# Population Genomics Reveals Molecular Determinants of Specialization to Tomato in the Polyphagous Fungal Pathogen *Botrytis cinerea*

Alex Mercier[1,2], Adeline Simon[1], Nicolas Lapalu[1], Tatiana Giraud[3], Marc Bardin[4], Anne-Sophie Walker[1†], Muriel Viaud[1†] and Pierre Gladieux[5*†]

[1] Université Paris-Saclay, INRAE, AgroParisTech, UMR BIOGER, 78850 Thiverval-Grignon, France

[2] Université Paris-Saclay, Orsay, France

[3] Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, 91400 Orsay, France

[4] UR0407 Pathologie Végétale, INRAE, 84143 Montfavet, France

[5] UMR BGPI, INRAE, CIRAD, Montpellier Supagro, Université de Montpellier, 34398 Montpellier, France

[*] Corresponding author: pierre.gladieux@inrae.fr

[†] contributed equally to this study

## Abstract

Many fungal plant pathogens encompass multiple populations specialized on different plant species. Understanding the factors underlying pathogen adaptation to their hosts is a major challenge of evolutionary microbiology, and it should help preventing the emergence of new specialized pathogens on novel hosts. Previous studies have shown that French populations of the grey mould pathogen *Botrytis cinerea* parasitizing tomato and grapevine are differentiated from each other, and have higher aggressiveness on their host-of-origin than on other hosts, indicating some degree of host specialization in this polyphagous pathogen. Here, we aimed at identifying the genomic features underlying the specialization of *B. cinerea* populations to tomato and grapevine. Based on whole genome sequences of 32 isolates, we confirmed the subdivision of *B. cinerea* pathogens into two genetic clusters on grapevine and another, single cluster on tomato. Levels of genetic variation in the different clusters were similar, suggesting that the tomato-specific cluster has not recently emerged following a bottleneck. Using genome scans for selective sweeps and divergent selection, tests of positive selection based on polymorphism and divergence at synonymous and non-synonymous sites and analyses of presence/absence variation, we identified several candidate genes that represent possible determinants of host specialization in the tomato-associated population. This work deepens our understanding of the genomic changes underlying the specialization of fungal pathogen populations.

**Keywords:** Host specialization; grey mould; gene content variation; selective sweeps; positive selection

## Introduction

Many fungal plant pathogens encompass multiple lineages, host races or *formae speciales* specialized on different plant species. Understanding the proximate (*i.e.* molecular) and ultimate (*i.e.* eco-evolutionary) factors underlying adaptation to hosts is a major goal for evolutionary microbiology, because emerging diseases are often caused by the appearance and spread of new pathogen populations specialized onto new hosts (Fisher *et al.* 2012; Stukenbrock & McDonald 2008). Evolutionary theory predicts that pathogen specialization should facilitate the emergence of new populations onto novel hosts, because specialization restricts encounters of potential mates within hosts and reduces the survival of offspring due to maladaptation of immigrants and hybrid offspring, thereby reducing gene flow between ancestral and emerging populations (Giraud *et al.* 2010; Nosil *et al.* 2005). The role of specialization as a barrier to gene flow is expected to be strong for pathogens mating within or onto their hosts because, for individuals evolving the ability to infect novel hosts, mating automatically becomes assortative with respect to host use, and reproductive isolation arises as a direct consequence of adaptive divergence (Gladieux *et al.* 2011; Servedio *et al.* 2011). Evolutionary theory also predicts that specialization, and the associated emergence of new populations, could be facilitated by the molecular basis of plant-pathogen interactions, because compatibility is often determined by a limited number of genes in the host and the pathogen, and selection is more efficient when it acts on a smaller number of genes (Giraud *et al.* 2010; Schulze-Lefert & Panstruga 2011). However, despite the apparent ubiquity of specialized fungal pathogens and the negative impact that specialized populations can have on food security and ecosystem health, the genomic features involved in host specialization

remain largely unknown. Acquiring knowledge about the genomic features underlying pathogen specialization can provide key insights into the mechanisms of specialization.

The ascomycete *Botrytis cinerea* is often presented as a textbook example of a polyphagous plant pathogen, parasitizing more than 1400 host plant species belonging to 580 genera (Elad *et al.* 2016). Previous population genetic studies reveal that *B. cinerea* is not strictly speaking a generalist pathogen, and that a more appropriate qualifier is "polyspecialist", i.e. a set of populations specialized to different host species. Indeed, with the exception of California (Ma and Michailides, 2005; Caseys et al, 2020; Soltis et al, 2019; Atwell et al, 2015), studies conducted in multiple regions of the world reveal that populations were structured (reviewed in Walker 2016), and recognize the host as the factor with the highest explanatory power for population structure in *B. cinerea*, ahead of geography. In France, our previous work revealed population subdivision in *B. cinerea,* with genetic differentiation between field populations infecting tomato (*Solanum lycopersicum*) and grapevine (*Vitis vinifera*), respectively (Walker *et al.* 2015). This population structure in *B. cinerea* was shown to be stable in time and was observed in multiple regions in France. Furthermore, this structure was later associated with differences in performance on the two hosts, with pathogens isolated from tomato being more aggressive on tomato than pathogens isolated from grapevine, and reciprocally (Mercier *et al.* 2019). Altogether, these data were consistent with a certain degree of specialization of *B. cinerea* populations onto these two host plants.

Here, we aimed to identify the molecular basis of host specialization in the *B. cinerea* tomato- and grapevine-associated populations, by addressing the following questions: (1) Can we confirm the genetic subdivision between *B. cinerea* populations from tomato and

grapevine using genomic data, and what is the degree of divergence between them? (2) Can we identify genes with footprints of positive selection and/or divergent selection in the genomes of populations specialized to different hosts, and what are their predicted functions? (3) Is there variation in gene content between *B. cinerea* populations associated with tomato and grapevine? To address these questions, we used a set of *B. cinerea* isolates collected on tomato and grapevine in different regions of France. We Illumina-sequenced their genomes and identified single nucleotide polymorphisms by mapping sequencing reads against a high-quality reference genome (van Kan *et al.* 2017). Because some tests of selection can be biased by population subdivision, while other tests are based on patterns of population differentiation, we first analyzed the population structure of *B. cinerea* collected on tomato and grapevine. To detect genes potentially involved in the specialization of *B. cinerea* to tomato, we searched for signatures of positive selection, by scanning genomes in the tomato population for selective sweeps, and by estimating the direction and intensity of selection using McDonald-Kreitman tests on coding sequences. Furthermore, we investigated signatures of divergent selection using genomic differentiation between populations, and we characterized variations in the presence/absence of predicted genes between populations collected on tomato and grapevine using *de novo* genome assemblies, gene prediction, and orthology analysis.

## Materials and methods

*Sample collection*

*Botrytis cinerea* samples were selected in a collection of isolates (i.e. single-spored mycelial colonies) originating from three regions of France (Champagne, Occitanie and Provence) and previously characterized using analyses of population structure based on microsatellite markers and pathogenicity tests (Walker *et al.* 2015; Mercier *et al.*, 2019). For each region and each host, we randomly selected three to nine isolates with high membership proportions ($q$>0.9) in the cluster matching their host of origin in a previous analysis of population structure based on microsatellite genotyping (Table 1; Mercier *et al.* 2019). Collection sites were 15 to 133 km apart within regions, and 204 to 722 km apart between regions (Supplementary Figure S1). The set of 32 isolates originated from the following hosts: (i) tomato (*Solanum lycopersicum*; fruits; 13 isolates), (ii) grapevine (*Vitis vinifera*; berries; 16 isolates), (iii) bramble (*Rubus fruticosus*; berries; two isolates) and (iv) hydrangea (*Hydrangea macrophylla*; flower buds; one isolate). Samples from tomato originated from plastic tunnels with sides opened (Occitanie region) or glasshouses (Provence and Champagne regions). Mycelia were cultured on malt-yeast-agar (MYA; 20 $g.L^{-1}$ malt extract, 5 $g.L^{-1}$ yeast extract, 15 $g.L^{-1}$ agar) at 23°C under continuous light until conidiation, and stored as conidial suspensions in glycerol 20% at -80°C until use.

**Table 1.** *Botrytis* isolates included in the study. Isolates from grapevine berries and tomatoes were collected in three regions of France between September 2005 and June 2007 for Champagne and Provence isolates (Walker *et al.* (2015), and between May and June 2009 for Occitanie isolates. Tomatoes in Occitanie were grown in

123 plastic tunnels with sides open, whereas those from the Champagne and Provence areas were grown in glass-

124 greenhouses. B05.10 is used as the reference isolate for genomic analysis (van Kan et al. 2016)

| Species | Isolate ID | Other ID | Host, cultivar | Region, city |
|---|---|---|---|---|
| *B. cinerea* | Vv1 | VC636 | *Vitis vinifera,* Pinot noir | Champagne, Hautvilliers |
| | Vv2 | VC095 | *Vitis vinifera,* Pinot noir | Champagne, Vandières |
| | Vv3 | VC621 | *Vitis vinifera,* Pinot noir | Champagne, Hautvilliers |
| | Vv4 | VC671 | *Vitis vinifera,* Pinot noir | Champagne, Hautvilliers |
| | Vv5 | VC224 | *Vitis vinifera,* Pinot meunier | Champagne, Courteron |
| | Vv6 | VC624 | *Vitis vinifera,* Pinot noir | Champagne, Hautvilliers |
| | Vv7 | ACBER342 | *Vitis vinifera,* Grenache | Provence, Berre |
| | Vv8 | ACBER356 | *Vitis vinifera,* Grenache | Provence, Berre |
| | Vv9 | ACBER358 | *Vitis vinifera,* Grenache | Provence, Berre |
| | Vv10 | ACSAR333 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv11 | ACSAR334 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv12 | ACSAR335 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv13 | ACSAR342 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv14 | ACSAR354 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv15 | ACSAR357 | *Vitis vinifera,* Grenache | Provence, Sarrians |
| | Vv16 | VC610 | *Vitis vinifera,* Pinot noir | Champagne, Hautvilliers |
| | Sl1 | VA714 | *Solanum lycopersicum,* Moneymaker | Champagne, Foissy-sur-Vanne |
| | Sl2 | VC800 | *Solanum lycopersicum,* Moneymaker | Champagne, Courceroy |
| | Sl3 | VC806 | *Solanum lycopersicum,* Moneymaker | Champagne, Courceroy |
| | Sl4 | AABER19 | *Solanum lycopersicum*, Alison | Provence, Berre |
| | Sl5 | ACBER304 | *Solanum lycopersicum*, Alison | Provence, Berre |
| | Sl6 | ACPIE306 | *Solanum lycopersicum,* Hipop | Provence, Pierrelatte |
| | Sl7 | ADPIE463 | *Solanum lycopersicum,* Hipop | Provence, Pierrelatte |
| | Sl8 | ADPIE475 | *Solanum lycopersicum,* Hipop | Provence, Pierrelatte |
| | Sl9 | 65_TT8 | *Solanum lycopersicum*, Brenda | Occitanie, Alenya |
| | Sl10 | 5_TT8 | *Solanum lycopersicum*, Brenda | Occitanie, Alenya |
| | Sl11 | 13_TT8 | *Solanum lycopersicum*, Brenda | Occitanie, Alenya |
| | Sl12 | 9_TT8 | *Solanum lycopersicum*, Brenda | Occitanie, Alenya |
| | Sl13 | 66_TT8 | *Solanum lycopersicum*, Brenda | Occitanie, Alenya |
| | Rf1 | VC902 | *Rubus fruticosus*, wild | Champagne, Foissy-sur-Vanne |

| | | | | |
|---|---|---|---|---|
| | Rf2 | VC399 | *Rubus fruticosus*, wild | Champagne, Courteron |
| | Hm1 | MSN-Bot 2556 | *Hydrangea macrophylla*, Leuchtfeuer | Anjou, Angers |
| | B05.10 | - | - | - |
| *B. fabae* | Bfab | MSN-Bot 2220 | *Vicia faba* | Region of Tunis (Tunisia) |

126 *Pathogenicity tests on tomato plants*

127 Isolates of *B. cinerea* collected from tomato or grape were cultivated on MYA medium in a
128 growth chamber (21°C, 14 hours light) for 14 days. Conidia were then washed with sterile
129 distilled water. The conidial suspension was filtered through a 30 µm mesh sterile filter to
130 remove mycelium fragments. The conidial concentration was determined with a
131 haemacytometer and adjusted to $10^6$ spores/mL. Seeds of tomato var. Clodano (Syngenta)
132 were sown in compost and transplanted after one week in individual pots. Plants were
133 grown in a glasshouse for 6 to 8 weeks where they received a standard commercial nutrient
134 solution once or twice a day, depending on needs. Plants had at least eight fully expanded
135 leaves when inoculated with a conidial suspension. Each isolate was inoculated on five
136 plants, from each of which two leaves were removed, leaving 5-10 mm petiole stubs on the
137 stems. Each pruning wound was inoculated with 10 µl of conidial suspension at $10^6$
138 conidia/mL. Inoculated plants were incubated in a growth chamber in conditions conducive
139 to disease development (21°C, 16h-photoperiod, 162 $\mu mol.s^{-1}.m^{-2}$, relative humidity > 80%).
140 Due to a growth chamber area that did not allow all the isolates to be tested at the same
141 time, nine series of pathogenicity tests were conducted with 10-12 isolates each, together
142 with the BC1 reference isolate collected in 1989 in a tomato glasshouse in Brittany
143 (Decognet *et al.* 2009). For each isolate, two to four independent repetitions of the
144 pathogenicity test were performed. Lesion sizes (in mm) were assessed daily between the $4^{th}$
145 and the $7^{th}$ day post-infection and the Area Under the Disease Progress Curve (AUDPC;
146 (Simko & Piepho 2012) was computed to take into account the kinetics of disease
147 development for each isolate. To compare the aggressiveness of isolates, an aggressiveness
148 index (AI), relative to the reference isolate BC1, was computed as follows: $AI_{isolate} =$

149 $100 \times (AUDPC_{isolate}/AUDPC_{BC1})$, where $AUDPC_{isolate}$ was the average AUDPC for a given
150 isolate and $AUDPC_{BC1}$ is the average AUDPC for the reference isolate BC1. Using the AI index
151 calibrating the AUDPC of a given isolate with that of the reference isolate BC1 in the same
152 test allows comparing isolates aggressiveness while taking into account the variability
153 occurring among assays (e.g. plant physiological state; (Leyronas *et al.* 2018). Because of
154 data non-normality, data were analysed using non-parametric tests. Three statistical tests
155 were carried out with STATISTICA: (1) we used a non-parametric analysis of variance (Kruskal-
156 Wallis test) to assess differences among isolates in terms of aggressiveness on grape and
157 tomato, considering the average values for each of the independent pathogenicity tests as
158 replications; (2) we used the Mann-Whitney U-test to compare the aggressiveness of isolates
159 from different hosts of origin (tomato vs. grape), considering the independent repetitions of
160 the pathogenicity test as blocks and the 13 isolates from tomato and 16 isolates from
161 grapevine as replicates; (3) we used a Kruskal-Wallis test to compare isolates from three
162 different clusters (see *Results* section) in terms of aggressiveness on tomato plants.
163

164 *DNA preparation and sequencing*

165 Isolates were cultivated for 48 h on MYA + cellophane medium at 23 °C in the dark and then
166 ground using a mortar and pestle in liquid nitrogen. DNA was extracted using a standard
167 sarkosyl procedure (Dellaporta *et al.* 1983). Paired-end libraries were prepared and
168 sequenced (2 x 100 nucleotides) on a HiSeq4000 Illumina platform at Integragen (Evry,
169 France). Sequencing coverage ranged from 58 to 305 X. Genomic data were deposited at SRA
170 under accession number PRJNA624742. Read quality was checked using FASTQC
171 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

172

173 *SNP calling and filtering*

174 SNPs were detected with the same workflow as described in (Zhong *et al.* 2017). Sequencing
175 reads were preprocessed with TRIMMOMATIC v0.36 (Bolger *et al.* 2014). Preprocessed reads
176 were mapped onto the *B. cinerea* B05.10 reference genome (van Kan *et al.* 2017) using BWA
177 v0.7.15 (Li & Durbin 2009). Aligned reads were filtered based on quality using SAMTOOLS v1.3
178 (Li *et al.* 2009) and PICARD TOOLS (http://broadinstitute.github.io/picard/) to remove
179 secondary alignments, reads with a mapping quality <30 and paired reads not at the
180 expected distance. SNP calling was performed with FREEBAYES v1.1 (Garrison & Marth 2012).
181 Further filtering was carried out using script VCFFiltering.py
182 (https://urgi.versailles.inra.fr/download/gandalf/VCFtools-1.2.tar.gz), following (Li 2014).
183 We kept only biallelic SNPs supported by more than 90% of aligned reads, detected outside
184 low-complexity regions or transposable elements (as identified in the reference isolate
185 B05.10: https://doi.org/10.15454/TFYH9N; Porquier *et al.* 2016) and with coverage lower
186 than twice the standard deviation from the mean depth coverage. The VCF file is available
187 on Zenodo (doi: 10.5281/zenodo.4293375).

188

189 *Population structure and demographic history*

190 We performed a principal component analysis based on biallelic SNPs using the python
191 library SCIKIT-ALLEL 1.3.2 (https://github.com/cggh/scikit-allel). We used the sNMF program to
192 infer individual ancestry coefficients in *K* ancestral populations. This program is optimized for
193 the analysis of large datasets and it estimates individual admixture coefficients based on

194 sparse non-negative matrix factorization, without assuming Hardy-Weinberg equilibrium
195 (Frichot et al., 2014). We used SPLITSTREE 4 (Huson & Bryant 2005) to visualize relationships
196 between genotypes in a phylogenetic network, with reticulations representing the
197 conflicting phylogenetic signals caused by recombination or incomplete lineage sorting. The
198 position of the root was determined using a *B. fabae* isolate as the outgroup. *Botrytis fabae*
199 is one of the closest known relatives of *B. cinerea* (Amselem *et al.* 2011; Walker 2016).
200 Summary statistics of genomic variation (segregating sites S, nucleotide diversity π,
201 Watterson's θ, Tajima's D) were estimated using EGGLIB 3.0 (https://egglib.org/) on 10kb
202 windows, excluding sites with more than 50% missing data and removing windows with
203 lseff<1000 (lseff is the number of analysed sites) or nseff<3 (nseff is the average number of
204 exploitable samples). Site frequency spectra were estimated using DADI 1.7.0 (GUTENKUNST *ET*
205 *AL.* 2009). Allelic richness and private allele richness were estimated with ADZE 1.0 (Szpiech
206 *et al.* 2008), using a generalized rarefaction approach to account for differences in sample
207 size among populations.

208 We employed the f3 statistic (Reich *et al.* 2009) to test for admixture based on shared
209 genetic drift, as implemented in the POPSTATS python script (Skoglund *et al.* 2015;
210 https://github.com/pontussk/popstats/). The *f3* statistic is used to test for admixture among
211 three populations Px, P1, P2. In the no-admixture case, the *f3* statistic measures the branch
212 length between Px and the internal node of the unrooted population tree (Px;P1,P2), and it
213 is therefore expected to be greater than zero. In the case Px has a mixed ancestry from P1
214 and P2, or populations closely related to them, the *f3* statistic is expected to be negative.
215 Significance was assessed by block jackknife by treating each chromosome as a block and

weighting each block by the number of SNPs. The standard error of the test statistic was used to define a Z-score.

We conducted historical demographic analysis with MSMC2 (Schiffels & Wang 2020; https://github.com/stschiff/msmc2), based on randomly generated pairs of pseudodiploid genomes (i.e. quadruplets of haploid genomes) and using time patterning of 12 segments (option -p 1*2+8*1+1*2). We generated 100 pairs of pseudodiploid genomes for populations T and G1, and only one pair for G2 for which sample size is four. We excluded 67 pairs for population T and 52 pairs for G1, because they generated spurious MSMC2 profile with extremely large estimates in the most ancient time intervals, which is a sign of overfitting. Changing the time patterning, as suggested in Schiffels & Wang (2020), did not remove the spurious demographic profiles (not shown). We used a mutation rate of 2e-8/bp/years (Gladieux *et al.* 2018; Latorre *et al.* 2020).

*Genome assembly and gene prediction*

Gene content was determined using two independent approaches. In the first approach, we used the read mapping coverage of the genes previously predicted in the reference genome B05.10 (van Kan *et al.* 2017). Read count per site for each gene and for each isolate was computed using SAMTOOLS BEDCOV, and normalized using EDGER (McCarthy *et al.* 2012; Robinson *et al.* 2010). Genes showing significant differences in mapping coverage across populations were identified with EDGER (adjusted p-value <0.05 and more than two-fold change). Putative gene duplication events (higher mapping coverage) and missing B05.10

237 genes (lack of mapping reads) were visually inspected in their genomic context using a
238 genome browser for validation.

239 The second approach was based on genome assembly and *de novo* gene prediction.
240 Illumina paired-reads were assembled using a combination of VELVET (Zerbino & Birney
241 2008), SOAPDENOVO and SOAPGAPCLOSER (Luo *et al.* 2012), as follows: (1) reads were trimmed
242 at the first N, (2) contigs were generated with several k-mer values using SOAPDENOVO, (3)
243 several VELVET assemblies were built using several k-mer values and as the input the trimmed
244 reads and all SOAPDENOVO contigs considered as "long reads ", (4) the assembly that
245 maximizes the criterion (N50*size of the assembly) was selected, (5) SOAPGAPCLOSER was run
246 on the selected assembly, and (6) Contigs completely included in other longer contigs were
247 deleted. Genomic regions mapping to transposable elements previously identified in *B.*
248 *cinerea* (https://doi.org/10.15454/TFYH9N; Porquier *et al.* 2016) were masked with
249 REPEATMASKER (http://www.repeatmasker.org) prior to gene prediction. Genes were
250 predicted using the FGENESH *ab initio* gene-finder (Solovyev *et al.* 2006;
251 http://www.softberry.com/berry.phtml), the program previously used to annotate the
252 reference genome (Amselem *et al.* 2011), and for which *Botrytis*-specific gene-finding
253 parameters were thus available. Completeness of the assembly and gene prediction were
254 evaluated with BUSCO using the Ascomycota gene set (Seppey *et al.* 2019). We then used
255 ORTHOFINDER (Emms & Kelly 2015) in order to define groups of orthologous sequences
256 (hereafter “groups of orthologs”) based on sequences of predicted genes translated into
257 protein sequences. To reduce the impact of incomplete gene prediction (*e.g.* truncated
258 genes in small contigs), groups of orthologs were then manually checked for the
259 presence/absence of the protein-encoding genes in the genomes using TBLASTN (Johnson *et*

260 *al.* 2008). Proteins were functionally annotated using INTERPROSCAN (Jones *et al.* 2014), and
261 SIGNALP (Almagro Armenteros *et al.* 2019) was used to predict secretion signal peptides.
262 Prediction of transmembrane helices in proteins was performed using TMHMM 2.0 (Krogh *et*
263 *al.* 2001; Sonnhammer *et al.*). As we detected traces of bacterial contamination in the
264 genome of the Sl13 isolate, this genome was not included in orthology analysis. However,
265 we have kept this genome for analyzes based on polymorphism, because bacterial reads
266 cannot map to the reference genome.
267

268 *Tests of positive selection based on polymorphism and divergence at synonymous and non-*
269 *synonymous sites*

270 We estimated the intensity and direction of selective pressures exerted on genes in
271 populations using the McDonald-Kreitman test based on polymorphism and divergence at
272 synonymous and non-synonymous sites (McDonald & Kreitman 1991). This test is based on
273 the number of nucleotide polymorphisms and substitutions in gene sequences, and assumes
274 that synonymous mutations are neutral. Pseudo-sequences for coding sequences of all
275 genes in the reference genome were generated using the VCF file and the reference
276 sequence. Synonymous and non-synonymous divergence was computed with CODEML (model
277 0; Yang 1997, 2007), using *B. fabae* as the outgroup. Synonymous and non-synonymous
278 polymorphism was computed using EGGLIB 3 (https://egglib.org/), filtering out sites with
279 more than 80% missing data and excluding filtered alignments with less than 10 codons or
280 four sequences. Samples Sl1, Sl2, Sl3, Sl13, Vv3, Vv5, Vv2, Vv4 and Vv6 were excluded
281 because they introduced missing data that prevented calculations of synonymous and non-
282 synonymous divergence. The neutrality index (NI), defined as $(P_N/D_N)/(P_S/D_S)$ (Stoletzki &

283 Eyre-Walker 2011), was computed for every gene, with $P_N$ and $D_N$ the numbers of non-
284 synonymous polymorphisms and substitutions, respectively, and $P_S$ and $D_S$ the numbers of
285 synonymous polymorphisms and substitutions, respectively. Pseudocounts of one were
286 added to each cell of the McDonald-Kreitman tables to ensure the NI was always defined
287 (*i.e.* no division by zero).
288
289 *Linkage disequilibrium and recombination*
290 We used POPLDDECAY (Zhang *et al.* 2019a) to measure linkage disequilibrium ($r^2$) as a function
291 of the distance between pairs of SNPs. POPLDDECAY was configured with a maximum distance
292 between SNPs of 300 kbp, a minor allele frequency of 0.005 and a maximum ratio of
293 heterozygous allele of 0.88. Recombination rates were estimated for each chromosome with
294 PAIRWISE in LDHAT version 2.2 (Auton & McVean 2007). Sites with missing data were excluded.
295
296 *Tests of positive selection based on linkage disequilibrium and the site frequency spectrum*
297 We searched for signatures of selective sweeps along genomes using three different
298 softwares, each implementing a different approach. The SWEED 3.0 software (Pavlidis *et al.*
299 2013) implements a composite likelihood ratio (*CLR*) test based on the SWEEPFINDER algorithm
300 (Nielsen *et al.* 2005), which uses the site frequency spectrum (SFS) of a locus to compute the
301 ratio of the likelihood of a hard selective sweep at a given position to the likelihood of a null
302 model hypothesis without selection. The CLR statistic was computed for each chromosomes
303 of each population using a grid size of 50 or 200 (grid size is the number of positions where
304 the likelihood is calculated), but only results for a grid size of 200 are presented because the
305 CLR profiles were highly similar between the two settings. Input files included both

306 “unfolded” SNPs (i.e. SNPs for which ancestral and derived states can be determined using
307 the allelic state of the outgroup) and “folded” SNPs (i.e. SNPs for which the outgroup had
308 missing data). Only biallelic sites with sample size greater than or equal to five were
309 included. The $nS_L$ method implemented in the NSL software (Ferrer-Admetlla *et al.* 2014)
310 detects hard and soft selective sweeps based on haplotype homozygosity. The nSL statistic
311 was computed for each chromosome of each population, including only biallelic sites with
312 sample size greater than or equal to five. Nine isolates (Sl1, Sl2, Sl3, Sl13, Vv3, Vv5, Vv2, Vv4,
313 Vv6) were excluded to reduce the proportion of missing data. The *hapFLK* method (Fariello
314 *et al.* 2013) implemented in the HAPFLK software is based on the original *FLK* method by
315 (Bonhomme *et al.* 2010), which detects signatures of selection from differentiation between
316 populations. This metric was used to test the null hypothesis of neutrality by contrasting
317 allele frequencies at a given locus in different populations. *hapFLK* extends the FLK method
318 to account for the haplotype structure in the sample, and the method is robust to the effects
319 of bottlenecks and migration. HAPFLK takes the number of cluster of haplotypes as a
320 parameter (K). To determine the number of clusters of haplotypes K, we ran FASTPHASE v1.4
321 (Scheet & Stephens 2006) and R package IMPUTEQ (Khvorykh & Khrunin 2020) on
322 polymorphism data for the largest chromosome BCIN01. For each population, we used
323 IMPUTEQ to generate five datasets with 10% of polymorphic positions masked, and for each
324 masked dataset we used FASTPHASE for imputing masked positions assuming clusters of
325 K=2,3...10 haplotypes and the following parameters: -T10 -C25 -H-1 -n -Z. We estimated
326 error using Estimate-Errors function in IMPUTEQ and the number of clusters that minimizes
327 the error was selected as the optimum. The HapFLK metric was computed individually on
328 each chromosome using as number of clusters K = 5, and nfit = 2. Accessory chromosomes

329 were not included in these analyses as they showed presence/absence polymorphism (see
330 *Results* section).

## 331 Results

332 *Whole-genome sequencing, population structure and demographic history*

333 Previous work using microsatellite data revealed differentiation between populations of *B.*
334 *cinerea* collected on tomato and grapevine (Mercier *et al.* 2019; Walker *et al.* 2015). To
335 investigate the genetic basis of specialization of the tomato- and grapevine-infecting
336 populations, we randomly selected for genome sequencing 32 isolates collected on tomato
337 (13 isolates), grapevine (16 isolates), *Rubus* (two isolates) and *Hydrangea* (one isolate) (Table
338 1). Isolates from *Rubus* and *Hydrangea* were previously shown to belong to generalist
339 populations (*i.e.* assigned to clusters found on all sampled hosts; Mercier *et al.* 2019). One
340 isolate of the sister species *B. fabae* was also sequenced and used as the outgroup.
341 Information about the sequenced isolates is summarized in Table 1. Alignment of sequencing
342 reads to the B05.10 reference genome (van Kan *et al.* 2017) followed by SNP calling
343 identified 249,084 high-quality SNPs.

344 In clustering analyses based on sparse nonnegative matrix factorization algorithms, as
345 implemented in the sNMF program (Frichot et al., 2014), the model with *K*=4 clusters was
346 identified as the best model based on cross-entropy (Supplementary Figure S2) and models
347 with K>4 did not identify well-delimited and biologically relevant clusters (Fig. 1A). At K=4,
348 one cluster was associated with tomato (hereafter referred to as "T" cluster), two clusters
349 were associated with grapevine (hereafter referred to as "G1" for the largest, and "G2" for
350 the smallest), and one cluster was formed by the isolates Rf1 and Hm1 from bramble (*R.*

351 *fruticosus*) and hydrangea (*H. macrophylla*) (Figure 1A). The isolate Rf2 collected on wild
352 blackberry displayed ancestry in multiple clusters, and the reference isolate B05.10 had
353 ancestry in the G2 and T clusters. No pattern of geographical subdivision was observed,
354 consistent with previous findings (Walker *et al.* 2015); Mercier *et al.,* 2019). The neighbor-
355 net network inferred with SPLITSTREE revealed three main groups, corresponding to the three
356 main clusters identified with sNMF, with cluster G2 (isolates Vv8, Vv9, Vv11 and Vv15) more
357 closely related to T than to G1 (Figure 1B). Reticulations in the neighbor-net network and
358 patterns of membership at K=2 and K=3 in the analysis with sNMF indicated that cluster G2
359 shared recent ancestry with clusters G1 and T. However, tests for admixture using the f3
360 statistic (Reich *et al.* 2009; Skoglund *et al.* 2015) did not support a scenario in which G2
361 derived from admixture between the two other clusters (Supplementary Table S1). The
362 principal component analysis corroborated the results of clustering analyses and the
363 neighbor-net network (Figure 1C). The first principal component separated isolates Hm1 and
364 Rf1 from the rest of the dataset. The second principal component individualized isolate Rf2,
365 as well as the three clusters T, G1 and G2. The third and fourth principal components
366 individualized cluster G2 and isolate Sl11, respectively. Together, analyses of population
367 subdivision revealed three clearly defined populations (two on grapevine and one on
368 tomato) and we therefore focused on these populations to identify the genes underlying
369 differences in host specialization.

370 On average across core chromosomes BCIN01 to BCIN16, nucleotide diversity $\pi$ and
371 Watterson's $\theta$ were comparable in the three populations (from $\pi$=0.0018/bp in G1 to $\pi$
372 =0.0030/bp in G2; from $\theta$=0.0017/bp in G1 to $\theta$=0.0027/bp in G2; Table 2; Supplementary
373 Table S2), although all comparisons were statistically significant except between populations

T and G2 for θ (two-tailed Wilcoxon signed-rank test, P-value>0.05). Allelic richness was slightly higher in G2 than in T (AR=1.104 vs AR=1.094), and lower in G1 (AR=1.078). Private allele richness was higher in G1 (PAR=0.085), than in T (PAR=0.073) and G2 (PAR=0.071) (Table 2).

Table 2. Summary statistics of genomic variation in three clusters of *Botrytis cinerea*

| Cluster | S | π | θ | AR | PAR | D | LD50 | ρ |
|---|---|---|---|---|---|---|---|---|
| T | 37763 | 0.0027 | 0.0026 | 1.094 (0.0002) | 0.073 (0.0002) | 0.309 | 8600 | 0.0108 |
| G1 | 26769 | 0.0018 | 0.0017 | 1.078 (0.0002) | 0.085 (0.0002) | 0.167 | 3100 | 0.0178 |
| G2 | 14307 | 0.0030 | 0.0027 | 1.104 (0.0003) | 0.071 (0.0002) | 1.470 | 16800 | 0.2047 |

S, number of segregating sites; π, nucleotide diversity per site; θ, Watterson's estimate of the population mutation parameter per site; ρ, population recombination parameter per site; AR, allelic richness (standard error of the mean); PAR, private allele richness (standard error of the mean); D, Tajima's D; H, Fay and Wu's standardised H; LD50, distance (in bp) at which linkage disequilibrium reaches half of its maximum value. Only core chromosomes BCIN01 to BCIN16 were included in calculations. Per site estimates of π, ρ and θ were computed by summing across chromosomes and dividing by number of sites covered. Tajima's D was computed across 10kb windows, and averaged. AR and PAR were computed using a generalized rarefaction approach and a standardized sample size of two haploid genomes.

The site frequency spectra estimated in populations G1 and T were U-shaped, indicating an excess of high-frequency derived alleles (Figure 2; population G2 was excluded because of too small a sample size), consistent with ongoing episodes of positive selection, mis-assignment of ancestral alleles or gene flow (Marchi & Excoffier 2020). Estimates of Tajima's D were positive but close to zero in clusters T and G1 (T: D=0.309; G1: D=0.167), indicating a slight deficit of low frequency variants (Table 2; Supplementary Table 2), consistent with

385 balancing selection or population contraction. In cluster G2, the estimated Tajima's D was
386 D=1.470 but the estimate was likely upwardly biased by the small sample size, because
387 Watterson's θ is underestimated when sample size is small. The distance at which linkage
388 disequilibrium (LD) decayed to 50% of its maximum was an order of magnitude longer in G2
389 (LDdecay50: 16,800bp) than in T (LDdecay50: 8600bp) and G1 (LDdecay50: 3100bp)
390 (Supplementary Figure S3). The recombination rate was higher in G2 (ρ=0.2047/bp) than in
391 G1 and T (G1: ρ=0.0178/bp; T: ρ=0.0108/bp).

392 The accessory chromosome BCIN18 showed no polymorphism in all three population and
393 the accessory chromosome BCIN17 showed no polymorphism in population G2
394 (Supplementary Table S2). However, it should be noted that coverage analysis also revealed
395 that these accessory chromosomes, which contain a reduced number of genes (23 and 19
396 respectively, van Kan *et al.* 2017), were distributed irrespectively of the host of origin:
397 BCIN17 was found present in all but three isolates (Sl9, Sl10 and Vv15), while BCIN18 was
398 only present in five isolates (Sl3, Sl5, Sl9, Vv9 and Vv11; Supplementary Table S3). In
399 population T, the accessory chromosome BCIN17 displayed a relatively high and positive
400 value of Tajima's D (D=1.37; Supplementary Table S2) and approximately twice as much
401 nucleotide diversity as in core chromosomes (π=0.0058/bp; Supplementary Table S2). In
402 population G1, accessory chromosome BCIN17 displayed a negative value of Tajima's D (D=-
403 0.845) and two orders of magnitude less nucleotide diversity than core chromosomes
404 (π=5.7e-5/bp; Supplementary Table S2). The differences in Tajima's D estimates for BCIN17
405 reflect the existence of two divergent haplotypes in T, but not in G1 (not shown).

406 To infer the population history of the three populations, we applied the multiple
407 sequentially Markovian coalescent model (MSMC2; Schiffels & Wang 2020). We generated

408 27 pairs of pseudodiploid sequences for population T, 36 pairs of pseudodiploid sequences
409 for G1, and one pair of pseudodiploid sequence for G2 (for which population size is only
410 four), to evaluate replication of the demographic histories inferred by MSMC2. The analysis
411 revealed a steady, and relatively ancient, decline of estimated population sizes in all three
412 populations (Figure 3), indicating population contraction and/or ancestral population
413 structure (Nielsen & Beaumont 2009).
414

415 *Isolates from the T population are more aggressive on tomato plants.*

416 To test whether isolates collected on tomato are more aggressive on their host of origin,
417 compared to isolates collected on grapevine, pathogenicity assays were performed on whole
418 tomato plants in controlled conditions. To assess differences among isolates in terms of
419 aggressiveness on grape and tomato, we used a non-parametric analysis of variance,
420 considering the average values for each of the independent pathogenicity tests as
421 replications. We found a significant isolate effect (Kruskal Wallis test, $H (28, N= 87) =69.89221$, $p < 0.0001$), consistent with the wide range of aggressiveness levels observed for
423 the 29 isolates.

424 To compare the aggressiveness on tomato of isolates from different hosts of origin
425 (tomato vs. grape), we tested for differences in the distribution of the aggressiveness index
426 between isolates originating from the two types of hosts, considering the independent
427 repetitions of the pathogenicity test as blocks and the 13 isolates from tomato and 16
428 isolates from grapevine as replicates. We observed a significant effect of the host of origin
429 (Figure 4, Mann-Whitney U test, $p < 0.0001$) and T isolates collected on tomato were on
430 average 2.7 times more aggressive on tomato plants than isolates collected on grapevine.

431 To compare the aggressiveness on tomato of isolates from three different clusters (T, 13
432 isolates; G1, 12 isolates; G2, 4 isolates), we tested whether the aggressiveness index of
433 isolates from different clusters originate from the same distribution. A Kruskall-Wallis test
434 rejected the null hypothesis that all clusters display the same median of the aggressiveness
435 index ($H(2, N= 87) =19.80083$, $p < 0.0001$). The T population was significantly different from
436 the G1 population but not from G2 (Kruskall-Wallis test, $p< 0.0001$ and $p = 0.37$,
437 respectively). G2 and G1 were not significantly different (Kruskall-Wallis test, $p = 0.57$).
438
439 *Gene content slightly differs between tomato- and grapevine-associated populations*
440 As variation in gene content can be involved in adaptation to novel hosts (Cummings *et al.*
441 2004; Inoue *et al.* 2017; Langridge *et al.* 2015), we sought to identify genes specific to the
442 tomato (T) and grapevine (G1 and G2) populations. We first explored the mapping coverage
443 of genes previously identified in the B05.10 reference to identify sets of genes that were
444 missing or showing duplication events (Supplementary Table S4). Five B05.10 genes were
445 identified as missing in the T population, including four consecutive genes in the
446 subtelomeric region of chromosome BCIN02 that could correspond to a secondary
447 metabolism gene cluster. Among these four genes, one is coding for a NRPS-like enzyme
448 similar to the protein MelA of *Aspergillus terreus* involved in the biosynthesis of an α-keto
449 acid dimer (Geib et al. 2016), two other genes encode putative biosynthetic enzymes (FAD-
450 binding and enoyl reductase domains). Mapping coverage of B05.10 genes also suggested
451 some possible duplication events in a subtelomeric region of chromosome BCIN08, with T
452 isolates showing approximately three times as many reads as the G1 and G2 isolates for the
453 four consecutive genes Bcin08g00060 to Bcin08g00090. This suggested that the

454 corresponding region of at least 25 kb would be in three copies in the genomes of the T
455 isolates. Among the four duplicated genes, two encode Carbohydrate-Active enzymes
456 (CAZymes) known as Plant Cell Wall Degrading Enzymes (PCWDEs) as they act on pectin
457 (Glycoside Hydrolase GH28) and hemicellulose or pectin side chains (GH43).

458 We also analysed the variation in gene content using a different approach that makes no
459 use of reference genome B05.10. We built *de novo* assemblies of the genomes of T, G1 and
460 G2 isolates. The genome assembly size of the 28 isolates ranged from 41 Mb to 42.5 Mb
461 (Supplementary Table S5), which was slightly smaller than the genome assembly size of the
462 B05.10 reference isolate (42.6 Mb). We then predicted genes *ab initio* using FGENESH. The
463 number of predicted genes ranged from 11,109 to 11,311 among genomes. To compare
464 gene content in the T, G1 and G2 populations, we used ORTHOFINDER to identify 12,319
465 groups of homologous sequences. The number of groups of homologs shared by pairs of
466 isolates within populations was higher that between populations (Supplementary Table S6).
467 By looking for orthologous groups that were present in at least 75% of the genomes of a
468 focal population and missing in other populations, we identified 21 G1-specific genes, a
469 single G2-specific gene, five genes specific to the G1 and G2 populations (those already
470 detected with the first approach described above), two genes missing specifically in the G1
471 population, and a single gene specific to the T population (Supplementary Table S7). This
472 latter gene was a GH71 glycoside hydrolase (OG0011490, an α-1,3-glucanase; (Lombard *et*
473 *al.* 2014) acting on fungal cell wall. Among the genes specific to G1, we found a GH10
474 glycoside hydrolase (in OG0011469; (Lombard *et al.* 2014) acting on plant cell wall (*i.e.*
475 hemicellulose). The proteins encoded by the other G1-specific genes had no functional
476 prediction though four of them shared a domain typical of metalloenzymes (IPR11249) with

477 putative peptidase activities and three other ones showed a versatile protein-protein
478 interaction motif involved in many functions (IPR011333). Two proteins with secretion
479 signal peptides were also found specific to G1 (OG0011305 and OG0011366), with
480 OG0011366 having a predicted function of interferon alpha-inducible protein-like (Rosebeck
481 & Leaman 2008) and a predicted transmembrane helix.

482 Together, these analyses revealed that the magnitude of gene content variation is limited
483 between *B. cinerea* populations, which emphasizes the need to investigate differences in
484 allelic content at shared genes for elucidating the genomic basis of host specialization.

485

486 *McDonald-Kreitman tests of positive selection identify genes related to virulence*

487 We investigated differences in the direction and intensity of natural selection driving the
488 evolution of gene sequences in the two populations with the greatest difference in terms of
489 quantitative pathogenicity on grape and tomato, which are also the two populations with
490 the largest sample size (G1 and T). More specifically, we searched for genes with signatures
491 of positive selection in both populations that also show high sequence divergence between
492 populations, or genes with signatures of positive selection in one population, but not in the
493 other population. The direction and intensity of selection was estimated using neutrality
494 indexes computed for each individual gene in each population based on McDonald-Kreitman
495 tables of polymorphisms and substitutions at synonymous and non-synonymous sites, using
496 *B. fabae* as the outgroup. The neutrality index is expected to be below one for genes under
497 positive selection (due to an excess of non-synonymous substitutions) and above one for
498 genes under negative selection (due to a deficit of non-synonymous polymorphisms). To
499 identify genes potentially involved in host specialization, we first selected genes showing low

500 values of the neutrality index in both populations (log [neutrality index] ≤-0.5), and high
501 values of the inter-population dN/dS ratio (dN/dS in the top 5% percentile; Supplementary
502 Figure S4). This analysis identified five genes: Bcin02p04900, Bcin07p02650, Bcin09p06530,
503 Bcin14p01690, Bcin09p02190 (Supplementary Table S8). Two genes (Bcin07p02650,
504 Bcin14p01690) code for glycosyl hydrolases of the GH5 family, a family that includes
505 enzymes acting on plant cell walls and enzymes act on fungal cell walls. One of the two
506 genes (Bcin14p01690), has a cellulose binding domain which strongly suggests a role as
507 PCWDE. Two genes (Bcin02p04900, Bcin09p02190) are involved in basic cell functions (a
508 DNA nuclease and a protein involved in ribosome biogenesis). The last gene (Bcin09p06530)
509 has no known domain.

510 We also selected genes showing low values of the neutrality index in one population (log
511 [neutrality index] ≤-0.5), and high values in the other (log [neutrality index]>=0)
512 (Supplementary Figure S4). This analysis identified 392 genes in G1 and 428 genes in T
513 (Supplementary Table S8). Functional enrichment analyses revealed contrasting results in
514 the G1 and T populations (Supplementary Table S8). In the G1 population, we identified a
515 significant two-fold enrichment in transporter-encoding genes among the 392 genes with
516 signatures of positive selection. These 28 transporters included many candidates with
517 putative roles in nutrition such as the transport of sugars and amino-acids (five genes of
518 each). Fifteen of them were proteins of the major facilitator superfamily (MFS). The MFS
519 transporter-encoding genes included five putative sugar transporters and ten unknown
520 transporters that could have roles in various processes, including obtaining nutrients from
521 the host, efflux of fungi-toxic compounds or the export of fungal phytotoxins (Hartmann *et*
522 *al.* 2018; Maruthachalam *et al.* 2011; Perlin *et al.* 2014). Finally, five of the 392 genes

523 encoded ATPase transporters including the BcPrm1 P-type Ca2+/Mn2+-ATPase that
524 mediates cell-wall integrity and virulence in *B. cinerea* (Plaza *et al.* 2015).

525 In the T population, the list of 428 genes with signatures of positive selection showed a
526 significant 2.5-fold enrichment in genes encoding for proteins involved in oxidative stress
527 response (eight genes). These genes encode enzymes that are able to detoxify reactive
528 oxygen species *i.e.* Glutathione-S-transferases (BcGST1, 9 and 24), the superoxide dismutase
529 BcSOD1, and two peroxidases (BcPRX8 and BcCCP1). In addition, the list of 428 genes also
530 included *BcatrO*, which encodes the transporter BcAtrO involved in the resistance to $H_2O_2$
531 (Pane *et al.* 2008).

532 A 2.5-fold enrichment was also observed for the genes coding for CAZYmes acting as
533 PCWDEs (ten genes) especially for those involved in the modification of hemicellulose (five
534 genes; Espino *et al.* 2010 ; Supplementary Table S8) such as the xylanase BcXyn11A that is
535 required for full virulence on tomato (Brito *et al.* 2006).

536

537 *Selective sweeps in regions encompassing genes encoding enzymes involved in carbohydrates*
538 *metabolism*

539 To identify genomic regions with signatures of selective sweeps, we conducted three
540 different genome scans, using different features of the data: i) hapFLK, which detects hard
541 and soft sweeps based on patterns of differentiation between clusters of haplotypes
542 between populations; ii) $nS_L$, which detects hard and soft sweeps based on the distribution
543 of fragment length between mutations and the distribution of the number of segregating
544 sites between pairs of chromosomes; iii) and SweeD's CLR, which detects hard sweeps based
545 on the site frequency spectrum. The *CLR* and $nS_L$ metrics are population-specific and were

546 computed for each population independently, while the *hapFLK* metric is $F_{ST}$-based and was
547 thus computed for populations G1 and T (Supplementary figure S5). We identified candidate
548 SNPs located in (hard or soft) selective sweeps as the SNPs that were in the top 5% of the
549 hapFLK statistic, but also in the top 5% of either the nSL or CLR statistic in a least one
550 population (Figure 5). In total, this approach identified 4,667 SNPs of which 1,300 were
551 localized in coding sequences, 256 in introns, 830 in untranslated transcribed regions, and
552 465 less than 1500bp upstream of coding sequences. These 2,851 SNPs corresponded to 351
553 genes, of which 15 were identified by SNPs in the top 5% of selective sweep metrics in both
554 populations, 200 by SNPs in the top 5% of selective sweep metrics in population T, and 175
555 by SNPs in the top 5% of selective sweep metrics in population G1 (Supplementary table S9).
556 Candidate genes in the selective sweep regions of each population included genes coding
557 for proteins with functions consistent with a role in infection, such as transporters, CAZymes,
558 putative effectors and two genes that are confirmed virulence factors. Indeed, one region
559 identified in the G1 population contains the gene encoding the Pectin Methyl Esterase
560 BcPme1 required for full virulence of *B. cinerea* on several host including grapevine (Valette-
561 Collet *et al.* 2003), and another region contains the gene encoding BcCgf1, a small secreted
562 protein that is essential for infection structure development (Zhang *et al.* 2020).
563 Nevertheless, it is unlikely that all genes in selective sweep regions have been direct targets
564 of positive selection, most of them being possible hitch-hikers. This could be the reason why
565 no significant functional enrichment was detected among these genes.
566 Finally, comparison of the lists of genes identified in the selective sweep regions and
567 those with signatures of positive selection according to McDonald-Kreitman tests identified
568 nine genes in common for the T population and four genes in common for the G1 population

569 (Supplementary table S9). This suggests that these genes were subjected to both recurrent
570 positive selection for amino-acid changes and to recent positive selection in populations of
571 *B. cinerea*. In addition, we can hypothesize that these genes may be the actual targets of
572 positive selection, and that surrounding candidate genes could be only hitch-hikers.
573 Functional annotations of the genes that show both recurrent positive selection for amino-
574 acid changes and recent positive selection signals further indicated a cutinase-encoding gene
575 (Bcin01g09430) in the T population and a sugar transporter encoding gene (Bcin16g00530) in
576 the G1 population, and also pointed out various other functions.
577

578 **Discussion**

579 *Genetic differentiation in* B. cinerea *between populations associated with grapevine and*
580 *tomato*

581 Genetic structure associated with the host of origin in *B. cinerea* has been extensively
582 investigated (reviewed in Walker 2016). Most of the studies based sufficient sampling size
583 (n>100) found significant population differentiation between populations of *B. cinerea* from
584 different hosts (e.g. in Chile, Tunisia, Hungary, United Kingdom). One noticeable exception is
585 the population of *B. cinerea* collected from various hosts in California (Ma & Michailides
586 2005), for which whole-genome sequencing data and cross-infectivity assays revealed small
587 differences in pathogenicity against different hosts and shallow population structure (Atwell
588 *et al.* 2015; Caseys *et al.* 2020; Soltis *et al.* 2019). In France, previous research concluded that
589 *B. cinerea* populations were differentiated according to some of their hosts, including
590 tomato, grapevine and, to a lesser extent, bramble (Fournier & Giraud 2008; Walker *et al.*

591 2015). In two recent studies comparing the aggressiveness of isolates coming from diverse
592 hosts, the disease severity caused by isolates from tomato was significantly greater than the
593 severity caused by isolates from grape or other crops, thus indicating that *B. cinerea*
594 populations parasitizing tomato were specialized to this host (Bardin *et al.* 2018; Mercier *et*
595 *al.* 2019). Here we show that populations parasitizing tomato and grapevine are subdivided
596 into three populations, two being associated with grapevine (G1 and G2) and one with
597 tomato (T). This pattern of population genetic structure differs from previous findings
598 (Walker et al., 2015), as three populations parasitizing grapevine had previously been
599 detected, but studies differ in terms of sampling and genotyping schemes (thousands of
600 SNPs vs 8 SSR markers, and an order of magnitude of difference in the size of sample sets).
601 The clear pattern of population subdivision found in our study also stands in sharp contrast
602 with the weak, host- and geography-independent population subdivision uncovered among
603 populations from various hosts (including tomato and grape) in California (Atwell *et al.* 2015;
604 Atwell *et al.* 2018). This difference in population structure between France and California
605 indicates that the factors leading to host-specific differentiation between *B. cinerea*
606 pathogens from grape and tomato do not operate everywhere. The differences in LD decay
607 (up to an order of magnitude longer in our study compared to Californian *B. cinerea*) and
608 nucleotide diversity (half as much in our study compared to Californian *B. cinerea*) also
609 suggest that the demographic history and population biology of the pathogen is contrasted
610 between the two regions.

611 Multiple factors can contribute to reduce gene flow between populations parasitizing
612 grapevine and tomato. A first possible factor limiting gene flow is adaptation to host. Mating
613 in *B. cinerea* occurs on the host after infection, between individuals that were thus

614 sufficiently adapted to infect the same host, which induces assortative mating with respect
615 to host use and reduce opportunities for inter-population crosses (Giraud 2006; Giraud *et al.*
616 2010; Giraud *et al.* 2008). Another factor possibly limiting gene flow between populations
617 infecting tomato and grape is habitat isolation (*i.e.* reduced encounters caused by mating in
618 different habitats). Tomatoes are grown in nurseries before being dispatched to the fields or
619 greenhouses, and this may generate habitat isolation if sexual reproduction in the pathogen
620 occurs in nurseries for the tomato-infecting population of *B. cinerea*. Such habitat isolation
621 may contribute to promote adaptation to tomato, by preventing the immigration of alleles
622 which are favorable for infection of non-tomato hosts but not favorable for infection of
623 tomato. Differences in the timing of epidemics are unlikely to contribute to this habitat
624 isolation, as the period of infection of greenhouse tomatoes runs from late to early winter,
625 which includes the period of infection of grape. The same goes for the location of epidemics,
626 since the sites studied were chosen because the two types of crops are grown nearby. A final
627 possibility to explain the lack of gene flow between *B. cinerea* populations from grape and
628 tomato is that the frequency of sexual reproduction might be lower in populations infecting
629 greenhouse tomatoes. Higher winter temperatures and the removal of plant residues in the
630 greenhouse represent conditions that are less conducive to sexual reproduction. However,
631 our estimates of LD decay and recombination rates are not consistent with a substantially
632 lower frequency of sexual reproduction in the population associated with tomato, compared
633 to the population associated with grape.
634

635 *Widespread signatures of selection along genomes*

636 We identified little variation in the gene content among T, G1 and G2 populations, with one
637 gene specific to T, 22 genes specific to G1 and five genes shared between G1 and G2 but not
638 T, suggesting that gene gain or loss is not the main process of adaptation to tomato. In
639 parallel to our analysis of presence/absence variation, our genome scans for positive
640 selection allowed us to pinpoint several genomic regions which may harbour determinants
641 of ecological differentiation between the population specialized to tomato and the
642 population parasitizing grapevine. In order to cover multiple time scales and different
643 signatures of positive selection, we used a variety of analytical approaches. The McDonald-
644 Kreitman test focuses on genes and detects repeated episodes of selective sweeps fixing
645 non-synonymous substitutions, thus generating a higher ratio of amino acid divergence to
646 polymorphism (Dn / Pn), relative to the ratio of silent divergence to polymorphism (Ds / Ps),
647 than expected under neutrality. The values of nucleotide diversity and Tajima's D measured
648 in the T population specialized in tomatoes were very close to the values measured for the
649 two other populations, which is not consistent with a very recent origin of this population
650 and justifies the use of the McDonald-Kreitman test. Genome scans for selective sweeps
651 detect more recent events, and by nature these methods can also detect genes that are not
652 directly the target of selection, but may have hitch-hiked due to physical linkage with sites
653 under positive selection. However, the LD decay values measured for the T population
654 remain moderate (9kb), and we used a combination of different selective sweep metrics to
655 substantially shorten the list of candidate genes, which should reduce the impact of genetic
656 hitch-hiking on our list of genes under recent positive selection. Despite using a more
657 conservative approach, we identified more selective sweeps than in the generalist

658 *Sclerotinia sclerotiorum* fungus (Derbyshire *et al.* 2019), or in the Californian generalist
659 population of *B. cinerea* (Soltis *et al.* 2019).
660

661 *Genes under positive selection*

662 We identified a number of genes showing signatures of positive selection using the
663 approaches discussed above and highlighting potential candidates for their role in host
664 specialization. Functional annotation of the *B. cinerea* genome and previous experimental
665 studies provided lists of genes involved in host-pathogen interaction and in other
666 developmental processes (Amselem *et al.* 2011; Nakajima & Akutsu 2014; Rodriguez-
667 Moreno *et al.* 2018). We used these published lists of genes to investigate whether some
668 specific biological processes were subjected to positive selection in the different
669 populations. Our data revealed that the five genes showing the strongest signatures of
670 selection in both the T and G1 populations, but also showing high sequence divergence
671 between the two populations, included two genes encoding for CAZymes with glycoside
672 hydrolase activity, which are potential PCWDEs. We also found that the 428 genes showing
673 the strongest signatures of selection in the T population with McDonald-Kreitman tests were
674 enriched both in genes coding for PCWDEs and in genes coding for enzymes involved in the
675 oxidative stress response. Notably, ten genes encoding secreted CAZymes targeting
676 compounds of the plant cell wall, *i.e.* cellulose and pectin (Amselem et al., 2011; (Lombard *et*
677 *al.* 2014), were found under positive selection. One of these ten genes encodes the xylanase
678 BcXyn11A that has previously been shown to be a virulence factor on tomato (Brito *et al.*
679 2006). Another one encodes a cutinase (Bcin01g09430) that was further detected in a
680 selective sweep region suggesting, recurrent and recent positive selection events. Additional

681 PCWDEs were found in other genomic regions identified as selective sweeps. Finally, our
682 comparative genomic analysis suggested that a subtelomeric region that contains two
683 PCWDEs acting on pectin and/or hemicellulose is duplicated in the T population.
684 Necrotrophic species have important repertoires of CAZymes especially those corresponding
685 to PCWDEs which are known to act as major virulence factors in fungi (Zhao *et al.* 2013;
686 Rodriguez-Moreno *et al.* 2018). The genome of the reference isolate of *B. cinerea* (B05.10)
687 revealed 118 PCWDEs (Amselem et al., 2011) and our data suggest that, within this
688 repertoire, some cellulases and pectinases may be of particular importance for the
689 degradation of tomato cell wall. SNPs within a pectinesterase gene were also associated with
690 virulence on tomato in a previous genome-wide association study (Soltis *et al.* 2019).

691 In addition to PCWDEs, the single gene that we found as present in the T population but
692 missing in the G1 and G2 populations encoded a CAZyme acting on the fungal cell wall, an α-
693 1,3-glucanase classified as a member of the GH71 family. In the fungal cell wall, α-1,3-glucan
694 is a major component that encloses the α-(1,3)-glucan-chitin fibrillar core. Because of its
695 external localization and specific composition, α-1,3-glucan of pathogenic fungi plays a major
696 role in infection-related morphology and host recognition (Beauvais *et al.* 2013; King *et al.*
697 2017). A dozen of genes of *B. cinerea* encode for enzymes of the GH71 family (Amselem et
698 al., 2011). The T-specific GH71 CAZyme might therefore have been retained in the T
699 population as a mean to specifically facilitate infection of tomato by modification of the
700 fungal cell wall resulting in adaptive morphological changes or impairment of host
701 recognition.

702 As mentioned above, the McDonald-Kreitman tests also indicated that eight genes coding
703 for enzymes that detoxify reactive oxygen species showed signatures of positive selection in

704 the T population. During infection, *B. cinerea* encounters an oxidative burst, an early host
705 response that results in the death of plant cells. This mechanism is used in turn by *B. cinerea*
706 to achieve full virulence but this also implies that the fungus has to resist to this toxic
707 environment. The fungal oxidative stress response system includes detoxifying enzymes such
708 as superoxide dismutases (SODs) that convert $O_2\cdot$ into the less toxic $H_2O_2$, and catalases,
709 peroxidases and peroxiredoxins that convert $H_2O_2$. Additional non-enzymatic mechanisms
710 include the oxidation of compounds such as glutathione (Heller & Tudzynski 2011). In
711 addition to the genes encoding SOD (BcSOD3), peroxidases (BcCCP1), peroxiredoxins
712 (BcPRX8) and other detoxifying enzymes sur as glutathione-S-transferases (BcGST1, 9 and
713 24), a gene encoding the transporter BcAtrO also showed a signature of positive selection in
714 the T population. Inactivation of this gene previously suggested that it would allow the efflux
715 of $H_2O_2$ and resistance this reactive oxygen species (Pane *et al.* 2008). Altogether, our data
716 suggest that the oxidative burst occurring in the *B. cinerea*/tomato interaction is particularly
717 challenging for the fungus.
718

719 *Concluding remarks*

720 We identified a population of *B. cinerea* specialized to tomato, which diverged from a
721 grapevine-associated population. Genome scans for selective sweeps and McDonald-
722 Kreitman tests revealed widespread signatures of positive selection that identified genes
723 that may contribute to the pathogen's adaptation to its tomato host. Candidate genes for
724 specialization to tomato were significantly enriched in those encoding cellulases, pectinases
725 and enzymes involved in the oxidative stress response, suggesting that the ability to degrade
726 the host cell wall and to cope with the oxidative burst are two key process in the *B.*

*cinerea*/tomato interaction. Our work sets the stage for future studies aiming to elucidate the phenotypic and fitness effects of the candidate genes for specialization of *B. cinerea* to tomato, for instance by knocking-out or replacing candidate genes for host specialization.

## Acknowledgements

AM was supported by a grant from the Doctoral School « Sciences du Végétal », Université Paris-Saclay. We are grateful to INRAE-LIPM Bioinformatics Platform (Jérôme Gouzy, Sébastien Carrère, Erika Sallet, Ludovic Legrand) and Genotoul Bioinformatics Platforms Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing computational support and resources. This work was supported by a grant overseen by the French National Research Agency (ANR) as part of the "Investments d'Avenir" Programme (LabEx BASC; ANR-11-LABX-0034) and by the INRAE department SPE. The BIOGER unit also benefits from the support of "Saclay Plant Science-SPS" (ANR-17-EUR-0007).

741 ## References


742 Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, *et al.* (2019) SignalP 5.0 improves signal peptide predictions
743 using deep neural networks. *Nature biotechnology* **37**, 420-423.
744 Amselem J, Cuomo CA, van Kan JA, *et al.* (2011) Genomic analysis of the necrotrophic fungal pathogens
745 *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS genetics* **7**, e1002230.
746 Atwell S, Corwin J, Soltis N, *et al.* (2015) Whole genome resequencing of Botrytis cinerea isolates identifies high
747 levels of standing diversity. *Frontiers in microbiology* **6**, 996.
748 Atwell S, Corwin JA, Soltis N, *et al.* (2018) Resequencing and association mapping of the generalist pathogen
749 Botrytis cinerea. *bioRxiv*, 489799.
750 Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. *Genome research* **17**,
751 1219-1227.
752 Bardin M, Leyronas C, Troulet C, Morris CE (2018) Striking similarities between Botrytis cinerea from non-
753 agricultural and from agricultural habitats. *Frontiers in plant science* **9**, 1820.
754 Beauvais A, Bozza S, Kniemeyer O, *et al.* (2013) Deletion of the α-(1,3)-glucan synthase genes induces a
755 restructuring of the conidial cell wall responsible for the avirulence of *Aspergillus fumigatus*. *PLoS*
756 *pathogens* **9**, e1003716.
757 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data.
758 *Bioinformatics* **30**, 2114-2120.
759 Bonhomme M, Chevalet C, Servin B, *et al.* (2010) Detecting selection in population trees: the Lewontin and
760 Krakauer test extended. *Genetics* **186**, 241-262.
761 Brito N, Espino JJ, González C (2006) The endo-β-1, 4-xylanase Xyn11A is required for virulence in Botrytis
762 cinerea. *Molecular Plant-Microbe Interactions* **19**, 25-32.
763 Caseys C, Shi G, Soltis N, *et al.* (2020) Quantitative interactions drive *Botrytis cinerea* disease outcome across
764 the plant kingdom. *bioRxiv*, 507491.
765 Cummings CA, Brinig MM, Lepp PW, van de Pas S, Relman DA (2004) Bordetella species are distinguished by
766 patterns of substantial gene loss and host adaptation. *Journal of bacteriology* **186**, 1484-1492.
767 Decognet V, Bardin M, Trottin-Caudal Y, Nicot PC (2009) Rapid change in the genetic diversity of *Botrytis*
768 *cinerea* populations after the introduction of strains in a tomato glasshouse. *Phytopathology* **99**, 185-
769 193.
770 Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: Version II. *Plant Molecular Biology*
771 *Reporter* **1**, 19-21.
772 Derbyshire MC, Denton-Giles M, Hane JK, *et al.* (2019) A whole genome scan of SNP data suggests a lack of
773 abundant hard selective sweeps in the genome of the broad host range plant pathogenic fungus
774 Sclerotinia sclerotiorum. *PloS one* **14**, e0214201.

775 Elad Y, Pertot I, Prado AMC, Stewart A (2016) Plant hosts of Botrytis spp. In: *Botrytis–the fungus, the pathogen*
776 *and its management in agricultural systems*, pp. 413-486. Springer.
777 Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically
778 improves orthogroup inference accuracy. *Genome biology* **16**, 157.
779 Espino JJ, Gutierrez-Sanchez G, Brito N, *et al.* (2010) The *Botrytis cinerea* early secretome. *Proteomics* **10**, 3020-
780 3034.
781 Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through
782 haplotype differentiation among hierarchically structured populations. *Genetics* **193**, 929-941.
783 Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective
784 sweeps using haplotype structure. *Molecular biology and evolution* **31**, 1275-1291.
785 Fisher MC, Henk DIA, Briggs CJ, *et al.* (2012) Emerging fungal threats to animal, plant and ecosystem health.
786 *Nature* **484**, 186.
787 Fournier E, Giraud T (2008) Sympatric genetic differentiation of a generalist pathogenic fungus, *Botrytis cinerea,*
788 on two different host plants, grapevine and bramble. *Journal of Evolutionary Biology* **21**, 122-132.
789 Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-*
790 *bio].*
791 Geib E, Markus Gressler M, Iuliia Viediernikova I, Falk Hillmann F, Ilse D. Jacobsen ID, Sandor Nietzsche S,
792 Christian Hertweck C, Matthias Brock M (2016) A Non-canonical Melanin Biosynthesis Pathway
793 Protects Aspergillus terreus Conidia from Environmental Stress.
794 *Cell Chemical Biology* **23**, 587-597.
795 Giraud T (2006) Speciation: Selection against migrant pathogens: the immigrant inviability barrier in pathogens.
796 *Heredity* **97**, 316-318.
797 Giraud T, Gladieux P, Gavrilets S (2010) Linking the emergence of fungal plant diseases with ecological
798 speciation. *Trends in ecology & evolution* **25**, 387-395.
799 Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME (2008) Speciation in fungi. *Fungal genetics and*
800 *biology : FG & B* **45**, 791-802.
801 Gladieux P, Guerin F, Giraud T, *et al.* (2011) Emergence of novel fungal pathogens by ecological speciation:
802 importance of the reduced viability of immigrants. *Molecular ecology* **20**, 4521-4532.
803 Gladieux P, Ravel S, Rieux A, *et al.* (2018) Coexistence of multiple endemic and pandemic lineages of the rice
804 blast pathogen. *MBio* **9**.
805 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history
806 of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**, e1000695.
807 Hartmann FE, McDonald BA, Croll D (2018) Genome-wide evidence for divergent selection between
808 populations of a major agricultural pathogen. *Molecular ecology* **27**, 2725-2741.
809 Heller J, Tudzynski P (2011) Reactive oxygen species in phytopathogenic fungi: signaling, development, and
810 disease. *Annual review of phytopathology* **49**, 369-390.

811 Huson DH, Bryant D (2005) Estimating phylogenetic trees and networks using SplitsTree 4. *Manuscript in*
812 *preparation, software available from www. splitstree. org.*
813 Inoue Y, Vy TTP, Yoshida K, *et al.* (2017) Evolution of the wheat blast fungus through functional losses in a host
814 specificity determinant. *Science* **357**, 80-83.
815 Johnson M, Zaretskaya I, Raytselis Y, *et al.* (2008) NCBI BLAST: a better web interface. *Nucleic acids research* **36**,
816 W5-W9.
817 Jones P, Binns D, Chang HY, *et al.* (2014) InterProScan 5: genome-scale protein function classification.
818 *Bioinformatics* **30**, 1236-1240.
819 Khvorykh GV, Khrunin AV (2020) imputeqc: an R package for assessing imputation quality of genotypes and
820 optimizing imputation parameters. *BMC bioinformatics* **21**, 1-14.
821 King R, Urban M, Lauder RP, *et al.* (2017) A conserved fungal glycosyltransferase facilitates pathogenesis of
822 plants by enabling hyphal growth on solid surfaces. *PLoS pathogens* **13**, e1006672.
823 Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a
824 hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-580.
825 Langridge GC, Fookes M, Connor TR, *et al.* (2015) Patterns of genome evolution that have accompanied host
826 adaptation in *Salmonella*. *Proceedings of the National Academy of Sciences of the United States of*
827 *America* **112**, 863-868.
828 Latorre SM, Reyes-Avila CS, Malmgren A, *et al.* (2020) Differential loss of effector genes in three recently
829 expanded pandemic clonal lineages of the rice blast fungus. *BMC biology* **18**, 1-15.
830 Leyronas C, Bardin M, Berthier K, *et al.* (2018) Assessing the phenotypic and genotypic diversity of *Sclerotinia*
831 *sclerotiorum* in France. *European journal of plant pathology* **152**, 933-944.
832 Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples.
833 *Bioinformatics (Oxford, England)* **30**, 2843-2851.
834 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*
835 **25**, 1754-1760.
836 Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
837 **25**, 2078-2079.
838 Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The Carbohydrate-Active enZymes
839 database (CAZy) in 2013. *Nucleic acids research* **42**, D490-495.
840 Luo R, Liu B, Xie Y, *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo
841 assembler. *GigaScience* **1**, 18.
842 Ma Z, Michailides TJ (2005) Genetic structure of Botrytis cinerea populations from different host plants in
843 California. *Plant disease* **89**, 1083-1089.
844 Marchi N, Excoffier L (2020) Gene flow as a simple cause for an excess of high-frequency-derived alleles.
845 *Evolutionary applications* **13**, 2254-2263.

846 Maruthachalam K, Klosterman SJ, Kang S, Hayes RJ, Subbarao KV (2011) Identification of pathogenicity-related
847 genes in the vascular wilt fungus *Verticillium dahliae* by *Agrobacterium tumefaciens*-mediated T-DNA
848 insertional mutagenesis. *Molecular Biotechnology* **49**, 209-221.
849 McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments
850 with respect to biological variation. *Nucleic acids research* **40**, 4288-4297.
851 McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652-
852 654.
853 Mercier A, Carpentier F, Duplaix C, *et al.* (2019) The polyphagous plant pathogenic fungus *Botrytis cinerea*
854 encompasses host-specialized and generalist populations. *Environmental microbiology*.
855 Nakajima M, Akutsu K (2014) Virulence factors of *Botrytis cinerea*. *Journal of General Plant Pathology* **80**, 15-23.
856 Nielsen R, Beaumont MA (2009) Statistical inferences in phylogeography. *Molecular ecology* **18**, 1034-1047.
857 Nielsen R, Williamson S, Kim Y, *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome*
858 *research* **15**, 1566-1575.
859 Nosil P, Vines TH, Funk DJ (2005) Perspective: reproductive isolation caused by natural selection against
860 immigrants from divergent habitats. *Evolution; international journal of organic evolution* **59**, 705-719.
861 Pane C, Rekab D, Firrao G, Ruocco M, Scala F (2008) A NOVEL GENE CODING FOR AN ABC TRANSPORTER IN
862 BOTRYTIS CINEREA (BOTRYOTINIA FUCKELIANA) IS INVOLVED IN RESISTANCE TO $H^+$ $O^+$. *Journal of*
863 *Plant Pathology*, 453-462.
864 Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps
865 in thousands of genomes. *Molecular biology and evolution* **30**, 2224-2234.
866 Perlin MH, Andrews J, Toh SS (2014) Essential letters in the fungal alphabet: ABC and MFS transporters and
867 their roles in survival and pathogenicity. *Advances in genetics* **85**, 201-253.
868 Plaza V, Lagües Y, Carvajal M, *et al.* (2015) bcpmr1 encodes a P-type Ca2+/Mn2+-ATPase mediating cell-wall
869 integrity and virulence in the phytopathogen Botrytis cinerea. *Fungal Genetics and Biology* **76**, 36-46.
870 Porquier A, Morgant G, Moraga J, *et al.* (2016) The botrydial biosynthetic gene cluster of *Botrytis cinerea*
871 displays a bipartite genomic structure and is positively regulated by the putative Zn(II)2Cys6
872 transcription factor BcBot6. *Fungal genetics and biology : FG & B* **96**, 33-46.
873 Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature*
874 **461**, 489-494.
875 Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression
876 analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140.
877 Rodriguez-Moreno L, Ebert MK, Bolton MD, Thomma BPHJ (2018) Tools of the crook- infection strategies of
878 fungal plant pathogens. *The Plant journal : for cell and molecular biology* **93**, 664-674.
879 Rosebeck S, Leaman DW (2008) Mitochondrial localization and pro-apoptotic effects of the interferon-inducible
880 protein ISG12a. *Apoptosis : an international journal on programmed cell death* **13**, 562-572.

881 Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data:
882 applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human*
883 *Genetics* **78**, 629-644.
884 Schiffels S, Wang K (2020) MSMC and MSMC2: the multiple sequentially markovian coalescent. In: *Statistical*
885 *Population Genomics*, pp. 147-166. Humana, New York, NY.
886 Schulze-Lefert P, Panstruga R (2011) A molecular evolutionary concept connecting nonhost resistance,
887 pathogen host range, and pathogen speciation. *Trends in plant science* **16**, 117-125.
888 Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness.
889 In: *Gene Prediction*, pp. 227-245. Springer.
890 Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: 'magic' but not rare?
891 *Trends in ecology & evolution* **26**, 389-397.
892 Simko I, Piepho H-P (2012) The area under the disease progress stairs: calculation, advantage, and application.
893 *Phytopathology* **102**, 381-389.
894 Skoglund P, Mallick S, Bortolini MC, *et al.* (2015) Genetic evidence for two founding populations of the
895 Americas. *Nature* **525**, 104-108.
896 Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes
897 and promoters. *Genome biology* **7**, S10.11-12.
898 Soltis NE, Atwell S, Shi G, *et al.* (2019) Interactions of tomato and Botrytis cinerea genetic diversity: Parsing the
899 contributions of host differentiation, domestication, and pathogen variation. *The Plant Cell* **31**, 502-
900 519.
901 Sonnhammer ELL, Von Heijne G, Krogh A A hidden Markov model for predicting transmembrane helices in
902 protein sequences **6**, 175-182.
903 Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Molecular biology and evolution* **28**, 63-70.
904 Stukenbrock EH, McDonald BA (2008) The origins of plant pathogens in agro-ecosystems. *Annual review of*
905 *phytopathology* **46**, 75-100.
906 Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to
907 combinations of populations. *Bioinformatics (Oxford, England)* **24**, 2498-2504.
908 Valette-Collet O, Cimerman A, Reignault P, Levis C, Boccara M (2003) Disruption of Botrytis cinerea pectin
909 methylesterase gene Bcpme1 reduces virulence on several host plants. *Molecular Plant-Microbe*
910 *Interactions* **16**, 360-367.
911 van Kan JA, Stassen JH, Mosbach A, *et al.* (2017) A gapless genome sequence of the fungus *Botrytis cinerea*.
912 *Molecular plant pathology* **18**, 75-89.
913 Walker A (2016) Diversity within and between species of *Botrytis*. In: *Botrytis - The Fungus, the Pathogen and*
914 *its Management in Agricultural Systems* (eds. Fillinger S, Elad Y), pp. 91-125.
915 Walker A, Gladieux P, Decognet V, *et al.* (2015) Population structure and temporal maintenance of the
916 multihost fungal pathogen *Botrytis cinerea*: causes and implications for disease management.
917 *Environmental microbiology* **17**, 1261-1274.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences* **13**, 555-556.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829.

Zhang C, Dong SS, Xu JY, He WM, Yang TL (2019a) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics (Oxford, England)* **35**, 1786-1788.

Zhang C, Dong SS, Xu JY, He WM, Yang TL (2019b) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786-1788.

Zhang MZ, Sun CH, Liu Y, *et al.* (2020) Transcriptome analysis and functional validation reveal a novel gene, BcCGF1, that enhances fungal virulence by promoting infection-related development and host penetration. *Molecular plant pathology* **21**, 834-853.

Zhao Z, Liu H, Wang C, Xu JR (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC genomics* **14**, 274.

Zhong Z, Marcel TC, Hartmann FE, *et al.* (2017) A small secreted protein in *Zymoseptoria tritici* is responsible for avirulence on wheat cultivars carrying the Stb6 resistance gene. *The New phytologist* **214**, 619-631.

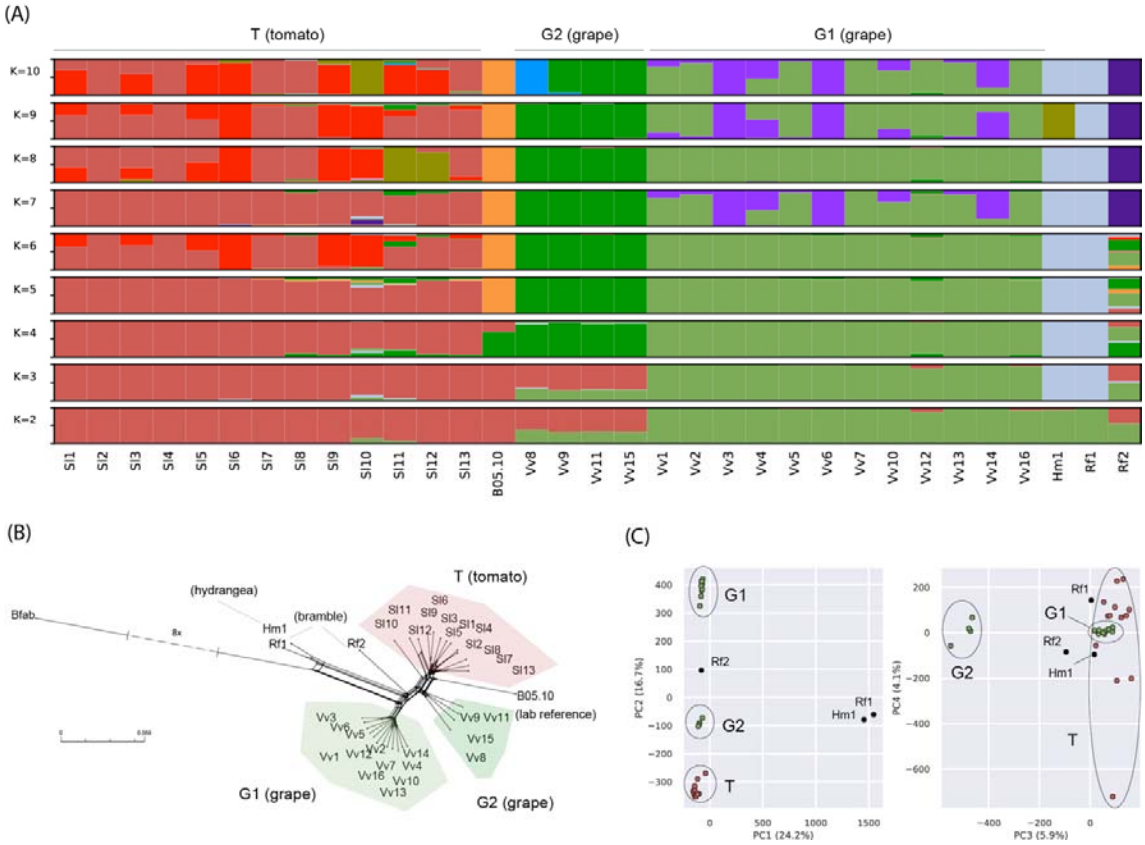


**Figure 1.** Population subdivision inferred based on SNPs identified in 32 isolates of *Botrytis cinerea* collected on tomato (red, Sl prefix, *Solanum lycopersicum*), grape (green, Vv prefix, *Vitis vinifera*), bramble (black, Rf prefix, *Rubus fruticosus*) and hydrangea (black, Hm prefix, *Hydrangea macrophylla*). (A) Ancestry proportions in K clusters, as estimated with the sNMF program. Each multilocus genotype is represented by a vertical bar divided into K segments, indicating membership in K clusters. (B) Neighbor-net phylogenetic network estimated with SPLITSTREE, with one isolate of *B. fabae* (Bfab) used as the outgroup. Reticulations indicate phylogenetic conflicts caused by recombination or incomplete lineage sorting. (C) Principal component analysis showing first four principal components PC1, PC2, PC3 and PC4. Isolate B05.10 in (A) and (B) is the reference genome for *B. cinerea* (van Kan et al. 2016)

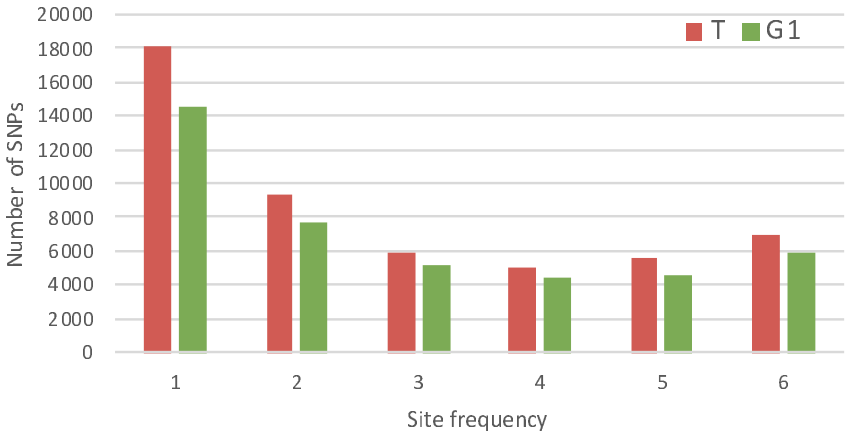


**Figure 2.** Site frequency spectra estimated in populations T2 and Vv1 based on 85,415 SNPs using the python package DADI, projecting sample sizes to seven haploid genomes. Projection consists in averaging over all possible re-samplings of the larger sample size data, thus biallelic positions with data for less than seven individuals are not included in calculations and not counted as SNPs.

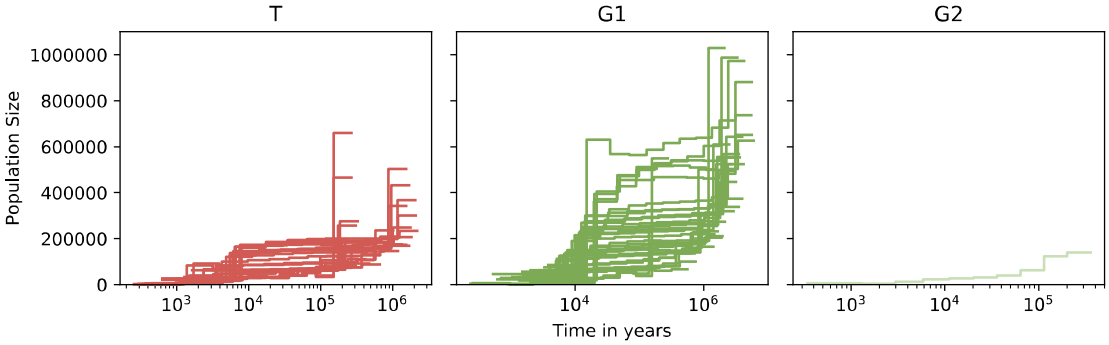


**Figure 3.** Demographic history of populations T, G1 and G2 of *B. cinerea*. Effective population size and time were estimated using the Multiple Sequential Markovian Coalescent method implemented in the MSMC2 program, based on 27 pairs of pseudodiploid sequences (i.e. four haploid sequences) for population T, 36 pairs of pseudodiploid sequences for G1, and one pair of pseudodiploid sequence for G2 (for which sample size was only four haploid genomes).

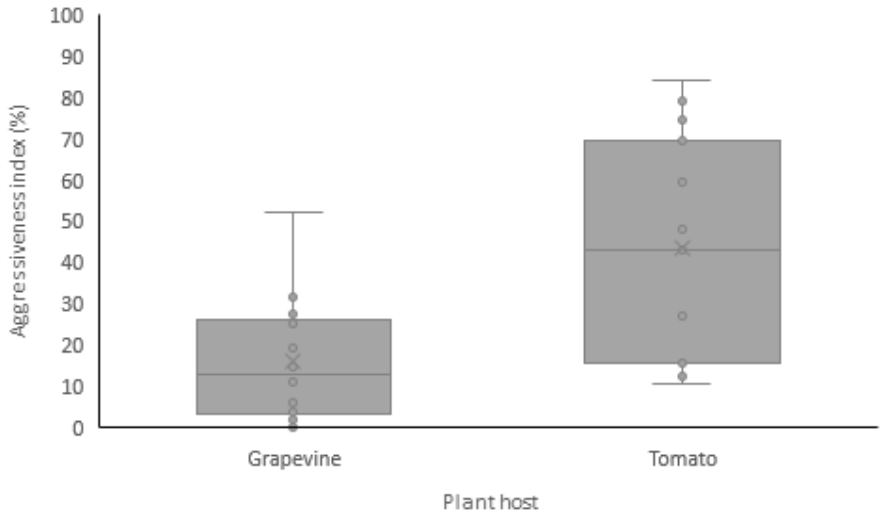


**Figure 4.** Boxplots representing the extent of aggressiveness (% relative to the reference isolate BC1) on tomato of the *Botrytis cinerea* isolates collected on grapevine (n=16) and tomato (n=13). For each boxplot, mean (crosses), median (horizontal lines), values of the aggressiveness index (circles), 25-75% quartiles, and maximum and minimum values are represented.

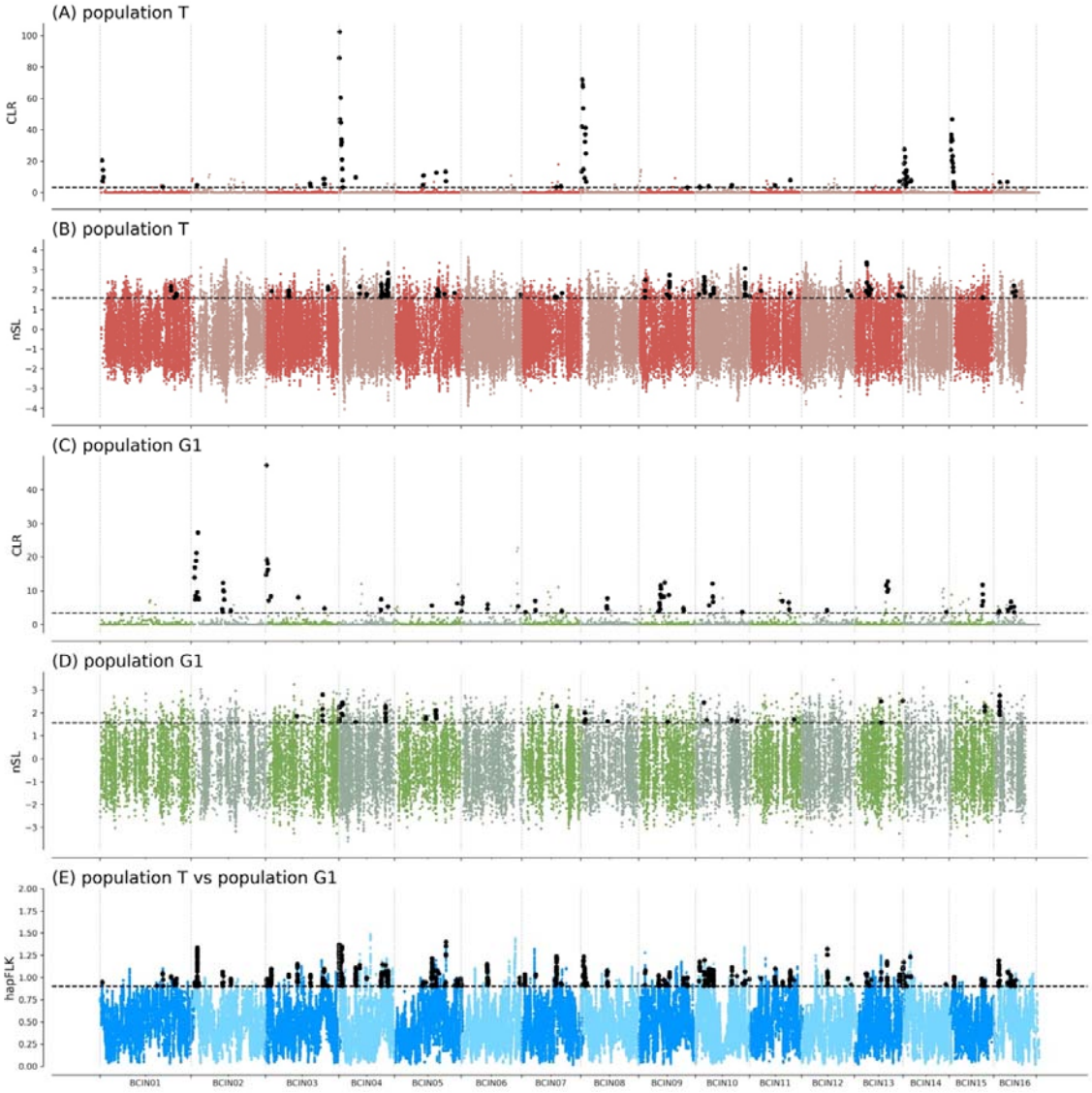


**Figure 5.** Genome scans for selective sweeps in the T and G1 populations of *Botrytis cinerea*, parasitizing tomato and grapevine, respectively. (A) and (C): Composite likelihood ratio (CLR) estimated using the SWEED software (Nielsen *et al.* 2005; Pavlidis *et al.*, 2013) in T and G1, respectively. (B) and (D) Number of segregating sites by length, $nS_L$, estimated using the NSL software (Ferrer-Admetlla *et al.* 2014) in T and G1, respectively. (E) hapFLK statistic estimated using the HAPFLK software (Bonhomme *et al.* 2010; Fariello *et al.* 2013). Horizontal dashed black lines represent the top 5%. Vertical dashed grey lines represent the boundaries of the 16 chromosomes and the chromosome names are set along the x axis. SNPs in chromosomes with names ending with odd numbers are represented in dark colors, while

SNPs in chromosomes with names ending with even numbers are represented in light colors. Black dots represent SNPs belonging to the following set: [(top 5% CLR population G1) ∪ (top 5% CLR population T) ∪ (top 5% nSL population G1) ∪ (top 5% nSL population T)] ∩ [top 5% hapFLK].