Datamining, Genetic Diversity Analyses, and Phylogeographic Reconstructions Redefine the Worldwide Evolutionary History of Grapevine Pinot gris virus and Grapevine berry inner necrosis virus

Jean-Michel Hily, Nils Poulicard, Thierry T. Candresse, Emmanuelle Vigne, Monique Beuve, Lauriane Renault, Amandine Velt, Anne-Sophie Spilmont, Olivier Lemaire

OPEN ACCESS
phytobiomesjournal.org
Phytobiomes Journal
A Transdisciplinary Journal of Sustainable Plant Productivity

RESEARCH

e-Xtra*

Datamining, Genetic Diversity Analyses, and Phylogeographic Reconstructions Redefine the Worldwide Evolutionary History of *Grapevine Pinot gris virus* and *Grapevine berry inner necrosis virus*

Jean-Michel Hily,[1,2,†] Nils Poulicard,[3] Thierry Candresse,[4] Emmanuelle Vigne,[2] Monique Beuve,[2] Lauriane Renault,[2] Amandine Velt,[2] Anne-Sophie Spilmont,[1] and Olivier Lemaire[2]

[1] IFV, Le Grau-Du-Roi, France
[2] Université de Strasbourg, INRAE, SVQV UMR-A 1131, F-68000 Colmar, France
[3] IRD, Cirad, Université Montpellier, IPME, Montpellier, France
[4] UMR 1332 Biologie du Fruit et Pathologie, INRAE, University of Bordeaux, CS 20032, 33882 Villenave d'Ornon Cedex, France

Accepted for publication 8 December 2019.

ABSTRACT

The recently described member of the genus *Trichovirus* grapevine Pinot gris virus (GPGV) has now been detected in most grape-growing countries. While it has been associated with severe mottling and deformation symptoms under some circumstances, it has generally been detected in asymptomatic infections. The cause(s) underlying this variable association with symptoms remain(s) subject to speculations. GPGV genetic diversity has been studied using short genomic regions amplified by RT-PCR but not so far at the pan-genomic level. In an attempt to gain insight into GPGV diversity and evolutionary history, a systematic datamining effort was performed on our own high-throughput sequencing (HTS) data as well as on publicly available sequence read archive files. One hundred new complete or near complete GPGV genomic sequences were thus obtained, together with 69 new complete genomes for the other grapevine-infecting *Trichovirus*, grapevine berry inner necrosis virus (GINV). Phylogenetic and diversity analyses revealed that both viruses likely have their origin in Asia and that China is the most probable country of origin of GPGV. However, despite their common taxonomy, origin, and host, these two trichoviruses display very distinct genetic features and evolutionary traits. GINV shows an important overall genetic diversity, and is likely evolving under a balancing selection in a very restricted region of the world. On the contrary, GPGV shows a worldwide distribution with a modest genetic diversity and presents a strong selective sweep pattern. Taken together, these results show how two closely related trichoviruses differ drastically in their evolutionary history and epidemiological success. Possible causes for these differences are discussed.

Keywords: center of origin, datamining, grapevine, metagenomics, phylogeography, plant pathology, trichovirus, virology

[†]Corresponding author: J.-M. Hily; jean-michel.hily@vignevin.com

The newly obtained sequences used in the present manuscript have been deposited as GenBank accession numbers MN458411 to MN458460 and BK011060 to BK011178.

Author contributions: Conceptualization: J.M.H.; methodology and investigation: J.M.H., N.P., M.B., A.V., and T.C.; software: J.M.H., N.P., and A.V.; data curation: L.R.; writing original draft: J.M.H.; writing review and editing: all authors; funding: J.M.H., E.V., O.L., and A.S.S.; and supervision: O.L., T.C., and A.S.S.

Funding: This work was supported by Institut National de la Recherche Agronomique (INRA) and Institut Français de la Vigne et du Vin (IFV), funded by the Plan National Dépérissement du vignoble (project VACCIVINE; French Ministry of Agriculture, class of 2017). A grant from Moët & Chandon, Comité Champagne, Bureau Interprofessionnel des Vins de Bourgogne and Comité Interprofessionnel des Vins d'Alsace was awarded to J.-M. Hily.

*The *e-Xtra* logo stands for "electronic extra" and indicates that supplementary material and supplementary tables are published online.

The author(s) declare no conflict of interest.

With the dawn of high-throughput sequencing (HTS) approaches for virome analysis in the early 2010s, grapevine has been shown to belong among the crops infected by the most virus species. There are now more than 70 virus species identified as infecting grapevine (Martelli 2017), and their number is still growing, as exemplified by the recent identification of two grapevine-infecting negative-sense RNA viruses (Diaz-Lara et al. 2019).

Among these viruses, two belong to the genus *Trichovirus*, grapevine Pinot Gris virus (GPGV) and grapevine berry inner necrosis virus (GINV). Their genomes are constituted of positive-sense RNAs with three open reading frames (ORFs) encoding for the replication-associated protein (REP), the movement protein (MP) and the coat protein (CP), in that order. Symptoms later associated with GPGV were reported for the first time in northern Italy in 2003 but the virus was only characterized 9 years later (Giampetruzzi et al. 2012). It has since been detected in most grapevine growing regions around the world: Europe (France, Germany, Czech Republic, Greece, Slovakia, Slovenia, Turkey, Spain, Portugal, and Georgia) (Bertazzon et al. 2016; Beuve et al. 2015; Gazel et al. 2016; Glasa et al. 2014; Pleško et al.

2014; Reynard et al. 2016), Asia (China, South Korea, and Pakistan) (Cho et al. 2013; Fan et al. 2016b; Rasool et al. 2017), and in "recent" grapevine growing areas such as Canada, the United States, Uruguay, or Australia (Al Rwahnih et al. 2016; Jo et al. 2015; Wu and Habili 2017; Xiao et al. 2016). An eastern European origin of the virus has been proposed following its detection by RT-PCR (Bertazzon et al. 2016). GPGV has been identified in many different grapevine cultivars (Bertazzon et al. 2016) and has also been shown to naturally infect some herbaceous hosts (Gualandri et al. 2017). The presence of GPGV is sometimes associated with leaf deformation, chlorotic mottling and stunting (grapevine leaf mottling and deformation [GLMD]). Yield and quality of berries may also be affected (Giampetruzzi et al. 2012). However, in most cases, GPGV infection appears to be asymptomatic. An early stop codon located in the MP coding sequence, shortening the MP from 375 to 369 amino acids (aa) (Saldarelli et al. 2014), has been suggested to have a link with symptomatology, but this notion has not yet been completely established and is still a matter of conjecture.

The other *Trichovirus* infecting grapevine, GINV, was first described in 1997 (Yoshikawa et al. 1997) due to its important economic impact in Yamanashi, Japan (Terai et al. 1993). Since then, GINV has only been reported from China (Fan et al. 2016a; Fan et al. 2017). Both viruses have been demonstrated to be transmitted through grafting, and by mites, in semicontrolled conditions and in the field for GPGV and GINV, respectively (Kunugi et al. 2000; Malagnini et al. 2016). No other mode of transmission is known in the genus *Trichovirus* (King et al. 2012).

So far, genetic studies on these two viruses have been performed on RT-PCR amplified partial genomic sequences (Bertazzon et al. 2017; Fan et al. 2017) or on a very limited number of complete genomic sequences (Tarquini et al. 2019). For GPGV, all studies confirmed a low genetic diversity, with sequence identity ≥97.2% (Tarquini et al. 2019). When focusing on MP/CP and RNA-dependent RNA polymerase genomic regions, three distinct GPGV genetic lineages tentatively separating virulent and latent variants have been reported (Saldarelli et al. 2014). As for GINV, phylogenetic analyses revealed the existence of three well defined clades, with one corresponding to Japanese isolates (Fan et al. 2017). In the present work, the genetic diversity and evolutionary patterns along the complete genome of these two viruses were analyzed using a very large corpus of sequences spanning a wide geographic range. As of 1 June 2019, only 26 GPGV and four GINV complete to near complete sequences, were available in the GenBank database (https://www.ncbi.nlm.nih.gov/genbank/). Here, an additional 100 GPGV genomes and 69 GINV genomes were recovered from our own HTS datasets as well as from data available in sequence read archives (SRA, https://www.ncbi.nlm.nih.gov/sra). These 169 sequences were obtained from 42 grapevine varieties coming from 10 different countries in Asia, Europe, and the Americas. By capitalizing on available and newly generated sequences, we investigated the genetic diversity and evolution pattern of both grapevine-infecting trichoviruses. Our study exposes the great dichotomy that separates the two viruses: arising from their diversity index, their evolutionary traits and their propagation around the world, in spite of seemingly having the same Asian origin. This work demonstrates the importance of datamining in order to better study the evolution of viruses, ultimately helping in deciphering their center of origin.

MATERIALS AND METHODS

High throughput RNA sequencing. Total RNAs were extracted from 100 mg of grapevine leaf tissue using the RNeasy Plant mini kit (Qiagen, Venlo, Netherlands), as per the manufacturer's recommendations. Post extraction, purity criteria ($A260_{nm}/A230_{nm}$ and $A260_{nm}/A280_{nm}$ both >1.8) and quality levels (RNA integrity number >8) were assessed with a Bioanalyzer (Agilent, Santa Clara, CA). Total RNAs were used to prepare cDNA libraries after a poly-A selection at the GeT-PlaGe Genotoul sequencing facility (INRA-Toulouse, France) and sequenced as paired-end 2 × 150 bp on a Hiseq 3000 (Illumina, San Diego, CA) following the manufacturer's instructions. Demultiplexing was performed by GeT-PlaGe Genotoul, using Bcl2fastq version 2.20.0.422 and allowing 0 barcode mismatches. Only reads above 70 nucleotides (nt) were kept after trimming and quality check.

HTS data analysis and datamining. A selection of grapevine RNA-seq and small RNA-seq data available in SRA (Nourinejhad Zarghani et al. 2018) was downloaded and screened for the presence of GPGV and GINV. The selection was based on variety, country, time of sampling, but not on the phytopathological status (i.e., symptoms…). GPGV detection was achieved following the same protocol for both proprietary RNAseq datasets and datamined SRA datasets, with all analyses being performed using the CLC Genomics Workbench 11.0 software (Aarhus, Denmark). First, reads that mapped to the *Vitis vinifera* genome (http://www.plantgdb.org/XGDB/phplib/download.php?GDB=Vv), and those corresponding to grapevine transcriptome were removed. The remaining reads were then mapped on viral references for GPGV and GINV (GPGV-SK30 [KF134123], GPGV-Mer [KM491305] and GINV NC_015220 [KU971246], respectively) using relaxed mapping stringency (0.5/0.7 corresponding to read length/similarity parameters) in order to take into account genome diversity within the two virus species. Datasets for which GPGV and GINV presence was detected in this way, were then de novo assembled using the following parameters: word size 21, bubble size 50, and minimal contig size 250 nt. GPGV and GINV contigs identified by mapping against the above reference genomes were then further extended by multiple rounds of residual reads mapping, until near complete genomes were obtained. All near complete genomes were confirmed using very stringent mapping parameters (0.97/0.98 length/similarity) and have been deposited as GenBank accession numbers MN458411 to MN458460 and BK011060 to BK011178.

Sequence analyses, genetic diversity, and recombination detection. Multiple sequence alignments and maximum likelihood-based phylogenetic trees were prepared using CLUSTALW (Thompson et al. 1994) implemented in MEGA7 and MEGAX software (Kumar et al. 2016), excluding the viral untranslated regions (UTRs). The best ML-fitted model for each sequence alignment was used, and nodes in phylogenetic trees and branch validity were evaluated by bootstrap analyses (100 replicates).

The diversity index (π), which is the average number of nucleotide substitutions per site between any two sequences in a multisequence alignment, and the variation of π along genomes were evaluated by sliding window analyses (length: 80, step size: 20) using DnaSP v.6.12.03 (Librado and Rozas 2009) and MegaX. A search for potential recombination signals for both GPGV and GINV was performed using all seven algorithms implemented in the RDP program v4.97 (Martin et al. 2015). The preloaded settings were used for each algorithm and only recombination events detected by four or more methods were considered.

Differences in nucleotide diversity of viral populations defined using different modalities were tested by analysis of molecular variance (AMOVA), as implemented in Arlequin v. 5.3.1.2 (Excoffier et al. 2005). AMOVA calculates the F_{ST} index explaining the between-groups fraction of total genetic diversity. The significance of these differences was evaluated by performing 1,000 sequence permutations.

Tajima's D (D_T) and sliding window analyses were conducted using DnaSP v. 6.12.03 (Librado and Rozas 2009) in order to distinguish the viral populations evolving randomly (per mutation-drift equilibrium;

$D_T = 0$) from those evolving under a nonrandom process ($D_T > 0$: balancing selection, sudden population contraction; $D_T < 0$: recent selective sweep, population expansion after a recent bottleneck).

Discrete phylogeographic analyses. In order to examine the degree of temporal signal in the GPGV dataset, we employed an exploratory linear regression approach (Duchêne et al. 2015; Murray et al. 2016). After discarding the recombinant sequences identified previously in the GPGV dataset and the sequence with unknown sampling location, a maximum likelihood phylogenetic tree was reconstructed on a 116 GPGV sequence dataset under the best evolutionary model (i.e., GTR+G+I) as implemented in MEGAX (Kumar et al. 2016). Then, TempEst v1.5.1 (Rambaut et al. 2016) was used to regress phylogenetic root-to-tip distances against sampling date using the root that minimized the residual mean squares.

In addition to the visual exploration, the significance of the temporal signal was evaluated by a date-randomization test. Thus, the mean rate and its 95% highest probability density (HPD) estimated with the observed sampled dates (using the Bayesian Evolutionary Analysis Sampling Trees BEAST v1.8.2 package, see below) were compared with a null distribution obtained by randomly permutating the tip dates 10 times (Firth et al. 2010). As previously described (Duchêne et al. 2015; Murray et al. 2016), the criterion for a significant temporal signal was that the 95% HPD for the rate estimate obtained with the observed sampled dates should not overlap with the 95% HPD for the estimate obtained with randomized sampling times.

The discrete phylogeographic inferences were generated on the dataset of full-length GPGV coding sequences with no recombinant sequences using BEAST 1.8.2 (Drummond et al. 2012) and the BEAGLE library (Ayres et al. 2011) to improve computational performance. BEAST uses Markov chain Monte Carlo (MCMC) integration to average over all plausible evolutionary histories for the data, as reflected by the posterior probability.

A Hasegawa-Kishino-Yano 85 (HKY85) substitution model was applied, with a discretized Γ and I distributions to model rate heterogeneity across sites and invariable site proportion, respectively. An uncorrelated relaxed molecular clock that models branch rate variation according to a log normal distribution (Drummond et al. 2006) was specified to accommodate among-lineage rate variation. The flexible nonparametric demographic skygrid prior was selected (Gill et al. 2012), using a cut-off to 90 years with 50 grid points.

Discrete phylogeographic inferences were estimated at the continental level (America, Asia, and Europe) using the continuous-time Markov chain (CTMC) process (Lemey et al. 2009) and with a Bayesian stochastic search variable selection (BSSVS). This method reconstructs the dispersion history between discrete locations and infers a posterior distribution of trees whose internal nodes are associated with an estimated ancestral location. MCMC analyses were run for 600 million, sampling every 100,000th and 10% iterations discarded as the chain burn-in. The maximum clade credibility (MCC) tree was obtained with TreeAnnotator 1.8.2 (Drummond et al. 2012) and convergence and mixing properties (e.g., based on effective sample sizes >200 for the parameters) were inspected using Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer).

RESULTS

Obtention of 169 complete genome sequences of grapevine-infecting trichoviruses from HTS data. From our own RNAseq datasets obtained from grapevine plants coming from French vineyards (Champagne, Alsace, and the south-eastern part of France), 50 near complete genomes (covering at least all open reading frames [ORFs], see below) of grapevine Pinot gris virus were de novo assembled (Supplementary Table S1). In addition, among the SRA files analyzed, 43 datasets were positive for the presence of GPGV, from which another 50 GPGV near complete genome sequences could be assembled (Supplementary Table S1). As of 1 June 2019, 26 complete or near complete GPGV genome sequences were available at NCBI and were included, generating the dataset of 126 near complete GPGV sequences that was used in the analyses described below (Supplementary Table S1). Eleven of these sequences lacked a maximum of 145 nt at the 3′ end of the genome (60 nt corresponding to the 3′ end of the CP gene and 85 nt of the 3′ UTR), so that this region was not included in the analyses. Overall, the dataset included isolates coming from 10 countries and from 36 grapevine varieties and one herbaceous plant (Silene) (Supplementary Table S1).

From the same collection of data, GINV was not identified from our RNA-seq data from French samples and was only detected during the SRA datamining. Remarkably, all GINV positive samples came from China. In total, 69 complete or near complete GINV genome sequences were obtained from 33 grapevine samples (Supplementary Table S1). In all, 10 of these 33 samples were coinfected with GPGV (Supplementary Table S1). In addition, four complete GINV genomes were available in the GenBank database (as of 1 June 2019) and were included in the analyses. Because some assembled sequences displayed very high genetic redundancy (having been assembled from transcriptome analyses replicates), only samples with different origin (i.e., coming from different locations, year, and/or variety) and displaying less than 99% identity were retained for analysis, so that a total of 39 GINV sequences obtained from 19 samples were finally analyzed (Supplementary Table S1).

GPGV and GINV genetic organization is highly conserved. The genetic organization of GPGV and GINV is fully conserved between all analyzed genomic sequences and is typical of members of the genus *Trichovirus*, within the family *Betaflexiviridae*. Three open reading frames (ORFs), surrounded by two UTRs are found, with ORF1 corresponding to the large replication-associated protein (REP) and ORF2 and ORF3 encoding the movement protein (MP) and the capsid protein (CP), respectively. Not considering the UTRs, and excluding GPGV-Goldfinger (KU508673), all GPGV sequences are identical in size (7,073 nt). GPGV-Goldfinger is 9 nt longer due to indels within ORF1 (Fig. 1).

Other than for GPGV-Goldfinger, ORF1 has the same size in all analyzed GPGV isolates (5,565 nt corresponding to a 1,855 aa protein). The Goldfinger ORF displays an early stop codon, shortening the protein by 10 aa. ORF2 seems to be more plastic and three different sizes are observed with the most common being 375 aa long. Six variants (three from France and three from China, IV6-I70-5-3, IV8-F85, IV8-F82, SRR5332107-GPGV, SRR5332108-GPGV, and SRR2845691-GPGV) encode a 370 aa protein and 13 variants have a predicted 369 aa MP (10 coming from Italy, 2 from China, and 1 from France). In all cases, the difference in ORF length is linked with previously described mutations introducing an early termination codon (Saldarelli et al. 2014). Two CP sizes of 195 and 193 aa were observed. The shorter protein is only found in the GPGV-BC1 sequence (KU194413) and is due to a single nucleotide insertion, 11 nucleotides upstream the regular stop codon, resulting in a frameshift leading to an early termination (data not shown).

In the case of GINV, excluding the 5′ and 3′ UTRs, two different genome sizes were also found, of 7,046 nt (23 isolates) and 7,055 nt (16 isolates), respectively. As for GPGV, the indels identified were all located in ORF1. The ORF1 of GINV was the most plastic, with sizes of 5,595, 5,601, and 5,604 nt (corresponding to 1,865, 1,867, and 1,868 aa, respectively). Interestingly, all shorter versions of

ORF1 were observed in isolates belonging to clades I and III (Fig. 2) and exhibiting a small ORF on the genome negative strand, 176 nt upstream of the MP start codon (Fig. 3). Except NC_015220 with its 1,050 nt (350 aa) ORF2, all other GINV isolates had an ORF2 of 1,047 nt (349 aa), much shorter than the corresponding GPGV ORF2. The CP ORF was fully conserved in size 585 nt (195 aa) between GINV isolates.

Overall, for GPGV isolates, identity percentages for either ORF1 or ORF3 (Supplementary Tables S2 and S3) were well above the accepted 72% (nt) and 80% (aa) species demarcation values for the family *Betaflexiviridae* (https://talk.ictvonline.org/ICTV/proposals/2015.011a-adP.A.v2.Betaflexiviridae_rev.pdf), confirming that all analyzed isolates belong to the GPGV species. For GINV ORF1, pairwise identity percentages ranged between 70.3 and 99.5% at the nucleotide level and between 76.5 and 99.8% at the amino acid level (Supplementary Tables S4 and S5). All comparisons yielding identity percentages below the species demarcation values involved the only isolate from Japan (NC_015220), while all other comparisons, involving isolates from China yielded values higher than 74.7%, above the species demarcation criteria based on the REP coding sequences. For the CP gene, pairwise identity percentages were all above 77.5% (nucleotide level) and 84.6% (amino acid level) and therefore clearly within the CP gene-based species boundary. It therefore appears that depending on the gene considered, different conclusions can be reached as to whether all GINV isolates belong to a single species or whether the Chinese and Japanese isolates should be considered as belonging to different species.

The genetic diversity of GPGV and the identification of a distinct Asian lineage. The overall genetic diversity of GPGV was analyzed using a dataset covering only the coding genome sequence, except the last 60 nt at the end of the CP gene (described previously), and ultimately consisting of a total of 7,010 aligned nucleotides. For all pairwise comparisons, the percentage of nucleotide

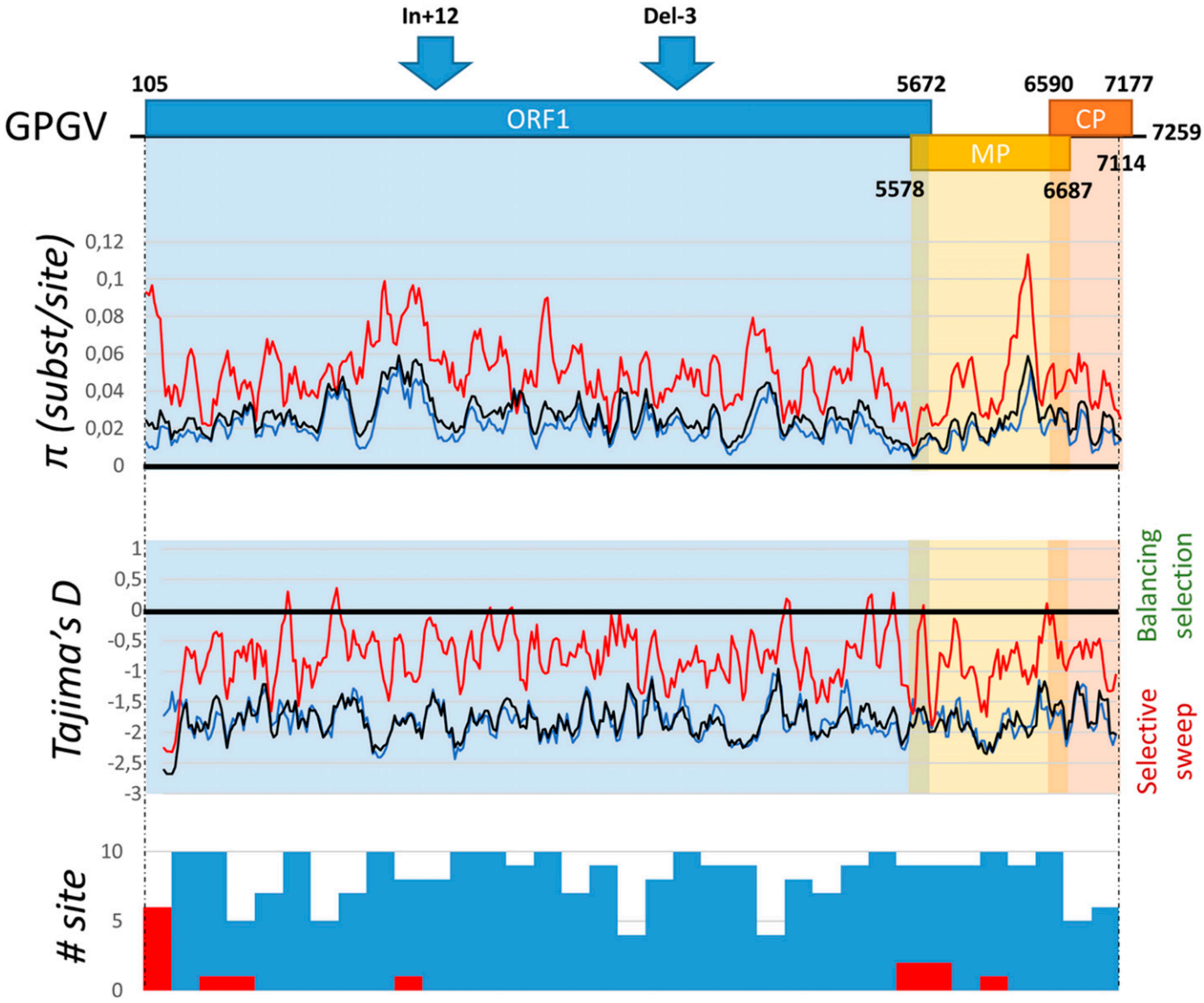

Fig. 1. Genetic diversity analyses of grapevine Pinot gris virus (GPGV) using a corpus of 126 sequences covering 7,010 nt of the genome. Graphics represent π (substitution/sites) along the genome, Tajima's *D* (D_T) for evolution study, comparing sequences from Asia (CN and PK, in red), Europe (FR, IT, DE, and SK, in blue) and the overall values (in black). Below is the representation of the number of sites, per windows of 10 sites along the genome, under selection for which D_T are at $P < 0{,}10$, 0.05, 0.01, and 0.001 for Europe and Asia (Supplementary Table S7). Arrows indicate GPGV-Goldfinger InDels locations.

identity was ≥91.6%. Interestingly, KU508673-GPGV-Goldfinger displayed an extreme divergence within the first 150 nt of the genome compared with other sequences, due to a stretch of 76 nt (61 to 137 from the ORF1-ATG), which was highly homologous to a fragment of the grapevine genome. The most likely explanation for this unusual situation is that the Goldfinger sequence was assembled by the mapping of siRNAs on the initial FR877530.1 GPGV reference sequence (Giampetruzzi et al. 2012) which also contained this inaccuracy but has since been corrected in GenBank (FR877530.2). However, we were unable to test or confirm this hypothesis, so the original KU508673-Goldfinger sequence was retained as such for further analyses.

Nucleotide diversity (π) was assessed and plotted along the genome. The GPGV sequences from Asia and, specifically, from

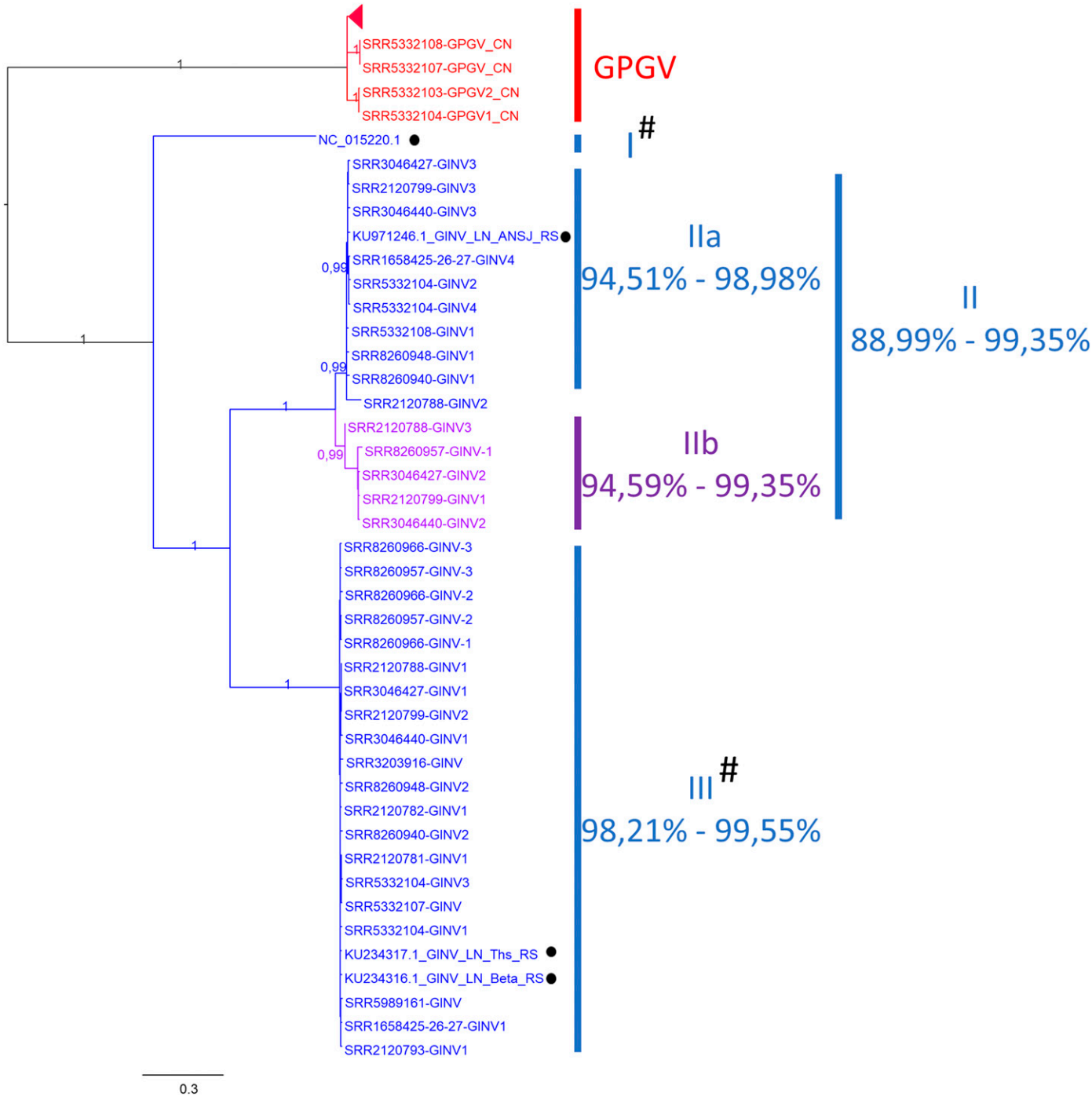

Fig. 2. Maximum-likelihood tree inferred from the sequences covering all open reading frame of grapevine inner berry necrosis virus (GINV). Thirty-nine sequences were used. Black dot (●) indicates sequences already available in NCBI, while all others were newly de novo assembled from high-throughput sequencing (HTS) datasets. Number at each node indicates bootstrap percentages based on 100 replicates. The scale bar corresponds to the number of substitutions per site. Grapevine Pinot gris virus (GPGV) sequences were used as outgroup (in red). A newly described subclade is highlighted in purple (IIb). Identity percentage for each clade is on the right. Details of the isolates are given in Supplementary Table S1. # indicates clade whose sequences display shorter open reading frame 1 (ORF1) and an additional ORF on the negative strand located at the 3′-end of ORF1 before the start of ORF2 (Fig. 3).

China present the highest genetic diversity ($\pi = 0.0498$ and 0.0510, respectively) (Table 1) compared with the GPGV overall genetic diversity ($\pi = 0.0263$) (Table 1). In addition, the genetic diversity in the Asian GPGV population was always greater than that of isolates from any other geographic area, at any point along the genome (Fig. 1 and Supplementary Fig. S1).

A phylogenetic analysis was performed (Fig. 4) in which sampling countries are presented by different colors. Although there are some exceptions, there is a clear tendency for isolates from the same geographic regions to cluster together, highlighting a potential geographical population structure. This assumption was confirmed by measuring genetic differentiation by geographic origin (fixation index, F_{ST}). When grouping isolates based on their sampling origin (by continent: Americas, 6 sequences; Asia, 18 sequences; Europe, 101 sequences or by region: France [FR], 56 sequences; Italy [IT], 39 sequences; Eastern Europe (regrouping Germany [GE] and Slovakia [SK]), 6 sequences; and China [CN] with 16 sequences), and analyzing the three continents and all five regions in a pairwise comparison, most F_{ST} were statistically significant (Table 2). When considering only the continents, the highest F_{ST} values always involved comparisons with Asia, with $F_{ST} = 0.2079$ ($P \leq 10^{-5}$) between Europe and Asia, and $F_{ST} = 0.1478$ ($P \leq 10^{-5}$) between Asia and the Americas. When considering narrower geographical regions of the world, similar conclusion were drawn, with all computations confirming a strong geographic structuration (Table 2). The lowest F_{ST} value (and the only one not statistically significant) was by comparing GPGV population between France and the Americas, indicating that these two populations are genetically indistinguishable when using the present dataset.

When focusing on Tajima's D values (D_T), the difference in predicted evolutionary scenario between continents, such as Asia and Europe, was striking (Fig. 1 and Table 1). Asian isolates displayed all along their genome D_T values close to 0 (with an average of $D_T = -0.9424$), suggesting that this population evolved under balancing selection (Fig. 1). This observation was especially true when focusing on sequences obtained from China (Supplementary Fig. S1 and Table 1), with an average D_T of –0.8573 and with only a few sites under strong selection located in the first 150 nt at the 5′-end of the sequence (Fig. 1). This 5′-end-specific effect is seemingly an artifact due to the aforementioned highly divergent region of KU508673-Goldfinger since it is no longer observed when excluding this sequence from the analysis (data not shown). On the other hand and for most other geographic regions with a minimum of 16 isolates analyzed (i.e., Europe, France, and Italy), D_T values were much lower, with the extreme being an average $D_T = -2.1001$ for France (Table 1). In addition, almost all sites along

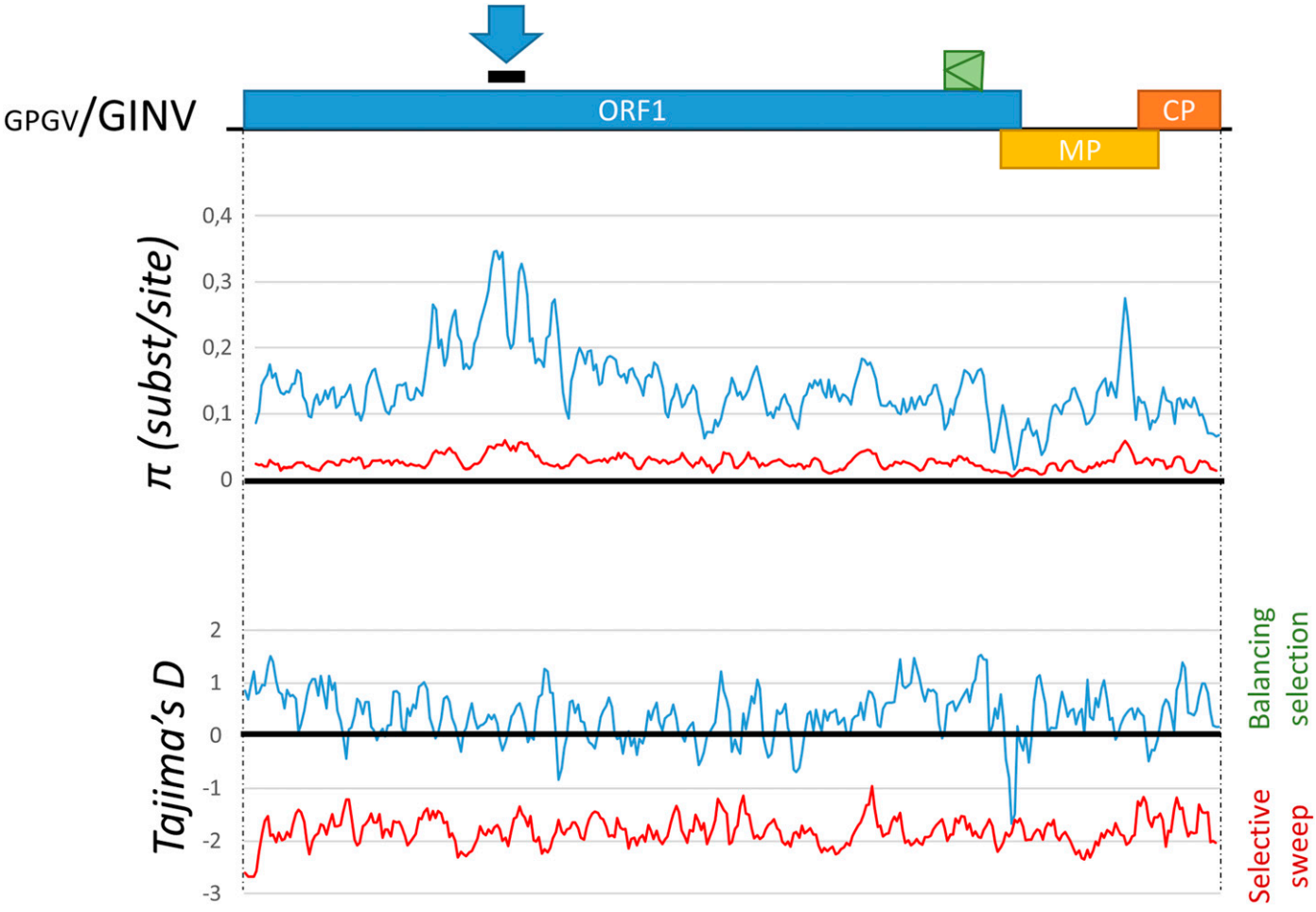

Fig. 3. Genetic diversity analyses of both grapevine berry inner necrosis virus (GINV) (blue) and grapevine Pinot gris virus (GPGV) (red) using a corpus of 39 and 126 sequences, respectively, covering all three open reading frames (ORFs). Graphics represent π (substitution/sites) along the genome and Tajima's D (D_T) for evolution study, comparing sequences from both viruses. Average of π and D_T are provided in Table 1. Arrow corresponds to all InDels location and the green square corresponds to the small ORF found in negative strand for GINV only. Sites under selection are listed in Supplementary Tables S7 and S8.

the genome were significantly under selection (Fig. 1 and Supplementary Table S7). These low D_T values and sites under selection suggest either a recent selective sweep or bottleneck of the GPGV populations in these regions, possibly resulting from the recent introduction of this virus in these areas.

The possible existence of recombination events in the analyzed GPGV sequences were evaluated using RDP4 (Table 3). In total, 10 sequences were identified as potential recombinants, accounting for a total of 22 breakpoints. The majority (13) of these events were located in ORF1, while only one was detected in the MP. A recombination hot-spot (8) was revealed at the end of the CP (nucleotides 6950 to 6980 from the ORF1-ATG). Because recombination can generate confounding effects in phylogeographic reconstructions (Ruths and Nakhleh 2005), recombinant isolates were removed from the dataset for further analyses.

Reconstruction of a phylogeographic evolution scenario for GPGV advocating for an Asian origin. The linear regression exploration of root-to-tip distances as a function of sampling time revealed a moderate temporal signal in the dataset (correlation coefficient = 0.41, $R^2 = 0.17$, $P = 5.93 \times 10^{-6}$; Supplementary Fig. S4). The presence of temporal signal was not confirmed consistently with the date-randomization tests implemented in BEAST, as some permutation results overlapped the real (nonpermutated) data (Supplementary Fig. S5). Hence, these results indicate that the date (but not the locations) have to be taken thereafter with caution.

By mapping the tip locations of the MCC tree in geographic space (Fig. 5, Supplementary Table S6), the most basal nodes of the phylogeny (i.e., from node 1 to node 2), support an Asian geographical origin of GPGV. This result is consistent with the topology of the ML tree obtained with the full GPGV dataset (126 sequences) (Fig. 4) and also with the higher diversity observed for GPGV in Asia (Table 1). The root of the MCC tree, represented by node 1, was dated at the 19th century (Supplementary Table S6). Then, the intervals between nodes 2 and 10 identified three independent GPGV transmissions from Asia to Europe that occurred during the mid-20th century. While the nodes 3 and 4 are not resolving the geographic locations ($P_{Asia} = 0.42$; $P_{Europe} = 0.58$ and $P_{Asia} = 0.41$; $P_{Europe} = 0.59$, respectively, Supplementary Table S6), the different introduction events are revealed as followed: i_1 (Fig. 5) is revealed by the interval between nodes 5 ($P_{Asia} = 0.76$; $P_{Europe} = 0.24$) and 6 ($P_{Asia} = 0.03$; $P_{Europe} = 0.97$), the i_2 event by the interval between nodes 2 ($P_{Asia} = 0.77$; $P_{Europe} = 0.23$) and 7 ($P_{Asia} = 0.00$; $P_{Europe} = 1.00$), and the i_4 event by the interval between nodes 2 and 9 ($P_{Asia} = 0.00$; $P_{Europe} = 1.00$) (Fig. 5 and Supplementary Table S6). After a phase of diversification and dispersion of the GPGV populations throughout Europe, two independent transmission events were identified from Europe to Asia in the late 20th century and in the early 21st century with node 8 (i_3, $P_{Asia} = 0.01$; $P_{Europe} = 0.99$) and with the interval between nodes 15 ($P_{Asia} = 0.00$; $P_{Europe} = 1.00$) and 16 ($P_{Asia} = 1.00$; $P_{Europe} = 0.00$) for i_8 (Fig. 5 and Supplementary Table S6). In parallel, at least three independent transmissions of GPGV from Europe to the New World were reported by the interval between nodes 10 ($P_{Europe} = 1.00$; $P_{NewWorld} = 0.00$) and 11 ($P_{Europe} = 0.00$; $P_{NewWorld} = 1.00$) for i_5, by node 12 ($P_{Europe} = 1.00$; $P_{NewWorld} = 0.00$) for i_6, and by node 14 ($P_{Europe} = 1.00$; $P_{NewWorld} = 0.00$) for i_7 (Fig. 5 and Supplementary Table S6).

The genetic diversity of GINV compared with that of GPGV. Compared with GPGV, the genetic diversity of GINV is much higher, with an average pairwise nucleotide diversity percentage as high as 28.4% (24.2% when removing the unique Japanese sequence and considering Chinese isolates only). A phylogenetic analysis using the complete genome (minus the UTRs) was performed (Fig. 2 and Supplementary Fig. S6), using GPGV as an outgroup. GINV sequences grouped into three well-supported clades, as previously described using partial genome sequences (Fan et al. 2017), with the stand-alone clade I represented by the sole Japanese GINV reference genome (NC_015220), and the two other clades (II and III) comprising 10+ sequences, all from China. Except for clade I, the isolates within each clade came from a minimum of eight different grapevine varieties. Clade II can be further separated into two subclades (Fig. 2), with clade IIb being newly described here. Within each clade/subclade, sequences were quite homogeneous, displaying a maximum diversity of 5.5%. However, genetic diversity is very high between clades (>23% interclade diversity).

When plotting the nucleotide diversity (π) along the genome, it is clear that genetic diversity is very different between the two

TABLE 1
Genetic diversity analyses of grapevine Pinot gris virus (GPGV) and grapevine berry inner necrosis virus (GINV) with number of sequences (*N*), π ± standard error (SE), and Tajima's *D* (D_T) with associated *P* values[a]

Virus	Population	*N*	π ± SE	D_T
GPGV	Overall	126	0.0263 ± 0.0007	–2.0690 ($P < 0.05$)
	Asia	18	0.0498 ± 0.0014	–0.9424 ($P > 0.10$)
	Europe	101	0.0211 ± 0.0006	–2.0909 ($P < 0.05$)
	Americas	6	0.0090 ± 0.0008	–0.3616 ($P > 0.10$)
	France, FR	56	0.0169 ± 0.0016	–2.1001 ($P < 0.05$)
	Italy, IT	39	0.0224 ± 0.0010	–1.5897 ($0.10 > P > 0.05$)
	Eastern Europe, EE	6	0.0151 ± 0.0009	–0.7799 ($P > 0.10$)
	China, CN	16	0.0510 ± 0.0014	–0.8573 ($P > 0.10$)
	RoW	110	0.0212 ± 0.0016	–2.1119 ($P < 0.05$)
GINV	Overall	39	0.1402 ± 0.0020	0.4074 ($P > 0.10$)
	China, CN	38	0.1329 ± 0.0020	1.6275 ($P > 0.10$)

[a] *N* corresponds to the number of sequences, π corresponds to the diversity index ± standard error, D_T corresponds to Tajima's *D* with associated *P* value, Asia corresponds to sequences from China and Pakistan, Eastern Europe corresponds to sequences from Germany and Slovakia, RoW (rest of world) corresponds to all sequences minus China. Asia is composed of sequences from China and Pakistan, Europe with sequences from France, Italy, and Eastern Europe, and the Americas with sequences from the United States, Canada, and Uruguay.

grapevine-infecting trichoviruses (Fig. 3). While π follows the same pattern along the genomes of GPGV and GINV, the average π value is much greater in the case of GINV (over fivefold; π-GINV = 0.1402 and π-GPGV = 0.0263). Removing the highly divergent Japanese reference sequence did not drastically change the outcome (π = 0.1329) (Supplementary Fig. S7). Only one major constrained region was detected in the GINV genome and was located in the overlapping region between ORF1 and ORF2.

A clear difference between GPGV and GINV was also observed when comparing Tajima's *D* values (Fig. 3), with GINV showing a

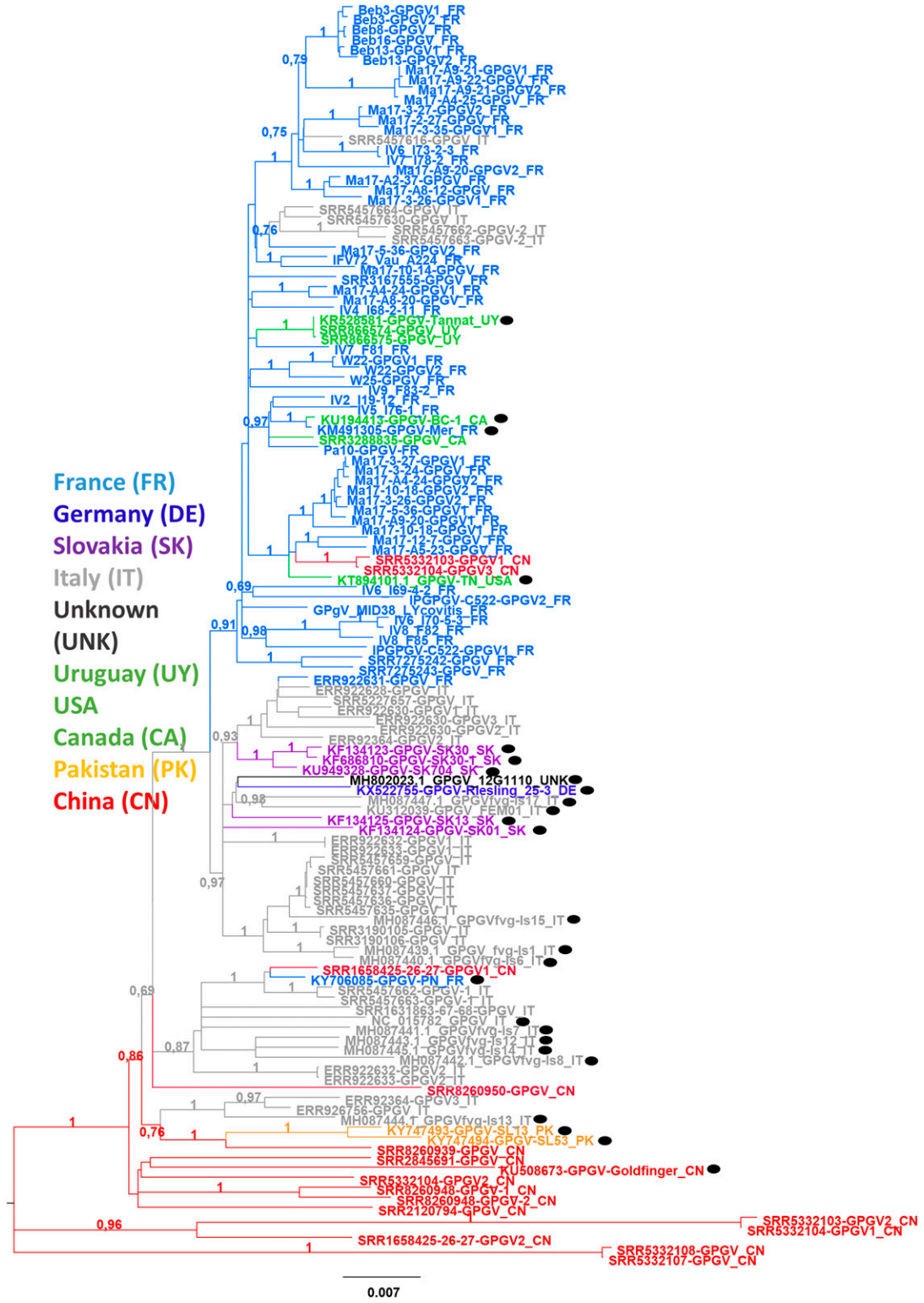

Fig. 4. Maximum-likelihood tree inferred from 126 sequences spanning 7,010 nt of grapevine Pinot gris virus genome. Black dot (●) indicates sequences already available in NCBI, while all others were newly de novo assembled from our own high-throughput sequencing datasets and from sequence read archive files. Number at each node indicates bootstrap percentages based on 100 replicates. The scale bar corresponds to the number of substitutions per site. Details of the isolates are given in Supplementary Table S1. Color code corresponds to sample collection location.

positive D_T value indicative of a balancing selection ($D_T = 0.4074$). The trend was even stronger after removing the unique divergent sequence from Japan ($D_T = 1.6275$) (Supplementary Fig. S7).

As for GPGV, a few intraspecies recombination events were detected using RDP4 (Table 3). Interestingly, a single interspecies recombination event was detected by six software of the RDP4 package, involving KU508673-Goldfinger, a GPGV sequence described as coming from China.

DISCUSSION

Knowledge on the genetic diversity of the two grapevine-infecting trichoviruses was so far mostly limited to the analysis of partial genomic sequences (Fan et al. 2017) or of a limited number of complete sequences (Tarquini et al. 2019). With the addition of 100 near complete sequences for GPGV, its known diversity increased from 2.8% (Tarquini et al. 2019) to up to 8.4% (this study). This result is mostly the consequence of the addition of novel, divergent isolate sequences retrieved from SRA files from Asia and, most prominently, from China. There is however no ambiguity that these divergent isolates belong to the GPGV species since their divergence level is below the various species discrimination criteria within the family *Betaflexiviridae* (https://talk.ictvonline.org/ICTV/proposals/2015.011a-adP.A.v2.Betaflexiviridae_rev.pdf) (Supplementary Tables S2 and S3).

Similarly, the addition of 69 complete GINV sequences greatly extends our knowledge of its diversity, in particular by providing a vision at the pan-genomic level. Although the existence of isolates forming two highly divergent clades from the reference Japanese isolate had been documented using partial sequence information (Fan et al. 2016a, 2017), the analyses presented here allowed the identification of a novel subclade from China (IIb) (Fig. 2). Whether the Chinese isolates should be considered as belonging to a separate virus species than the sole Japanese isolate, or as representing divergent strains of GINV is not easy to decide. Indeed, different answers are obtained when considering the two genes used to discriminate species in the family *Betaflexiviridae*. When comparing the CP gene (or the encoded protein), identity values fall unambiguously within the species boundary (Supplementary Table S5). However, when considering ORF1 and the REP protein (Supplementary Table S4), both nucleotide and amino acid identity values fall outside the species boundary. There are already several cases within the same family *Betaflexiviridae* (e.g., *Vitivirus* and *Foveavirus*), in which different conclusions can be reached depending on the gene considered. And the reverse situation has been described in the case of Asian prunus viruses 1, 2, and 3, with the REP sequences suggesting that they belong to the same species but the CP sequences arguing for different species (Marais et al. 2016).

With this work, we show that datamining, by providing access to a wide range of virus isolates, may be useful for describing the genetic diversity of a virus and for attempting to identify its center of origin. Although samples, for which SRA data are available, might not include all parts of the world and lack biological and symptomatology information, the results presented here show that a dataset covering a wide range of varieties from different grapevine growing regions could be assembled for both GPGV and GINV. While GPGV was identified for the first time in a vineyard located in the northern part of Italy (Giampetruzzi et al. 2012), it has since then been suggested that its origin lies in Eastern European countries, where the virus has been widely detected (Bertazzon et al. 2016). Contrary to these reports, we show here that the probable origin center of GPGV is located in Asia, with China being the most likely country of origin. This finding is of course influenced by the current dataset and might evolve with the acquisition of new sequences from other countries or other hosts in the future. Indeed, most Chinese isolates have a basal phylogenetic position (Fig. 4), even in trees rooted with other trichoviruses (Supplementary Fig. S8). In addition, isolates from China form a population with very distinctive characteristics all converging toward the hypothesis that China is the center of origin, such as (i) the highest diversity index ($\pi = 0.0510$, compared with the average $\pi = 0.0263$); (ii) a fairly neutral evolution pattern (DT-CN = –0.8573 and DT-CNMOG = –0.5405 after removal of outgroups, Supplementary Fig. S2) while GPGV populations from other geographic areas display signatures of a recent selective sweep or bottleneck (i.e., compatible with the hypothesis of recent introductions); and (iii) F_{ST} values indicating a significant genetic differentiation between the Chinese population and those from other regions of the world (Table 2). This Asian origin hypothesis for GPGV was supported by modeling the discrete location transitioning between Asia, Europe, and the Americas while the date to the most common ancestor was estimated in the middle of the 19th century. In addition, intercontinental jumps were identified between Asia and Europe and timed to the middle of the 20th century (Fig. 5). Introduction using a direct path or other routes cannot be confirmed confidently. However, introduction dates fit within the range of the resurgence of big national grapevine breeding programs in Europe, such as in France. With a large acreage of low quality *Vitis* spp. hybrids being removed in the early 1960s, the French wine industry mainly focused on clonal selection of *V. vinifera* (Reynolds 2015). In addition, many research programs were initiated, often dedicated to create cultivars resistant to fungal diseases through the successive introgression of factors from Asiatic wild-type *Vitaceae*

TABLE 2
Measurements of geographic population's differentiation (fixation index, F_{ST}), and associated statistics (*P* value)[a]

	Overall sequence (7,010 nt)	
Populations	F_{ST}	*P*
Europe versus Asia	0.2079	<0.0000
Europe versus Americas	0.0533	0.03604
Asia versus Americas	0.1478	<0.0000
FR versus IT	0.1788	<0.0000
FR versus CN	0.2762	<0.0000
FR versus Americas	0.0596	0.05405
FR versus EE	0.2470	<0.0000
IT versus CN	0.1534	<0.0000
IT versus Americas	0.1670	<0.0000
IT versus EE	0.0950	0.00901
CN versus Americas	0.1498	0.00901
CN versus EE	0.1317	0.01802
Americas versus EE	0.3509	0.00901
CN versus RoW	0.2163	<0.0000

[a] Three analyses were performed according to geographical regions comparing continents: Europe (101 sequences), Asia (18 sequences), and Americas (6 sequences); five regions in the world with France (FR), Italy (IT), and Eastern Europe (EE); and comparing China (CN) with the rest of the world (RoW). Fixation index, F_{ST}, and associated statistics (*P* value).

(http://observatoire-cepages-resistants.fr/wp-content/uploads/2017/11/fiche-OPECST-creation-france.pdf). Similar programs were started in Germany (Töpfer et al. 2011) and around Europe, promoting and accelerating exchanges of genetic material and resistance resources in the middle of the 20th century. A second wave of intercontinental jumps/introductions has been identified in the late 20th century and the early 21st century between Europe and Asia, and between Europe and the New World. These introductions could correspond to the increase in production in specific areas around the world, triggering important flux and importation of plant materials between producing countries. For example, as much as 4.5 million cuttings were exported from France to China in 1998 (http://www.fao.org/3/x6897e/x6897e05.htm), which could match with both "reintroductions" of GPGV in China with European variants (Fig. 5, Supplementary Table S6). The other 'reintroduction' in Asia was probably due to material importation by scientists (see comments associated with the datasets SRR5332103, SRR5332104, SRR5332107, and SRR5332108 in NCBI, mentioning that the two GPGV-infected varieties, Cabernet Franc and Merlot, initially originated from Livourne [Libourne] and the Loire Valley, France. Unfortunately, importing dates were not provided). This "reintroduction" hypothesis is supported by the fact that most other variants from Asia are divergent and basal isolates that mainly infect popular and important grapevine varieties in Asia such as Goldfinger, Summer Black, Gui fei me gvi, and Bai-Ji-Xin. As for the introduction of European variants of GPGV in the New World (i.e., the Americas), it is probably due to the constant transformation of vineyard with new plantings of international elites and well-accepted cultivars originating from the Old World (e.g., California's Zinfandel from Croatia and Italy, Argentina's Malbec and Chile's Carmenère, Cabernet Sauvignon, and Merlot from France). Increase of land dedicated to the wine industry has been exponential between 1990 and 2000 in the United States, exemplified by the Sonoma county (Merenlender 2000). In South America, a complete "wine revolution" took by storm many countries, such as Chile, Argentina in the 1980s and later on in Brazil and Uruguay.

TABLE 3
Recombination events were detected in grapevine Pinot gris virus (GPGV) and grapevine berry inner necrosis virus (GINV) using Recombination Detection Program package (RDP v.4.97) with RDP (R), GeneConv (G), Bootscan (B), MaxChi (M), Chimaera (C), Siscan (S), and 3Seq (3) software being included[a]

Virus species	Recombinant name	Breaking point location		Parents				RDP	
		Beginning	End	Major	% Similarity	Minor	% Similarity	Maximum clade corrected	Software
GPGV	KU508673-Goldfinger	6973	144	SRR8260948-GPGV1	96.60%	SRR5332103-GPGV2	–	6.09E-37	R,G,B,M,C,S,3
		[6867–end]	[128–248]						
	MH087445-ls14	6966	2187	MH087441-ls7	98.90%	MH087443-ls12	100.00%	3.34E-12	R,G,B,M,C,S,3
		[6772–end]	[2025–2275]						
	SRR1658425-26-27-GPGV2	4935	6910	SRR1658425-26-27-GPGV1	96.90%	SRR5332103-GPGV2	97.70%	1.07E-06	R,B,M,C,S,3
		[4807–4976]	[6888–end]						
	MH087442-ls8	7007	2122	SRR5457616-GPGV	–	MH087443-ls12	100.00%	3.97E-14	R,G,B,M,C,S
		[6895–end]	[1982–2190]						
	MH087447-ls17	7007	792	KU312039-FEM01	98.60%	MH087443-ls12	99.90%	1.00 E-09	R,G,M,S,3
		[6940–end]	[634–1031]						
	SRR1658425-26-27-GPGV2	738	2003	SRR5457664-GPGV	–	SRR1658425-26-27-GPGV1	100.00%	1.07 E-12	R,G,B,M,C,S,3
		[634–1028]	[1872–2029]						
	SRR5457662-GPGV2	6981	830	SRR5457630-GPGV	99.20%	SRR5457662-GPGV1	100.00%	4.38E-06	R,G,B,M,C,S,3
		[6858–end]	[676–914]						
	MH087441-ls7	6953	764	SRR1658425-26-27-GPGV1	98.00%	MH087439-ls1	100.00%	1.15E-04	R,G,B,M,S
		[6768–end]	[542–962]						
	MH087445-ls14	2186	5031	MH087443-ls12	–	MH087447-ls17	99.90%	3.03E-05	G,B,M,S
		[2030–2726]	[4477–5084]						
	ERR922632-GPGV2	4569	6097	NC_015782	97.70%	ERR922633-GPGV1	99.70%	2.16E-09	G,M,C,S
		[4412–4890]	[5895–6192]						
	ERR922633-GPGV2	4569	6089	NC_015782	97.70%	ERR922633-GPGV1	99.70%	3.30E-04	G,B,M,C,S
		[4411–4886]	[5891–6189]						
GINV	SRR2120788-GINV2	60	2495	SRR2120800-GINV3	99.80%	SRR2120788-GINV3	–	7.99E-102	R,G,B,M,C,S,3
		[0–72]	[2458–2516]						
	SRR2120788-GINV3	60	2495	SRR3046428-GINV2	99.40%	SRR2120800-GINV3	99.70%	2.00E-100	R,G,B,M,C,S,3
		[0–95]	[2458–2516]						
	SRR8260966-GINV1	3063	6783	SRR8260959-GINV1	100.00%	SRR8260968-GINV1	100.00%	1.27E-07	G,M,C,S,3
		[2596–3242]	[6270–end]						
	SRR8260966-GINV2	2942	6782	SRR8260968-GINV1	99.80%	SRR8260968-GINV2	99.90%	2.54E-06	M,C,S,3
		[2149–3194]	[6434–end]						
GPGV-GINV	KU508673-Goldfinger	7020	130	ERR922630-GPGV3	96.80%	SRR5332104-GINV1	–	7.37E-53	R,G,B,M,C,3
		[7005–end]	[127–149]						

[a] Information on major and minor parents are provided as well as breakpoint locations after removal of the untranslated regions.

This flourishing of the wine industry with the flux of European material importation would fit with the time windows given by the software. In addition, strong material exchange between Europe and the Americas was highlighted by the lowest fixation index between the two regions of the world, which was also the unique F_{ST} value that was not statistically supported (Table 2).

While both viruses display a highly conserved genomic organization, with minimal variation in genome and ORF sizes, the extent and structure of GINV and GPGV diversities are quite different. Phylogenetic analysis revealed that GINV shows a limited number of fairly tight (<5.5% intraclade diversity) but very distinct clusters (>23% interclade diversity). While also showing

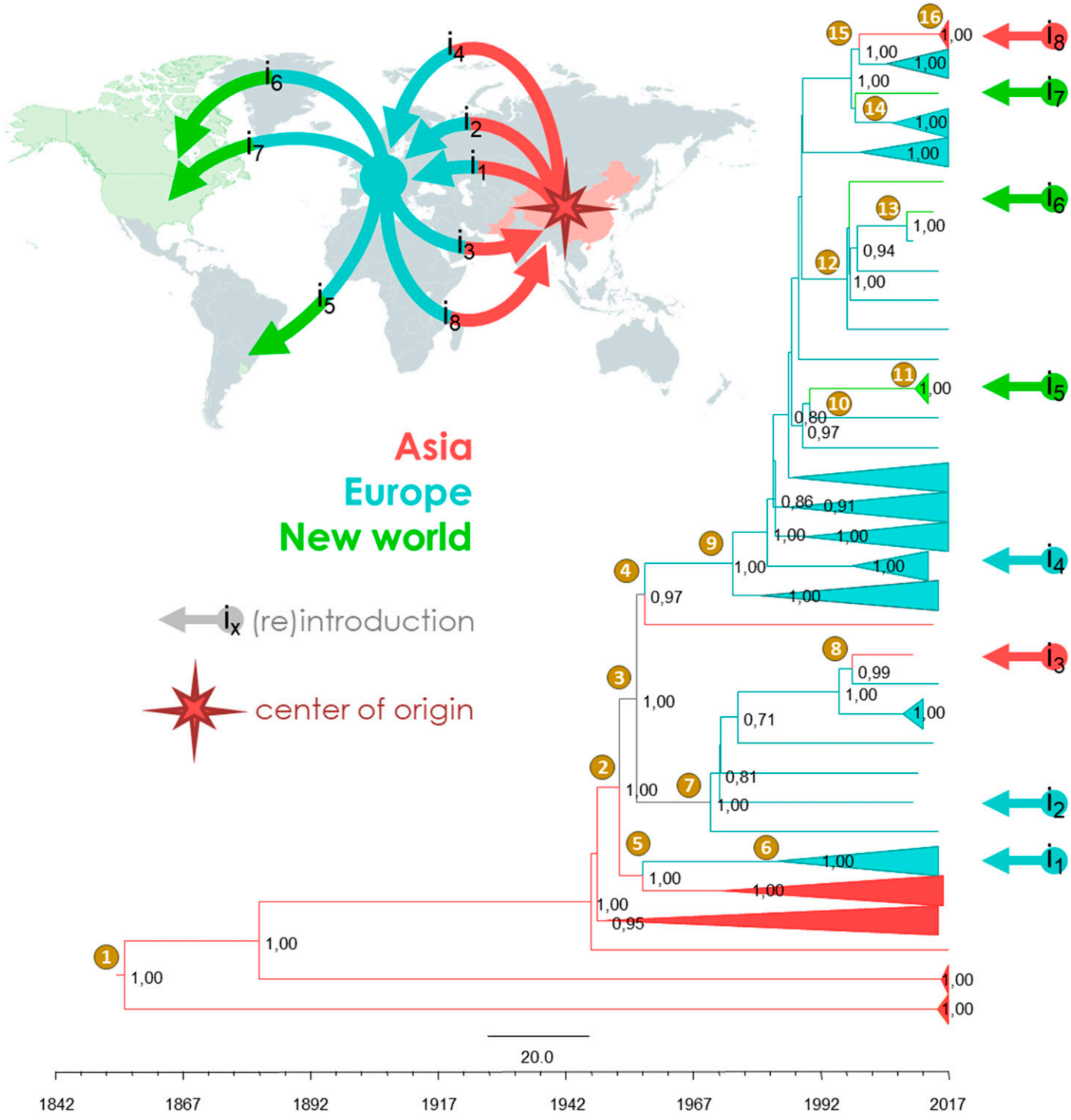

Fig. 5. Evolutionary history of grapevine Pinot gris virus (GPGV). The maximum clade credibility (MCC) tree was reconstructed from the full-length coding sequences of GPGV of 116 isolates. The values on the nodes correspond to their posterior probabilities. The bottom axis gives the timeframe (by date, from 1842 to 2017) of GPGV diversification. The geographic origin of sequences is indicated by colors (Asia, red; Europe, blue; and New world, green). The arrows (labeled from i_1 to i_8) show the branches demonstrating the movement of the GPGV between continents and the color indicates where the GPGV was (re)introduced. Sixteen nodes of interest (supported by posterior probabilities above 0.80) were annotated with their time to the most common ancestors and their 95% highest probability density intervals (Supplementary Table S6). Projection on the world map of the GPGV dispersion, according to the MCC tree. The center of origin of GPGV and the events of introduction in Asia, Europe, and New world are indicated.

some very distinct basal Asian clades (described previously), most GPGV isolates are found to cluster in a large clade with some evidence of geographical structuring and little intraclade diversity (<2%). This distinct structuration of viral diversity is paralleled by the very different geographical distributions of the two viruses. While GPGV seems to be found essentially everywhere in the world where grapevine is grown, the occurrence of GINV is so far to be restricted to Asia (China and Japan). Such differences for two closely related viruses that share the same host and have a similar mite-borne transmission mechanism are quite remarkable. If, as suggested by the evolutionary scenario reconstructed above, the presence of GPGV outside Asia is due to intercontinental transfers likely involving grapevine propagation materials, why has GINV not similarly spread and is not found outside Asia? A few plausible reasons could explain such differences. The first possibility that comes to mind is that given the symptoms it causes, quarantine and other controls measures could have successfully restricted GINV spread via contaminated grapevine materials, but that these measures could have been largely inefficient in the case of GPGV given that its infections are frequently asymptomatic if not latent. Indeed, in most countries where its presence has been described, GPGV has often been associated with infection showing no visible symptoms. The second hypothesis is based on the host range and transmission properties. GPGV is known to infect not only its natural host *Vitis vinifera*, but also other herbaceous plants, such as *Silene latifolia* and *Chenopodium album* (Gualandri et al. 2017), which are frequently encountered in vineyards around the world. To our knowledge, GINV is strictly restricted to grapevine. The ability of GPGV to infect and multiply in these herbaceous species may increase its persistence and spread, ultimately serving as reservoir(s) for the virus for further vector transmission to neighboring grapevines. In addition, both viruses share the same eriophyid vector species (Kunugi et al. 2000; Malagnini et al. 2016), *Colomerus vitis*, a monophagous mite known to feed only on grapevine. So far, no direct proof of field transmission for GPGV has been obtained, and its spread has been only suggested on the basis of aggregated patterns of GPGV symptoms in vineyards (Malossini et al. 2015). However, as GPGV infects other species than *V. vinifera*, this could potentially widen its vector panel, increasing its ability to be spread within a vineyard and ultimately accelerating its epidemic development. The possible existence of vectors of the virus other than *C. vitis* need to be explored. So far, all reports concerning GLMD disease are quite alarming and suggest it could pose a serious threat to the grapevine industry around the world. However, more work is needed if we want to attribute GLMD symptoms to specific GPGV variants. All these findings would provide tools and support material for stakeholders and decision-makers whether to include, or not, GPGV (or particular variants) in the certification scheme, which is not the case to date.

Conclusion. With this work, we demonstrate the importance of datamining to study the genetic diversity of two trichoviruses. In the case of GPGV, the assembled dataset allows us to propose a scenario describing its dispersion history over the world from an Asian center of origin (Fig. 5). Important plant material movements over the years can track down most major introductions of GPGV in Europe and in the New World. These dissemination risks need to be considered when designing control strategies, not only for GPGV, but also for viruses in general, especially for plants that can be easily propagated by cuttings. Such a strategy could likely involve the implementation of a certification scheme using for example unbiased HTS-based detection methodologies, which has proven quite sensitive for viral detection.

ACKNOWLEDGMENTS

We thank Martin Drucker for valuable comments and proofreading of the manuscript and the IRD i-Trop HPC located in Montpellier for providing HPC resources that have contributed to the research results reported within this article.

LITERATURE CITED

Al Rwahnih, M., Golino, D., and Rowhani, A. 2016. First report of *Grapevine Pinot gris virus* infecting grapevine in the United States. Plant Dis. 100:1030.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., and Suchard, M. A. 2011. BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61:170-173.

Bertazzon, N., Filippin, L., Forte, V., and Angelini, E. 2016. Grapevine Pinot gris virus seems to have recently been introduced to vineyards in Veneto, Italy. Arch. Virol. 161:711-714.

Bertazzon, N., Forte, V., Filippin, L., Causin, R., Maixner, M., and Angelini, E. 2017. Association between genetic variability and titre of *Grapevine Pinot gris virus* with disease symptoms. Plant Pathol. 66:949-959.

Beuve, M., Candresse, T., Tannières, M., and Lemaire, O. 2015. First report of *Grapevine Pinot gris virus* (GPGV) in grapevine in France. Plant Dis. 99:293.

Cho, I. S., Jung, S. M., Cho, J. D., Choi, G. S., and Lim, H. S. 2013. First report of *Grapevine Pinot gris virus* infecting grapevine in Korea. New Dis. Rep. 27:10.

Diaz-Lara, A., Navarro, B., Di Serio, F., Stevens, K., Hwang, M. S., Kohl, J., Vu, S. T., Falk, B. W., Golino, D., and Al Rwahnih, M. 2019. Two novel negative-sense RNA viruses infecting grapevine are members of a newly proposed genus within the family Phenuiviridae. Viruses 11:685.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969-1973.

Duchêne, D. A., Duchêne, S., Holmes, E. C., and Ho, S. Y. W. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol. Biol. Evol. 32:2986-2995.

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

Fan, X., Hong, N., Zhang, Z., Yang, Z., Ren, F., Hu, G., Li, Z., Zhou, J., Dong, Y., and Wang, G. 2016a. Identification of a divergent variant of grapevine berry inner necrosis virus in grapevines showing chlorotic mottling and ring spot symptoms. Arch. Virol. 161:2025-2027.

Fan, X. D., Dong, Y. F., Zhang, Z. P., Ren, F., Hu, G. J., Li, Z. N., and Zhou, J. 2016b. First report of *Grapevine Pinot gris virus* in grapevines in China. Plant Dis. 100:540.

Fan, X. D., Zhang, Z. P., Ren, F., Hu, G. J., Zhou, J., Li, Z. N., Wang, G., and Dong, Y. 2017. Occurrence and genetic diversity of *Grapevine berry inner necrosis virus* from grapevines in China. Plant Dis. 101:144-149.

Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C., and Rambaut, A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. 27:2038-2051.

Gazel, M., Caglayan, K., Elçi, E., and Öztürk, L. 2016. First report of *Grapevine Pinot gris virus* in grapevine in Turkey. Plant Dis. 100:657.

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., and Saldarelli, P. 2012. A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in cv. Pinot gris. Virus Res. 163:262-268.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. 2012. Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. Mol. Biol. Evol. 30:713-724.

Glasa, M., Predajňa, L., Komínek, P., Nagyová, A., Candresse, T., and Olmos, A. 2014. Molecular characterization of divergent grapevine Pinot gris virus isolates and their detection in Slovak and Czech grapevines. Arch. Virol. 159:2103-2107.

Gualandri, V., Asquini, E., Bianchedi, P., Covelli, L., Brilli, M., Malossini, U., Bragagna, P., Saldarelli, P., and Si-Ammour, A. 2017. Identification of herbaceous hosts of the *Grapevine Pinot gris virus* (GPGV). Eur. J. Plant Pathol. 147:21-25.

Jo, Y., Choi, H., Kyong Cho, J., Yoon, J.-Y., Choi, S.-K., and Kyong Cho, W. 2015. In silico approach to reveal viral populations in grapevine cultivar Tannat using transcriptome data. Sci. Rep. 5:15841.

King, A. M. Q., Adams, M. J., Cartens, E. B., and Lefkowitz, E. J. 2012. Family–Betaflexiviridae. Pages 920-941 in: Virus Taxonomy. Elsevier, San Diego.

Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33: 1870-1874.

Kunugi, Y., Asari, S., Terai, Y., and Shinkai, A. 2000. Studies on the grapevine berry inner necrosis virus disease, 2: Transmission of grapevine berry inner necrosis virus by the grape erineum mite, *Colomerus vitis* in Yamanashi. [Japan] Bull. Yamanashi Fruit Tree Experiment Stn. Jpn. 10:57-63.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. 2009. Bayesian phylogeography finds its roots. PLOS Comput. Biol. 5:e1000520.

Librado, P., and Rozas, J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452.

Malagnini, V., de Lillo, E., Saldarelli, P., Beber, R., Duso, C., Raiola, A., Zanotelli, L., Valenzano, D., Giampetruzzi, A., Morelli, M., Ratti, C., Causin, R., and Gualandri, V. 2016. Transmission of grapevine Pinot gris virus by *Colomerus vitis* (Acari: Eriophyidae) to grapevine. Arch. Virol. 161: 2595-2599.

Malossini, U., Bianchedi, P., Beber, R., Credi, R., Saldarelli, P., and Gualandri, V. 2015. Spread of GPGV-associated disease in two vineyards in Trentino (Italy). Pages 212-213 in: 18th Congress of ICVG, Ankara, Turkey.

Marais, A., Faure, C., and Candresse, T. 2016. New insights into Asian *Prunus* viruses in the light of NGS-based full genome sequencing. PLoS One 11: e0146420.

Martelli, G. P. 2017. An Overview on Grapevine Viruses, Viroids, and the Diseases They Cause. Pages 31-46 in: Grapevine Viruses: Molecular Biology, Diagnostics and Management. B. Meng, G. P. Martelli, D. A. Golino, and M. Fuchs, eds. Springer International Publishing, Cham.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1:1-5.

Merenlender, A. M. 2000. Mapping vineyard expansion provides information on agriculture and the environment. Calif. Agric. 54:7-12.

Murray, G. G. R., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., Holmes, M. A., Rambaut, A., and Welch, J. J. 2016. The effect of genetic structure on molecular dating and tests for temporal signal. Methods Ecol. Evol. 7:80-89.

Nourinejhad Zarghani, S., Hily, J. M., Glasa, M., Marais, A., Wetzel, T., Faure, C., Vigne, E., Velt, A., Lemaire, O., Boursiquot, J. M., Okic, A., Ruiz-Garcia, A. B., Olmos, A., Lacombe, T., and Candresse, T. 2018. Grapevine virus T diversity as revealed by full-length genome sequences assembled from high-throughput sequence data. PLoS One 13:e0206010.

Pleško, I. M., Marn, M. V., Seljak, G., and Žežlina, I. 2014. First report of *Grapevine Pinot gris virus* infecting grapevine in Slovenia. Plant Dis. 98:1014.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2:vew007.

Rasool, S., Naz, S., Rowhani, A., Golino, D. A., Westrick, N. M., Farrar, K. D., and Al Rwahnih, M. 2017. First report of *Grapevine Pinot gris virus* infecting grapevine in Pakistan. Plant Dis. 101:1958.

Reynard, J. S., Schumacher, S., Menzel, W., Fuchs, J., Bohnert, P., Glasa, M., Wetzel, T., and Fuchs, R. 2016. First report of *Grapevine Pinot gris virus* in German vineyards. Plant Dis. 100:2545.

Reynolds, A. G. 2015. Grapevine breeding in France–A historical perspective. Pages 65-76 in: Grapevine Breeding Programs for the Wine Industry. A. G. Reynolds, ed. Woodhead Publishing, Elsevier.

Ruths, D., and Nakhleh, L. 2005. Recombination and phylogeny: Effects and detection. Int. J. Bioinform. Res. Appl. 1:202-212.

Saldarelli, P., Giampetruzzi, A., Morelli, M., Malossini, U., Pirolo, C., Bianchedi, P., and Gualandri, V. 2014. Genetic variability of *Grapevine Pinot gris virus* and its association with grapevine leaf mottling and deformation. Phytopathology 105:555-563.

Tarquini, G., De Amicis, F., Martini, M., Ermacora, P., Loi, N., Musetti, R., Bianchi, G. L., and Firrao, G. 2019. Analysis of new grapevine Pinot gris virus (GPGV) isolates from Northeast Italy provides clues to track the evolution of a newly emerging clade. Arch. Virol. 164:1655-1660.

Terai, Y., Kunigi, Y., and Yanase, H. 1993. A new virus disease, grapevine berry inner necrosis with natural spread in Japan. In: Extended Abstracts 11th Meeting ICVG. P. Gugerli, ed. Federal Agricultural Research Station of Changins, Montreux, Switzerland.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Töpfer, R., Hausmann, L., Harst, M., Maul, E., and Zyprian, E. 2011. New Horizons for Grapevine Breeding. Pages 79-100 in: Methods in Temperate Fruit Breeding, Vol. 5, Special Issue 1. H. Flachowsky and M.-V. Hanke, eds. Global Science Books, Ltd., Ikenobe, Japan.

Wu, Q., and Habili, N. 2017. The recent importation of Grapevine Pinot gris virus into Australia. Virus Genes 53:935-938.

Xiao, H., Shabanian, M., McFadden-Smith, W., and Meng, B. 2016. First report of *Grapevine Pinot gris virus* in commercial grapes in Canada. Plant Dis. 100:1030.

Yoshikawa, N., Iida, H., Goto, S., Magome, H., Takahashi, T., and Terai, Y. 1997. Grapevine berry inner necrosis, a new trichovirus: Comparative studies with several known trichoviruses. Arch. Virol. 142:1351-1363.