Cause and Effectors: whole genome comparisons reveal shared but rapidly evolving effector sets among host-specific plant-castrating fungi

William C Beckerson, Ricardo Rodriguez de La Vega, Fanny E Hartmann, Marine Duhamel, Tatiana Giraud, Michael Perlin

Cause and Effectors: whole genome comparisons reveal shared but rapidly evolving effector sets among host-specific plant-castrating fungi

William C. Beckerson [1†], Ricardo C. Rodríguez de la Vega [2†], Fanny E. Hartmann [2], Marine Duhamel[2], Tatiana Giraud [2*], Michael H. Perlin [1#*]

[1]Department of Biology, Program on Disease Evolution, University of Louisville, Louisville, KY 40292, USA

[2]Ecologie Systématique et Evolution, Bâtiment 360, Univ. Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, 91400 Orsay, France

[†]These authors contributed equally to this work and should be considered as co-first authors.

[*] These authors jointly supervised the work

#To whom correspondence should be addressed: michael.perlin@louisville.edu

Running Title: Rapidly-evolving effectors of *Microbotryum* species

Abstract Word Count: 240

Text Word Count: 6,466

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, that 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 few species-specific proteins (42-75) 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. Compared 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*; RIP (repeat-induced point mutations) had no role in effector diversification as SPs were not closer to transposable elements 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, that 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.

Key words Functional Proteomics, Effectors, Small Secreted Proteins, Host Specificity, Fungal Pathogens

Importance summary

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 more 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.

Introduction

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 must not only be able to colonize the host, but must also 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, 4, 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 composition 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 intra-specific 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. Repeat-induced point mutations (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). RIP indeed acts 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). Intra-specific 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 specificity of the various *Microbotryum* species to their corresponding host plants has 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 mainly differed in their secretomes 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 relative, *M. violaceum* var *paradoxa*. We sought to identify sets of core secreted 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 consistent with plant local adaptation (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 could have been facilitated by RIP.

Results

Overview of Microbotryum *predicted 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 (Figure 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.

FIGURE 1

Over 85% of the predicted SPs were clustered into 453 orthologous groups, 225 comprising exclusively predicted SPs (645 SPs), henceforth called “SP-only”, and 239 in which

at least one member was not predicted as SP (298 SPs), henceforth called "SP-mixed" (Figure 2). Over two thirds of the predicted SPs belonged to orthologous groups with genes in all three species (753 predicted SPs in 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 with Yates correction $p \leq 0.0002$). We classified as "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 will be referred to as "monoSP" hereafter (Figure 2 and Supplemental File SF1).

FIGURE 2

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; Figure 3a and Supplemental File SF1). Initial screening of secretomes showed a 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 percentage of genes without identified Pfam domains was 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* (Figure

149 3b). This trend was further observed when analyzing the subset of core SPs (Figure 3 and
150 supplemental file SF1).
151
152 FIGURE 3
153
154 *Signal peptide clusters and yeast secretion trap results*
155 The clustering of the signal peptides of predicted SPs resulted in 280 groups with two or more
156 sequences at 75% sequence identity (823 sequences out of the 1091 predicted SPs). The signal
157 peptides tested here together with the four previously tested (19) are representative of the signal
158 peptides of 28 predicted SPs in the three *Microbotryum* species under study (Figure 4). To test
159 whether the predicted secretion signals can indeed direct secretion, we used an invertase-
160 deficient mutant of *Saccharomyces cerevisiae.* Such mutants can grow on glucose but not on
161 sucrose unless transformed with a plasmid containing the invertase gene with a functional
162 secretion signal, which allows the invertase to cleave extracellular sucrose into glucose and
163 fructose in the medium. Cells of the invertase-deficient mutant SEY6120 of *S. cerevisiae* were
164 transformed with pYST-0 vectors containing each tested signal peptide region upstream and in-
165 frame with the invertase gene. As evidenced by the ability of their respective secretion signals to
166 allow SEY6120 to grow on medium containing sucrose as the sole carbon source, all 9 predicted
167 secreted proteins that have been tested so far using yeast secretion trap have been confirmed to
168 be secreted (Figure 4 and reference19). Interestingly, protein 12964 from *M. violaceum var*
169 *paradoxa*, was originally filtered out of our list of predicted SPs, due to the prediction that it is
170 GPI-anchored to the membrane. Nevertheless, in this assay using only the secretion signal of the
171 protein, invertase was secreted, suggesting that our conservative approach to estimating secretion

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

FIGURE 4

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 compared to *M. violaceum* var *paradoxa* (median 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 $p < 7e\text{-}7$ for all three pairwise between-species comparisons, Figure 5). Out of the 150 single-copy orthologous groups with a 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 non-synonymous substitutions than synonymous substitutions. Likelihood ratio tests comparing models with or 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, supplemental file SF2). Similarly, we identified 74 out of 118 monoSP orthologous groups with codons exhibiting dN/dS values above one, among which multiple test-corrected likelihood ratio tests revealed 21 orthologous groups evolving under positive selection. Selection tests on the 314 control orthologous groups of similar lengths as SPs returned 20 groups evolving under positive

196 selection. Core secretome and monoSP orthologous groups were found enriched in proteins with
197 signs of positive selection (Fisher's exact text p = 0.02505 for core versus control and p <
198 0.00048 for monoSP versus control; ssupplemental files SF1 and SF2). We found nine core and
199 fourteen monoSP orthologous groups under positive selection with hits in the Pfam-A database
200 (supplemental file SF1), among which pectinesterase (PF01095.19) and chitin deacetlyase
201 (PF01522.21) have been implicated in fungal biotrophy, potentially for the manipulation of host
202 development (18, 26). Glycosyl hydrolases (GHs) (PF00295.17 and PF00704.28) were found in
203 the core and monoSP orthologous groups, despite an overall paucity of GHs represented among
204 *M. lychnidis-dioicae* genes (18). Enzymes of these particular families are interesting due to their
205 ability to hydrolyze pectin, a process important in both pathogenic and saprophytic fungi life
206 stages (27).
207
208 *Intraspecific comparisons of* Microbotryum *predicted secretomes*
209 We further investigated footprints of positive selection using McDonald–Kreitman (MK) tests
210 that compare the amount of variation within a species (polymorphism) to the divergence between
211 species (substitutions) at two types of sites, synonymous and non-synonymous. A ratio of
212 nonsynonymous to synonymous polymorphism within species lower than the ratio of
213 nonsynonymous to synonymous differences between species indicates positive selection (28).
214 We performed three pairwise species comparisons between *M. violaceum* var *paradoxa*, *M.*
215 *lychnidis-dioicae* and *M. silenes-dioicae,* using 148 core, 115 monoSP and 314 control
216 orthologous groups. We used population genomics data from 20, 18, and 4 isolates from *M.*
217 *lychnidis-dioicae*, *M. silenes-dioicae*, and *M. violaceum* var *paradoxa*, respectively (22, 29, 30;
218 supplemental table ST1). Figure 5A shows the locations where the isolates were sampled. The

MK tests indicated signatures of within-species positive selection in eight core secretome orthologous groups and fifteen monoSP orthologous groups (supplemental file SF3). Out of the 23 orthologous groups with signatures of positive selection detected using MK tests, six were also detected to evolve under positive selection in the SELECTON analysis (supplemental file SF1). 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 non-synonymous 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 only found 29 Lamole *M. lychnidis-dioicae* SPs without a corresponding 1318 *M. lychnidis-dioicae* gene (12 predicted SPs in 10 orthologous groups and 17 species/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 vs 2642/12277 all genes, Chi-square with Yates correction p < 1e-11). We found few predicted SPs within genome regions showing presence/absence polymorphism within species as analyzed previously (21) in both *M. lychnidis-dioicae* Lamole (five) and *M. silenes-dioicae* (two). Substitutions, on the other hand, were more frequent between *M. lychnidis-dioicae* Lamole and

241 *M. lychnidis-dioicae* 1318 strains in predicted SPs than in control genes (Wilcox rank sum test
242 with continuity correction p = 2.537e-05, Figure 5C and supplemental file SF4).
243
244 FIGURE 5
245
246 *Genomic context of predicted SPs*
247 In contrast to some other plant pathogenic fungi with effectors frequently located in repeat-rich
248 regions, we did not find genes encoding predicted SPs to be significantly closer to transposable
249 elements than other genes (Figure 6) and found no evidence for genome compartmentalization
250 into AT-rich or GC-rich regions in any of the three genomes analyzed, extending previous
251 observations (18). We nevertheless estimated the frequency of sites potentially affected by the
252 RIP-like mechanism reported in *Microbotryum* fungi, targeting TTG and CAA trinucleotides.
253 We calculated a RIP index that takes values above one when there is an excess of TTG and CAA
254 trinucleotides over the corresponding target sites not affected by RIP (TCG and CGA),
255 controlling for local sequence composition (see Methods). The coding regions of predicted SPs
256 did not show any significant excess of RIP-affected trinucleotides, regardless of whether the
257 orthologous groups showed signs of positive selection (Figure 6). Our RIP-index measure was
258 negatively correlated with distance to transposable elements (TEs), indicating RIP leakage to
259 TE-neighboring regions. The RIP index was not correlated with the ratio between non-
260 synonymous and synonymous substitutions (Figure 6), indicating that the RIP-like mechanism
261 does not play a significant role in the diversification of genes under positive selection in
262 *Microbotryum* fungi.
263

FIGURE 6

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 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 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 genes, median log2FC range -7.35 – 4.15), medium (117 genes, median log2FC range 0.0 – 1.8), high (29 genes, median log2FC range 9.19 – 12.40), and no change (248 genes, median log2FC 0) average gene expression across infection stages. We found no major changes in 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 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 = 1.8E-3 and 1.1E-3, respectively; Figure 7). In line with the pattern observed across all predicted SPs, we could infer the function 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 (supplemental files SF1 and SF5).

FIGURE 7

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 compared to 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 (31-47%), 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 in 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 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 be causing divergence in effectors between differentiated populations of pathogens. In fact, the populations from South 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. In contrast with several crop pathogens (e.g., 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 in *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 core secreted 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 pathogens (27, 41, 42), rather than in biotrophic fungi. On the other hand, the up-regulation of many carbohydrate active enzyme genes related to cell wall degradation was also seen in both wheat stem and poplar rust, *P. graminis* and *M. larici-populina*, respectively (43). In the case of *M. lychnidis-dioicae*, GH28 polygalacturonase domain-containing proteins were up-regulated 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 lifestyle of anther-smut fungi.

Future research with *Microbotryum* will utilize these findings to better understand the function of the most promising SP candidates, by identifying their targets within each host. Such research geared towards 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 among the most highly expressed across infection stages. Its predicted protein product (residues 21-156) 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 narrowing 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 different locally-adapted genetic clusters, involved little 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 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 for 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, 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 (Figure 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 a SP, the candidates must also have passed stringent cutoff values for secretion, listed in Figure 1, for at least three of the following four tests: a predicted secretion signal by TargetP (53), a D-score 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 SP belonging to orthologous groups, we distinguished between species-specific, two- or three-way orthologous groups (i.e., predicted as SP in a single, in two or in three species, respectively) and between orthologous groups composed exclusively by predicted SP (SP-only) and those containing at least one gene model not predicted as SP (SP-mixed). 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 as "accessory secretome" 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 SP in all species; Figure 2). Together, the core and accessory secretomes make up the "pan-secretome", i.e., the full set of predicted SP in all species considered.

Pfam domain annotation

We searched Pfam release 32 (56) against the translated gene models of all predicted SP 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 (supplemental file SF1).

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 (non default options: -c 0.75 -l 5). We tested if 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. 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 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 *Eco* RI and *Not* I 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$^-$) *Sacchromyces cerevisiae* strain (SEY 6210 - MATαleu2-3, 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^{TM} from Zymo Research. Cells were then suspended in water and spread onto synthetic drop (SD) out, SD/-Leu (Clontech) selection plates with 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 and 10 µL of undiluted, 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 empty pYST-0 vector were used as negative controls. Plasmid DNA was extracted from the positive clones and used to retransform *E. 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 SP. 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, hereafter abbreviated as 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

447 selection was more or less frequent in SPs compared to other (non-SP) genes, we performed the
448 same test in 314 randomly picked single-copy three-way orthologous groups without predicted
449 SP and with the same length distribution as predicted SPs. The evolutionary model M8, in which
450 a proportion of sites are drawn from a category with dN/dS ratio greater than one, i.e., allowing
451 for sites undergoing positive selection, was tested against M8a, in which no site is allowed to
452 have a dN/dS ratio larger than one, i.e., does not allow for positive selection, using a likelihood
453 ratio test with one degree of freedom to determine the statistical probability that the genes evolve
454 under positive selection (61). We adjusted chi-squared p-values using Bonferroni's correction for
455 multiple testing in R considering 582 tests.

456 We also performed McDonald–Kreitman (MK) tests to infer the existence of positive
457 selection (28). MK tests contrast levels of polymorphism and divergence to test for a departure
458 from neutrality in terms of non-synonymous substitutions (i.e., rapid amino-acid changes) while
459 controlling for gene-specific mutation rates. MK tests estimate α, the fraction of amino acid
460 substitutions that were driven by positive selection. To analyze within-species polymorphism, we
461 used genome sequences previously obtained with Illumina paired-end sequencing technology for
462 populations of the three focal species *M. lychnidis-dioicae*, *M. silenes-dioicae* and *M. violaceum*
463 var *paradoxa* (22, 29, 30). We downloaded raw data publicly available from the NCBI Short
464 Read Archive (SRA) under the BioProject IDs PRJNA295022, PRJNA269361 and
465 PRJEB16741. Four major genetic clusters were identified in Europe in *M. lychnidis-dioicae* (22),
466 and we only considered strains belonging to the largest cluster in North Western Europe so that
467 population subdivision does not bias selection inferences. A list of the isolates used in the
468 analysis is presented in supplemental table ST1. We processed the raw genome data of 18 *M.*
469 *silenes-dioicae*, 20 *M. lychnidis-dioicae*, and four *M. violaceum* var *paradoxa* isolates to build

pseudo-alignments 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 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 pseudo-alignments 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 Rpackage (68) to perform MK tests.

With these tools, we performed three comparisons. We tested for positive selection comparing polymorphism and divergence of 148 core secretome and 115 monoSP orthologous groups for (1) *M. violaceum* var *paradoxa* against *M. lychnidis-dioicae* and *M. silenes-dioicae* strains; (2) *M. silenes-dioicae* against *M. violaceum* var *paradoxa* strains; and (3) *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 the pairwise species comparison (1), 67 orthologous groups in the pairwise species comparison (2) and 67 orthologous groups in the pairwise species comparison (3), due to lack of synonymous or non-synonymous polymorphism. We performed the same three pairwise

493 comparisons with 314 genes from the control group described above. No neutrality index or α
494 value could be computed for 30, 99 and 84 in the control pairwise comparisons (1), (2) and (3),
495 respectively. We assessed significance of positive selection for genes having a neutrality index
496 inferior to 1 and a positive α value using a Fisher test (p-value < 0.05).
497
498 *Footprints of RIP (repeat-induced point mutations)*
499 We investigated the extent of RIP-like footprints in *Microbotryum* genomes with a per-gene RIP-
500 index defined as the ratio of *t* over *n* (RIP-index=*t/n*), with *t* being the sum of TTG and CAA
501 trinucleotides (forward and reverse potentially RIP-affected targets; 24) divided by the sum TCG
502 and CGA (forward and reverse non RIP-affected targets), and *n* being the sum of all other non-
503 target trinucleotides [ACG]TG and CA[CGT] divided by the sum of [ACG]CG and CG[CGT], to
504 control for contextual sequence composition. A RIP-index greater than one thus represents an
505 excess of potentially RIPed sites controlling for the base composition. We compared the
506 distribution of per-gene RIP-index values between genes predicted to encode SPs and those not
507 predicted to encode SPs (non-SPs), and considering whether or not the genes belonged to
508 orthologous groups undergoing positive selection.
509
510 *Genomic landscape analyses*
511 We used OcculterCut v1.1 (69) to determine if *Microbotryum* genomes harbored AT-rich
512 regions. Contigs suspected to contain mitochondrial sequences were removed from the
513 assemblies prior to the analysis using the mito_filter.sh script, available as part of the
514 OcculterCut distribution (https://sourceforge.net/projects/occultercut). Transposable elements
515 locations for *M. lychnidis-dioicae* and *M. silenes-dioicae* were retrieved from a previous study

516 (21) and predicted in *M. violaceum* var *paradoxa* using the same TE centroid sequence database
517 (21). Distance to TE was parsed with bedtools (70).
518
519 *Intraspecific secretome comparison between* M. lychnidis-dioicae *isolates from differentiated*
520 *populations*
521 For analyzing the genome-wide intraspecific variation in secretomes, a second genome
522 (assembly GCA_003121365.1) of *M. lychnidis-dioicae* isolated in Olomouc, Czech Republic,
523 and abbreviated as *M. lychnidis-dioicae* 1318, was analyzed (21). We used blastp and
524 orthAgogue to obtain high-scoring pairs between gene models of *M. lychnidis-dioicae* 1318 and
525 the entire gene model set analyzed previously (30) and re-ran the mcl algorithm. We then parsed
526 the extended orthologous groups to identify the *M. lychnidis-dioicae* 1318 gene models
527 homologous to the *M. lychnidis-dioicae* Lamole SPs identified in this work. We compared the
528 frequency of synonymous and non-synonymous single nucleotide substitutions in codon-based
529 pairwise alignments of *M. lychnidis-dioicae* Lamole and *M. lychnidis-dioicae* 1318 genes
530 corresponding to the core secretome or to the non-SP control single-copy orthologous groups.
531 Per-site substitution numbers were calculated as the sum of substitutions divided by the length of
532 the nucleotide alignment.
533
534 *Analysis of gene expression level across infection stages and mating conditions*
535 We retrieved gene expression data across *M. lychnidis-dioicae* Lamole infection stages on *Silene*
536 *latifolia* and phytol-induced mating conditions from previous studies (18, 31, 32) as average log2
537 fold change (log2FC) against the mated (non-infection) condition (n=2-4 for each of the eight
538 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's adjusted p-value lower than 1e-5 in at least one condition were considered. Clustering and plotting was performed in R with the heatmap.2 function of the gplots package using 10 bins for colouring the log2FC 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 R Core Team, 2019). Final layout of the figures was produced with Inkscape version 0.92.3.

ACKNOWLEDGMENTS. This work was supported by European Research Council [GenomeFun 309403 grant] to [TG]; by National Institute of Health (NIH) [sub-award #OGMB131493C1] to [MHP] from [P20GM103436] to [Nigel Cooper, PI]; and by the 2016-2017 STEM Chateaubriand Fellowship to [WCB]. The contents of this work are solely the responsibility of the authors and do not represent the official views of the NIH.

Author Contributions: WCB, TG, and MHP conceived the project. WCB, RCRdlV, TG, and MHP wrote the paper with the input of all authors. WCB performed initial bioinformatic analysis and tested candidate signal peptides. RCRdlV analyzed the data and prepared the final draft. FEH analyzed population genomics data. MD analyzed RIP footprints. WCB, FEH, and RCRdlV generated the figures and tables.

6. REFERENCES

1.Sánchez-Vallet A, et al. 2018. The genome biology of effector gene evolution in filamentous plant pathogens. *Annu Rev Phytopathol.* 56:21-40

2. Vienne DM, et al. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytol*. 198: 347-385

3. Lanver D, et al. 2017. *Ustilago maydis* effectors and their impact on virulence. *Nat Rev Microbiol*. 15: 409-421

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(6): 499-512

5. Whisson SC, et al. 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*. 450: 115-118

6. Albersheim R, Anderson AJ. 1971. Proteins from plant cell walls inhibit polygalacturonases secreted by plant pathogens. *P Natl Acad Sci USA*. 68(8): 1815-1819

7. Jones JDG, Dangl JL. 2006. The plant immune system. *Nature*. 444: 323-329

8. Meile L, et al. 2018. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to *Septoria tritici* blotch. *New Phytologist*. 219(3): 1048-1061

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(8):932-41. doi: 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(11):e1001180. doi: 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. doi: 10.1186/1471-2148-8-100. PMID:18371215

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.

13. Perlin MH, et al. 1997. Molecular approaches to differentiate subpopulations or *formae speciales* of the fungal phytopathogen *Microbotryum violaceum*. *Int J Mol Sci.* 158 (5): 568-574

14. Le Gac M, Hood ME, Giraud T. 2007. Evolution of reproductive isolation within a parasitic fungal species complex. *Evolution*. 61-7: 1781-1787

15. Hood ME, et al. 2010. Distribution of the anther-smut pathogen *Microbotryum* on species of the Caryophyllaceae. *New Phytol*. 187: 217-229

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.

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(3):668-679. doi: 10.1111/nph.14125. Epub 2016 Aug 8. PMID:27500396

18. Perlin MH, et al. 2015. Sex and parasites: genomic and transcriptomic analysis of *Microbotryum lychnidis-dioicae*, the biotrophic and plant-castrating anther smut fungus. *BMC Genomics*. 16(461): 1-24.

19. Kuppireddy VS, et al. 2017. Identification and initial characterization of effectors of an anther smut fungus and the potential host target proteins. *Int J Mol Sci,* 18: 2489.

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(4):683-98. doi: 10.1111/j.1420-9101.2009.01702.x. Epub 2009 Feb 18. PMID:19228274

21. Hartmann FE, Rodríguez de la Vega RC, Brandenbrug J-T, Carpentier F, Giraud T. 2018. Gene presence-absence polymorphism in castrating anther-smut fungi: recent gene gains and phylogeorgraphic structure. *Genome Biol Evol.* 10(5):1298-1314

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(7):2041–2062.

23. Aguileta G, et al. 2010. Finding candidate genes under positive selection in non-model species: examples of genes involved in host specialization in pathogens. *Mol Ecol*. 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.

25. Hood ME, Katawczik M, Giraud T. 2005. Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. *Genetics*. 170(3):1081-1089. Epub 2005 May 23.

26. Juge N. 2006. Plant protein inhibitors of cell wall degrading enzymes.Trends Plant Sci. 2006 Jul;11(7):359-67. Epub 2006 Jun 13. Review. PMID:16774842

27. Sprockett DD, Piontkivska H, Blackwood CB. 2011. Evolutionary analysis of glycosyl hydrolase family 28 (GH28) suggests lineage-specific expansions in necrotrophic fungal pathogens.
Gene. 479(1-2):29-36. doi: 10.1016/j.gene.2011.02.009 . Epub 2011 Feb 25.

28. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351:652–654

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(3):809–816.

30. Branco S, et al. 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nature Communications*. 9:2000

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.

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: Genes, Genomes, Genetics| 8:2 505-518; 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.

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(1):5

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

691 36. Plissonneau C, Stürchler A, Croll D. 2016. The evolution of orphan regions in genomes of a
692 fungal pathogen of wheat. *MBio*. 7(5). pii: e01231-16
693
694 37. Hartmann FE, Croll D. 2018. Distinct trajectories of massive recent gene gains and losses in
695 populations of a microbial eukaryotic pathogen. *Mol Biol Evol*. 34(11):2808-2822
696
697 38. Brown NA, Hammond-Kosack KE. Secreted biomolecules in fungal plant pathogenesis. In:
698 Gupta VK, Mach RL, Sreenivasaprasad S, editors. Fungal Biomolecules: Sources, Applications
699 and Recent Developments. Oxford, UK: John Wiley & Sons, Ltd; 2015. pp. 263–310.
700
701 39. Monod M, Capoccia S, Léchenne B, Zaugg C, Holdom M, Jousson O. 2002. Secreted
702 proteases from pathogenic fungi. *International Journal of Medical Microbiology* 292(5-6):405-
703 19 DOI: 10.1078/1438-4221-00223
704
705 40. Keyhani NO. 2018. Lipid biology in fungal stress and virulence: Entomopathogenic fungi.
706 *Fungal Biol*. 122(6):420-429. doi: 10.1016/j.funbio.2017.07.003. Epub 2017 Jul 19.
707
708 41. Reis H, Pfiffi S, Hahn M. 2005. Molecular and functional characterization of a secreted
709 lipase from *Botrytis cinerea*. *Mol Plant Pathol*. 6(3):257-67. doi: 10.1111/j.1364-
710 3703.2005.00280.x.
711
712 42. Gacura MD, Sprockett DD, Heidenreich B, Blackwood CB. 2016. Comparison of pectin-
713 degrading fungal communities in temperate forests using glycosyl hydrolase family 28 pectinase
714 primers targeting Ascomycete fungi. *J Microbiol Methods*. 2016 Apr;123:108-13. doi:
715 10.1016/j.mimet.2016.02.013. Epub 2016 Feb 17.
716
717 43. Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, et al.
718 2011.Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl*
719 *Acad Sci*. 108:9166–71.
720
721 44. Wang F, Sun X, Shi X, Zhai H, Tian C, Kong F, Liu B, Yuan X. 2016. A global analysis of
722 the polygalacturonase gene family in soybean (*Glycine max*). *PLoS One*. 11(9):e0163012. doi:
723 10.1371/journal.pone.0163012. eCollection 2016. PMID:27657691
724

725 45. Branco S, Badouin H, Rodríguez de la Vega RC, Gouzy J, Carpentier F, Aguileta G,
726 Siguenza S, Brandenburg JT, Coelho MA, Hood ME, Giraud T. 2017. Evolutionary strata on
727 young mating-type chromosomes despite the lack of sexual antagonism. *Proc Natl Acad Sci*
728 *USA*. 114(27):7067-7072. doi: 10.1073/pnas.1701658114. Epub 2017 Jun 19. PMID:28630332

729
730 46. Gibson AK, Refrégier G, Hood ME, Giraud T. 2014. Performance of a hybrid fungal
731 pathogen on pure-species and hybrid host plants. *Int J Plant Sci*. 175(6): 724-730
732

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(21):4521-32. doi: 10.1111/j.1365-294X.2011.05288.x. Epub 2011 Oct 4. PMID:21967446

48. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*. 8: 785-786

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(3): 567-580

50. Käll L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide predition – the Phobius web server. *Nucleic Acids Res*. 35: W429-32

51. Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*. 23;9: 392

52. Brameier M, Krings A, MacCallum RM. 2007. NucPred – predicting nuclear localization of proteins. *Bioinformatics*. 23(9): 1159-60

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(4): 1005-1016

54. Dongen S, Graph clustering by flow simulation. PhD thesis, University of Utrecht, May 2000. Available at 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(5): 734-735

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(D1):D427-D432.

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(5):680-2.

58. Lee SJ and Rose JK. 2012. A yeast secretion trap assay for identification of secreted proteins from eukaryotic phytopathogens and their plant hosts. *Methods Mol Biol,* 835:519-30

59. Yang Z, Nielsen R, Goldman N, Perdersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155(1): 431-449

60. Doron-Faigenboim A, et al. 2005. Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*. 21: 2101–2103.

61. Stern A, et al. 2007. Advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res.* 35: W506–11.

62. Martin M. 2011. Cutadapt removes adapter sequences from highthroughput sequencing reads. *EMBnet.journal* 17(1):10–12.

63. Langmead B, Trapnell C, PopM, Salzberg SL. 2009. Ultrafast and memory efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10(3):R25.

64. McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.

65. Danecek P, et al. 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.

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.

68. Pfeifer B, Wittelsbürger U, Ramos-Onsins S E, Lercher M J. 2014. POPGENOME: an efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, 31, 1929–1936

69. Testa AC, Oliver AP, Hane JK. 2016. OculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biol Evol.* 8(6):2044-64. doi: 10.1093/gbe/evw121.

70. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28.

71. R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL 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.

Figure Legends

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 (criterion 1). The resulting proteins were then screened for transmembrane segments (criteria 2-3) and for conflicting cellular localization (criteria 4-6). Candidate secretory peptides were retained for further analysis if they passed all first six criteria (criteria 1-6) plus at least three out of four additional signal peptide prediction cutoffs (criteria 7-10). Each column corresponds to a species, each box to the criteria employed and the numbers to the translated gene models that passed the criteria above.

Fig. 2. Comparison between the secretomes from three *Microbotryum* species. A) Key to the phylogenetic profile of predicted secreted protein (SP) and non-SP homologs with examples for the orthologous group terminology used in this study. Cladogram on the left shows the phylogenetic relationships of the three species. In the SP-only orthologous groups at left, with the light green background, 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 were 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 at right, with the yellow background, 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 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 the 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, also including species-specific genes. Each area is annotated with six-cell blocks with the number and proportion of predicted SPs in SP-only and SP-mixed orthologous groups, respectively, colored following 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 orthologous groups (right). Rows in the blocks correspond to *M. lychnidis-dioicae, M. silenes-dioicae*, *M. violaceum* var *paradoxa*, from top to bottom. Venn diagram was obtained with BioVenn (72). Abbreviations for all panels: a2way, accessory SP two-way orthologous groups; Core, orthologous groups in which all members are predicted as SP 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 a SP; SP-only, orthologous groups in which all genes are predicted as encoding SPs.

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 SP 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-secretome with broken lines. Shaded area corresponds to predicted SPs larger than 250 amino-acids (Large SP in the figure). *Microbotryum* species abbreviations are as in Figure 2.

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 following 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, diluted 10x or 100x) to better distinguish differences, if any. B) Amino acid sequences and species range 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 Figure 2. *Microbotryum* species abbreviations are as in Figure 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.

887 **Fig. 5. Inter- and intra-specific comparisons of *Microbotryum* secretomes.** A) Sampling
888 locations of the isolates used in this study. B) Distribution of pairwise percentage of amino-acid
889 sequence identity between predicted SPs and background orthologous genes from *M. lychnidis-*
890 *dioicae*, *M. silenes-dioicae* and *M. violaceum* var *paradoxa*. C) Quantile-quantile (main) and
891 violin (inset) plots of substitution numbers per site between two strains of *M. lychnidis-dioicae*
892 from Lamole, Italy (MvSl-Lamole), and from Olomouc, Czech Republic (MvSl-1318). The
893 shaded area at the bottom right zooms into the low divergence zone of the quantile-quantile plot.
894 The straight lines correspond to a 45 degree reference line (i.e., points would fall close to this
895 line if the two data sets have the same distribution). *Microbotryum* species abbreviations in A
896 and B are as in Figure 2.
897
898 **Fig. 6. Investigation of the impact of RIP (repeat-induced point mutations) on gene**
899 **diversification among species.** A) Principal component analysis (PCA) of gene copies
900 according to their trait value for six variables : (i) their annotation as binary variable, i.e.
901 encoding secreted protein SP (genes colored in red) or non-SP (in grey), (ii) their length in bp as
902 continuous variable, (iii) the species they belong to as category variable (MvSl: *Microbotryum*
903 *lychnidis-dioicae,* MvSd: *M. silenes-dioicae,* MvSp: *M. violaceum* var *paradoxa*), (iv) their
904 distance to the nearest transposable element as continuous variable (TE distance), (v) their RIP
905 index as continuous variable (RIP-affected gene noted as triangles and non RIP-affected genes as
906 circles) and (vi) the detection of positive selection (genes with dark colors) or the lack of positive
907 selection (light colors) as binary variable. The projection of the variables is plotted as arrows in
908 the space defined by the first (PC1) and second (PC2) components and the percentage of the total
909 variance explained by each principal component is provided in brackets. The arrows representing

the variable projection were scaled for better visualization (6-fold magnification). The contribution of the variables to principal components is shown in a correlation plot (upper right). B) TE distance, dN/dS (synonymous substitutions over non-synonymous substitutions) and RIP index distribution of predicted SPs (red contour) or non-SPs (grey contour) in the three species (area colored according to species). Distance to TE was transformed as log10 bp distance; dN/dS was calculated within orthologous groups. The boxplots represent the median (center line), the 25th percentile and 75th percentiles (box bounds), 1.5 times the distance between the 25th and the 75th percentiles (whiskers), and points being the outliers.

Fig. 7. Relative expression of *Microbotryum lychnidis-dioicae* genes across infection stages on flower structures. Heatmap of average gene expression (n=2-4) across infection stages in flower structures (32) and mating conditions (31) as log2 fold change against a non-infection 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 log2 fold change median values as follows: low, -6; no-change, 0; medium, 1.36; high, 12. Sidebar represents the annotation of the genes following the color scheme on the left. Pie charts detail the proportion of SP (core and monoSP) and non-SP (control) genes in each expression profile cluster. 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.

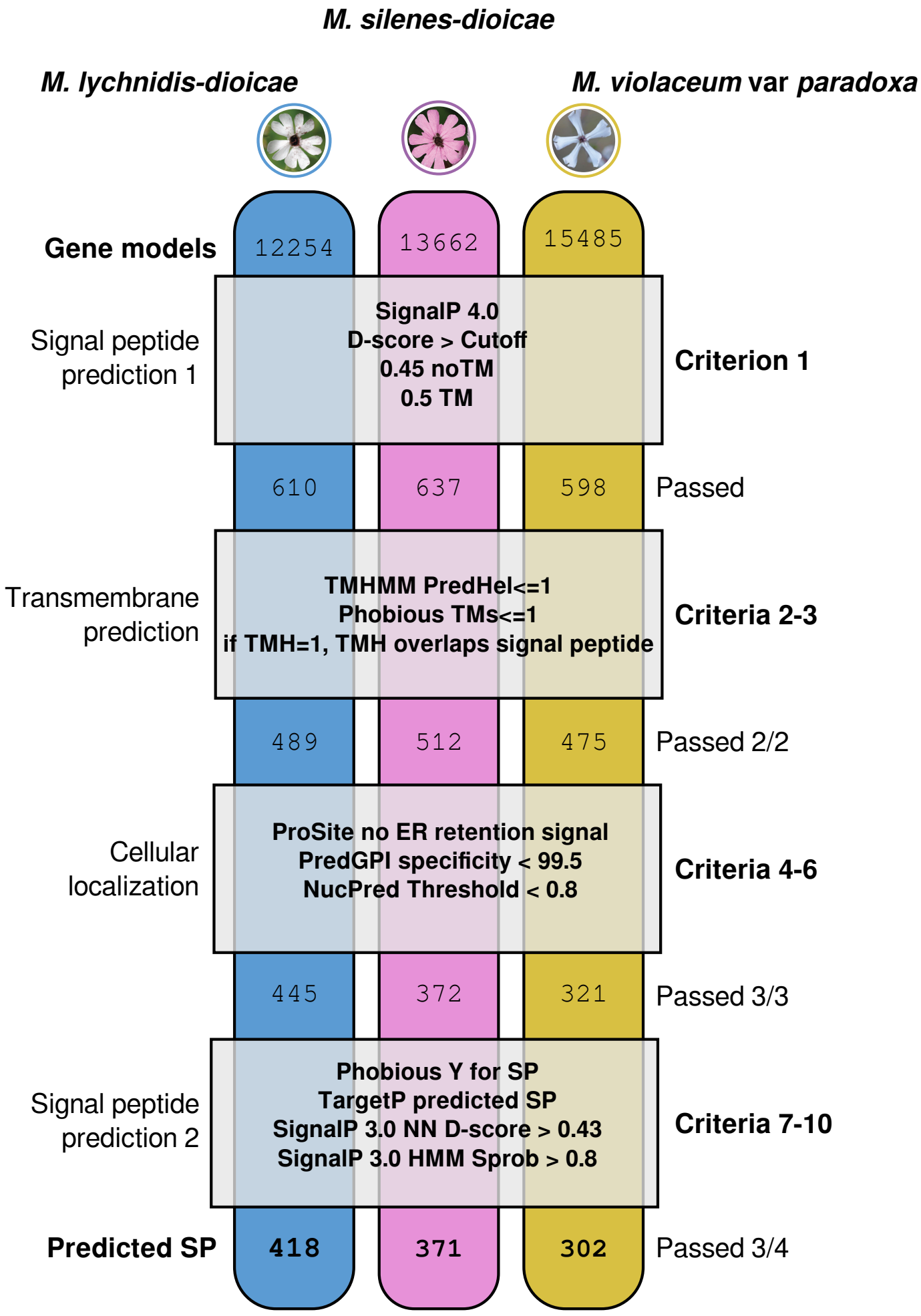
Supplementary materials

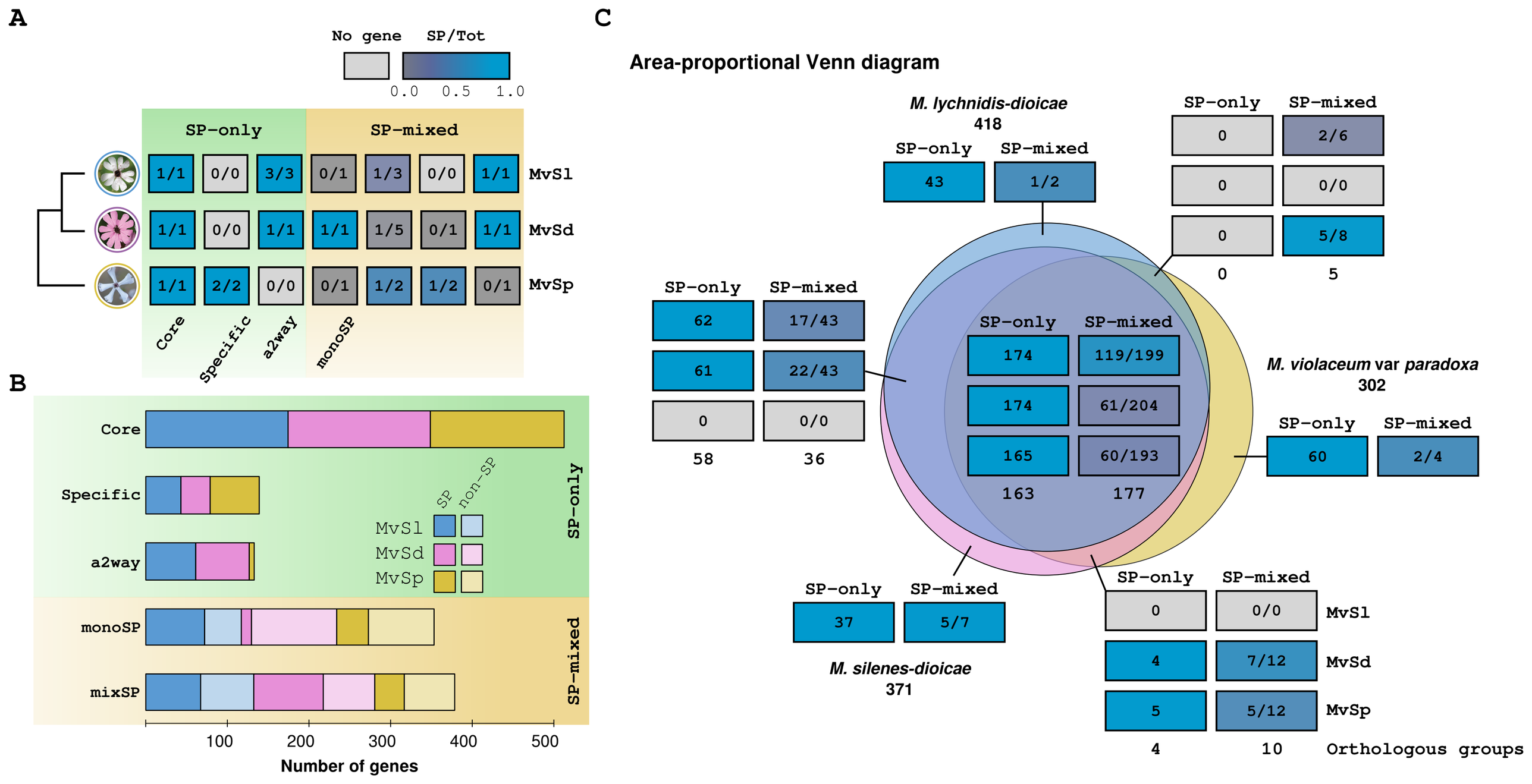
Supplemental file SF1: Full annotation of predicted gene models three *Microbotryum* species. Tab separated file. Columns: 1, gene ID; 2, predicted SP (SPr1) or not (non-SP); 3, Orthologous group ID (xxAg*, gene model was not clustered into an orthologous group); 4, annotation class (AnnotR1); 5, protein length; 6, signal-peptide length (lengthSP); 7, average dN/dS ratio; 8, positive selection (YES/NO); 9, best Pfam hit code; 10, distance to the nearest transposable element; 11, RIP index.

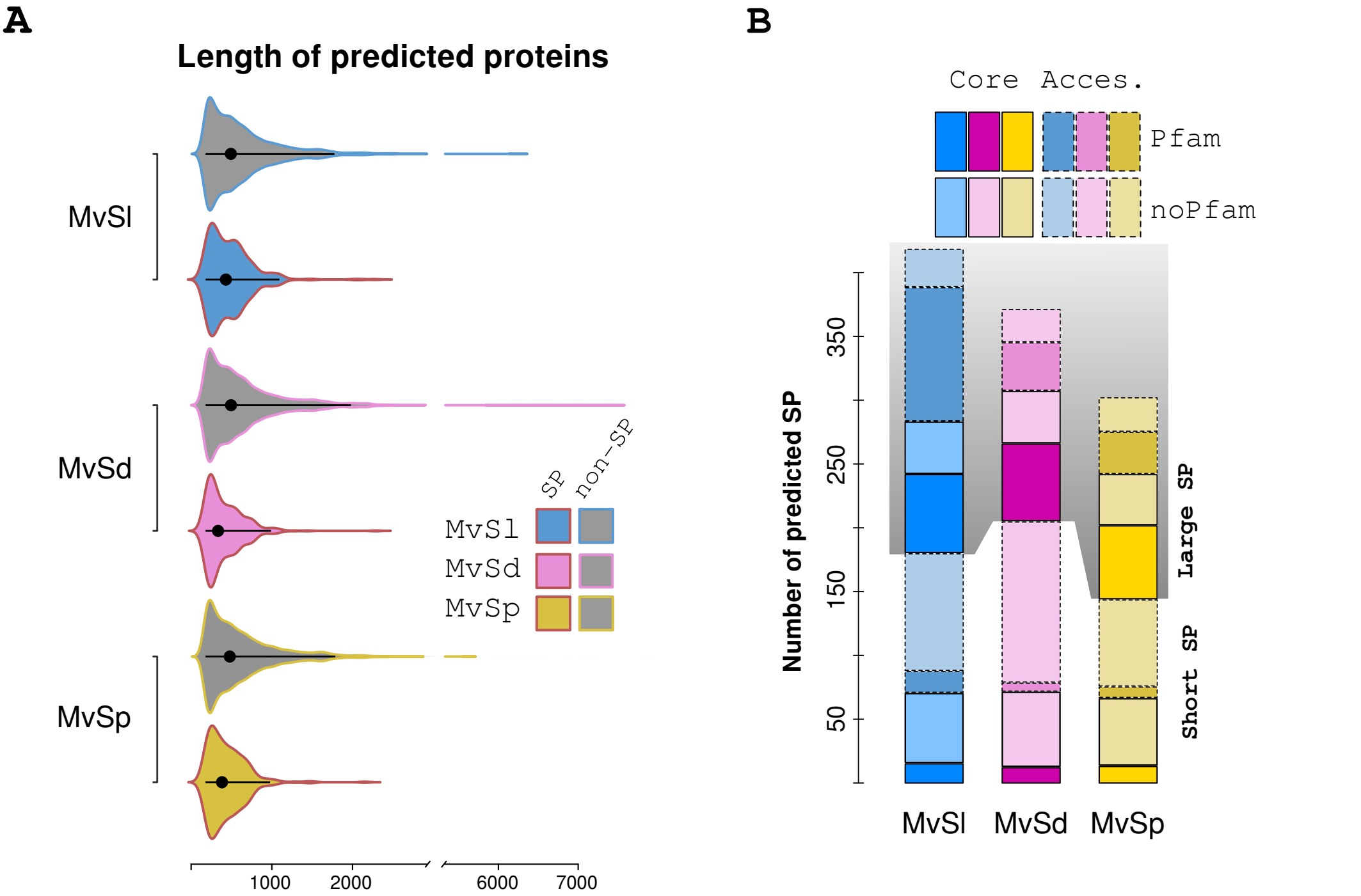
Supplemental file SF2: Interspecific selection tests (SELECTON) on three *Microbotryum* species. Tab separated file. Columns: 1, Orthologous group code (Agogue); 2, annotation class (monoR1, coreR1, contR1); 3, log likelihood M8; 4, log likelihood M8a; 5, average (AVG) dN/dS; 6, likelihood ratio test (LRT); 7, Bonferroni-adjusted p-value; 8, positive selection Y(es)/NO.

Supplemental file SF3: Intra-specific selection tests (MK-tests) in three *Microbotryum* species. Tab separated file: Columns: 1, Orthologous group code (Agogue); 2, annotation class (AnnotR1); 3, predicted SP in species MvSl (100), MvSd (010), MvSp (001), all three (111) or none (000); 4, MK test performed; 5, non-synonymous polymorphisms in population 1 (P1_nonsyn); 6 non-synonymous polymorphisms in population 2 (P2_nonsyn); 7 synonymous polymorphisms in population 1 (P1_syn); 8, synonymous polymorphisms in population 2 (P2_syn); 9, non-synonymous substitutions between species 1 and 2 (D_nonsyn); 10, synonymous substitutions between species 1 and 2 (D_syn); 11, neutrality index; 12, alpha parameter; 13, Fisher's adjusted p-value.

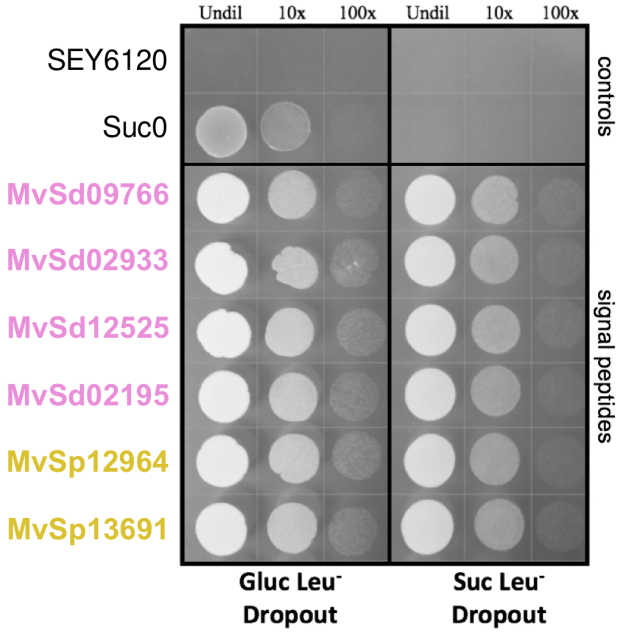
955 Supplemental file SF4: Per-gene substitutions between *Microbotryum lychinidis-dioicae* strains
956 Lamole and 1318. Tab separated file. Columns: 1, Orthologous group code (Agogue); 2,
957 annotation class (AnnotR1); 3, codon alignment length (alnL); 4, non-synonymous
958 polymorphisms (PN); 5, synonymous polymorphisms (PS); 6, absolute distance (PN+PS)/alnL.
959
960 Supplemental file SF5: Normalized expression of *Microbotryum lychinidis-dioicae* Lamole
961 genes across infection stages and mating conditions. Tab separated file. Columns: 1, Gene ID; 2-
962 6, log2FC across infection stages (32); 7-11, Benjamini-Hochberg's corrected p-values (a.k.a.
963 FDR) across infection stages; 12-14, log2FC across mating conditions (31); 15-17, Benjamini-
964 Hochberg's corrected p-values across mating conditions.
965
966 Supplemental Table ST1: Isolates of *Microbotryum* species and accession numbers of population
967 genomics data. Spreadsheet. Columns: A, Sample ID; B, Fungal species; C, Host species; D,
968 BioProject ID; E, sequence read archive accession ID.
969
970


M. silenes-dioicae
M. lychnidis-dioicae
M. violaceum var paradoxa
Gene models
12254
13662
15485
Signal peptide prediction 1
SignalP 4.0
D-score > Cutoff
0.45 noTM
0.5 TM
Criterion 1
610
637
598
Passed
Transmembrane prediction
TMHMM PredHel<=1
Phobious TMs<=1
if TMH=1, TMH overlaps signal peptide
Criteria 2-3
489
512
475
Passed 2/2
Cellular localization
ProSite no ER retention signal
PredGPI specificity < 99.5
NucPred Threshold < 0.8
Criteria 4-6
445
372
321
Passed 3/3
Signal peptide prediction 2
Phobious Y for SP
TargetP predicted SP
SignalP 3.0 NN D-score > 0.43
SignalP 3.0 HMM Sprob > 0.8
Criteria 7-10
Predicted SP
418
371
302
Passed 3/4


A
No gene
SP/Tot
0.0
0.5
1.0
SP-only
SP-mixed
1/1
0/0
3/3
0/1
1/3
0/0
1/1
MvSl
1/1
0/0
1/1
1/1
1/5
0/1
1/1
MvSd
1/1
2/2
0/0
0/1
1/2
1/2
0/1
MvSp
Core
Specific
a2way
monoSP
B
Core
Specific
a2way
monoSP
mixSP
SP
non-SP
MvSl
MvSd
MvSp
SP-only
SP-mixed
100
200
300
400
500
Number of genes
C
Area-proportional Venn diagram
M. lychnidis-dioicae
418
SP-only
SP-mixed
43
1/2
M. violaceum var paradoxa
302
SP-only
SP-mixed
0
2/6
0
0/0
0
5/8
0
5
M. silenes-dioicae
371
SP-only
SP-mixed
62
17/43
61
22/43
0
0/0
58
36
SP-only
SP-mixed
174
119/199
174
61/204
165
60/193
163
177
SP-only
SP-mixed
60
2/4
SP-only
SP-mixed
37
5/7
SP-only
SP-mixed
0
0/0
MvSl
4
7/12
MvSd
5
5/12
MvSp
4
10
Orthologous groups


A
Length of predicted proteins
MvSl
MvSd
MvSp
SP
non-SP
MvSl
MvSd
MvSp
1000
2000
6000
7000
B
Core Acces.
Pfam
noPfam
Number of predicted SP
350
250
150
50
Large SP
Short SP
MvSl
MvSd
MvSp

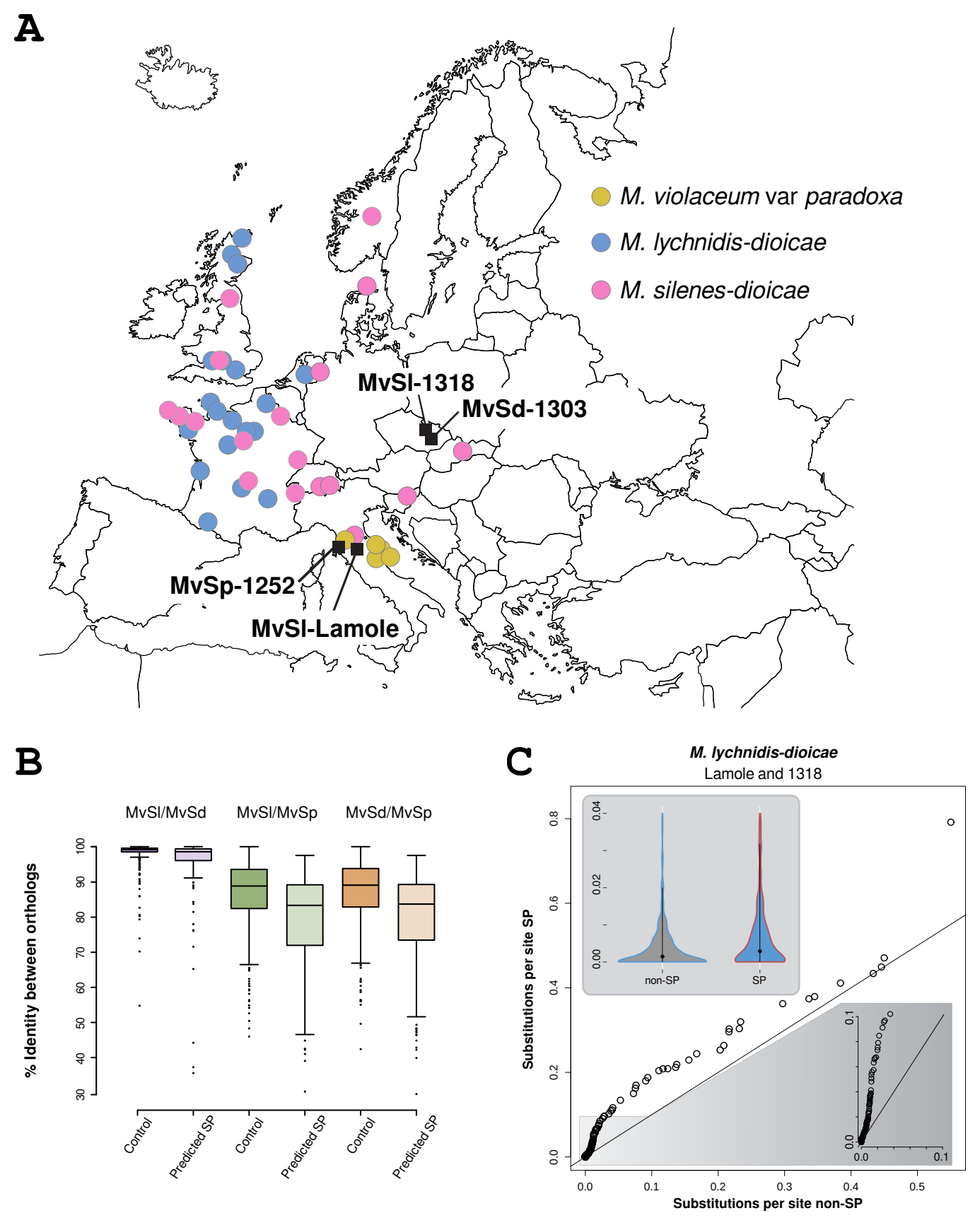
A

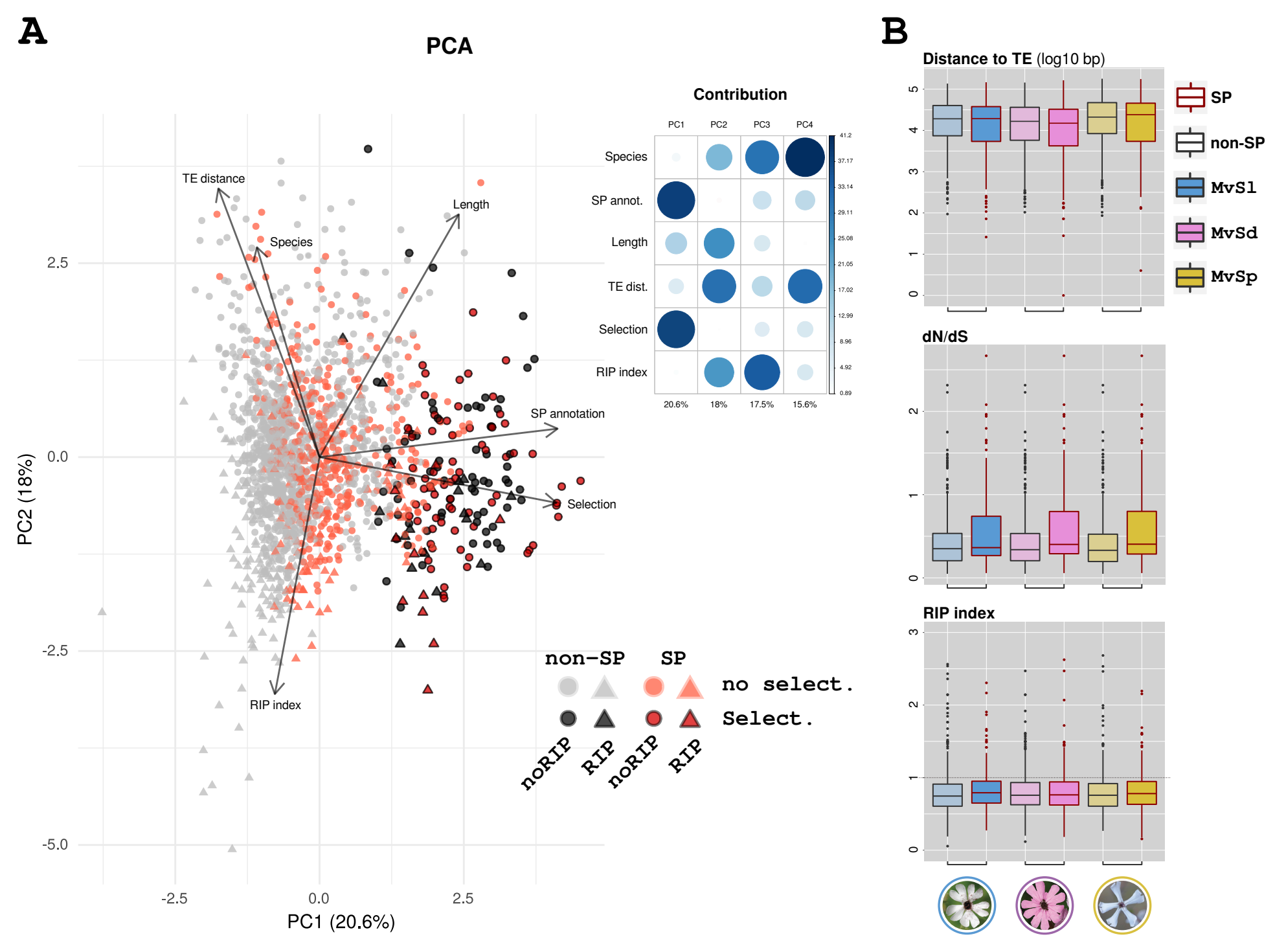


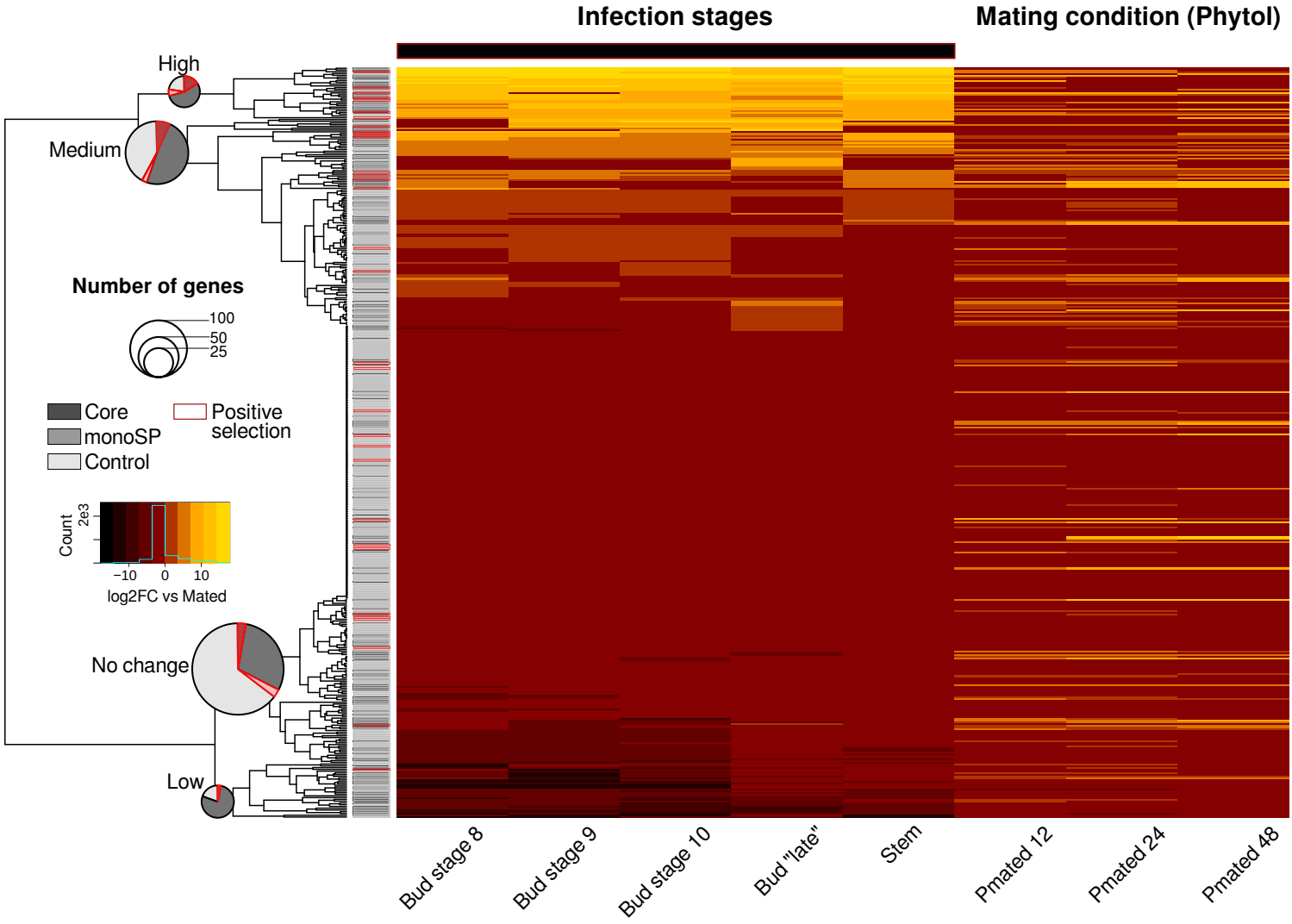
B

Sequence	SP/gene count MvSl	MvSd	MvSp	Code in A	SP range
MKMLFPIACFLFLLAETFLTWEKASSL	2/2	1/2	0	02195	mixed OG
MKLLAIAVAVVAMRVAASQAT	1/1	1/1	0	02933	CNV
MRLLFAITFSLAVCMIHAL	1/1	1/1	0	09766	CNV
MKLSTLILTLLVGSSIAVAA	1/1	1/1	0	12525	CNV
MVSKLLGALDLFFPLSRALAD	0	0	0/1	12964	non-SP
MRFSMLIPVASLIATVIGG	1/1	1/1	1/1	13691	Core
MKYSLVFVALVVIATRIVSALAA*	2/2	2/2	1/1	NA	mixed OG
MLLKLTITLIVALLVLNVSAL*	1/1	1/1	1/1	NA	Core
MMRSLIKLLVLFTAVSVALAN*	1/1	1/1	1/1	NA	Core
MWTSSIVQAALLFAVIVLYSSPVVAWAF*	1/1	3/3	1/1	NA	CNV

*Kuppireddy et al., 2017


A
M. violaceum var paradoxa
M. lychnidis-dioicae
M. silenes-dioicae
MvSl-1318
MvSd-1303
MvSp-1252
MvSl-Lamole
B
MvSl/MvSd
MvSl/MvSp
MvSd/MvSp
% Identity between orthologs
30
40
50
60
70
80
90
100
Control
Predicted SP
Control
Predicted SP
Control
Predicted SP
C
M. lychnidis-dioicae
Lamole and 1318
Substitutions per site SP
Substitutions per site non-SP
0.0
0.2
0.4
0.6
0.8
0.1
0.3
0.5
0.00
0.02
0.04
non-SP
SP


A
PCA
TE distance
Species
Length
SP annotation
Selection
RIP index
PC1 (20.6%)
PC2 (18%)
Contribution
PC1
PC2
PC3
PC4
Species
SP annot.
Length
TE dist.
Selection
RIP index
20.6%
18%
17.5%
15.6%
non-SP
SP
no select.
Select.
noRIP
RIP
noRIP
RIP
B
Distance to TE (log10 bp)
SP
non-SP
MvSl
MvSd
MvSp
dN/dS
RIP index


Infection stages
Mating condition (Phytol)
High
Medium
No change
Low
Number of genes
100
50
25
Core
monoSP
Control
Positive selection
Count
2e3
−10
0
10
log2FC vs Mated
Bud stage 8
Bud stage 9
Bud stage 10
Bud "late"
Stem
Pmated 12
Pmated 24
Pmated 48