Cause and Effectors: Whole-Genome Comparisons Reveal Shared but Rapidly Evolving Effector Sets among Host-Specific Plant-Castrating Fungi

William C. Beckerson,[a] Ricardo C. Rodríguez de la Vega,[b] Fanny E. Hartmann,[b] Marine Duhamel,[b] Tatiana Giraud,[b] Michael H. Perlin[a]

[a]Department of Biology, Program on Disease Evolution, University of Louisville, Louisville, Kentucky, USA
[b]Ecologie Systématique et Evolution, Université Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, Orsay, France

ABSTRACT Plant pathogens utilize a portfolio of secreted effectors to successfully infect and manipulate their hosts. It is, however, still unclear whether changes in secretomes leading to host specialization involve mostly effector gene gains/losses or changes in their sequences. To test these hypotheses, we compared the secretomes of three host-specific castrating anther smut fungi (*Microbotryum*), two being sister species. To address within-species evolution, which might involve coevolution and local adaptation, we compared the secretomes of strains from differentiated populations. We experimentally validated a subset of signal peptides. Secretomes ranged from 321 to 445 predicted secreted proteins (SPs), including a few species-specific proteins (42 to 75), and limited copy number variation, i.e., little gene family expansion or reduction. Between 52% and 68% of the SPs did not match any Pfam domain, a percentage that reached 80% for the small secreted proteins, indicating rapid evolution. In comparison to background genes, we indeed found SPs to be more differentiated among species and strains, more often under positive selection, and highly expressed *in planta*; repeat-induced point mutations (RIPs) had no role in effector diversification, as SPs were not closer to transposable elements than background genes and were not more RIP affected. Our study thus identified both conserved core proteins, likely required for the pathogenic life cycle of all *Microbotryum* species, and proteins that were species specific or evolving under positive selection; these proteins may be involved in host specialization and/or coevolution. Most changes among closely related host-specific pathogens, however, involved rapid changes in sequences rather than gene gains/losses.

IMPORTANCE Plant pathogens use molecular weapons to successfully infect their hosts, secreting a large portfolio of various proteins and enzymes. Different plant species are often parasitized by host-specific pathogens; however, it is still unclear whether the molecular basis of such host specialization involves species-specific weapons or different variants of the same weapons. We therefore compared the genes encoding secreted proteins in three plant-castrating pathogens parasitizing different host plants, producing their spores in plant anthers by replacing pollen. We validated our predictions for secretion signals for some genes and checked that our predicted secreted proteins were often highly expressed during plant infection. While we found few species-specific secreted proteins, numerous genes encoding secreted proteins showed signs of rapid evolution and of natural selection. Our study thus found that most changes among closely related host-specific pathogens involved rapid adaptive changes in shared molecular weapons rather than innovations for new weapons.

KEYWORDS functional proteomics, effectors, small secreted proteins, host specificity, fungal pathogens

Host specialization is a phenomenon well documented in many fungal pathogen/plant host systems (1), which most often occurs through host shifts (2). The ability to infect a new host is determined by the protein-protein interactions that occur at the pathogen-host interphase. For pathogens to be successful, they not only must be able to colonize the host but also must work around a gauntlet of host defense responses, as well as manipulate the host to their advantage. Pathogens accomplish these ends through the deployment of many secreted effectors (3–5).

It has been understood for several decades that plant pathogens utilize secreted effectors to infect their hosts (1, 6), including the maize pathogen member of the smut fungi *Ustilago maydis* (3). To defend against these pathogens, plants continuously evolve to recognize pathogen-associated molecular patterns and trigger a variety of immune responses (7). Reciprocally, there is an ongoing selective pressure for plant pathogens to adapt to their host by developing new effectors or otherwise alter the compositions of their secretomes to evade detection and find new ways to manipulate the host to their advantage. Secretomes can thus evolve rapidly, not only during host shift events but also due to intraspecific coevolution (8). It is, however, still unclear whether changes in secretomes leading to host specialization and local adaptation primarily involve effector gene gains/losses or changes in their sequences. A repeat-induced point mutation (RIP) is a fungal defense mechanism against transposable elements that has been suggested to play a role in effector diversification in fungi harboring effectors in regions rich in repetitive elements (9, 10). RIPs indeed act via mutations of repeated sequences at specific target sites and can "leak" on neighbor genes (9, 10).

Host specialization following host shift is particularly common in the fungal pathogen species complex *Microbotryum violaceum* (11). *Microbotryum* species are basidiomycete smut fungi that complete their life cycle in the anthers of their respective host plants, replacing the pollen with their own fungal spores (12). Originally described as a single species, these anther smuts are now understood to represent a complex of species (13, 14), most being highly specific to particular species of the Caryophyllaceae family, also known as "pinks" (15). Intraspecific coevolution has also been suggested to occur based on local adaptation patterns, where host plants were more resistant to their local sympatric anther smut pathogen than to those from geographically distant populations of the same species (16, 17).

To infect their hosts, *Microbotryum* fungi, like many other plant pathogens, employ an array of effector proteins to block plant immune response and otherwise manipulate the host during infection (18, 19). While the specificities of the various *Microbotryum* species to their corresponding host plants have been extensively described (14, 15, 20), the molecular basis for host specialization and coevolution within the complex has just recently begun to be explored (21–23). Understanding the changes that have occurred in the secretomes of these host-specific species will broaden our understanding of the mechanisms behind coevolution, host shifts, and emergent diseases. Furthermore, *Microbotryum* species offer a unique model system to study host shifts and specialization, with multiple host-specific and closely related pathogens (24), which is not often the case in agriculturally propagated crops.

To test whether host-specific or locally adapted closely related pathogens differed in their secretomes mainly by gene gains/losses or by rapid evolution of shared effectors, we compared the secretomes of three *Microbotryum* species, two sister species, *M. lychnidis-dioicae* and *M. silenes-dioicae*, and a more distantly related species, *M. violaceum* var. *paradoxa*. We sought to identify sets of secreted core proteins (i.e., orthologous genes encoding secreted proteins shared by all species) that likely play a major role in the pathogenicity of the species complex as a whole. We also sought to identify species-specific effectors and effectors evolving under positive selection and highly expressed *in planta*, thus perhaps involved in host specificity. To further our understanding of coevolution and local adaptation, we compared the secretomes of two *M. lychnidis-dioicae* strains collected from geographically distant populations belonging to distinct genetic clusters that have shown contrasted infection patterns
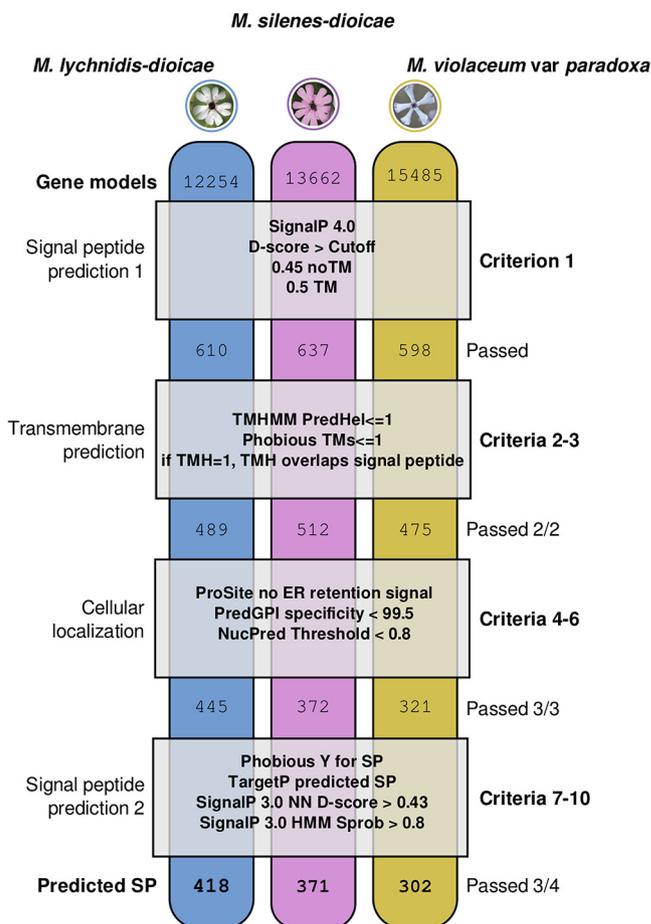
FIG 1 Procedural framework for predicting secreted proteins in three *Microbotryum* species. The genomes for the three fungal species (*M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa*) were first screened to identify putative secreted proteins (SPs) (criterion 1). The resulting proteins were then screened for transmembrane (TM) segments (criteria 2 to 3) and for conflicting cellular localizations (criteria 4 to 6). Candidate secretory peptides were retained for further analysis if they passed all of the first six criteria (criteria 1 to 6) plus at least three out of four additional signal peptide prediction cutoffs (criteria 7 to 10). Each column corresponds to a species, each box corresponds to the criteria employed, and the numbers correspond to the translated gene models that passed the criteria above. PredHel, number of predicted transmembrane helices by N-best; TMH, number of predicted transmembrane helices; ER, endoplasmic reticulum; PredGPI, prediction of glycosylphosphatidylinositol anchored; NucPred, prediction of nuclear localization; Y, yes (predicted to be secreted by Phobius); Sprob, SignalP HMM secretion probability.

consistent with local adaptation of plants (17). We also investigated whether the most frequent changes among host-specific species or locally adapted clusters involved mostly the gain/loss of secreted proteins or the diversification of shared proteins. As RIP-like footprints have been detected in *Microbotryum* fungi (25), we also tested whether sequence divergence in genes under positive selection and/or in genes encoding secreted proteins may have been facilitated by RIPs.

RESULTS

Overview of predicted *Microbotryum* secretomes. Analysis of the three *Microbotryum* secretomes revealed inventories of SPs of similar sizes in all three species. Initial prediction identified around 600 genes with signal peptides in each species (Fig. 1). Utilizing sequence-based criteria of cellular localization and secretory signals, we kept 302, 371, and 418 SPs in *M. violaceum* var. *paradoxa*, *M. silenes-dioicae*, and *M. lychnidis-dioicae*, respectively, for further analysis.

Over 85% of the predicted SPs were clustered into 453 orthologous groups, 225 comprising exclusively predicted SPs (645 SPs), henceforth called the SP-only group,
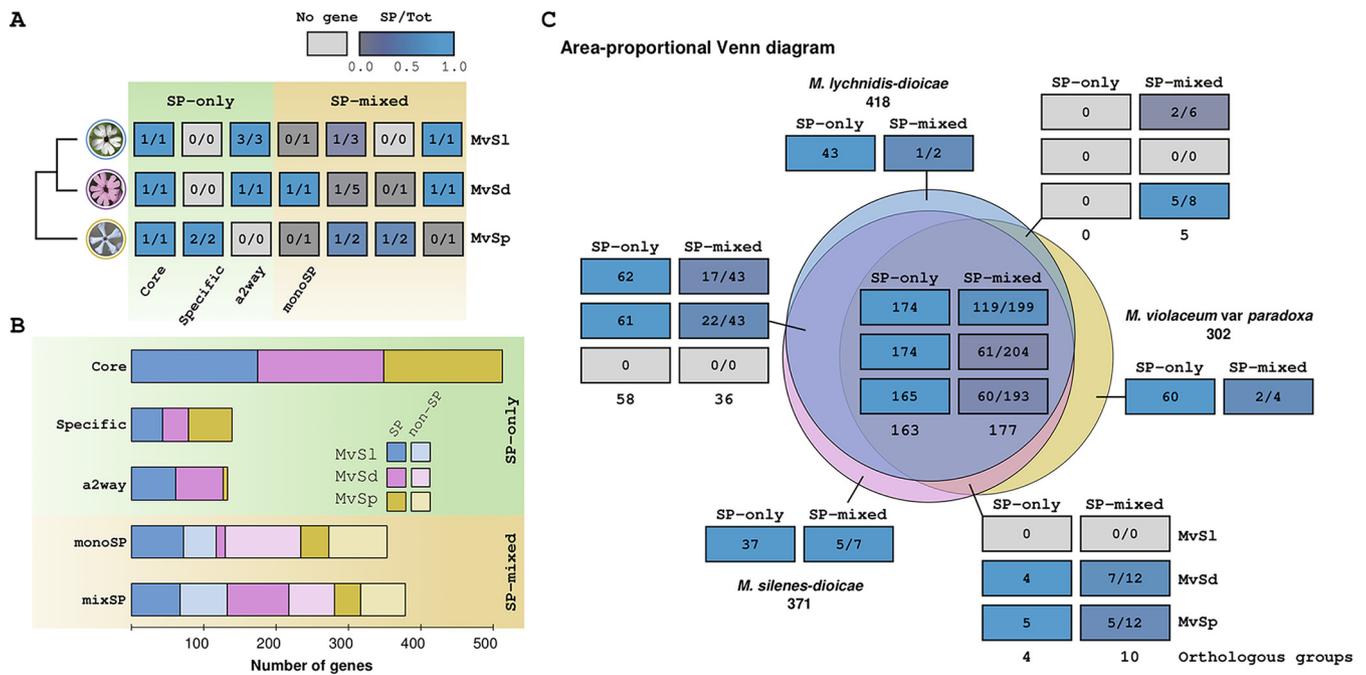
FIG 2 Comparison between the secretomes from three *Microbotryum* species. (A) Key to the phylogenetic profile of predicted SP and non-SP homologs, with examples for the orthologous group terminology used in this study. The cladogram on the left shows the phylogenetic relationships of the three species. In the SP-only orthologous groups (with the light-green background) at the left, all genes are predicted as secreted. In the core secretome, all three species have at least one predicted SP. In the species-specific orthologous groups, predicted SPs are represented in a single species (i.e., paralogous genes). In the accessory two-way (a2way) groups, one species did not have any ortholog in our reconstruction. In the SP-mixed orthologous groups (with the yellow background) at the right, not all orthologs were predicted as secreted; for example, in the monoSP group, a single species had predicted secreted proteins in the mono-copy orthologous group. The box color key corresponds to the ratio of predicted SPs over the total (Tot) number of genes in a given orthologous group per species, with a gradient from blue, when all orthologs in all three species are predicted as secreted, to dark gray, when no ortholog is predicted as secreted. Pale-gray boxes represent missing genes in a given orthologous group. (B) Stacked bar plots of gene counts in the different categories described in panel A, with the same terminology; light colors correspond to non-SP homologs of predicted SPs. (C) Area-proportional Venn diagram of predicted SP and non-SP homologs, as well as species-specific genes. Each area is annotated with six cell blocks, with the number/proportion of predicted SPs in the SP-only and SP-mixed orthologous groups, respectively, colored according to the same gradient as in panel A. Numbers at the bottom of the blocks correspond to the number of SP-only (left) or SP-mixed (right) orthologous groups. Rows in the blocks correspond to *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa*, from top to bottom. The Venn diagram was obtained with BioVenn (72). Abbreviations: a2way, accessory SP two-way orthologous groups; Core, orthologous groups in which all members are predicted as SPs and with at least one gene in each species; mixSP, orthologous groups with both SP and non-SP genes, not including monoSP; monoSP, orthologous groups with one gene in each species but with a single predicted SP; MvSl, *M. lychnidis-dioicae*; MvSd, *M. silenes-dioicae*; MvSp *M. violaceum* var. *paradoxa*; SP-mixed, orthologous groups with at least one gene not predicted as encoding an SP; SP-only, orthologous groups in which all genes are predicted as encoding SPs.

and 239 in which at least one member was not predicted as an SP (298 SPs), henceforth called the SP-mixed group (Fig. 2). Over two-thirds of the predicted SPs belonged to orthologous groups with genes in all three species (753 predicted SPs in the 163 SP-only and 177 SP-mixed groups). Further, 190 predicted SPs belonged to orthologous groups shared by only two species. Only 148 SPs (i.e., 14% of the total) had no ortholog in two of the species and were therefore classified as species-specific SPs (62 in *M. violaceum* var. *paradoxa*, 44 in *M. lychnidis-dioicae*, and 42 in *M. silenes-dioicae*). Predicted SPs were significantly depleted in species-specific genes in all three species (chi-square test with Yates correction, $P \leq 0.0002$). We classified as the "core secretome" 47% of the predicted SPs (513 genes belonging to 163 SP-only orthologous groups with members in all three species). In 118 SP-mixed orthologous groups with single-copy members in all three species, secretion signals were predicted in the orthologs of a single species, orthologs being non-SPs in the two other species; such orthologous groups are referred to as "monoSPs" here (Fig. 2 and Data Set S1 in the supplemental material).

The majority of SPs for each species were smaller than the median length of all predicted proteins in the three species (57%, 68%, and 65% of SPs were smaller than 361 amino acids for *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa*, respectively) (Fig. 3A and Data Set S1). Initial screening of secretomes
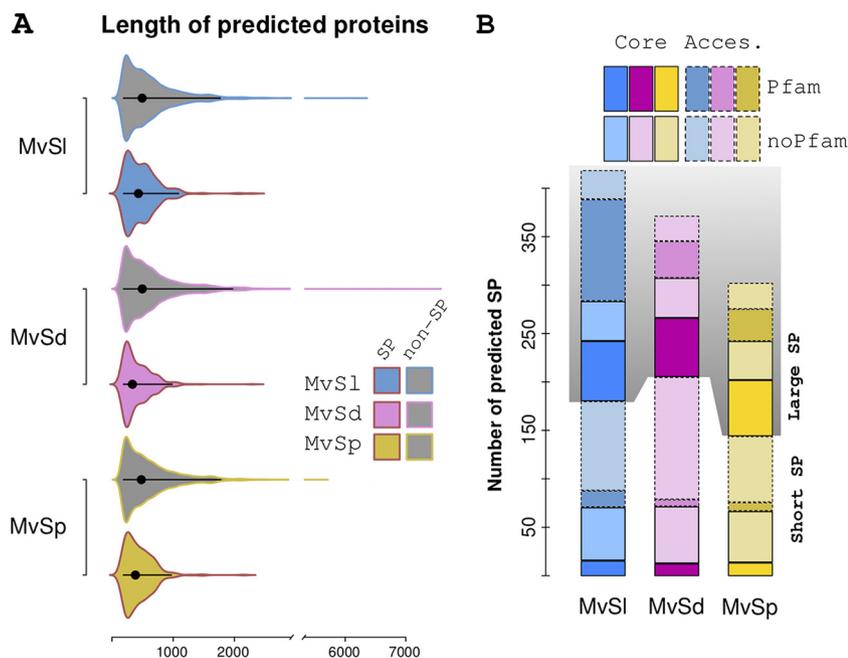
FIG 3 Overview of predicted SP (secreted protein) and non-SP homologs. (A) Length distribution of predicted SPs (area colored by species) and non-SPs (gray area with outline colored by species) in the three species. Black bars and large black dots indicate the range containing 95% of the points and the median, respectively. (B) Pfam screening results for predicted SPs in each of the three species. Stacked bars show the number of predicted SPs with (dark colors) and without (light colors) hits among Pfam-A models. Predicted SPs from the core secretome are boxed with a continuous line, and those from the accessory (Acces.) secretome are boxed with broken lines. The shaded area corresponds to predicted SPs larger than 250 amino acids (Large SP). *Microbotryum* species abbreviations are as defined in the legend of Fig. 2.

showed high percentages of SPs without known Pfam domains, i.e., 52.1% in *M. lychnidis-dioicae*, 67.9% in *M. silenes-dioicae*, and 62.3% in *M. violaceum* var. *paradoxa*. The percentages of genes without identified Pfam domains were even higher for predicted SPs smaller than 250 amino acids, i.e., 81.7% in *M. lychnidis-dioicae*, 88.9% in *M. silenes-dioicae*, and 84.0% in *M. violaceum* var. *paradoxa* (Fig. 3B). This trend was further observed when we analyzed the subset of core SPs (Fig. 3 and Data Set S1).

Signal peptide clusters and yeast secretion trap results. The clustering of the signal peptides of predicted SPs resulted in 280 groups with two or more sequences at 75% sequence identity (823 sequences out of the 1,091 predicted SPs). The signal peptides tested here together with the four previously tested (19) are representative of the signal peptides of 28 predicted SPs in the three *Microbotryum* species under study (Fig. 4). To test whether the predicted secretion signals can indeed direct secretion, we used an invertase-deficient mutant of *Saccharomyces cerevisiae*. Such mutants can grow on glucose but not on sucrose unless transformed with a plasmid containing the invertase gene with a functional secretion signal, which allows the invertase to cleave extracellular sucrose into glucose and fructose in the medium. Cells of the invertase-deficient mutant SEY6120 of *S. cerevisiae* were transformed with pYST-0 vectors containing each tested signal peptide region upstream and in frame with the invertase gene. As evidenced by the ability of their respective secretion signals to allow SEY6120 to grow on medium containing sucrose as the sole carbon source, all 9 predicted secreted proteins that have been tested so far using a yeast secretion trap have been confirmed to be secreted (Fig. 4 and see reference 19). Interestingly, protein 12964 from *M. violaceum* var. *paradoxa* was originally filtered out of our list of predicted SPs, due to the prediction that it is anchored with glycosylphosphatidylinositol (GPI) to the membrane. Nevertheless, in this assay using only the secretion signal of the protein, invertase was secreted, suggesting that our conservative approach for estimating
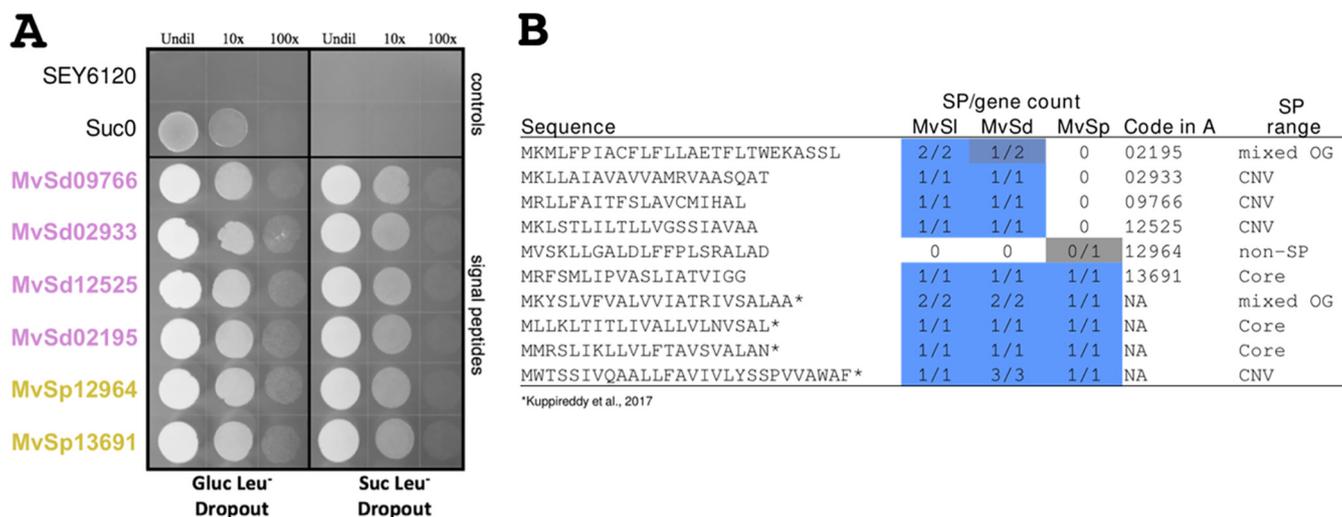
FIG 4 Experimental validation of predicted signal peptides. (A) Yeast secretion trap analysis of a subset of putative secreted proteins from *Microbotryum silenes-dioicae* and *M. violaceum* var. *paradoxa*. The invertase-deficient mutant SEY6120 of *Saccharomyces cerevisiae* is shown in the top row and represents a negative control on medium containing sucrose as the sole carbon source. SEY6120 cells transformed with the pYST-0 vector without a signal peptide upstream of the invertase gene is shown in the second row. Such cells are able to grow on the glucose –Leu dropout medium but not when sucrose is the sole carbon source. The SEY6120 cells in the subsequent six rows are transformed with a construct in which the signal peptide region corresponding to the putative secreted protein ID listed on the left of the row is fused to the truncated *SUC2* gene. If the signal peptide allows secretion, then the transformed *S. cerevisiae* cells are able to grow on sucrose as the sole carbon source. Different dilutions of cells were made (undiluted [Undil], diluted 10×, or diluted 100×) to better distinguish differences, if any. (B) Amino acid sequences and species ranges of signal peptides tested here and in a previous study (19). Cells under the "SP/gene count" columns follow the same color scheme as in Fig. 2. *Microbotryum* species abbreviations are as defined in the legend of Fig. 2. The signal peptide with the code 12964 in panel A corresponds to a protein from *M. violaceum* var. *paradoxa* predicted to be GPI anchored to the membrane. NA, not applicable; OG, orthologous groups; CNV, copy number variation.

secretion may initially filter out membrane proteins with potential functional components outside the fungal cell.

Interspecies comparison of *Microbotryum*-predicted secretomes. As expected due to their phylogenetic placement, the orthologous proteins of *M. silenes-dioicae* and *M. lychnidis-dioicae* were more similar (median identity, 98.7%) than either of the two sister groups to *M. violaceum* var. *paradoxa* (medians, 86.9% for *M. lychnidis-dioicae*/*M. violaceum* var. *paradoxa* and 87.1% for *M. silenes-dioicae*/*M. violaceum* var. *paradoxa*). Orthologous SPs, including those belonging to the core secretome, were significantly less similar to one another than control non-SPs from single-copy orthologous groups of similar lengths (Wilcoxon rank sum test with continuity correction, $P < 7e-7$ for all three pairwise between-species comparisons) (Fig. 5). Out of the 150 single-copy orthologous groups with an SP predicted in each of the three species, i.e., most of what we call the core secretome (leaving out 13 single-copy orthologous groups with more than one gene in at least one species), we identified 92 groups with codons exhibiting more nonsynonymous substitutions than synonymous substitutions. Likelihood ratio tests comparing models with and without positive selection indicated that the model with positive selection was significantly more likely in 18 of these groups (Bonferroni multiple-test-corrected P value, <0.05) (Data Set S2). Similarly, we identified 74 out of 118 monoSP orthologous groups with codons exhibiting ratios of nonsynonymous to synonymous evolutionary changes (*dN/dS* ratios) above 1, among which multiple-test-corrected likelihood ratio tests revealed 21 orthologous groups evolving under positive selection. Selection tests of the 314 control orthologous groups with lengths similar to those of SPs returned 20 groups evolving under positive selection. Core secretome and monoSP orthologous groups were found enriched in proteins with signs of positive selection (Fisher's exact text, $P = 0.02505$ for core versus control and $P < 0.00048$ for monoSP versus control) (Data Sets S1 and S2). We found 9 core and 14 monoSP orthologous groups under positive selection with hits in the Pfam-A database (Data Set S1), among which pectinesterase (PF01095.19) and chitin deacetylase (PF01522.21) have been implicated in fungal biotrophy, potentially for the manipulation of host
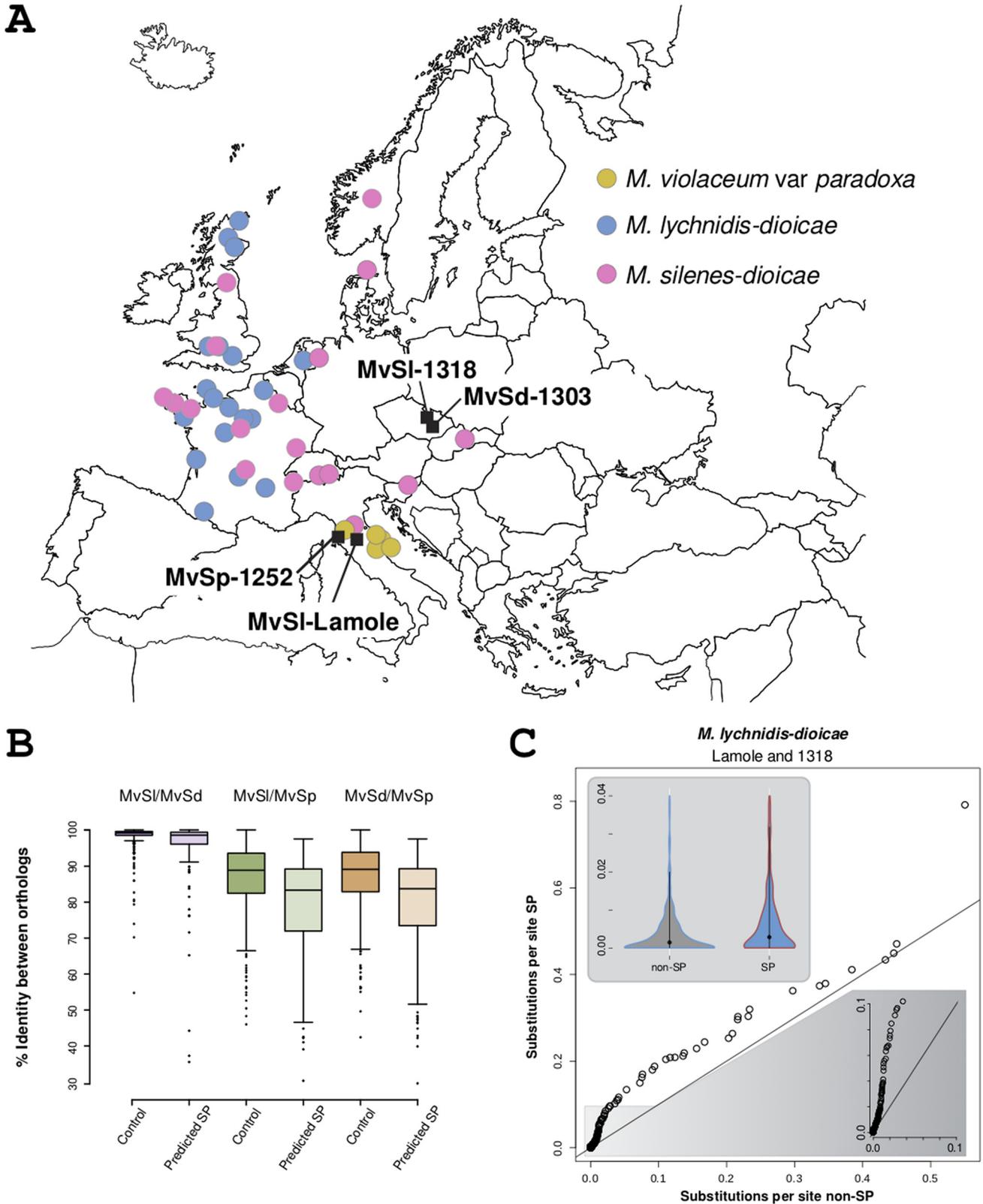
FIG 5 Inter- and intraspecific comparisons of *Microbotryum* secretomes. (A) Sampling locations of the isolates used in this study. (B) Distribution of pairwise percentages of amino acid sequence identity between predicted SPs and background orthologous genes from *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa*. (C) Quantile-quantile (main) and violin (inset) plots of substitution numbers per site between two strains of *M. lychnidis-dioicae* from Lamole, Italy (MvSl-Lamole), and from Olomouc, Czech Republic (MvSl-1318). The shaded area at the bottom right zooms into the low-divergence zone of the quantile-quantile plot. The straight lines correspond to a 45-degree reference line (i.e., points would fall close to this line if the two data sets have the same distribution). *Microbotryum* species abbreviations in panels A and B are as defined in the legend of Fig. 2.

development (18, 26). Glycosyl hydrolases (GHs) (PF00295.17 and PF00704.28) were found in the core and monoSP orthologous groups, despite an overall paucity of GHs represented among *M. lychnidis-dioicae* genes (18). Enzymes of these particular families are interesting due to their ability to hydrolyze pectin, a process important in both pathogenic- and saprophytic-fungus life stages (27).

Intraspecific comparisons of *Microbotryum*-predicted secretomes. We further investigated footprints of positive selection using McDonald-Kreitman (MK) tests that compare the amount of variation within a species (polymorphism) to the divergence between species (substitutions) at two types of sites, synonymous and nonsynonymous. A ratio of nonsynonymous to synonymous polymorphisms within species lower than the ratio of nonsynonymous to synonymous differences between species indicates positive selection (28). We performed three pairwise species comparisons between *M. violaceum* var. *paradoxa*, *M. lychnidis-dioicae*, and *M. silenes-dioicae*, using 148 core, 115 monoSP, and 314 control orthologous groups. We used population genomics data from 20, 18, and 4 isolates from *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa*, respectively (22, 29, 30) (Table S1). Figure 5A shows the locations where the isolates were sampled. The MK tests indicated signatures of within-species positive selection in 8 core secretome orthologous groups and 15 monoSP orthologous groups (Data Set S3). Out of the 23 orthologous groups with signatures of positive selection detected using MK tests, 6 were also detected to evolve under positive selection in the Selecton analysis (Data Set S1). Five orthologous groups were found undergoing intraspecific positive selection in all three comparisons. Intraspecific selection tests on control non-SP orthologous groups revealed that 11 underwent positive selection. While core SPs showed no excess of fixed nonsynonymous polymorphisms, monoSPs were enriched in genes evolving under within-species positive selection (15 out of 115 monoSPs versus 11 out of 314 non-SP genes; Fisher's exact test, $P = 0.0008147$).

When we compared two well-assembled *M. lychnidis-dioicae* genomes, those of the Lamole and 1318 strains, originating from two differentiated populations maladapted to their sympatric hosts (17), we found only 29 Lamole *M. lychnidis-dioicae* SPs without a corresponding 1318 *M. lychnidis-dioicae* gene (12 predicted SPs in 10 orthologous groups and 17 species- or strain-specific SPs). In addition, we found 11 orthologous groups for which gene model counts were different between the 1318 and Lamole *M. lychnidis-dioicae* strains. The ratio of SP-containing orthologous groups with gene count polymorphisms between *M. lychnidis-dioicae* strains was significantly smaller than the genome-wide ratio (21/357 SPs versus 2,642/12,277 total genes; chi-square with Yates correction, $P < 1e-11$). We found few predicted SPs within genome regions showing a presence/absence polymorphism within species, as analyzed previously (21), in both *M. lychnidis-dioicae* Lamole (five) and *M. silenes-dioicae* (two). Substitutions in predicted SPs, on the other hand, were more frequent between *M. lychnidis-dioicae* Lamole and *M. lychnidis-dioicae* 1318 strains than substitutions in control genes (Wilcox rank sum test with continuity correction, $P = 2.537e-05$) (Fig. 5C and Data Set S4).

Genomic context of predicted SPs. Unlike with some other plant-pathogenic fungi with effectors frequently located in repeat-rich regions, we did not find genes encoding predicted SPs to be significantly closer to transposable elements than other genes (Fig. 6) and found no evidence for genome compartmentalization into AT-rich or GC-rich regions in any of the three genomes analyzed, extending previous observations (18). We nevertheless estimated the frequency of sites potentially affected by the RIP-like mechanism reported in *Microbotryum* fungi, targeting TTG and CAA trinucleotides. We calculated an RIP index that takes values above 1 when there is an excess of TTG and CAA trinucleotides over the corresponding target sites not affected by RIP (TCG and CGA), controlling for local sequence composition (see Materials and Methods). The coding regions of predicted SPs did not show any significant excess of RIP-affected trinucleotides, regardless of whether the orthologous groups showed signs of positive selection (Fig. 6). Our RIP index measure was negatively correlated with the distance to transposable elements (TEs), indicating RIP leakage to TE-neighboring
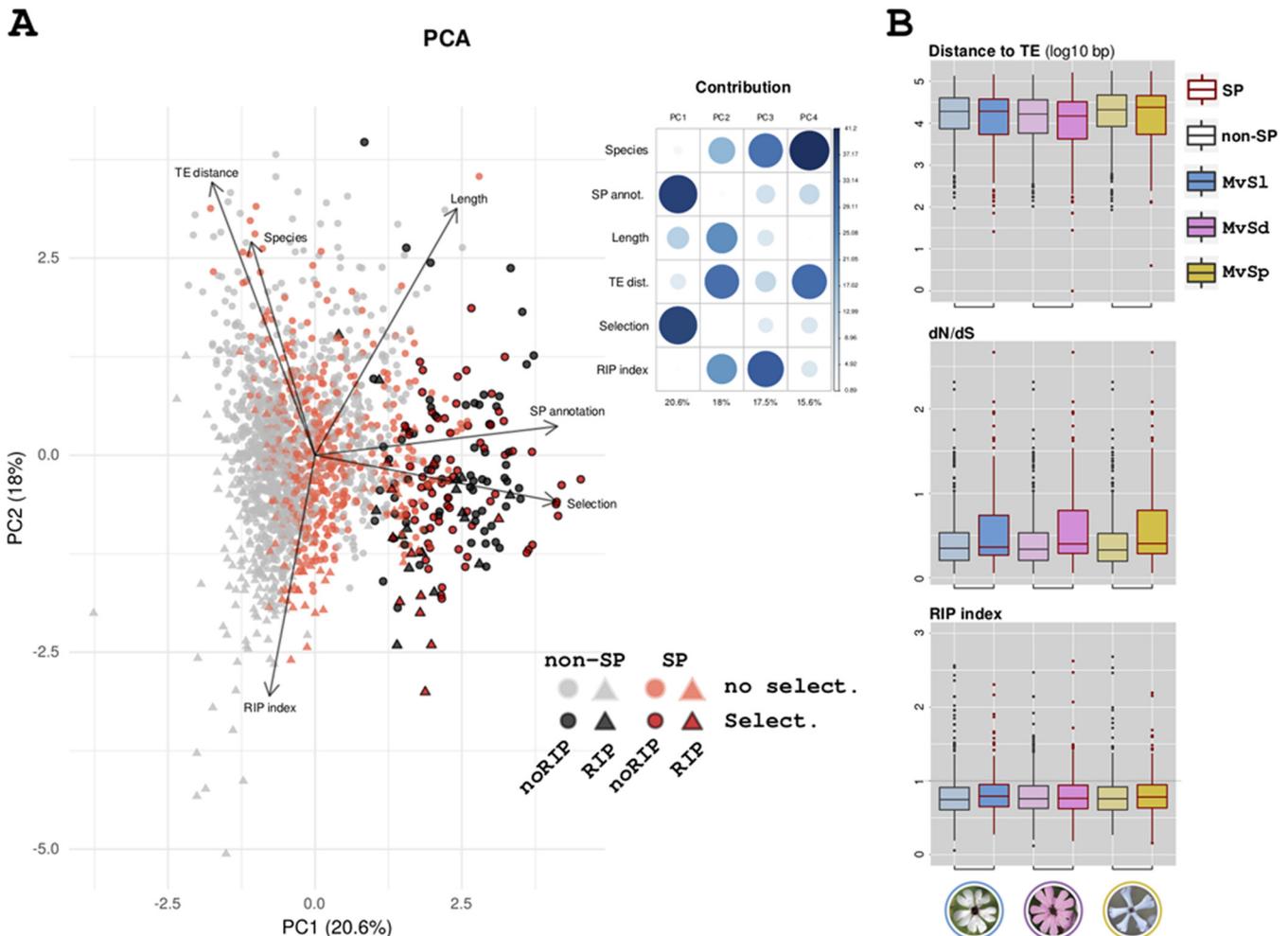
FIG 6 Investigation of the impact of RIPs (repeat-induced point mutations) on gene diversification among species. (A) Principal-component analysis (PCA) of gene copies according to their trait values for six variables: (i) their annotation (annot.) as a binary variable, i.e., encoding SPs (genes colored in red) or non-SPs (in gray); (ii) their length in base pairs as a continuous variable; (iii) the species to which they belong as a category variable (MvSl, *Microbotryum lychnidis-dioicae*; MvSd, *M. silenes-dioicae*; MvSp, *M. violaceum* var. *paradoxa*); (iv) their distance to the nearest transposable element as a continuous variable (TE dist.); (v) their RIP index as a continuous variable (RIP-affected genes are triangles, and non-RIP-affected genes are circles); and (vi) the detection of positive selection (genes with dark colors) or the lack of positive selection (light colors) as a binary variable. The projection of the variables is plotted as arrows in the space defined by the first (PC1) and second (PC2) components, and the percentage of the total variance explained by each principal component is provided in parentheses. The arrows representing the variable projections were scaled for better visualization (6-fold magnification). The contribution of the variables to principal components is shown in a correlation plot (upper right). no select., no selection. (B) TE distance, *dN/dS* ratio (synonymous substitutions over nonsynonymous substitutions), and RIP index distributions of predicted SPs (red contour) or non-SPs (gray contour) in the three species (areas are colored according to species). The distance to the TE was transformed as the log$_{10}$ base pair distance, and the *dN/dS* ratio was calculated within orthologous groups. The boxplots represent the median (center line), the 25th percentile, the 75th percentile (box boundaries), and 1.5 times the distance between the 25th and the 75th percentiles (whiskers); points are the outliers.

regions. The RIP index was not correlated with the ratio of nonsynonymous to synonymous substitutions (Fig. 6), indicating that the RIP-like mechanism does not play a significant role in the diversification of genes under positive selection in *Microbotryum* fungi.

Expression of predicted SPs across infection stages. We focused our analysis on *M. lychnidis-dioicae* Lamole genes expressed in at least one of the five infection stages or under three mating conditions for which we retrieved expression data (18, 31, 32). Among the 2,840 genes fulfilling this condition, we found 135 and 58 predicted SPs from the single-copy core and monoSP orthologous groups, respectively, and compared their expression profiles to those of 232 genes from the non-SP control group (same length distribution but not predicted as potential effectors). Hierarchical clustering of expression profiles across infection stages grouped the genes into low (31
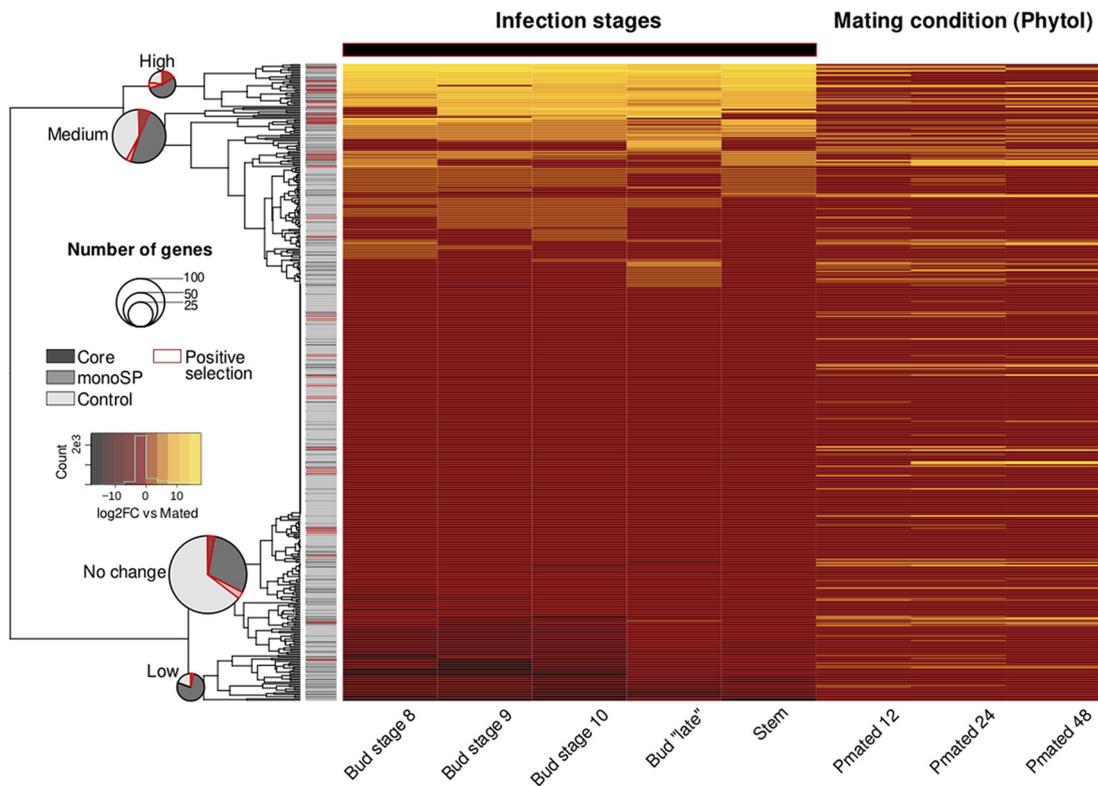
FIG 7 Relative expression of *Microbotryum lychnidis-dioicae* genes across infection stages on flower structures. Heatmap of average gene expression ($n = 2$ to 4) across infection stages in flower structures (32) and under mating conditions (31) as a $\log_2$ fold change against a noninfection condition (mating on phytol, "Pmated"). Hierarchical clustering based on mean row values across the infection stages (horizontal black bar) distinguish four expression profiles with average $\log_2$ fold change median values as follows: low, −6; no change, 0; medium, 1.36; and high, 12. The sidebar represents the annotation of the genes according to the color scheme on the left. Pie charts detail the proportions of SP (core and monoSP) and non-SP (control) genes in each expression profile cluster. The pie chart area is proportional to the number of genes in each expression profile cluster. Red shades and outlines indicate genes with signatures of positive selection.

genes, median $\log_2$ fold change [FC] range, −7.35 to 4.15), medium (117 genes, median $\log_2$ FC range, 0.0 to 1.8), high (29 genes, median $\log_2$ FC range, 9.19 to 12.40), and no-change (248 genes, median $\log_2$ FC, 0) average gene expression profiles across infection stages. We found no major changes in the expression of core, monoSP, or non-SP genes across three mating conditions. Predicted SPs from the core orthologous groups were enriched among genes with high or low levels of average expression across infection stages, respectively (19 and 18 out of 135 core SPs compared with 7 and 6 out of 232 control genes; Fisher's two-tailed exact test, $P = 1.8\text{E}{-}3$ and $1.1\text{E}{-}3$, respectively) (Fig. 7). In line with the pattern observed across all predicted SPs, we could infer the functions of only 14 core and 7 monoSP genes with either high or low average expression. Glycosyl hydrolases, often involved in pathogenesis (27), were among the most common hits (Data Sets S1 and S5).

DISCUSSION

Microbotryum secretomes appeared as largely shared among species, i.e., with few gene gains/losses. Instead, we found SPs to be rapidly evolving, as these were more differentiated among species and more often under positive selection than non-SP genes, indicating that many SPs likely evolved under diversifying selection among species parasitizing different hosts. Such rapid evolution was also indicated by the low percentage of SPs matching Pfam domains (31 to 47%), a percentage that decreased to less than 20% for the small secreted proteins. Such a finding regarding the lack of identifiable Pfam domains of a substantial proportion of SPs is consistent with previous reports of other smut pathogens and is a hallmark of secreted effectors involved in host

specificity (33). Diversifying selection in *Microbotryum* SPs is likely due to coevolution within species, local adaptation, or specialization to different hosts, involving rapid changes in the sequences of secreted proteins to avoid detection in the plant and, more generally, to counteract evolving host defenses. Such a hypothesis is reinforced by the finding that SPs under positive selection were more often highly expressed *in planta* than non-SP genes. Although we found few species-specific SPs or SPs with copy number variation, these accessory SPs may also be involved in coevolution, local adaptation, and/or host specialization (34, 35).

The results from the intraspecific comparison between the two *M. lychnidis-dioicae* strains shed further light on coevolution and local adaptation. We indeed found SPs to be more differentiated than non-SPs between two strains from genetically differentiated populations. These findings further support the idea that coevolutionary pressures may cause divergence in effectors between differentiated populations of pathogens. In fact, the populations from southern and eastern Europe were genetically differentiated in both *M. lychnidis-dioicae* and its host plant *Silene latifolia*, and the plant showed local adaptation to the fungus (17), indicating the occurrence of coevolution. Gene presence/absence polymorphisms in *M. lychnidis-dioicae*, corresponding to the pathogen and host phylogeographic structure (21), and numerous selective sweeps across the genome (22) further supported the existence of coevolution. Unlike with several crop pathogens (36, 37), neither presence/absence polymorphisms nor selective sweep regions were enriched in predicted SPs, even though nearly 10% of SPs were found located within recent selective sweeps of *M. lychnidis-dioicae*, which suggests recent adaptive events involving some SPs.

The identification of a set of shared and conserved SPs, i.e., the 126 core secretome orthologous groups without positive selection, was also interesting, providing a starting point to search for effectors that play a central role in the common pathogenicity traits of these fungi, e.g., the effectors that allow the fungi to migrate to the plant anthers, to induce stunted ovary and pseudoanther development in female flowers, and to eliminate and replace host pollen with fungal spores. The observed differential expression of secreted core proteins further narrows the search for these central effectors and points to sets of genes within the secretome that may play other central roles in the fungal life cycle, including the secretion of extracellular enzymes for carbon source metabolism. Indeed, phosphatases, peptidases, lipases, and glycosidases accounted for half of the Pfam annotations of core secretome orthologous groups with no signs of positive selection (20 out of 38). While such enzymes are clearly associated with fungal pathogens (38–40), they are often found in animal (38, 39) and necrotrophic plant (27, 41, 42) pathogens, rather than in biotrophic fungi. On the other hand, the upregulation of many active carbohydrate enzyme genes related to cell wall degradation was also seen in both wheat stem and poplar rust, *Puccinia graminis* and *M. larici-populina*, respectively (43). In the case of *M. lychnidis-dioicae*, GH28 polygalacturonase domain-containing proteins were upregulated during infection and were among the proteins with signs of positive selection enriched in the core secretome and monoSP orthologous groups. Since polygalacturonase is required for the pathway implicated in pollen dehiscence (44), this is consistent with a fundamental role for such enzymes in the pathogenic lifestyles of anther smut fungi.

Future research with *Microbotryum* will utilize these findings to better understand the functions of the most promising SP candidates, by identifying their targets within each host. Such research geared toward identifying the targets of secreted effectors from *M. lychnidis-dioicae* in its corresponding host plant *Silene latifolia* has already made progress (19). For instance, we identified here MvSl-1064-A1-R4_MC02g04003 as part of the core secretome undergoing diversifying selection across species. We also found its transcript to be among the most highly expressed across infection stages. Its predicted protein product (residues 21 to 156) has been shown to interact with two host proteins in yeast two-hybrid assays (19). Extension of such work to analyze candidate effectors herein identified through *in silico* studies should add new insights into their relevance in host preference and the evolution of the *Microbotryum* species complex. By narrow-

ing down the genomes and identifying prime candidates that are likely to play a major role in the pathogen's life cycle, this work helps to bridge the gap between the quickly expanding availability of *Microbotryum* genomes (24, 30, 45) and the emerging cellular and molecular biology work being done to understand the role of effectors in this system (19).

More generally, this study showed that the molecular changes that lead to different host ranges between closely related plant pathogens, or between different locally adapted genetic clusters, involved few gene gains/losses in their secretome but instead rapid evolution of shared secreted proteins. This represents a significant advance in our understanding of pathogen evolution and may contribute to understanding host shifts and emergent diseases.

MATERIALS AND METHODS

Comparative genomics. To analyze the relationship between various predicted effectors, we performed genomic analyses on the following available genomes, obtained using Pacific Bioscience (PacBio) single-molecule real-time sequencing: GCA_900015465.1 for *M. lychnidis-dioicae* Lamole a_1 (Italy) (45), GCA_900015495.1 for *M. violaceum* var. *paradoxa* from *Silene paradoxa* 1252 a_1 (30), and QPIF00000000 for *M. silenes-dioicae* 1303 a_2 (45). These genomes were selected for comparison due to their relationship to one another; *M. lychnidis-dioicae* strains and *M. silenes-dioicae* are sister species, able to infect one another's host in the greenhouse, although they do so to a lesser degree than their natural host (46) and very little in natural populations (47), while *M. violaceum* var. *paradoxa* serves as an outgroup, unable to infect either of the sister species' hosts or vice versa (20).

In total, we used eight sequence-based prediction tools to identify potential effectors by searching each genome for genes with hallmarks of secretion and without conflicting cellular localization predictions. The initial list of putative secreted proteins (SPs) were generated by running the entire genomes through SignalP 4.0 (48). In order to increase the stringency of this analysis, the SPs must then have passed the following criteria to rule out potential localization or retention in various membranes within or on the cell, which is similar to the previously published protocol for *M. lychnidis-dioicae* (18). Potential transmembrane domains were predicted with TMHMM (49) and Phobius (50). Only gene models with none or a single transmembrane domain prediction overlapping the signal peptide prediction were considered further (18, 48). Prosite was used to screen for predicted endoplasmic reticulum retention signals, while PredGPI (51) was used to screen for potential glycosylphosphatidylinositol anchors, and NucPred (52) was used to screen for nuclear localization signals in the predicted protein (Fig. 1).

Gene models predicted to be secreted and without conflicting localization predictions (i.e., negative for transmembrane domains, endoplasmic reticulum retention, GPI anchoring, and nuclear localization) were further screened using additional criteria to identify strong predictive footprints of secretion in the signal peptide region. To qualify as an SP, the candidates must also have passed stringent cutoff values for secretion, listed in Fig. 1, for at least three of the following four tests: a predicted secretion signal by TargetP (53), a D score (neural network predictors) of greater than 0.43 for the neural network (NN), a secretion probability of greater than 0.8 for the hidden Markov model (HMM) from SignalP3.0, and predicted secretion by Phobius.

We searched the resulting putative SPs among the orthologous groups reconstructed previously (30). Briefly, the orthologous groups were obtained using mcl (54) to cluster high-scoring blastp matches between all gene models predicted in 15 haploid genomes from eight *Microbotryum* species, previously parsed with orthAgogue (55). We classified a predicted SP as a species-specific SP if there was no ortholog in two of the species being considered. For predicted SPs belonging to orthologous groups, we distinguished between species-specific, two- or three-way orthologous groups (i.e., predicted as an SP in a single species or in two or three species, respectively) and between orthologous groups composed exclusively of predicted SPs (SP-only group members) and those containing at least one gene model not predicted as SP (SP-mixed group members). We defined the core secretome as the full set of predicted SPs belonging to SP-only three-way orthologous groups (i.e., present and predicted as SPs in all three species). Conversely, we defined the accessory secretome as the predicted SPs that were either species specific or belonged to SP-mixed or two-way SP-only orthologous groups (i.e., were not present in all species or not predicted as SPs in all species) (Fig. 2). Together, the core and accessory secretomes make up the pan-secretome, i.e., the full set of predicted SPs in all species considered.

Pfam domain annotation. We searched Pfam release 32 (56) against the translated gene models of all predicted SPs and their homologs with hmmsearch from the HMMER 3.1b1 suite (http://hmmer.org). Hits with an E value smaller than 1e–3 were considered significant. The results were then categorized by size as well as presence/absence of a predicted Pfam domain (see Data Set S1 in the supplemental material).

Signal peptide clustering and experimental validation. We clustered the predicted signal peptide sequences with CD-HIT (57), allowing for up to five amino acid differences (nondefault options, -c [sequence identity threshold] 0.75 -l [length of throw_away_sequences] 5). We tested whether predicted signal peptides could direct the secretion of the Suc2 invertase employing a yeast-based secretion trap method (19, 58). Six signal peptide-encoding sequences, as determined by SignalP 4.1 software, were amplified by PCR. A standard PCR cycle was used, with initial denaturation set at 94°C for 4 min and 35 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 30 s and a final extension time of 5 min at 72°C. The

purified fragments were then subcloned into a TOPO vector using an Invitrogen TOPO TA cloning kit and subjected to restriction digestion with EcoRI and NotI enzymes. The digested fragments were then purified and cloned into the pYST-0 vector, upstream and in frame with an invertase coding sequence, *SUC2*. The presence of each signal peptide encoded in frame with the *SUC2* coding region was confirmed by DNA sequencing (Eurofins, Louisville, KY).

Invertase-deficient (*suc2*-negative) *Saccharomyces cerevisiae* strain SEY 6210 (*MATα leu2-3,112 ura3-52 his-Δ200 trp1-Δ901 lys2-801 suc2-Δ9 GAL*) cells were transformed with the constructs using the Frozen-EZ yeast transformation II kit from Zymo Research. Cells were then suspended in water and spread onto dropout selection plates with synthetic defined medium (SD) lacking Leu (SD/–Leu) (Clontech) and either sucrose as the sole carbon source or glucose as a control. Resulting colonies from the sucrose plates were grown overnight in 3 ml of SD/–Leu broth with sucrose; 10-μl undiluted samples, 10-fold dilutions, and 100-fold dilutions were spotted onto SD/–Leu with glucose or sucrose as the carbon source and incubated for 2 days at 30°C. Clones harboring functional signal peptides with the reconstituted invertase activity were able to grow on sucrose as the sole carbon source. Untransformed mutant yeast strain SEY 6210 and transformed SEY 6210 cells with the empty pYST-0 vector were used as negative controls. Plasmid DNA was extracted from the positive clones and used to retransform *Escherichia coli*. The constructs were again checked for the presence of signal peptide sequence by DNA sequencing (Eurofins, Louisville, KY).

Tests for positive selection. We focused our selection analysis on single-copy three-way orthologous groups with one or three predicted SPs. We found 163 three-way SP-only orthologous groups, among which 150 were single-copy orthologous groups (i.e., single-copy three-way SP-only orthologous groups or single-copy core secretome). Furthermore, 118 single-copy orthologous groups retained a single predicted SP after annotation (i.e., single-copy three-way SP-mixed orthologous groups from the accessory secretome, here abbreviated monoSP). As a first method to test for positive selection, we compared evolutionary codon models M8 and M8a (59) on 150 core and 118 monoSP single-copy orthologous groups using Selecton (60). To check whether positive selection was more or less frequent in SP genes than in other (non-SP) genes, we performed the same test on 314 randomly picked single-copy three-way orthologous groups without predicted SPs and with the same length distribution as predicted SPs. The evolutionary model M8, in which a proportion of sites is drawn from a category with a *dN/dS* ratio greater than 1 (i.e., it allows for sites undergoing positive selection), was tested against M8a, in which no site is allowed to have a *dN/dS* ratio larger than 1 (i.e., it does not allow for positive selection), using a likelihood ratio test with 1 degree of freedom to determine the statistical probability that the genes evolve under positive selection (61). We adjusted chi-square *P* values using Bonferroni's correction for multiple testing in R with 582 tests.

We also performed McDonald-Kreitman (MK) tests to infer the existence of positive selection (28). MK tests contrast levels of polymorphism and divergence to test for a departure from neutrality in terms of nonsynonymous substitutions (i.e., rapid amino acid changes) while controlling for gene-specific mutation rates. MK tests estimate α, the fraction of amino acid substitutions that were driven by positive selection. To analyze within-species polymorphism, we used genome sequences previously obtained with Illumina paired-end sequencing technology for populations of the three focal species, *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa* (22, 29, 30). We downloaded raw data publicly available from the NCBI Sequence Read Archive (SRA) under the BioProject accession numbers PRJNA295022, PRJNA269361 and PRJEB16741. Four major genetic clusters were identified in Europe in *M. lychnidis-dioicae* (22), and we considered only strains belonging to the largest cluster in northwestern Europe so that population subdivision does not bias selection inferences. A list of the isolates used in the analysis is presented in Table S1. We processed the raw genome data of 18 *M. silenes-dioicae*, 20 *M. lychnidis-dioicae*, and 4 *M. violaceum* var. *paradoxa* isolates to build pseudoalignment sequences of gene coding sequences within each species using as reference genomes the assemblies reported in GCA_900015465.1 for *M. lychnidis-dioicae*, GCA_900120095.1 for *M. silenes-dioicae*, and GCA_900015485.1 for *M. violaceum* var. *paradoxa*. First, reads were trimmed for quality (length, >50; quality base, >10) using the Cutadapt v1.12 software (62). We mapped Illumina reads against the reference genomes of each species using bowtie2 v2.1.0 (63) and filtered for PCR duplicates using Picard tools (http://broadinstitute.github.io/picard). We realigned reads, called for single nucleotide polymorphisms (SNPs), and filtered them for quality, high genotyping rate (>90%), and minor allele frequency (>10%) using GATK version 3.7 (64) and vcftools version 0.1.13 (65) as described previously (21, 30). We built pseudoalignment sequences of gene coding sequences from the VCF file produced by GATK using a customized script. For each strain, reference nucleotides were replaced by their variants in the reference sequence. We used MUSCLE (66) and TranslatorX (67) to perform codon-based alignments of gene coding sequences among and between species. We used the MKT() and get.MKT() functions in the PopGenome R package (68) to perform MK tests.

With these tools, we performed three comparisons. We tested for positive selection comparing polymorphisms and the levels of divergence of 148 core secretomes and 115 monoSP orthologous groups for (i) *M. violaceum* var. *paradoxa* against *M. lychnidis-dioicae* and *M. silenes-dioicae* strains, (ii) *M. silenes-dioicae* against *M. violaceum* var. *paradoxa* strains, and (iii) *M. lychnidis-dioicae* against *M. violaceum* var. *paradoxa* strains. We excluded from the analyses genes having multiple (paralogous) copies. No neutrality index or α value could be computed for 27 orthologous groups in pairwise species comparison i, 67 orthologous groups in pairwise species comparison ii, and 67 orthologous groups in pairwise species comparison iii, due to a lack of synonymous or nonsynonymous polymorphisms. We performed the same three pairwise comparisons with 314 genes from the control group described above. No neutrality index or α value could be computed for 30, 99, and 84 orthologous groups in the

control pairwise comparisons i, ii, and iii, respectively. We assessed the significance of positive selection for genes having a neutrality index inferior to 1 and a positive α value using a Fisher test (P value < 0.05).

Footprints of RIPs. We investigated the extent of RIP-like footprints in *Microbotryum* genomes with a per-gene repeat-induced point mutation (RIP) index defined as the ratio of t over n (t/n), with t being the sum of TTG and CAA trinucleotides (number of potentially RIP-affected forward and reverse targets, 24) divided by the sum of TCG and CGA trinucleotides (non-RIP-affected forward and reverse targets), with n being the sum of all other nontarget trinucleotides, [ACG]TG and CA[CGT], divided by the sum of [ACG]CG and CG[CGT], to control for contextual sequence composition. A RIP index greater than 1 thus represents an excess of potential RIP sites controlling for the base composition. We compared the distributions of per-gene RIP index values between genes predicted to encode SPs and those not predicted to encode SPs (non-SPs), considering whether or not the genes belonged to orthologous groups undergoing positive selection.

Genomic landscape analyses. We used OcculterCut v1.1 (69) to determine if *Microbotryum* genomes harbored AT-rich regions. Contigs suspected of containing mitochondrial sequences were removed from the assemblies prior to the analysis using the mito_filter.sh script, available as part of the OcculterCut distribution (https://sourceforge.net/projects/occultercut). Transposable element (TE) locations for *M. lychnidis-dioicae* and *M. silenes-dioicae* were retrieved from a previous study (21) and predicted in *M. violaceum* var. *paradoxa* using the same TE centroid sequence database (21). The distance to TE was parsed with bedtools (70).

Intraspecific secretome comparison of *M. lychnidis-dioicae* isolates from differentiated populations. For analyzing the genome-wide intraspecific variation in secretomes, a second genome (assembly GCA_003121365.1) of *M. lychnidis-dioicae* isolated in Olomouc, Czech Republic, and abbreviated as *M. lychnidis-dioicae* 1318, was analyzed (21). We used blastp and orthAgogue to obtain high-scoring pairs between gene models of *M. lychnidis-dioicae* 1318 and the entire gene model set analyzed previously (30) and reran the mcl algorithm. We then parsed the extended orthologous groups to identify the *M. lychnidis-dioicae* 1318 gene models homologous to the *M. lychnidis-dioicae* Lamole SPs identified in this work. We compared the frequencies of synonymous and nonsynonymous single nucleotide substitutions in codon-based pairwise alignments of *M. lychnidis-dioicae* Lamole and *M. lychnidis-dioicae* 1318 genes corresponding to the core secretome or to the non-SP control single-copy orthologous groups. Per-site substitution numbers were calculated as the sum of substitutions divided by the length of the nucleotide alignment.

Analysis of gene expression level across infection stages and mating conditions. We retrieved gene expression data across *M. lychnidis-dioicae* Lamole infection stages under *Silene latifolia* and phytol-induced mating conditions from previous studies (18, 31, 32) as the average $\log_2$ fold change ($\log_2$ FC) against the mated (noninfection) condition (2 to 4 experiments were performed for each of the eight conditions analyzed). We obtained the one-to-one gene model correspondences between long- and short-read sequencing-based assemblies of the same *M. lychnidis-dioicae* Lamole strain as best reciprocal hits with blastp. We focused our analysis on predicted SPs from the core and monoSP orthologs, using gene models from the control set described above for comparisons. Only genes with a Benjamini-Hochberg adjusted P value lower than 1e−5 under at least one condition were considered. Clustering and plotting were performed in R with the heatmap.2 function of the gplots package, using 10 bins for coloring the $\log_2$ FC values and clustering by mean values per row. Pie charts were generated with the pie function of R base.

Plotting, statistical tests, and figures. Unless otherwise stated, all plots and statistical tests were performed in R version 3.6.1 (71). The final layout of the figures was produced with Inkscape version 0.92.3.

Data availability. Raw data of Illumina paired-end sequencing for populations of the three focal species, *M. lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var. *paradoxa* (22, 29, 30), were downloaded from the NCBI Sequence Read Archive (SRA) under the BioProject accession numbers PRJNA295022, PRJNA269361, and PRJEB16741. A list of the isolates used in this analysis along with their accession numbers is presented in Table S1. All RNA sequence data for this study are available in NCBI under BioProject accession number PRJNA246470.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02391-19.

TABLE S1, XLSX file, 0.01 MB.
DATA SET S1, XLSX file, 2.3 MB.
DATA SET S2, TXT file, 0.03 MB.
DATA SET S3, TXT file, 0.1 MB.
DATA SET S4, TXT file, 0.02 MB.
DATA SET S5, TXT file, 0.7 MB.

REFERENCES

1. Sánchez-Vallet A, Fouché S, Fudal I, Hartmann FE, Soyer JL, Tellier A, Croll D. 2018. The genome biology of effector gene evolution in filamentous plant pathogens. Annu Rev Phytopathol 56:21–40. https://doi.org/10.1146/annurev-phyto-080516-035303.

2. Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. New Phytol 198:347–385. https://doi.org/10.1111/nph.12150.

3. Lanver D, Tollot M, Schweizer G, Lo Presti L, Reissmann S, Ma L-S, Schuster M, Tanaka S, Liang L, Ludwig N, Kahmann R. 2017. *Ustilago maydis* effectors and their impact on virulence. Nat Rev Microbiol 15:409–421. https://doi.org/10.1038/nrmicro.2017.33.

4. Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB. 2010. Plants versus pathogens: an evolutionary arms race. Funct Plant Biol 37:499–512. https://doi.org/10.1071/FP09304.

5. Whisson SC, Boevink PC, Moleleki L, Avrova AO, Morales JG, Gilroy EM, Armstrong MR, Grouffaud S, van West P, Chapman S, Hein I, Toth IK, Pritchard L, Birch PRJ. 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. Nature 450:115–118. https://doi.org/10.1038/nature06203.

6. Albersheim R, Anderson AJ. 1971. Proteins from plant cell walls inhibit polygalacturonases secreted by plant pathogens. Proc Natl Acad Sci U S A 68:1815–1819. https://doi.org/10.1073/pnas.68.8.1815.

7. Jones JDG, Dangl JL. 2006. The plant immune system. Nature 444:323–329. https://doi.org/10.1038/nature05286.

8. Meile L, Croll D, Brunner PC, Plissonneau C, Hartmann FE, McDonald BA, Sánchez-Vallet A. 2018. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to *Septoria tritici* blotch. New Phytol 219:1048–1061. https://doi.org/10.1111/nph.15180.

9. Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, Balesdent MH, Rouxel T. 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol Plant Microbe Interact 22:932–941. https://doi.org/10.1094/MPMI-22-8-0932.

10. Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ. 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathog 6:e1001180. https://doi.org/10.1371/journal.ppat.1001180.

11. Refregier G, Le Gac M, Jabbour F, Widmer A, Shykoff JA, Yockteng R, Hood ME, Giraud T. 2008. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. BMC Evol Biol 8:100. https://doi.org/10.1186/1471-2148-8-100.

12. Schäfer AM, Kemler M, Bauer R, Begerow D. 2010. The illustrated life cycle of *Microbotryum* on the host plant *Silene latifolia*. Botany 88:875–885. https://doi.org/10.1139/B10-061.

13. Perlin MH. 1997. Molecular approaches to differentiate subpopulations or *formae speciales* of the fungal phytopathogen *Microbotryum violaceum*. Int J Mol Sci 158:568–574. https://doi.org/10.1086/297469.

14. Le Gac M, Hood ME, Giraud T. 2007. Evolution of reproductive isolation within a parasitic fungal species complex. Evolution 61:1781–1787. https://doi.org/10.1111/j.1558-5646.2007.00144.x.

15. Hood ME, Mena-Alí JI, Gibson AK, Oxelman B, Giraud T, Yockteng R, Arroyo MTK, Conti F, Pedersen AB, Gladieux P, Antonovics J. 2010. Distribution of the anther-smut pathogen *Microbotryum* on species of

the Caryophyllaceae. New Phytol 187:217–229. https://doi.org/10.1111/j.1469-8137.2010.03268.x.

16. Kaltz O, Gandon S, Michalakis Y, Shykoff JA. 1999. Local mal-adaptation in the anther-smut fungus *Microbotryum violaceum* to its host plant *Silene latifolia*: evidence from a cross-inoculation experiment. Evolution 53:395–407. https://doi.org/10.2307/2640776.

17. Feurtey A, Gladieux P, Hood ME, Snirc A, Cornille A, Rosenthal L, Giraud T. 2016. Strong phylogeographic co-structure between the anther-smut fungus and its white campion host. New Phytol 212:668–679. https://doi.org/10.1111/nph.14125.

18. Perlin MH, Amselem J, Fontanillas E, Toh SS, Chen Z, Goldberg J, Duplessis S, Henrissat B, Young S, Zeng Q, Aguileta G, Petit E, Badouin H, Andrews J, Razeeq D, Gabaldón T, Quesneville H, Giraud T, Hood ME, Schultz DJ, Cuomo CA. 2015. Sex and parasites: genomic and transcriptomic analysis of *Microbotryum lychnidis-dioicae*, the biotrophic and plant-castrating anther smut fungus. BMC Genomics 16:461–424. https://doi.org/10.1186/s12864-015-1660-8.

19. Kuppireddy VS, Uversky VN, Toh SS, Tsai MC, Beckerson WC, Cahill C, Carman B, Perlin MH. 2017. Identification and initial characterization of effectors of an anther smut fungus and the potential host target proteins. Int J Mol Sci 18:e2489. https://doi.org/10.3390/ijms18112489.

20. Vienne DM, Refrégier G, Hood ME, Guigue A, Devier B, Vercken E, Smadja C, Deseille A, Giraud T. 2009. Hybrid sterility and inviability in the parasitic fungal species complex *Microbotryum*. J Evol Biol 22:683–698. https://doi.org/10.1111/j.1420-9101.2009.01702.x.

21. Hartmann FE, Rodríguez de la Vega RC, Brandenburg J-T, Carpentier F, Giraud T. 2018. Gene presence-absence polymorphism in castrating anther-smut fungi: recent gene gains and phylogeographic structure. Genome Biol Evol 10:1298–1314. https://doi.org/10.1093/gbe/evy089.

22. Badouin H, Gladieux P, Gouzy J, Siguenza S, Aguileta G, Snirc A, Le Prieur S, Jeziorski C, Branca A, Giraud T. 2017. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. Mol Ecol 26:2041–2062. https://doi.org/10.1111/mec.13976.

23. Aguileta G, Lengelle J, Marthey S, Chiapello H, Rodolphe F, Gendrault A, Yockteng R, Vercken E, Devier B, Fontaine MC, Wincker P, Dossat C, Cruaud C, Couloux A, Giraud T. 2010. Finding candidate genes under positive selection in non-model species: examples of genes involved in host specialization in pathogens. Mol Ecol 19:292–306. https://doi.org/10.1111/j.1365-294X.2009.04454.x.

24. Hartmann FE, Rodríguez de la Vega RC, Carpentier F, Gladieux P, Cornille A, Hood ME, Giraud T. 2019. Understanding adaptation, coevolution, host specialization, and mating system in castrating anther-smut fungi by combining population and comparative genomics. Annu Rev Phytopathol 57:431–457. https://doi.org/10.1146/annurev-phyto-082718-095947.

25. Hood ME, Katawczik M, Giraud T. 2005. Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. Genetics 170:1081–1089. https://doi.org/10.1534/genetics.105.042564.

26. Juge N. 2006. Plant protein inhibitors of cell wall degrading enzymes. Trends Plant Sci 11:359–367. https://doi.org/10.1016/j.tplants.2006.05.006.

27. Sprockett DD, Piontkivska H, Blackwood CB. 2011. Evolutionary analysis of glycosyl hydrolase family 28 (GH28) suggests lineage-specific expan-

sions in necrotrophic fungal pathogens. Gene 479:29–36. https://doi
.org/10.1016/j.gene.2011.02.009.

28. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654. https://doi.org/10
.1038/351652a0.

29. Whittle CA, Votintseva A, Ridout K, Filatov DA. 2015. Recent and massive expansion of the mating-type-specific region in the smut fungus *Microbotryum*. Genetics 199:809–816. https://doi.org/10.1534/genetics.114
.171702.

30. Branco S, Carpentier F, Rodríguez de la Vega RC, Badouin H, Snirc A, Le Prieur S, Coelho MA, de Vienne DM, Hartmann FE, Begerow D, Hood ME, Giraud T. 2018. Multiple convergent supergene evolution events in mating-type chromosomes. Nat Commun 9:2000. https://doi.org/10
.1038/s41467-018-04380-9.

31. Toh SS, Chen Z, Schultz DJ, Cuomo CA, Perlin MH. 2017. Transcriptional analysis of mating and pre-infection stages of the anther smut, *Microbotryum lychnidis-dioicae*. Microbiology 163:410–420. https://doi.org/10
.1099/mic.0.000421.

32. Toh SS, Chen Z, Rouchka EC, Schultz DJ, Cuomo CA, Perlin MH. 2018. *Pas de deux*: an intricate dance of anther smut and its host. G3 (Bethesda) 8:505–518. https://doi.org/10.1534/g3.117.300318.

33. Jones DAB, Bertazzoni S, Turo CJ, Syme RA, Hane JK. 2018. Bioinformatic prediction of plant-pathogenicity effector proteins of fungi. Curr Opin Microbiol 46:43–49. https://doi.org/10.1016/j.mib.2018.01.017.

34. Plissonneau C, Hartmann FE, Croll D. 2018. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol 16:5. https://doi.org/10
.1186/s12915-017-0457-4.

35. Schuster M, Schweizer G, Kahmann R. 2018. Comparative analyses of secreted proteins in plant pathogenic smut fungi and related basidiomycetes. Fungal Genet Biol 112:21–30. https://doi.org/10.1016/j.fgb
.2016.12.003.

36. Plissonneau C, Stürchler A, Croll D. 2016. The evolution of orphan regions in genomes of a fungal pathogen of wheat. mBio 7:e01231-16. https://doi.org/10.1128/mBio.01231-16.

37. Hartmann FE, Croll D. 2017. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. Mol Biol Evol 34:2808–2822. https://doi.org/10.1093/molbev/msx208.

38. Brown NA, Hammond-Kosack KE. 2015. Secreted biomolecules in fungal plant pathogenesis, p 263–310. *In* Gupta VK, Mach RL, Sreenivasaprasad S (ed), Fungal biomolecules: sources, applications and recent developments. John Wiley & Sons, Ltd, Oxford, United Kingdom.

39. Monod M, Capoccia S, Léchenne B, Zaugg C, Holdom M, Jousson O. 2002. Secreted proteases from pathogenic fungi. Int J Med Microbiol 292:405–419. https://doi.org/10.1078/1438-4221-00223.

40. Keyhani NO. 2018. Lipid biology in fungal stress and virulence: entomopathogenic fungi. Fungal Biol 122:420–429. https://doi.org/10.1016/
j.funbio.2017.07.003.

41. Reis H, Pfiffi S, Hahn M. 2005. Molecular and functional characterization of a secreted lipase from *Botrytis cinerea*. Mol Plant Pathol 6:257–267. https://doi.org/10.1111/j.1364-3703.2005.00280.x.

42. Gacura MD, Sprockett DD, Heidenreich B, Blackwood CB. 2016. Comparison of pectin-degrading fungal communities in temperate forests using glycosyl hydrolase family 28 pectinase primers targeting Ascomycete fungi. J Microbiol Methods 123:108–113. https://doi.org/10
.1016/j.mimet.2016.02.013.

43. Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feau N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van de Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev IV, Szabo LJ, Martin F. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci U S A 108: 9166–9171. https://doi.org/10.1073/pnas.1019315108.

44. Wang F, Sun X, Shi X, Zhai H, Tian C, Kong F, Liu B, Yuan X. 2016. A global analysis of the polygalacturonase gene family in soybean (*Glycine max*). PLoS One 11:e0163012. https://doi.org/10.1371/journal
.pone.0163012.

45. Branco S, Badouin H, Rodríguez de la Vega RC, Gouzy J, Carpentier F, Aguileta G, Siguenza S, Brandenburg JT, Coelho MA, Hood ME, Giraud T. 2017. Evolutionary strata on young mating-type chromosomes despite

the lack of sexual antagonism. Proc Natl Acad Sci U S A 114:7067–7072. https://doi.org/10.1073/pnas.1701658114.

46. Gibson AK, Refrégier G, Hood ME, Giraud T. 2014. Performance of a hybrid fungal pathogen on pure-species and hybrid host plants. Int J Plant Sci 175:724–730. https://doi.org/10.1086/676621.

47. Gladieux P, Guérin F, Giraud T, Caffier V, Lemaire C, Parisi L, Didelot F, Le Cam B. 2011. Emergence of novel fungal pathogens by ecological speciation: importance of the reduced viability of immigrants. Mol Ecol 20:4521–4532. https://doi.org/10.1111/j.1365-294X.2011
.05288.x.

48. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786. https://doi.org/10.1038/nmeth.1701.

49. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10
.1006/jmbi.2000.4315.

50. Käll L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 35:W429–W432. https://doi.org/10.1093/nar/gkm256.

51. Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. BMC Bioinformatics 9:392. https://doi.org/10.1186/1471-2105-9-392.

52. Brameier M, Krings A, MacCallum RM. 2007. NucPred—predicting nuclear localization of proteins. Bioinformatics 23:1159–1160. https://doi
.org/10.1093/bioinformatics/btm066.

53. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016. https://doi.org/10.1006/jmbi.2000.3903.

54. Dongen S. 2000. Graph clustering by flow simulation. PhD thesis. University of Utrecht, Utrecht, Netherlands. https://micans.org/mcl/.

55. Ekseth OK, Kuiper M, Mironov V. 2014. orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics 30:734–735. https://doi.org/10.1093/bioinformatics/btt582.

56. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10
.1093/nar/gky995.

57. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26: 680–682. https://doi.org/10.1093/bioinformatics/btq003.

58. Lee SJ, Rose JK. 2012. A yeast secretion trap assay for identification of secreted proteins from eukaryotic phytopathogens and their plant hosts. Methods Mol Biol 835:519–530. https://doi.org/10.1007/978-1
-61779-501-5_32.

59. Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

60. Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E, Pupko T. 2005. Selecton: a server for detecting evolutionary forces at a single amino-acid site. Bioinformatics 21:2101–2103. https://doi.org/10.1093/bioinformatics/
bti259.

61. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. 2007. Advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35:W506–W511. https://doi.org/10.1093/nar/gkm382.

62. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10
.14806/ej.17.1.200.

63. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25.

64. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://
doi.org/10.1101/gr.107524.110.

65. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFTOOLS. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

66. Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi
.org/10.1093/nar/gkh340.

67. Abascal F, Zardoya R, Telford M. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38:W7–13. https://doi.org/10.1093/nar/gkq291.

68. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31:1929–1936. https://doi.org/10.1093/molbev/msu136.

69. Testa AC, Oliver AP, Hane JK. 2016. OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes. Genome Biol Evol 8:2044–2064. https://doi.org/10.1093/gbe/evw121.

70. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.

71. Core Team R. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

72. Hulsen T, de Vlieg J, Alkema W. 2008. BioVenn—a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9:488. https://doi.org/10.1186/1471-2164-9-488.