Population Genome Sequencing of the Scab Fungal Species Venturia inaequalis, Venturia pirina, Venturia aucupariae and Venturia asperata

Bruno Le Cam,*1 Dan Sargent,† Jérôme Gouzy,‡ Joëlle Amselem,§ Marie-Noëlle Bellanger,* Olivier Bouchez,** Spencer Brown,†† Valérie Caffier,* Marie De Gracia,* Robert Debuchy,** Ludovic Duvaux,* Thibaut Payen,* Mélanie Sannier,* Jason Shiller,* Jérôme Collemare,* and Christophe Lemaire*

*UMR 1345 Institut de Recherche en Horticulture et Semences (IRHS), INRA, Université d’Angers, Agrocampus-Ouest, SFR 4207 QuaSaV, 49071, Beaucouzé, France, †Research and Innovation Centre, Fondazione Edmund Mach, via Mach 1, 38010, San Michele All’Adige, TN, Italy, ‡Laboratoire Interactions Plantes Micro-organismes [LIPM], Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France, §Unité de Recherche Génomique Informatique [URGI], INRA, Université Paris-Saclay, 78026, Versailles, France, **US1426, INRA, GeT-PlaGe, Genotoul, Castanet-Tolosan, France, ††UMR 8298, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, 3584 CT, Utrecht, The Netherlands

ORCID IDs: 0000-0001-5695-4557 (J.G.); 0000-0001-7124-3454 (J.A.); 0000-0003-2328-5769 (O.B.); 0000-0003-0960-0312 (L.D.); 0000-0001-8152-6642 (J.C.)

ABSTRACT The Venturia genus comprises fungal species that are pathogens on Rosaceae host plants, including V. inaequalis and V. asperata on apple, V. aucupariae on sorbus and V. pirina on pear. Although the genetic structure of V. inaequalis populations has been investigated in detail, genomic features underlying these subdivisions remain poorly understood. Here, we report whole genome sequencing of 87 Venturia strains that represent each species and each population within V. inaequalis. We present a PacBio genome assembly for the V. inaequalis EU-B04 reference isolate. The size of selected genomes was determined by flow cytometry, and varied from 45 to 93 Mb. Genome assemblies of V. inaequalis and V. aucupariae contain a high content of transposable elements (TEs), most of which belong to the Gypsy or Copia LTR superfamilies and have been inactivated by Repeat-Induced Point mutations. The reference assembly of V. inaequalis presents a mosaic structure of GC-equilibrated regions that mainly contain predicted genes and AT-rich regions, mainly composed of TEs. Six pairs of strains were identified as clones. Single-Nucleotide Polymorphism (SNP) analysis between these clones revealed a high number of SNPs that are mostly located in AT-rich regions due to misalignments and allowed determining a false discovery rate. The availability of these genome sequences is expected to stimulate genetics and population genomics research of Venturia pathogens. Especially, it will help understanding the evolutionary history of Venturia species that are pathogenic on different hosts, a history that has probably been substantially influenced by TEs.

KEYWORDS Venturia Fusciadium formae specialis transposable elements effectors apple scab apple pear

Fungal species from the Venturia Sacc. genus (anamorph is Fusciadium Bonord.; Beck et al. 2005) are pathogenic on important Rosaceae host crops including apple and pear (Sivanesan, 1977). They are responsible for scab diseases which dramatically reduce the economic value of fruits by causing black lesions on the skin (Bowen et al. 2011). Taxonomically, scab fungi belong to the Dothideomycetes class and to the order Venturiales within the Venturiaceae family (Zhang et al. 2011).

Scab caused by Venturia inaequalis Cooke (Wint.) is the main fungal disease on apple, control of which may require over 20 chemical treatments each year. V. inaequalis has been extensively studied as a model for pathogens on woody hosts, particularly because it is amenable to in vitro crossing (Bénaouf and Parisi, 2000). V. inaequalis has been described on several genera of Rosaceous including Malus, Pyracantha, Eriobotrya and Sorbus (Sivanesan 1977; Jones and Aldwinckle 1990).
In some cases, host specificity occurs; for instance strains from *Pyracantha* are not able to infect *Malus* and vice-versa, which is referred to as *formae specialis f. sp pyracanthae* and *f. sp pomi*, respectively (Le Cam et al. 2002). Worldwide sampling of *V. inaequalis* strains has identified divergent populations, mostly reflecting their hosts: wild apple (*Malus sieversii*, *M. orientalis*, *M. sylvestris*, *M. floribunda*) and domesticated apple (*Malus x domestica*) trees (Gladieux et al. 2008; Gladieux et al. 2010a; Gladieux et al. 2010b; Leroy et al. 2013; Ebrahimi et al. 2016; Leroy et al. 2016). Certain populations have experienced a secondary contact with subsequent gene flow, resulting in the introduction from a wild population into the agricultural compartment of strains that are virulent on resistant cultivars carrying the *Rvi6* resistance gene (Guérin et al. 2007; Lemaire et al. 2016). On apple, scab can also be caused by another *Venturia* sp., *V. asperata* initially described as a saprophyte (Samuels and Sivanesan, 1975), and recently reported as a pathogen on apple trees (Caffier et al. 2012). Other *Venturia* species cause scab damages on their respective Rosaceae hosts, *V. pirina* (Aderh.) on European pear (*Pyrus communis*), *V. nashicola* on Asian pear (*Pyrus pyrifolia*), and *V. carpophila* on apricot (Crous et al. 2007; Sivanesan 1977). Other scab fungi are pathogenic on various tree hosts, including *Fusarium effusum* on pecan, *F. oleaginum* on olive, *V. populina* on poplar, *V. salicicarpa* on willow and *V. fraxini* on ash, among many others (Sivanesan 1977; Zhang et al. 2011; Young et al. 2018). Despite being economically important pathogens, the genetic basis of the interactions between scab fungi and their respective hosts remains poorly understood (Singh et al. 2005; Steiner and Oerke 2007; Kucheryava et al. 2008; Bowen et al. 2009; Mesarich et al. 2012; Cooke et al. 2014; Shiller et al. 2015; Deng et al. 2017).

Genome sequences are already available for *V. inaequalis f. sp. pomi*, *V. inaequalis f. sp. eriobotrya*, *V. pirina* and *F. effusum*, and they revealed that *V. inaequalis* strains tend to have a higher repeat content compared to other species, with up to 47% of their genome sequence covered by repetitive elements (Shiller et al. 2015; Bock et al. 2016; Deng et al. 2017; Passey et al. 2018). These studies highlighted in *Venturia* genomes the expansion of putative effector gene families, genes that are known to determine plant-fungus interactions (Shiller et al. 2015; Deng et al. 2017). The high repeat content and the employed Illumina sequencing strategy resulted in assemblies with between 1,012 and 1,680 scaffolds (Shiller et al. 2015; Deng et al. 2017) and up to 231 contigs using the PacBio technology (Passey et al. 2018). Genetic studies need higher quality genome assemblies and genomic population studies require re-sequencing of many strains from diverse populations.

Here, we report the PacBio genome assembly of the *V. inaequalis* EU-B04 reference strain and re-sequencing of 79 strains from two *formae specialis* and from diverse populations within *V. inaequalis f. sp. pomi*. We also present re-sequencing of *V. pirina* strains and whole genome sequencing of *V. aucupariae* and *V. asperata*. Availability of these genome sequences will greatly facilitate genetic and genomic population studies to better understand the evolution of *Venturia* pathogens and their interactions with different hosts.

**MATERIALS AND METHODS**

**Origin of Venturia strains**

All 87 *Venturia* strains used in this study are described in Table S1. Forty strains of *V. inaequalis* sampled in Kazakhstan on wild apple *M. sieversii* were used, of which half belong to the wild CAM (Central Asian Mountains) population and the other half to the domestic CAP (Central Asian Plains) population as described in Gladieux et al. (2010b). Other *V. inaequalis* strains were isolated from *M. x domestica* cultivars that do not carry the *Rvi6* resistance gene (6 domEU and 2 domASIA), from *M. x domestica* cultivars carrying the *Rvi6* resistance gene (5 domRvi6), from ornamental *M. floribunda* (4 FLO), from wild *M. sylvestris* (6 SYL) and *M. orientalis* (6 ORI), from *Pyracantha* (5 PYR), from *E. japonica* (5 LOQ) and from *Sorbus aucuparia* (1). In addition, strains of *V. pirina* (3), *V. aucupariae* (1) and *V. asperata* (3) were isolated from *Pyrus communis* (pear), *S. aucuparia* and *M. x domestica*, respectively (Table S1).

All strains were grown at 18°C on cellophane membranes displayed on PDA-YE (Potato-Dextrose Agar (BD Difco) supplemented with 0.3% Yeast Extract (Sigma-Aldrich)) plates with a 12h/12h light/night regime. They are stored at -20°C on dried cellophane membranes in the *Venturia* collection of the Institut de Recherche en Horticulture et Semences (IRHS, Beaucouzé, France).

**Genomic DNA extraction, library construction and DNA sequencing**

High molecular weight genomic DNA from *V. inaequalis* EU-B04 strain was extracted from mycelium grown on PDA-YE covered with cellophane membrane using the method described in Cheeseman et al. (2014) (extraction with CTAB followed by purification on an isopycnic gradient), except that ethidium bromide was used instead of DAPI for the purification of genomic DNA on a cesium chloride gradient. Ethidium bromide was preferred because it resulted in a single band of genomic DNA, while DAPI resulted in the dispersion of genomic DNA in several bands of different AT contents.

Genomic DNA quantity and quality were validated at the GeT-PlaGe core facility (Toulouse, France) using Qubit (Thermofisher), Nanodrop (Thermofisher), and Fragment Analyzer (Advanced Analytical) before purification on 0.5X AMPure PB beads and shearing at 40 kb using the Megaruptor system (Diagenode). Libraries were prepared according to the “20-kb Template Preparation Using BluePippin (Sage Science) Size-Selection System (15-kb size cutoff)” PACBIO protocol using 10 µg of sheared genomic DNA and the SMRTBell template Prep Kit. DNA fragments from 10 kb to 90 kb were selected. The sequencing run was performed on a PacBio RSII sequencer with the OneCellPerWell Protocol (OCPW) on C4 P6 chemistry and 360 min movies. Another PacBio sequencing was performed at the Cold Spring Harbor core genomics facility (CSHL, NY, USA); genomic DNA was sheared and purified, and a long-insert library was prepared following the protocol described in Quail et al. (2012). The library was sequenced on four SMRT cells using the C3 P6 chemistry using 2x 60 min movies.

Genomic DNA of all other *Venturia* strains was extracted from mycelium grown on PDA-YE plates using Möller’s protocol (Möller et al. 1992). Libraries for CAM and CAP strains were prepared using SPRworks- Fragment Library System I (Beckman Coulter) and sequenced at GATC Biotech (Konstanz, Germany) on HiSeq 2000 using TruSeq Multiplex Sequencing Primer and TruSeq PE (2x50bp)
(Illumina, San Diego, CA). Libraries for the remaining strains were prepared using the TruSeq Nano kit (Illumina, San Diego, CA) and sequenced at the GeT-PlaGe core facility (Toulouse, France) using SBS v3 (2x100 pb) or v4 (2x125 pb) reagent kits and using HiSeq 2000 or HiSeq 2500. For library construction, genomic DNAs were fragmented by sonication and sequencing adapters were ligated. Libraries were amplified with eight PCR cycles; their quality was assessed using a Bioanalyzer (Agilent) and quantity measured by quantitative PCR using the Kapa Library Quantification Kit (KapaBiosystems).

**Genome assembly**

For the PacBio genome, the PBcR wgs8.3rc2 assembly pipeline (Berlin et al. 2015) was used to perform read correction and assembly. The reads had high sequencing depth and quality of raw data allowed modifying several parameters to increase stringency: i) only raw reads longer than 5 kb were considered; ii) 50x of corrected reads were collected (assembleCoverage = 50); iii) the runCA part of the pipeline was run with a minimum overlap length of 1 kb (ovlMinLen = 1000); and iv) the unitigger bogart was tuned to consider overlaps with a very low error rate only (utgErrorRate = 0.0075 and utgMergeErrorRate = 0.04). Spurious contigs that were fully included in larger contigs were removed using a script based on ncbi-megablast to perform the all-vs-all analysis, combined with additional filters (minimum identity of 93%, minimum overlap of 500 nt and maximum overhang length of 500 nt). Contaminant contigs were identified using ncbi-blastx search against the NR database (downloaded in May 2016) and manual inspection. A total of 46 contigs were removed (maximum length of 113 kb) based on best hits with bacterial proteins confirmed with a GC percent analysis. Finally, the genome was polished with quiver (Chin et al. 2013) using stringent alignment cutoffs for pbalign (–minLength 3000–maxHits 1), ensuring a mapping as specific as possible. Assembly of other genomes were performed by two algorithms: SOAP-deNovo2 (Zerbino and Birney 2008) and Velvet (Luo et al. 2012) adapted for de novo genome assembly using next-generation sequencing (NGS) short reads.

Genome statistics were determined using Quast 2.3 (Gurevich et al. 2013). The GC content and distribution in each genome was analyzed using OccultorCut (Testa et al. 2016). Completeness of genome assemblies was assessed using BUSCO v1.1b1 (Simão et al. 2015) and the fungal BUSCO profile (busco.ezlab.org/v1/files/fungi_buscos.tar.gz). Fungal BUSCO genes are expected to be core genes, i.e., they are present in most fungal species. Thus, the absence of BUSCO genes in a genome assembly is indicative of an assembly that is likely not complete. The program was run in “genome assembly” mode using the metaparameters of Fusarium graminearum for AUGUSTUS gene prediction.

**SNP discovery, false discovery analysis and population structure of Venturia spp.**

For all strains, genome sequencing reads were cleaned using Trimmomatic 0.36 (Bolger et al. 2014) before mapping on the EU-B04 PacBio reference genome using BWA-MEM (Li and Durbin 2009). For species other than V. inaequalis, read mapping was refined with Stampy 1.0.29 (Lunter and Goodson 2011) using the parameter “–bamkeepgoodreads” and the parameter “–substitutionrate” was set to 0.15, 0.075, and 0.07 for V. asperata, V. pirina and V. aucupariae, respectively. Multisample SNP discovery was performed using the Unified Genotyper module of GATK 3.6 (McKenna et al. 2010), applying indel realignment and duplicate removal according to GATK Best Practices recommendations (DePristo et al. 2011; Van der Auwera et al. 2013). Instead of applying any posterior quality score recalibration and standard hard filtering, we performed a SNP false discovery rate (FDR) analysis making use of clones present in our dataset (Table S1). The FDR was estimated by simply counting the number of SNPs for each pair of clones. All steps but the Trimmomatic, Stampy and FDR analysis were integrated in a pipeline set using the TOGGLE toolbox (Monat et al. 2013). In order to assess the discrimination power of our resulting true positive SNP dataset, we computed a PCA of all the non-clonal Venturia inaequalis strains using PLINK v1.90b3.44 (Purcell et al. 2007).

**RNA extraction and sequencing**

V. inaequalis EU-B04, V. pirina VP36 and V. asperata 2349 strains were grown on seven different agar media: PDA-YE, apple medium (10% organic apple sauce in water), and Vogel media (Vogel, 1956) without carbon or nitrogen source, or supplemented with 1% glucose or 1% saccharose (with and without 80 μM zebulinarin). Total RNAs were extracted using the kit Nucleosip RNA Plant (Macherey-Nagel, Düren, Germany), according to the manufacturer’s recommendations. Total RNAs from complex media and from defined medium were combined in two respective pools. Paired-end 2x150 bp libraries were prepared using the Illumina TruSeq Stranded mRNA kit (Illumina, San Diego, CA) and sequenced at the GeT-PlaGe core facility (Toulouse, France) using a paired-end read length of 2x150 pb with the Illumina HiSeq3000 sequencing kits and using HiSeq 3000. Libraries were prepared according to Illumina’s protocols on a Tecan EVO200 liquid handler. Messenger RNAs were selected using poly-T beads, fragmented to generate double stranded cDNA; and then sequencing adaptors were ligated. Libraries were amplified with ten PCR cycles. Their quality was assessed using a Fragment Analyzer (Advanced Analytical) and quantity measured by quantitative PCR using the Kapa Library Quantification Kit (KapaBiosystems).

**Flow cytometry**

Analyses were performed according to Marie and Brown (1993). About 1 cm² of mycelium grown on PDA-YE plate was scraped and resuspended in 500 μL of cold nuclear buffer (45 mM MgCl2, 30 mM sodium citrate, 20 mM MOPS pH 7, 0.02% Triton X-100, 1% (w/v) cold polyvinyl pyrrolidone, 5 mM sodium metabisulfit, 0.05 mg/mL RNase (Roche, Basel, Switzerland)). A ligule of Arabidopsis thaliana cv. Columbia-0 was added as an internal standard. Samples were cut with a razor blade, filtered with 30 μm CellTrics paper (Sysmex Partec, Kobe, Japan) and stained with 50 μg/mL propidium iodide (Fluka). After 5 min incubation at 4°C, cytometric analysis was performed on 20,000 nuclei using a CyFlo3 cytometer (Sysmex Partec, Kobe, Japan). Intensity ratio between 1C fungal nuclei and 2C (=0.33 pg, corresponding to 326 Mbp) A. thaliana nuclei was determined to measure the size of fungal genomes (Dolezel et al. 2003).

**Gene and transposable element prediction and annotation**

Genes of the EU-B04 reference genome were predicted using the Braker1 pipeline (Hoff et al. 2016). RNA-alignments to the EU-B04 genome from all the in vitro conditions described above were used to provide extrinsic hints for gene prediction in this pipeline. Genes in all other genomes were predicted using AUGUSTUS version 3.2.3 (Stanke et al. 2008) utilizing the species training parameter files created in the Braker1 pipeline above.

Secreted proteins in the predicted proteome were annotated with SignalP version 4.1 (Petersen et al. 2011) using default parameters. Small-secreted proteins (SSPs) were identified as secreted proteins with a maximum of 300 amino acids, containing no Glycosylphosphatidylinositol (GPI) anchor motif according to Big-PI (default parameters; Eisenhaber
et al. 2004) and PredGPI (False-Positive rate lower than 1e-3; Pierleoni et al. 2008), and no transmembrane domain in the mature form according to TMHMM (default parameters; Krogh et al. 2001).

The REPET pipeline (urgi.versailles.inra.fr/Tools/REPET) was used for the detection, classification (TEedenovo; Flutre et al. 2011; Hoede et al. 2014) and annotation (TEannot; Quesneville et al. 2005) of transposable elements (TEs). The TDedenovo pipeline was used to search for repeats in *V. inaequalis* EU-B04 genome sequence. The first step used Blaster with the following parameters: identity > 90%, HSP (high-scoring segment pairs) length >100 bp and <20 kb, e-value ≤ 1e-300. HSPs were clustered using three different methods: Piler (Edgar and Myers, 2005), Grouper (Quesneville et al. 2003) and Recon (Bao and Eddy, 2002). The MAP software (Huang, 1994) was used to make multiple alignments of the 20 longest members of each cluster containing at least 3 members and build a consensus. Consensus sequences were then classified on the basis of their structure and similarities relative to Repbase Update (v17.11) (Jurka et al. 2005) and PFAM domain library v2 (Finn et al. 2014), before redundancy removal with Blaster and Matcher as performed in the TDedenovo pipeline. Consensus sequences with no known structure or similarity were classified as “unknown”.

**Phylogenetic species tree**

All strains of *V. pirina*, *V. asperata* and *V. aucupariae*, and one isolate from each population of *V. inaequalis* were used to build a phylogenetic tree. The *Dothideomycetes Zymoseptoria tritici* IPO323 and *Alternaria alternata SRC1KrK2f* (genome.jgi.doe.gov) were used as outgroups. The DNA sequence of β-tubulin *TUB1* (KOG1375), eukaryotic initiation factor 3, subunit b *eIF3b* (KOG2314), phosphoadenylphosphosulfate reductase *PAP* (KOG0189) and ribosomal protein S9 *RPS9* (KOG1753) genes were retrieved in all genomes and each were aligned using MUSCLE (Edgar, 2004). Regions with low quality alignment were removed using Gblocks (default parameters; Castresana, 2000) prior to concatenation of the four alignments. A maximum-likelihood tree was built with MEGA6 (Tamura et al. 2011) using the best-fit Tamura-Nei substitution model with 5 gamma distribution categories and 500 bootstrap replicates.

**Data availability**

All genome sequencing data generated in this study are accessible (assembly fasta files) in GenBank under the BioProject number PRJNA407103. Supplemental material available at FigShare: https://doi.org/10.25387/g3.7346714.

**RESULTS AND DISCUSSION**

**Phylogeny of Venturia species**

Phylogenetic relationships between the different *Venturia* species were determined using the *Dothideomycete* species *Zymoseptoria tritici* and *Alternaria alternata* as outgroup to the *Venturiales* order (Zhang et al. 2011; Bock et al. 2016). The maximum likelihood tree shows that *V. asperata* is basal to the other *Venturia* species (Figure 1A). *V. aucupariae*...
has diverged more recently from \textit{V. inaequalis} than \textit{V. pirina}. Phylogenetic relationships among \textit{V. inaequalis} strains from different populations cannot be accurately resolved, probably because of gene flow between populations (Gladieux \textit{et al.} 2010b; Lemaire \textit{et al.} 2016; Leroy \textit{et al.} 2016).

\textbf{Genome assemblies}

The assembly size of the different \textit{V. inaequalis} strains varies greatly from 30 to 73 Mb, with a median of 7,435 scaffolds (from 66 to 31,301), an N50 median of 55.7 kb and an estimated sequencing coverage ranging from 8.8x to 150x (Figure 1B and Table S1), consistent with previous sequencing of \textit{V. inaequalis} strains (Shiller \textit{et al.} 2015; Deng \textit{et al.} 2017; Passey \textit{et al.} 2018). As expected, the best assembly result was obtained for the EU-B04 reference strain with PacBio sequencing (66 scaffolds, N50 = 2,978 kb). Genomic DNA library construction likely explains the variability in assembly size as certain protocols might have excluded long repetitive regions. It is especially the case of CAM and CAP strains that both exhibit a genome size half the size of genomes of strains found on apple trees. In this case, library construction kits included PCR amplification that might have filtered out the major part of AT-rich regions (Aird \textit{et al.} 2011), a bias already reported in other fungi harboring an AT-rich genome (Oyola \textit{et al.} 2012). The actual genome size of \textit{V. inaequalis} strains is over 80 Mb as measured using flow cytometry (average size of 83.1 ± 5.0 Mb; Figure 1B and Table S1). The measured genome sizes suggest that scaffolds in the PacBio assembly are separated by extremely long stretches of repetitive elements. The average assembly size for \textit{V. aucupariae}, \textit{V. pirina} and \textit{V. asperata} strains is 52, 38 and 34 Mb, respectively (Figure 1B and Table S1).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Genome features of the \textit{Venturia inaequalis} EU-B04 reference strain. Twenty-nine out of 66 scaffolds are shown, representing 95% of the total assembly. Scaffold number is indicated on the outer circle. Tracks show (A) Single-Nucleotide Polymorphism (SNP) density between isolates (black) and clones (yellow); (B) GC content (%), the orange line indicate 50%; (C) Gene density according to Braker1 gene prediction, magenta circles indicate the localization of small secreted proteins (SSPs). (D) transposable element density according to REPET analysis. (E) RIP (Repeat Induced Point mutation) index. (F) close-up on scaffold 1 showing SNP densities and GC content as described in (A) and (B) tracks, respectively, the dotted blue line indicating the GC content threshold used by OcculterCut (Testa \textit{et al.} 2016) to separate GC-equilibrated and AT-rich regions; the latter are highlighted with gray shading. (G) Difference in di-nucleotide frequency between GC-equilibrated and AT-rich regions in genomes of all isolates as determined by OcculterCut. Negative and positive values indicate over-representation in GC-equilibrated and AT-rich regions, respectively. Values for the EU-B04 strain reference assembly are indicated in yellow.}
\end{figure}
petitive elements were sequenced. However, flow cytometry determined genome sizes of $63.6 \pm 1.6$ Mb, $70.6 \pm 2.7$ Mb and $46.1 \pm 1.2$ Mb for *V. aucupariae*, *V. pirina* and *V. asperata*, respectively (Figure 1C and Table S1). The former two species appear to contain a significant amount of repetitive elements that could not be deduced from their genome assemblies likely because Illumina sequencing tends to collapse repeat regions (Shiller et al. 2015; Deng et al. 2017).

The genome structure of *V. inaequalis* is also a mosaic of GC-equilibrated and AT-rich regions as shown for the EU-B04 reference isolate (Figure 2). Indeed, the GC-content distribution in all genomes exhibits a bimodal pattern typical of genomes with large proportions of AT-rich regions (Figure 1C; Testa et al. 2016). These regions cover on average 45.3% of *V. inaequalis* genome assemblies with a size over 60 Mb, while they cover on average 32.5%, 6.6% and 9.9% of *V. aucupariae*, *V. pirina* and *V. asperata* assembled genomes, respectively (Table S1). Considering the measured genome sizes, these values are likely underestimated and AT-rich regions may represent from 30% in *V. asperata* to 50% in *V. pirina*. The GC content of AT-rich regions in Venturia species ranges from 33.5 to 41.3% (Table S1), which is similar to the 33.8% reported for *Leptosphaeria maculans* (Testa et al. 2016), another species with a mosaic genome structure of GC-equilibrated and AT-rich regions (Rouxel et al. 2011).

Despite the difference in assembly size, the BUSCO assessment of gene content indicates a similar and very high level of completeness for most strains (median of 98% complete single copy genes). This result also indicates that most of the predicted genes are located in GC-equilibrated regions, which represent the most well sequenced and assembled in our dataset. Six strains show a retrieval rate of BUSCO genes lower than 90% (Table S1), meaning that genome sequencing and/or assembly quality is very low and these genomes should be used with caution in further analyses.

**SNP discovery and population structure of Venturia inaequalis**

Using multiple sequence alignments of the identified BUSCO genes, we computed genetic distances between isolates and identified six pairs of strains and one triplet with very low divergence, suggesting they were clones (2225/2446, 2226/2450, 2447/2458, 2474/2475, 2478/2480, 186/2269 and 2503/2504/2263). Consistently, fewer SNPs were discovered between these pairs compared to other strains from the same populations (Table 1). We could use this information to perform a SNP false discovery analysis because we expect *de novo* mutations to be rare between clones of *V. inaequalis* as it reproduces sexually each year (Bowen et al. 2011). Our data suggests that 25–50% of SNPs identified without masking genomes are false positives (Table 1). About 90% of SNPs identified between clones are located throughout the genome (Table 1, Figure 2A, 2D and 2F). Reads that originate from repetitive elements sharing high sequence similarity are likely to be inaccurately mapped on the reference genome, which could explain the vast majority of SNPs observed between clones. In contrast, GC-equilibrated regions appear to contain a higher proportion of SNPs discovered between non-clonal strains, which are enriched on certain scaffolds (Table 1, Figure 2A and 2F). Masking AT-rich regions as determined by OcculterCut eliminated between 88.7% and 96.6% of false positive SNPs for a saving of 16.5–48.4% of the initial SNPs between non-clonal strains (Table 1). The proportion of retained SNPs reached 52% in the global dataset of 1.27 million SNPs identified between all strains. The fact that GC-equilibrated regions still contain

| Table 1 | Single Nucleotide Polymorphism (SNP) false discovery analysis in *Venturia inaequalis* populations. Isolates originate from populations isolated on loquat (LOQ), firethorn (PYR), Malus sylvestris (SYL) and on Malus sieversii in Central Asian Mountains (CAM) and Central Asian Plains (CAP) |  |
|-----------------------------------|-----------------------------------|-----------------------------------|
| **LOQ Population** | **CAM population** | **SYL population** | **CAP population** |
| 2503/2504 | 2475/2476 | 2479/2480 |
| 2504/2505 | 2454/2456 | 2479/2480 |
| **LOQ** | **CAM** | **SYL** | **CAP** |
| **Total number of SNPs** | **SNPs in AT-rich region** | **SNPs in GC-equilibrated regions** | **FDR (a)** |
| 130,507 | 333,955 | 39.1% | 94,845 | 341,581 | 27.8% | 125,654 | 522,134 | 24.1% |
| 125,196 | 257,874 | 48.5% | 120,622 | 393,845 | 30.6% | 113,284 | 369,946 | 34.4% |
| 133,955 | 341,581 | 28.8% | 120,670 | 393,845 | 30.6% | 113,284 | 369,946 | 34.4% |
| **FDR (b)** | **FDR (c)** | **FDR (d)** | **FDR (e)** | **FDR (f)** | **FDR (g)** | **FDR (h)** | **FDR (i)** | **FDR (j)** |
| 4.8% | 12% | 24% | 36% | 48% | 60% | 72% | 84% | 96% |

(a) FDR: False Discovery Rate.
(b) GC-equilibrated and AT-rich regions (51.4% and 48.6% of the EU-B04 reference genome, respectively) were determined using OcculterCut (Testa et al. 2016).
(c) indicate clonal isolates.
3.4–11.3% of false positive SNPs is likely explained by the expansion of gene families through gene duplication in the genome of *Venturia* species (Table 1; Shiller et al. 2015; Deng et al. 2017). This observation suggests that masking AT-rich regions or transposable elements alone often may be not enough to reach an acceptable FDR. As reported here, sequencing two clones or twice the same individuals easily allows detecting regions with high level of FDRs and is of great help to define hard filtering and/or SNP recalibration settings, especially for highly repetitive or polyploid genomes of sexual organisms.

Once cleared of the false positive SNPs, our dataset allowed a clear discrimination of the *V. inaequalis* strains according to their host of origin as shown by a principle component analysis (Figure 3). This result is highly congruent with our phylogenetic analysis (Figure 1A) and with previous studies (Gladieux et al. 2010a).

### Gene predictions

Most of *V. inaequalis*, *V. pirina* and *V. aucupariae* strains contain around 11,600 predicted genes, while *V. asperata* strains contain fewer predicted gene models, around 10,400 (Figure 1D and Table S1). Virulence of plant pathogens relies on the secretion of small-secreted proteins (SSPs) that are often rich in cysteine residues (Lo Presti et al. 2015). The number of predicted SSPs of less than 300 amino acids ranges between 527 and 756 in *V. inaequalis* and *V. pirina* strains, representing between 4.8% and 6.1% of the predicted proteome (Figure 1D and Table S1). The PacBio assembly of the *V. inaequalis* EU-B04 isolate certainly explains the much higher number of predicted SSPs and percentage of the proteome. Despite a similar total number of predicted genes, the *V. aucupariae* isolate contains fewer SSPs (461; Figure 1D). In addition to a slightly reduced gene complement, *V. asperata* strains contain fewer predicted SSPs that represent less than 4% of the predicted proteome (between 395 and 427; Figure 1D and Table S1). Detailed analysis of the *V. inaequalis* and *V. pirina* predicted secretome identified specific SSPs putatively involved in host specificity and recognition by apple resistant genes (Deng et al. 2017). Similarly, in our report, differences in gene complement might explain the host range of the different species, especially the host specialization of *V. inaequalis* formae specialis and the low virulence of *V. asperata* on apple (Caffier et al. 2012). In the PacBio assembly, genes tend to be located within GC-equilibrated regions (Figure 2B and 2C). Although SSPs of *V. inaequalis* were reported closer to repetitive elements compared to other genes (Deng et al. 2017), SSPs do not tend to be located within long AT-rich regions as observed in *L. maculans* (Figure 2C and 2D; Rouxel et al. 2011). Further detailed analyses of specific gene families, especially SSPs, carbohydrate degrading enzymes and secondary metabolite gene clusters will provide new insights about the pathogenicity and host specificity of *Venturia* species.

### Invasion of transposable elements

The repetitive element content across all genomes was annotated using the REPET pipeline (Figure 1B; Table S1). The transposable element (TE) content covers 55.7% of the reference EU-B04 genome assembly, of which 66.4% are Class I LTR element, mainly assigned to Gypsy or Copia superfamilies (37.5% and 26.3%, respectively); and 9.5% are Class II elements, mostly belonging to the TIR superfamily (8.2%) (Figure S1). Consistent with the exclusion of repeat content from CAM and CAP populations DNA libraries, only a median of 8.5% of the assemblies are covered with TEs (Figure 1B and Table S1). In other *V. inaequalis* genomes, the median coverage of TEs is 41.7% (Figure 1B and Table S1). TE coverage in the *V. aucupariae* assembly is similar, with 17.4% (Figure 1B and Table S1). In contrast, TE coverage is very limited in *V. pirina* (4.1% on average) and *V. asperata* (3.4% on average), consistent with the exclusion of repeat content from genomic assemblies of the former and the limited number of TEs in the latter (Figures 1B and Table S1). The repetitive content of *V. asperata* strains differs from others as it does not contain Helitron and LINE elements, and is almost devoid of Copia elements (Figure S1 and Table S2). Gypsy and Copia TEs are predominant in all other *Venturia* strains. Notably Copia elements and class II TEs, especially TIR elements, appear to have greatly expanded after the divergence with *V. asperata* (Table S2 and Figure S1). Hierarchical clustering of the coverage of TE models annotated by REPET suggests recent expansion of 32 Copia and Gypsy models in all *Venturia* species (Table S3 and Figure S2); particularly two models (RLG-comp_VeinUTG-B-G309-Map3 and noCat_VeinUTGB-R23-Map20) exhibits high coverage and are likely still actively
transposing (Table S3 and Figure S2). A second cluster of 31 models might represent ancestral expansion of currently inactive TEs and a third cluster contains models with limited coverage (Figure S2). A few elements appear to be actively multiplying in certain genomes. For example, one Gypsy element (RLG-comp_VeinUTG-B-G1005-Map14) exhibits higher coverage in V. asperata, V. aucupariae and, to a lesser extent, in V. pirina compared to V. inaequalis strains (Table S3 and Figure S2), while another Gypsy element (RLG-comp_VeinUTG-B-R34-Map20) and a LARD element (RXX-LARD_VeinUTG-B-G646-Map11) show highest coverage in V. inaequalis 2199 and V. inaequalis B04 strains, respectively (Table S3 and Figure S2). In the V. inaequalis B04 reference genome, TEs are mostly located within AT-rich regions (Figure 2B and 2D), but many appear to be located in close vicinity to genes as reported for SSPs (Deng et al. 2017). In fungi, TE activity can be suppressed by Repeat-Induced Point mutation (RIP), a process that induces in repetitive DNA cytosine to thymine transition mutations during heterokaryon formation prior to meiosis (Watters et al. 1999). As such, RIP is a mechanism that produces AT-rich regions. Evaluation of RIP in Venturia strains showed that most TEs have been mutated through RIP (Figure 2E and Table S1). AT-rich regions in all strains are enriched in TpA di-nucleotides (Figure 2G), a feature typical of RIP in non-coding regions (Testa et al. 2016). Detailed analysis of the dynamics of TEs will determine their impact on the different interactions between Venturia species and their diverse hosts.

Conclusions
Strains of the Venturia species complex are responsible for important diseases on Rosaceae fruit trees, as such they constitute an environmental and public health concern due to the massive use of fungicides necessary for their control. Our study provides valuable genomic resources for those interested in identifying molecular basis of Venturia species adaptation to their host and in mapping traits of interest (virulence, aggressiveness, resistance to fungicides etc...). In addition, these genomes will be used for comparative genomic studies to develop diagnostic molecular tools and to understand the evolution of genomes in relation to their hosts and in particular during the domestication of the apple tree.

ACKNOWLEDGMENTS

The project was funded by Plant Health Department of INRA (ESCAPADES Project) and by the Agence Nationale pour la Recherche (Bioadapt ANR Gandalf ANR-12-ADAP-0009 and Fungisochore ANR-09-GENM-028 projects) and by the French Region Pays de la Loire (ROAD MOVIE project). JS and TP research was conducted in the framework of the regional program “Objectif Végétal, Research, Education and Innovation in Pays de la Loire”, supported by the French Region Pays de la Loire, Angers Loire Métropole and the European Regional Development Fund. LD was supported by the University of Angers, MDG was supported by Plant health Department of INRA and Region Pays de La Loire. We thank Dr. P. Gladiex (INRA, Montpellier, France), Dr. XG Zhang (Shandong University, Taian, China), Dr. H Ishii (Kobe University, Kobe, Japan), Dr. Tim Roberts (East Malling Research, East Malling, England), Dr. A Kollar (JKI, Dossenheim, Germany) and Dr. E Dapena (SERIDA, Villaviciosa, Spain), Dr. F. Laurens (INRA, Angers, France) and Catherine Peix for collecting and providing strains or infected apple leaves.

This work was performed in collaboration with the GeT core facility (Toulouse, France; get.genotoul.fr), and was supported by France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

The present work has benefited from the Cytoometry facility of Imagerie - Gif. (www.i2bcparis-saclay.fr), member of IBISA (www.ibisa.net), supported by “France-BioImaging” (ANR-10-INBS-04-01), and the Labex “Saclay Plant Science” (ANR-11-IDEX-0003-02).

LITERATURE CITED

Aird, D., M. G. Ross, W. S. Chen, M. Denielsson, T. Fennell et al., 2011 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12: R18. https://doi.org/10.1186/gb-2011-12-2-r18

Bao, Z., and S. Eddy, 2002 Automated de novo identification of repeat families and putative determinants of host range. BMC Genomics 18: 339. https://doi.org/10.1186/s12864-017-3699-1

Bénaouf, G., and L. Parisi, 2000 Genetics of host-pathogen relationships between Venturia inaequalis races 6 and 7 and Malus species. Phytopathology 90: 236–242. https://doi.org/10.1094/PHYTO.2000.90.3.236

Berlin, K., S. Koren, C. S. Chin, J. P. Drake, J. M. Landolin et al., 2015 Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33: 623–630. Erratum: 1109.

Bock, C. H., C. Chen, F. Yu, K. L. Stevenson, and B. W. Wood, 2016 Draft genome sequence of Fusarium effusium, cause of pecan scab. Stand. Genomic Sci. 11: 36. https://doi.org/10.1186/s40793-016-0114-8

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bowen, J. K., C. H. Mesarich, J. Rees-George, W. Cui, A. Fitzgerald et al., 2009 Candidate effector gene identification in the ascomycete fungal phytopathogen Venturia inaequalis by expressed sequence tag analysis. Mol. Plant Pathol. 10: 431–448. https://doi.org/10.1111/j.1364-3703.2009.00543.x

Bowen, J. K., C. H. Mesarich, V. G. Bus, R. M. Beresford, K. M. Plummer et al., 2011 Venturia inaequalis: the causal agent of apple scab. Mol. Plant Pathol. 12: 105–122. https://doi.org/10.1111/j.1364-3703.2010.00656.x

Caffrier, V., B. Le Cam, P. Expert, M. Devaux, M. Giraud et al., 2012 A new scab-like disease on apple caused by the formerly saprotroph fungus Venturia asperata. Plant Pathol. 61: 915–924. https://doi.org/10.1111/j.1365-3059.2011.02583.x

Castresana, J., 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Cheeseman, K., J. Ropars, P. Renault, J. Dupont, J. Gouzy et al., 2014 Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat. Commun. 5: 2876. https://doi.org/10.1038/ncomms3876

Chin, C. S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake et al., 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10: 563–569. https://doi.org/10.1038/nmeth.2474

Cooke, I. R., D. Jones, J. K. Bowen, C. Deng, P. Faou et al., 2014 Proteogenomic analysis of the Venturia pirina (pear scab fungus) secretome reveals potential effectors. J. Proteome Res. 13: 3635–3644. https://doi.org/10.1021/pr500176c

Crous, P. W., K. Schubert, U. Braun, G. S. de Hoog, A. D. Hocking et al., 2007 Opportunistic, human-pathogenic species in the Herpotrichiellaceae are phenotypically similar to saprobic or phytopathogenic species in the Venturiaceae. Stud. Mycol. 58: 185–217. https://doi.org/10.3114/sim.2007.58.07

Deng, C. H., K. M. Plummer, D. A. B. Jones, C. H. Mesarich, and J. Shiller 2017 Comparative analysis of the predicted secretomes of Rosaceae scab pathogens Venturia inaequalis and V. pirina reveals expanded effector families and putative determinants of host range. BMC Genomics 18: 339. https://doi.org/10.1186/s12864-017-3699-1
Purcell, B., K. Neale, L. Todd-Brown, M. A. R. Thomas, D. Ferreira et al., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575. https://doi.org/10.1086/519795
Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent. Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341. https://doi.org/10.1186/1471-2164-13-341
Quesnelle, H., D. Nouaud, and D. Anxolabéhère, 2003. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae Genomes. J. Mol. Evol. 57: 550–559. https://doi.org/10.1007/s00239-003-0007-2
Quesnelle, H., D. Nouaud, and D. Anxolabéhère, 2003. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae Genomes. J. Mol. Evol. 57: 550–559. https://doi.org/10.1007/s00239-003-0007-2
Rouxel, T., J. Grandaubert, J. K. Hane, C. Hoede, A. P. van de Wouw et al., 2011. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nat. Commun. 2: 202. https://doi.org/10.1038/ncomms1189
Samuels, G. J., and A. Sivanesan, 1975. Venturia asperata sp. nov. and its Fusicladium state on apple leaves. N. Z. J. Bot. 13: 645–652. https://doi.org/10.1080/0028825X.1975.10430352
Shiller, J. A. P. Van de Wouw, A. P. Taranto, J. K. Bowen, and D. Dubois, 2015. A large family of AvrLm6-like genes in the apple and pear scab pathogens, Venturia inaequalis and Venturia pirina. Front. Plant Sci. 6: 980. https://doi.org/10.3389/fpls.2015.00980
Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
Singh, P., M. Piotrowski, and A. E. Gau, 2005. Purification and partial characterization of an extracellular melanoprotein from the fungus Venturia inaequalis. Z. Naturforsch. C 60: 109–115. https://doi.org/10.1515/znc-2005-1-220
Sivanesan, A., 1977. The Taxonomy and Pathology of Venturia Species. Bibliotheca Mycologica, Vaduz, Liechtenstein.
Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637–644. https://doi.org/10.1093/bioinformatics/btn013
Steiner, U., and E. C. Oerkeme, 2007. Localized melanization of appressoria is required for pathogenicity of Venturia inaequalis. Phytopathology 97: 1222–1230. https://doi.org/10.1094/PHYTO-97-10-1222
Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28: 2731–2739. https://doi.org/10.1093/molbev/msr121
Testa, A. C., R. P. Oliver, and J. K. Hane, 2016. OcculterCut: A comprehensive survey of AT-rich regions in fungal genomes. Genome Biol. Evol. 8: 2044–2064. https://doi.org/10.1093/gbe/eww121
Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, and G. Del Angel, 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43: 11–33.
Vogel, H. J., 1956. A convenient growth medium for Neurospora crassa. Microbial Genetics Bulletin 13: 42–47.
Watters, M. K., T. A. Randall, B. S. Margolin, E. U. Selker, and D. R. Stadler, 1999. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in Neurospora. Genetics 153: 705–714.
Young, C. A., C. H. Bock, N. D. Charlton, C. Mattupalli, N. Krom et al., 2018. Evidence for Sexual Reproduction: Identification, Frequency, and Spatial Distribution of Venturia effusa (Pecan Scab) Mating Type Idiomorphs. Phytopathology 108: 837–846. https://doi.org/10.1094/PHYTO-07-17-0233-R
Zerbino, D. R., and E. Birney, 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829. https://doi.org/10.1101/gr.074492.107
Zhong, Y., P. W. Crous, C. L. Schoch, A. H. Bakhali, L. D. Guo et al., 2011. A molecular, morphological and ecological re-appraisal of Venturiidae - a new order of Dothideomycetes. Fungal Divers. 51: 249–277. https://doi.org/10.1007/s13225-011-0141-x

Communicating editor: D. Baltrus