sativa*	Ca 07	France	E	na	Incidence
20963	Ca 730	'Sauvage Marron'	*C. sativa*	Ca 07	France	E	na	Incidence
20969	Ca 381	-	*C. sativa*	Ca 07	France	E	na	Incidence
20970	Ca 381	-	*C. sativa*	Ca 07	France	E	na	Incidence
20972	Ca 151	'Bouche Rouge'	*C. sativa*	Ca 07	France	E	na	Incidence
FRlC1224A	Ca 74	'Maraval'	*C. crenata x C. sativa*	na	France	Ctifl	MT269853 (complete genome)	HTS
LC1224H	na	na	*Q. rubra*	na	France	Ctifl	MT339586; MT339542	CRA
T32018	Ca 74	Maraval	*C. crenata x C. sativa*	na	France	Ctifl	MT339587; MT339543	CRA
LCA552	Ca 74	Maraval	*C. sativa*	na	France	Ctifl	MT339588; MT339544	CRA
LCA584	Ca 74	Maraval	*C. sativa*	na	France	Ctifl	MT339589; MT339545	CRA
LC1224F	Ca 74	Maraval	*C. crenata x C. sativa*	na	France	Ctifl	MT339590; MT339546	CRA
ITUmito39	unknown	Marrone	*C. sativa*	na	Italy	I	MT261366 (complete genome)	HTS
ITCh2	unknown	Marrone	*C. sativa*	na	Italy	I	MT270674; MT270683	Incidence
ITCh3	unknown	Marrone	*C. sativa*	na	Italy	I	MT270667 [f]	Incidence
ITCh4	unknown	Marrone	*C. sativa*	na	Italy	I	MT270668; MT270684	Incidence
ITCh8	unknown	Marrone	*C. sativa*	na	Italy	I	MT270685 [g]	Incidence
ITCh10	unknown	Marrone	*C. sativa*	na	Italy	I	MT270669; MT270678	Incidence
ITCh11	unknown	Marrone	*C. sativa*	na	Italy	I	MT270664; MT270679	Incidence
ITCh12	unknown	Marrone	*C. sativa*	na	Italy	I	MT270665; MT270680	Incidence
ITCh14	unknown	Marrone	*C. sativa*	na	Italy	I	MT270666; MT270681	Incidence
ITCh20	unknown	Marrone	*C. sativa*	na	Italy	I	MT270670; MT270675	Incidence
ITChAs3	unknown	Marrone	*C. sativa*	na	Italy	I	MT270671; MT270676	Incidence
ITChAs4	unknown	Marrone	*C. sativa*	na	Italy	I	MT270672 [f]	Incidence
ITChAs5	unknown	Marrone	*C. sativa*	na	Italy	I	MT270677 [g]	Incidence
ITChAs8	unknown	Marrone	*C. sativa*	na	Italy	I	MT270673; MT270682	Incidence
ITChAs9	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh1	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh5	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh6	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh7	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh9	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh13	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh14	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh15	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh20	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh21	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence

ITCh22	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh23	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh24	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh25	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh27	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh28	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh29	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh30	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh31	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh32	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh33	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh35	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh36	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh37	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh38	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh39	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh41	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh42	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh43	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh44	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh46	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh47	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh48	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh49	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh50	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh51	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh52	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh54	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh55	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh56	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh58	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh59	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence
ITCh60	unknown	Marrone	*C. sativa*	na	Italy	I	na	Incidence

[a] Isolates from asymptomatic trees are indicated in bold; isolates with doubtful symptoms are underlined; isolates from asymptomatic cultivars with typical symptoms visible on rootstock regrowths are indicated in bold italic.

[b]; hybrid SC : hybrid between *C. sativa* and *C. crenata*, regardless the knowledge about which are the female and male parents.

[c] : Five plots have been sampled (A, Port, E, I, Ctifl)

[d] : Accession numbers are relative to the sequences obtained with both PCR detection assays (Ch-Bad-1466F/Ch-Bad-1800R and Ch-Bad-5860F/Ch-Bad-6109R)

[e:] Isolates were included either in the HTS analysis, or in the incidence analysis, or in the causal relationship analysis (CRA)

[f]: Only the fragment amplified with Ch-Bad-1466F/Ch-Bad-1800R was sequenced

[g]: Only the fragment amplified with Ch-Bad-5860F/Ch-Bad-6109R was sequenced

[h] 'Mengshankui' (*C. mollissima* cultivar) seedling

[i] *C. mollissima* seedling

Supplemental TABLE S2. Primers used for genome circularity confirmation and for molecular detection of chestnut mosaic virus by PCR

Primer	Nucleotide sequence (5' to 3')	Genome coordinates	Amplicon size	Isolates detected
Ch-Bad-6976F	CCCGAGCCATTTAC ACTTCACAAC	6,976-6,999	436 nt	FRlc1224A
Ch-Bad-252R	TCACTCATCTACCTC ACACGCTC	252-230		
Ch-Bad-6481F	GAAATTGAATTGGA AGGAAGA	6,481-6,501	1,007 nt	ITumito39
Ch-Bad-325R	TCAGATCAGCAAAC TCGAAC	344-325-		
Ch-Bad-1466F	TATCAGCACTACAG TGAACAACC	1,466-1,488	335 nt	polyvalent
Ch-Bad-1800R	GTCATGACGCAAAC TTGGAA	1,800-1,781		
Ch-Bad-5860F	AGTATGTAAATGG GCACCGTTC	5,857-5,878	232 nt	polyvalent
Ch-Bad-6109R	GTTGATCCATCGCA CTCTTG	6,109-6,090		

Supplemental TABLE S3. Datamining of publicly available chestnut HTS data for chestnut mosaic virus sequences

Type of data[a]	Dataset	*Castanea* species/cultivar	Country	Length of assembled sequence (nt)	% mapped reads[b]	% nt identity with FRlc1224A
EST	GO917001	*C. mollissima*, BX3, clone Vanuxum	USA	436	na	99%
WGA	JRKL01079565	*C. mollissima*, cv Vanuxem	USA	7,164 (full lenght)	na	99%
RNA-Seq	SRX4015368	*C. mollissima*	USA	7,161 (full lenght)	0.004%	97.4%
RNA-Seq	SRX001805	*C. mollissima*, cv Vanuxem	USA	5,889 (scaffold)	0.04%	95-100%
GBS	SRX5144434	*C. mollissima*	China	5,933 (scaffold)	0.06%	98.7% (average)
GSB	SRX5144449	*C. mollissima,* clone Vanuxem	China	6,396 (scaffold)	0.04%	97.5% (average)
GBS	SRX5145051	*C. mollissima,* cv Vanuxem	China	5,873 (scaffold)	0.02%	99% (average)
GBS	SRX5145050	*C. mollissima*	China	6,680 (scaffold)	0.01%	99% (average)
GBS	SRX51445044	*C. mollissima*	China	6,183 (scaffold)	0.01%	97.4% (average)
GBS	SRX51444621	*C. mollissima*	China	5,875 (scaffold)	0.02%	95.6% (average)
GBS	SRX51444448	*C. mollissima*	China	2,948 (scaffold)	0.04%	99% (average)
GBS	SRX51444446	*C. mollissima*	China	5,873 (scaffold)	0.04%	99% (average)

GBS	SRX5825095	*C. dentata*	USA	1,857 (scaffold)	0.0036%	89.2% (average)
GBS	SRX5145077	*C. dentata*	USA	6,068 (scaffold)	0.01%	90.9% (average)

[a] EST: expressed sequence tags; WGA: whole genome assembly; GBS: genotyping by sequencing

[b] Reads mapped to the genomic sequence of chestnut mosaic virus (French isolate)