Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates

Helene Badouin, Pierre Gladieux, Jerome Gouzy, Sophie Siguenza, Gabriela Aguileta, Alodie Snirc, Stéphanie Le Prieur, Celine B Jeziorski, Antoine Branca, Tatiana Giraud

MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13976

SPECIAL ISSUE: MICROBIAL LOCAL ADAPTATION

Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates

H. BADOUIN,* P. GLADIEUX,*† J. GOUZY,‡§ S. SIGUENZA,‡§ G. AGUILETA,* A. SNIRC,* S. LE PRIEUR,* C. JEZIORSKI,¶** A. BRANCA* and T. GIRAUD*

*Ecologie Systématique Evolution, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91400 Orsay, France, †UMR BGPI, Campus International de Baillarguet, INRA, 34398 Montpellier, France, ‡ Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, INRA, 31326 Castanet-Tolosan, France, §Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, CNRS, 31326 Castanet-Tolosan, France, ¶Genotoul, GeT-PlaGe, INRA Auzeville, 31326 Castanet-Tolosan, France, **UAR1209, INRA Auzeville, 31326 Castanet-Tolosan, France*

Abstract

Identifying the genes underlying adaptation, their distribution in genomes and the evolutionary forces shaping genomic diversity are key challenges in evolutionary biology. Very few studies have investigated the abundance and distribution of selective sweeps in species with high-quality reference genomes, outside a handful of model species. Pathogenic fungi are tractable eukaryote models for investigating the genomics of adaptation. By sequencing 53 genomes of two species of anther-smut fungi and mapping them against a high-quality reference genome, we showed that selective sweeps were abundant and scattered throughout the genome in one species, affecting near 17% of the genome, but much less numerous and in different genomic regions in its sister species, where they left footprints in only 1% of the genome. Polymorphism was negatively correlated with linkage disequilibrium levels in the genomes, consistent with recurrent positive and/or background selection. Differential expression in planta and in vitro, and functional annotation, suggested that many of the selective sweeps were probably involved in adaptation to the host plant. Examples include glycoside hydrolases, pectin lyases and an extracellular membrane protein with CFEM domain. This study thus provides candidate genes for being involved in plant–pathogen interaction (effectors), which have remained elusive for long in this otherwise well-studied system. Their identification will foster future functional and evolutionary studies, in the plant and in the anther-smut pathogens, being model species of natural plant–pathogen associations. In addition, our results suggest that positive selection can have a pervasive impact in shaping genomic variability in pathogens and selfing species, broadening our knowledge of the occurrence and frequency of selective events in natural populations.

Keywords: arms race, co-evolution, effectors, GC content, linked selection, *Microbotryum violaceum*

Received 28 May 2016; revision received 15 December 2016; accepted 19 December 2016

Introduction

Adaptation has been studied extensively and successfully on phenotypes and candidate genes. We are just beginning to be able to investigate the general processes underlying adaptation at the whole-genome level and the importance of adaptation in shaping genomic patterns. Reaching a new level of understanding of the genomic processes of adaptation requires in particular the following questions to be addressed: How many and what kinds of genes are important in adaptation?

Correspondence: Tatiana Giraud, Fax: +33 1 69 15 46 97; E-mail: tatiana.giraud@u-psud.fr

How are they distributed along genomes? Are major adaptation events frequent or recent enough to be detected in genomes? What is the influence of intrinsic genomic features and genomic architecture on local genomic diversity? Do introgressions between species play a major role in adaptation? There are still little data to evaluate these questions beyond a handful of model species (e.g. Svetec *et al.* 2009; Tian *et al.* 2009; Rubin *et al.* 2010, 2012; Hancock *et al.* 2011; Granka *et al.* 2012; Jones *et al.* 2013; Morris *et al.* 2013; Udpa *et al.* 2014; Haasl & Payseur 2016). In particular, very few studies have addressed the abundance and distribution of selective sweeps in genomes with a high-quality reference, with a few exceptions, such as in humans, *Drosophila*, maize, *Arabidopsis* or malaria parasites (Nair *et al.* 2003; Clark *et al.* 2004; Lamason *et al.* 2005; Hernandez *et al.* 2011; Sattath *et al.* 2011; Haasl & Payseur 2016). There is therefore a need for studies on more diverse organisms for eventually drawing generalities on the forces shaping diversity along genomes and for understanding the frequency and distribution of selective sweeps along genomes. Most genome scans published so far aiming at detecting footprints of selection in nonmodel organisms have used measures of differentiation between populations in different environments, whose reliability has been questioned (Haasl & Payseur 2016). The pitfalls and possible confounding effects of these methods based on differentiation, such as a possible coupling of genetic incompatibility barriers with local adaption (Bierne *et al.* 2011), do not apply to the methods designed to detect selective sweeps using multiple full genomes based on site frequency spectra. These, however, have been applied to very few organisms so far (Haasl & Payseur 2016), as they require a high-quality reference genome.

Fungi present great potential as tractable models for studying adaptation and introgression, and have great agronomic, medical, industrial and ecological importance (Anderson *et al.* 2004; Stajich *et al.* 2009; Giraud *et al.* 2010; Gladieux *et al.* 2014). Although fungi are regarded as microbes, they share many similarities with animals in terms of evolution, forming the Opisthokonta clade with them. The inferences drawn from fungi can therefore provide information that can be extended to the processes of genomic adaptation in eukaryotes. In the study of eukaryotic adaptive divergence, fungi are good models, presenting many experimental advantages (Stajich *et al.* 2009; Gladieux *et al.* 2014; Stukenbrock & Croll 2014), such as small genomes and high abundance of complexes of sibling species adapted to different hosts or habitats. In fact, fungi have started to be successfully used to study adaptation (Ellison *et al.* 2011; Gladieux *et al.* 2014), although there have been few studies so far analysing multiple fungal genomes within species for detecting footprints of selection, beyond a few exceptions (e.g. Fraser *et al.* 2010; Neafsey *et al.* 2010; Ellison *et al.* 2011; Branco *et al.* 2017).

Microbotryum fungi, causing anther-smut disease in Caryophyllaceae (Hood *et al.* 2010), are particularly attractive models for investigating the genomic basis of adaptation. *Microbotryum* pathogens represent some of the best studied plant pathogens in natural ecosystems (Bernasconi *et al.* 2009). Anther-smut fungi castrate their host plants, by producing their spores in the anthers, in place of pollen and inducing ovary abortion. Theoretical works suggest that pathogenic, and in particular castrating, lifestyles stand among the most favourable features for observing footprints of adaptation in genomes, because of the arms race they foster (Ashby & Gupta 2014; Tellier *et al.* 2014). *Microbotryum* fungi undergo an obligate sex event before each new plant colonization (Giraud *et al.* 2008a), which is also a favourable condition to detect selective sweeps in genomes.

Studies in which different populations of *Microbotryum lychnidis-dioicae* were used to inoculate different populations of its host *Silene latifolia* revealed variation in infection success, suggestive of co-evolution and evolutionary arms race (Kaltz *et al.* 1999; Feurtey *et al.* 2016). This variation was quantitative, with no gene-for-gene relationship detected in this system (Carlsson-Granér 1997; Kaltz *et al.* 1999). Little is known, however, so far about the genes involved in the interaction, neither from the fungi nor from the plant side. In particular, the effectors involved in the interaction between *Microbotryum* fungi and their host plants are unknown. Effectors are typically small secreted molecules facilitating infection by suppressing or evading plant basal immunity, while others manipulate host factors (Win *et al.* 2012; Presti *et al.* 2015). Effectors can correspond to a wide range of functions, and some are recognized by the plant and trigger defence mechanisms. Population genomic analyses of *Microbotryum* resequencing data provide a unique opportunity to identify candidate genetic factors underlying co-evolutionary interactions with their host plants and their genomic distribution, and to characterize the genome-wide patterns of divergence associated with the maintenance in sympatry of highly specialized *Microbotryum* species.

Here, we focused on the two most-studied species of anther-smut fungi, the sister species *M. lychnidis-dioicae* and *Microbotryum silenes-dioicae*, parasitizing *S. latifolia* and *Silene dioica*, respectively; their divergence has been estimated to have taken place ca. 420 000 years ago (Gladieux *et al.* 2011). In the laboratory, hybrids between *M. lychnidis-dioicae* and *M. silenes-dioicae* are viable and fertile (Van Putten *et al.* 2003; Le Gac *et al.*

2007; de Vienne *et al.* 2009; Gibson *et al.* 2012; Gibson *et al.* 2014) and both fungal species can infect both host plants (de Vienne *et al.* 2009; Gibson *et al.* 2014). In natural populations, hybrids have been detected, although they were rare (Gladieux *et al.* 2011), despite largely overlapping geographical distributions in Europe (Vercken *et al.* 2010) and lack of increased premating reproductive isolation in sympatry (Refrégier *et al.* 2010). These previous studies using a dozen of microsatellite markers could, however, probably only detect early-generation hybrids and could not address the question of whether some genomic regions were more or less permeable to persisting introgression. Studies within the *Microbotryum* species complex using microsatellites have furthermore detected a strong geographical population subdivision, especially in *M. lychnidis-dioicae*, revealing footprints of ancient glacial refugia, as three genetic clusters were found, in western Europe, Italy and eastern Europe, respectively (Vercken *et al.* 2010). Little admixture has been found between clusters based on microsatellites (Vercken *et al.* 2010; Feurtey *et al.* 2016).

The specific questions addressed here were therefore the following: (i) What is the degree of genome-wide long-term introgression between *M. lychnidis-dioicae* and *M. silenes-dioicae*? What is the degree of genome-wide gene flow between geographical clusters within *M. lychnidis-dioicae*? (ii) Are there footprints of selective sweeps in the genomes of each of the anther-smut fungi *M. lychnidis-dioicae* and *M. silenes-dioicae*? What proportion of the genome do they affect? Are they clustered in particular genomic locations? Do they involve the same genomic regions in the two *Microbotryum* species? (iii) Can we identify candidate effectors based on genomic regions carrying signatures of selective sweeps, for characterization in future functional analyses? To address these questions, we sequenced the genomes of multiple individuals of *M. lychnidis-dioicae* (hereafter called MvSl) and of *M. silenes-dioicae* (hereafter called MvSd) sampled across Europe, as well as one individual of an outgroup species, *Microbotryum coronariae* parasitizing *Lychnis flos-cuculi* (Table S1, Supporting information). We also used the high-quality reference genome available for *M. lychnidis-dioicae* infecting *S. latifolia* (Badouin *et al.* 2015), as well as whole transcriptome data available in *M. lychnidis-dioicae* (Fontanillas *et al.* 2015; Perlin *et al.* 2015). Because tests of selection can be biased by strong genetic subdivision (Huber *et al.* 2014), we first analysed the population structure of these species, for running test within panmictic populations. In order to detect genes potentially involved in host–pathogen co-evolution or adaptation to different host plants, we searched for selective sweeps and for genes with high rates of nonsynonymous substitutions. Tests of selection were run both genome-wide and focusing on candidate genes possibly involved in host interaction, that is secreted proteins and genes upregulated in planta during infection (Perlin *et al.* 2015). For candidate genes detected using genome-wide scans for signatures of positive selection, we checked for differential expression in planta as preliminary evidence for a potential role in molecular interactions with the host plant. We investigated the genome-wide effect of selection by testing whether polymorphism was negatively correlated with levels of linkage disequilibrium (LD) along the genome, as expected in cases of recurrent selective sweep or background selection. We also evaluated the possible influence of genomic features such as gene density and GC content on the distribution of diversity along genomes.

Materials and methods

Sample collection, DNA preparation and sequencing

Samples were collected in different locations in Europe (Table S1, Supporting information). We sequenced a single individual per locality, as previous studies showed little variability at this scale (Vercken *et al.* 2010; Gladieux *et al.* 2011). We excluded individuals identified as hybrids in previous studies using microsatellites, as these were likely early-generation hybrids that would therefore provide little information on long-term gene flow. Diploid spores from anthers of a single flower were spread on Petri dishes on potato dextro agar (PDA) medium at 23 °C under artificial light for a few days. On nutritive media, the diploid spores undergo meiosis and then the resulting haploid sporidia replicate clonally. A given flower bears diploid spores from a single individual (López-Villavicencio *et al.* 2007). Therefore, the harvested haploid sporidia on PDA represented thousands of meiotic products of a single diploid individual. For six strains, a single haploid clone of a given mating type was isolated for its genome to be sequenced as controls for artefactual heterozygosity (Table S2, Supporting information). For DNA extraction, cells were harvested from PDA medium and stored at −20 °C until use. Most DNAs were extracted using the following method: cells were resuspended in a CTAB buffer, frozen in liquid nitrogen and then crushed with glass beads to break cell walls. Samples were lysed at 60 °C during 4 h with an RNase treatment. DNA was purified with a solution of chloroform–isoamyl alcohol (24:1), precipitated in isopropanol and washed twice with ethanol 70%. The dry pellet was resuspended in deionized water (20–40 μL). A few DNAs (from the strains MvS-100-3, MvSl-IOA, MvSd-IT02, MvSd-932 and MvSd-1034) were extracted

using the Macherey-Nagel NucleoSpin Soil kit #740780.250 following the manufacturer's instructions and resuspended in deionized water (100 μL), as this method was found more rapid and yielding similar DNA quantities and qualities, as well as similar genome sequence qualities. DNA quality was assessed by measuring ratio of 230/260 and 280/260 nm with a NanoDrop 2000 spectrophotometer (Thermo Scientific), and double-strained DNA concentration was measured with a Qubit 2.0 fluorometer. Preparation of DNA libraries and sequencing were performed by Eurofins or at the INRA Genotoul platforms, between which no differences in genome qualities were observed. Paired-end libraries of 2 × 100 bp fragments with an insert size of 300 bp were prepared with Illumina TruSeq Nano DNA Library Prep Kits, and sequencing was performed on a HiSeq2000 Illumina sequencer, at 100× coverage on average. We sequenced the 27-Mb genomes of 34 individuals of MvSl and 19 individuals of MvSd from across Europe, as well as one individual of an outgroup species, *Microbotryum coronariae* (MvLf), using Illumina HiSeq2000 paired-end reads (Fig. 1, Table S1,

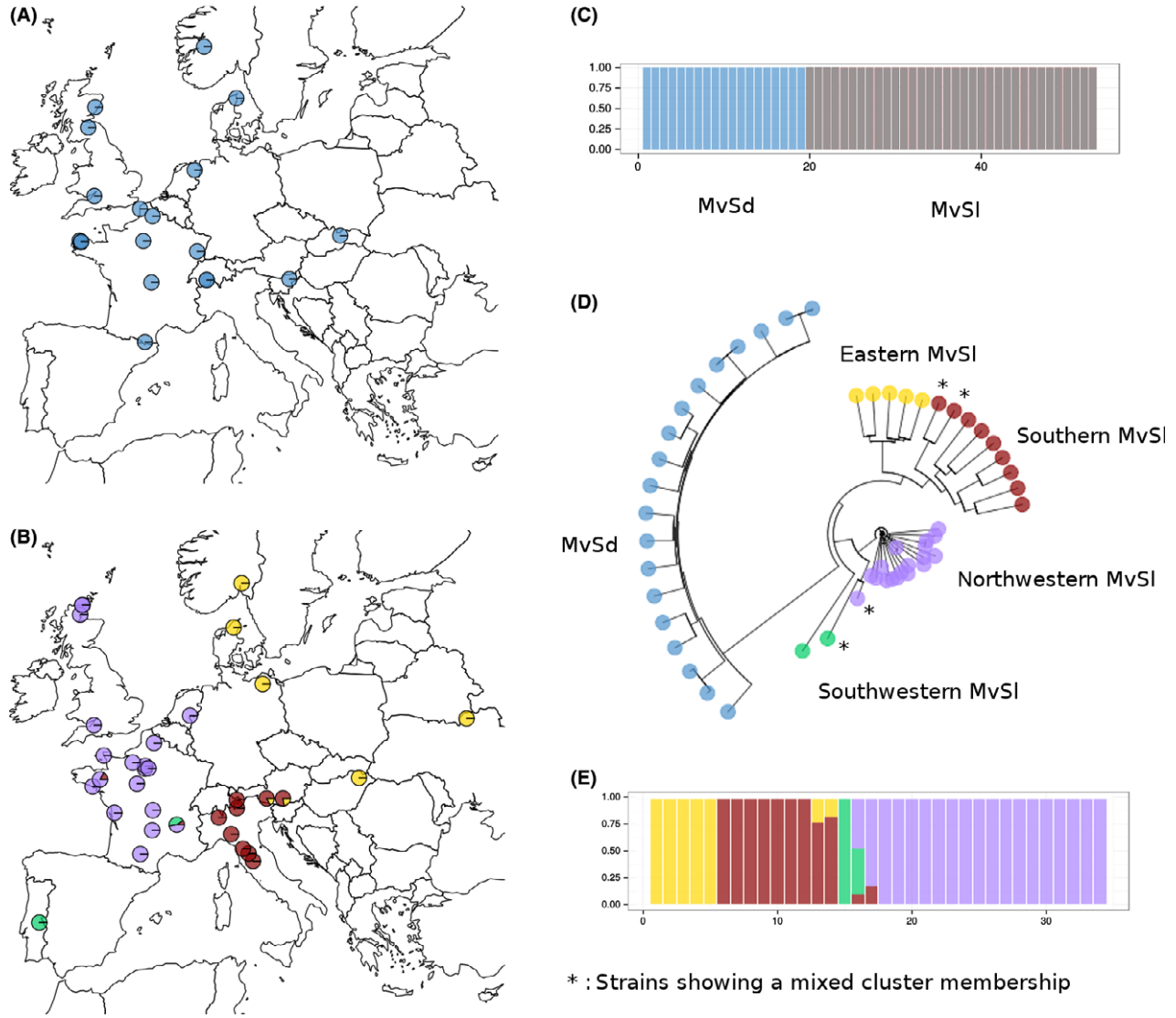

Fig. 1 Population genetic structure in *Microbotryum lychnidis-dioicae* (MvSl) and *Microbotryum silenes-dioicae* (MvSd). (A) Location of sampled individuals in MvSd, in which no genetic subdivision could be detected. (B) Location of sampled individuals in MvSl, with cluster membership for $K = 4$ represented by colours. (C) Membership proportions of MvSl and MvSd individuals, represented as vertical bars, in $K = 2$ clusters. (D) Unrooted neighbor-joining dendrogram representing the genetic distance between MvSl and MvSd individuals. (E) Membership proportions of MvSl individuals, represented as vertical bars, in four clusters. [Colour figure can be viewed at wileyonlinelibrary.com]

Supporting information). The 53 genomes yielded a total number of 18.5–77.4 million reads of 100 bp.

Reads mapping, SNP calling and filtering

Reads were mapped against the high-quality reference genome of the *Microbotryum lychnidis-dioicae* p1A1 Lamole (CLDV0100001-22), with finished, ungapped chromosomes or chromosomal arms, sequenced using the Pacific Biosciences SMRT technology (Badouin *et al.* 2015). We used the glint software (T. Faraut & E. Courcelle, unpublished; http://lipm-bioinfo.toulouse.inra.fr/download/glint/) with parameters set as follows: minimum length of the high-scoring pair hsp ≥90, with ≤5 mismatches, no gap allowed, only best-scoring hits taken into account. The percentage of paired-end reads mapped in a proper pair, that is in the expected orientation and distance, was 40% for the outgroup and ranged otherwise from 59.1% to 94.9% (Table S2, Supporting information). Pairs that were not properly mapped were removed. Because the mating-type chromosomes in *Microbotryum* exhibit suppressed recombination on 90% of their lengths (Hood 2002; Hood *et al.* 2013; Badouin *et al.* 2015), tests for selective sweeps cannot be applied. Therefore, only autosomal data were included in the present study.

Variants were called with VARSCAN v2.3 (Koboldt *et al.* 2012) separately in each strain with the following criteria: coverage ≥20, variant reads ≥10, average quality ≥30, minor allele frequency ≥0.3. This later threshold was chosen not too high to take into account a possible drift among haploid genotypes during the growth on Petri dishes of the haploid sporidia resulting from meioses of a diploid individual. Low complexity regions were masked, SNPs close to indels were removed with the *VarScan filter* function, and SNPs showing a significant strand bias were excluded (Fisher's exact test, *P*-value 0.05). For each strain and each chromosome, we masked positions whose coverage was higher than five times the peak of the Gaussian distribution, with per-base coverage computed with BAMTOOLS (Barnett *et al.* 2011) and coverage histograms generated with BEDTOOLS (Quinlan & Hall 2010). Only biallelic SNPs were retained in analyses, with a maximum rate of missing data of 0.2 per cluster. This threshold of a maximum rate of 0.2 of missing data was applied to all sites when normalizing statistics variation.

To assess the overall quality of SNP calling, we mapped 300× of Illumina paired reads resequencing data of the Lamole reference strain on the corresponding reference genome obtained using the Pacific Biosciences SMRT technology. We analysed the proportion of heterozygous SNPs in the haploid strains for assessing the reliability of heterozygous sites, focusing on autosomes (Table S3, Supporting information).

Genetic subdivision

To analyse genetic subdivision, we selected all autosomal SNPs. We performed a discriminant analysis of principal components (DAPC) with the R package ADEGENET (Jombart 2008). FASTSTRUCTURE (Raj *et al.* 2014) was run for $K = 1$ to $K = 10$ (ten runs per K value). Although *Microbotryum* fungi are highly selfing, previous worked showed that outcrossing and effective recombination occur frequently enough for Bayesian clustering algorithms to be able to accurately identify genetic clusters (Giraud *et al.* 2005; Vercken *et al.* 2010; Gladieux *et al.* 2011). We also built a dendrogram with the neighbor-joining method, using the *nj* function of the R package APE, that used the neighbor-joining algorithm (Saitou & Nei 1987) with default parameters (Paradis *et al.* 2004). To find potentially admixed strains that were not detected by these methods, we analysed the percentage of heterozygous SNPs of each strain in regard to its geographical location. In subsequent population genetics analyses, we excluded potentially admixed strains.

Inference of demographic history and recombination rates

We used the python package DADI (Gutenkunst *et al.* 2009) to infer the demographic history of MvSl and MvSd. The method implemented in DADI infers demographic parameters based on a diffusion approximation to the site frequency spectrum (SFS). Unlike coalescent approaches to demographic inference, which are based on simulations, the diffusion method enables the use of efficient optimization methods to fit a demographic model to the observed SFS. We analysed the MvSl and MvSd data sets separately, as the method cannot handle more than three population samples. The compared models included scenarios of population size changes, strict isolation, continuous postdivergence gene flow, ancient migration and secondary contact, assuming or not heterogeneous migration rates across the genome. Because MvSd and MvSl shared very few polymorphisms (Table S4, Supporting information), we inferred ancestral alleles for these polymorphic sites using, for each species, the alleles of the other species. This allowed to orient more sites than using the outgroup *M. violaceum s.l.* infecting *L. flos-cuculi* (MvLf), for which the mapping quality was lower. The agreement between the orientation inferences obtained by using MvLf or the other sister species was very good (Table S5, Supporting information). We restricted our data set to

include only SNPs that could be confidently oriented, that is for which all outgroup genomes shared the same allele, thus inferred as the ancestral state. Nineteen and six divergence models were considered for MvSl and MvSd, respectively (Table S6, Supporting information). For each model, we ran the numerical optimization at least 20 times with different starting parameter values to ensure convergence. We used the simplex method for numerical optimization, which is the routine recommended for optimization starting far from the true optimum. We assessed the model's goodness-of-fit by maximizing the model likelihood and visual inspection of the residuals between the site frequency spectra generated by the inferred model and the real data (joint-SFS). We used 100 parametric bootstraps to estimate parameters uncertainties and for computing thresholds of significance of the tests for selective sweeps (see below Section 'Genome scans for selective sweeps'). Full 25.2-Mb genomes could not be simulated using available coalescent simulators due to memory constraints. We therefore simulated genomes as combinations of multiple 1-Mb fragments, which was the maximum length that could be simulated and was much larger than the extent of LD. Data sets were simulated using the coalescent simulator msms (Ewing & Hermisson 2010) based on the best-fitting model and estimated parameters. Command lines are provided in Supporting information.

Recombination rates were estimated using the INTERVAL program in LDHAT version 2.2 (Auton & McVean 2007) (Table S7, Supporting information). Data sets for INTERVAL were prepared based on VCF files using custom scripts. Singletons and sites with missing data were excluded and the reversible-jump Markov chain Monte Carlo scheme implemented in INTERVAL was run for 5e6 iterations, with a block penalty of 10, samples taken every 5000 iterations and a burn-in phase of 5e5. For each cluster, LDHAT's program COMPLETE was used to generate likelihood look-up tables for INTERVAL with the population-scaled mutation rate estimated as total nucleotide diversity (π per pb estimated on 100-kb nonoverlapping windows, see Section 'Summary statistics of genomic variation'), and the population-scaled recombination rate ranging from 0 to 100 with a increment of 0.5. Results of INTERVAL were summarized using STAT, in LDHAT, discarding the first 100 samples as burn-in.

Genome scans for selective sweeps

Selective sweeps were searched for using SWEED (Pavlidis *et al.* 2013), which implements a composite likelihood ratio (CLR) test based on the SWEEPFINDER algorithm (Nielsen *et al.* 2005). The CLR uses the variation of the whole or derived SFS of a whole contig to compute the ratio of the likelihood of a selective sweep at a given position to the likelihood of a null model without selective sweep. The null hypothesis relies on the SFS of the whole-genome sequence rather than on a standard neutral model, which makes it more robust to demographic events such as population expansions (Nielsen *et al.* 2005; Pavlidis *et al.* 2013). CLRs were computed every 10, 50 or 100 kb along each contig, using SNPs oriented as described above and whole sequences, not only coding DNA sequence (CDS). To determine the significance of the test, we computed the distribution of CLRs across the genome in 100 data sets simulated under the best neutral demographic scenario (see Section 'Inference of demographic history and recombination rates'). Setting a significance threshold for the deviation of the SFS based on simulated data sets under a neutral demographic model allows further controlling for the impact of demographic events on genomes, such as bottleneck or expansion, that can mimic effects of selection, which gives conservative inferences on the occurrence selective sweeps (Nielsen *et al.* 2005; Pavlidis *et al.* 2013). We concatenated chromosomal fragments simulated with msms for parametric bootstrapping (see Section 'Inference of demographic history and recombination rates') to reconstitute a hundred sets of complete contigs, and computed CLRs along each contig. The 0.95 quantile was used as a significance threshold. Consecutive outlier positions were considered as belonging to a single selective sweep and 5000 bp were added at the flanks of each outlier region. We also computed CLR every 50 and 100 kb and merged consecutive outlier positions by adding 25 000 and 50 000 kb around each outlier position, respectively.

All graphics generated for illustrating these results and other analyses were conducted with circos (Krzywinski *et al.* 2009) or R 3.1.0 (R Core Team 2014).

Summary statistics of genomic variation

Statistics of population genetics were computed with EGGLIB V2 (Mita & Siol 2012) and LIBSEQUENCE (Thornton 2003) in CDSs. We kept only homozygous positions and treated all individuals as haploids, given the very low levels of heterozygosity of the strains retained for these analyses, resulting from high selfing rates in these fungi. We only run these analyses on the largest clusters identified using the assignment analyses, that is the three main clusters found in MvSl and a single cluster in MvSd. After excluding transposable elements and genes that did not pass quality filters (i.e. at least 90% of sites with <20% missing data), between 7059 and 7594 genes were included in analyses depending on the cluster considered. The number of CDS with at least

one SNP ranged between 3191 in MvSl eastern cluster and 5009 CDS in MvSl southern cluster (4894 in northwestern MvSl and 3958 in MvSd). We computed the following statistics within each cluster: total, synonymous and nonsynonymous nucleotide diversities (π, π_S and π_N), Watterson's estimator of θ (Watterson 1975) (θ_W, θ_{WS} and θ_{WN}), Tajima's D (Tajima 1989) and the standardized Fay and Wu statistic H (Zeng *et al.* 2006), using the same method to orient SNPs as in the Section 'Inference of demographic history and recombination rates' above. We also measured differentiation between populations with F_{ST} (Hudson *et al.* 1992), and divergence as the number of fixed differences between populations per kbp. All comparisons and correlations were performed using R version 3.1.0 (R Core Team 2014), in nonoverlapping windows of 50 and 100 kb. Windows with <5 genes or 10 genes in 50 or 100 kb, respectively, were not taken into account. Diversity and fixed divergence were normalized by the number of genotyped sites, defined as the number of sites that passed filters for at least 80% of the individuals in a given cluster or pair of clusters (see detail of filters in Section 'Reads mapping, SNP calling and filtering' section above). To produce input files for EGGLIB, VCF files were converted with custom python scripts (available upon request) to pseudo-alignments in fasta format, where the reference sequence was substituted with the variant nucleotides for each strain.

Linkage disequilibrium was computed as r^2, the coefficient of correlation between a pair of SNPs, with RSQ (Thornton 2003), excluding singletons SNPs. To compute a mean LD for each window, 10 SNPs were selected randomly from each window of 100 kb and 5 SNPs for each window of 50 kb, and the mean r^2 was computed. This was averaged across 10 random selections to get a value of mean LD per window. LD decay with physical distance was evaluated by fitting the observed r^2 values to the decay function (Hill & Weir 1988) with a nonlinear model.

Detection of genes under selection and analysis of gene categories

Functional annotation and detection of specific gene categories. Genes were assigned to functional categories using INTERPROSCAN V5 (Zdobnov & Apweiler 2001). Pfam annotations were used to detect genes of the major facilitator superfamily (PF07690) and sugar transporters (PF00083). Both categories correspond to transmembrane proteins highly represented in *Microbotryum*, and that may play a role in pathogenesis (Perlin *et al.* 2015). Putative secreted proteins were identified as proteins carrying a signal peptide using PHOBIUS (Käll *et al.* 2004) or SIGNALP V4.1 (Petersen *et al.* 2011) and without any detected transmembrane structures (TMHMM, Krogh *et al.* 2001).

Expression data. A previous study (Perlin *et al.* 2015) generated RNA-seq data in three different conditions: in vitro cultures, in low and rich nutrient conditions, respectively, and in vivo late infection, corresponding to infection of flower buds by the fungi. Analysis of differential expression (Perlin *et al.* 2015) yielded a list of 1432 genes that were differentially expressed between at least two conditions, including 307 genes upregulated during infection compared to both in vitro conditions, 208 upregulated in low nutrient in vitro and 59 upregulated in rich nutrients [false discovery rate (FDR) < 0.001]. We built a correspondence table between the transcripts assembled by the Broad Institute used by Perlin *et al.* (2015), and the gene models of high-quality reference genome sequenced using the Pacific Biosciences SMRT technology (Badouin *et al.* 2015). For this goal, transcripts were mapped using GMAP (Wu & Watanabe 2005) on the high-quality reference genome with default parameters, and the corresponding gene models were retrieved using the IntersectBed program of the bedtools suite (Quinlan & Hall 2010).

Detection of genes evolving under selection in specific gene categories. We excluded genes annotated as putative transposable elements or genes with fewer than 90% of sites passing filters from analyses of selection on specific gene categories. One-tailed Student's tests (FDR < 0.05) were used to detect an increase in pN/pS (i.e. the normalized ratio of the proportions of nonsynonymous over synonymous polymorphisms) or dN/dS (i.e. normalized ratio of the number of nonsynonymous over synonymous differences) in specific gene categories. We also performed McDonald and Kreitman tests and computed the number of nonsynonymous and synonymous polymorphism and divergence (P_N, P_S, D_N and D_S). Fisher's exact tests were performed and FDR correction applied on P-values to correct for multiple testing. P_N, P_S, D_N and D_S and the average number of synonymous and nonsynonymous sites (Nei & Gojobori 1986) were computed with egglib and libsequence, and the pN/pS and dN/dS ratios were computed with a custom python script.

The neutrality index, defined as $NI = (D_N/D_S)/(P_N/P_S)$, was computed by adding one pseudo-count to each class of mutation in the contingency table to assure that NI was defined for all genes. We performed an enrichment analysis for categories of interest (genes upregulated in planta, secreted proteins, major facilitator superfamily) on the tails of the NI distribution (0.05 and 0.95 quantiles), with one-tailed Fisher's exact tests.

FDR corrections were applied to correct for multiple testing.

Determinants of genetic diversity

To assess the influence of several genomic traits on the patterns of genetic diversity along sliding windows, we measured total and synonymous diversity with the π and π_S statistics, and tested several explanatory variables: density in CDS, GC content, mean LD and d*N*/d*S*. Mean LD was used as a proxy for the recombination rate, as LD has been shown to be strongly negatively correlated with recombination rate, including in some fungi (Croll *et al.* 2015). Using mean LD or estimates of recombination rate obtained with LDHAT yielded similar results (not shown). For the MvSl southern and eastern clusters, background LD was very strong due to the small numbers of strains, so we used the mean LD in the northwestern cluster for all MvSl clusters, assuming similar patterns of recombination rates along the genome among clusters. Correlations between genetic diversity and those variables were first assessed with Spearman's correlation tests. We also performed multiple linear regressions. For this, a log transformation was applied for improving the normality of the LD and d*N*/d*S* distributions. Normality and homoscedasticity of the residuals were visually assessed. Using the θ statistics instead of π yielded similar results.

Results

Genomic diversity and population subdivision

After filtering, we obtained 203 347 biallelic polymorphic positions within MvSl and 30 296 within MvSd. Because MvSl and MvSd exhibit high rates of selfing (Hood & Antonovics 2004), we expected low rates of heterozygous SNPs in autosomes in diploid genomes, except in recently admixed individuals. The percentages of heterozygous SNPs per strain in fact ranged from 4% to 9% for most strains (Table S3, Supporting information) and were similar between diploid and haploid genomes. This suggested that most heterozygous sites were artefacts due to repetitive DNA or recently duplicated regions, which was supported by the higher coverage of heterozygous than homozygous SNPs in most strains (Table S3, Supporting information). A previous study similarly found unexpected heterozygous SNPs despite stringent filtering in highly inbred nematodes (Rödelsperger *et al.* 2014). Therefore, we only considered homozygous SNPs in subsequent analyses and treated each individual as haploid. The only exceptions to the general low heterozygosity level were three strains in which heterozygosity reached 14%, 34% and 38%, respectively, and in which the coverage of homozygous and heterozygous SNPs was similar (Table S3, Supporting information), suggesting that these strains resulted from recent admixture.

We inferred patterns of population subdivision and admixture in MvSl and MvSd as they might bias inferences about selection. The Bayesian clustering algorithms implemented in FASTSTRUCTURE detected no interspecific hybrids and no population subdivision within MvSd (Fig. 1). Relative divergence between species (measured by F_{ST}) was uniformly high throughout the genome, indicating a lack of genomic regions more permeable to gene flow between species (Fig. S1; Table S8, Supporting information). Within MvSl, FASTSTRUCTURE showed a clear geographical pattern of subdivision, with well-separated clusters in northwestern, eastern and southern Europe (Fig. 1), consistent with previous findings (Vercken *et al.* 2010). Four strains from Austria or France displayed mixed ancestry (Fig. 1E). One of them was among the three strains displaying the highest numbers of heterozygous SNPs. However, the two other highly heterozygous strains did not appear admixed in FASTSTRUCTURE analyses, even when including heterozygous sites in the analysis (not shown). Nevertheless, the relatively high level of heterozygosity, combined with their intermediate geographical location between the northwest and southwest clusters, suggested that these strains might also result from recent admixture. The four heterozygous and/or admixed strains were therefore not considered in subsequent analyses. A DAPC and a distance-based dendrogram supported a population subdivision similar to that inferred by Bayesian approach (Fig. 1 and Fig. S2, Supporting information). More details are given about genetic subdivision patterns in Appendix S1, Fig. S3 (Supporting information).

Mean nucleotide diversity per kb pair ranged from 0.02 in MvSd to 1.10 in the southern cluster of MvSl (Table S9, Supporting information). Pairwise F_{ST} values between MvSl clusters were high (0.56–0.74, Table S8, Supporting information), and there were few shared polymorphisms and many fixed differences (Table S4, Supporting information), supporting previous inferences of low levels of gene flow between clusters.

Genome scans for selection

We used genome scans to detect selective sweeps in the two largest clusters, MvSd and the northwestern cluster of MvSl, by looking for genomic regions where the allelic frequency spectrum deviated from neutrality, taking into account the demographic history and the genome-wide allelic frequency spectrum. The results of the

demographic simulations used to assess the significance of the deviation of the SFS from neutrality, and thus to detect outliers to the expected distribution under neutrality, are given in Appendix S1, Fig. S4 and Table S10 (Supporting information). Derived SFS for each species and cluster, for simulated data sets and models fitted to simulated data sets, and joint-SFS and per-gene distributions of θ, π and Tajima's *D* are presented in Figs S5–S7 (Supporting information). When CLRs were calculated every 10 kb, we detected as many as 208 selective sweeps in the northwestern MvSl cluster, scattered throughout the genome of 27 Mb, spanning 20.8 kb on average and covering 16.9% of the genome (Figs 2 and 3, Table S11 and Table S12, Supporting information). In MvSd, we detected 19 selective sweeps (Figs 2, and 4, Table S13, Supporting information), with a mean size of 13.5 kb, covering 1.0% of the genome, with 23.9% overlap with the sweeps in the northwestern MvSl cluster. Computing the CLRs every 50 or 100 kb gave very similar results in terms of proportion of the genome affected by selective sweeps (Table 1), but detected fewer selective sweeps (45 for 100-kb windows in the northwestern MvSl cluster, and 3 in MvSd, Table 1), suggesting that the some selective sweeps could be counted twice when using smaller windows.

Genomic regions showing footprints of selective sweeps were enriched in genes and had a lower abundance of repeated elements than the rest of the genome. As expected given the construction of the CLR based on the SFS deviations from neutral expectations, the excess of derived low- and high-frequency variants in selective sweeps was also reflected in lower Tajima's *D* and Fay and Wu's *H* values than the rest of the genome (Table S11, Supporting information, Figs 4 and 5). Genomic regions showing footprints of selective sweeps were not enriched in any specific Gene Ontology terms; for instance, they were not enriched in genes encoding putatively secreted proteins or upregulated in planta. However, there were between 5.84 and 33.3 genes per sweep in average for MvSl and 3.74 and 27.3 per sweep for MvSd depending on the window size (Table 1 and Table S13, Supporting information), and it is unlikely that all the genes in these regions have been direct targets of positive selection. Most of them may instead have hitchhiked with another gene under selection; looking at enrichment patterns considering all the putative functions within swept regions may therefore have little power due to the noise introduced by hitchhiking genes. In order to identify candidate genes involved in interaction with the host, we examined the expression patterns and putative functions of the genes located at the centre of the sweeps (four examples are illustrated in Fig. 5, and Tables S12 and S13, Supporting information). The genes in genomic regions showing footprints of selective sweeps and with putative functions and expression patterns making them good candidates for being involved in interactions with the host included genes encoding extracellular membrane proteins with a cysteine-rich CFEM domain (present in effectors in several pathogens, Perlin *et al.* 2015), genes encoding oligopeptide transporters and major facilitators (both from expanded gene families in MvSl and involved in nutrient uptake, Perlin *et al.* 2015), a secreted lipase (another expanded gene family in the MvSl genome, Perlin *et al.* 2015; probably involved in plant cuticle penetration), glyoxal oxidases (required for the switch to filamentous growth and pathogenicity in *Ustilago maydis*, Leuthner *et al.* 2005) and glycoside hydrolases (also known to be involved in pathogenicity in some cases, Ma *et al.* 2015). Genes of unknown function upregulated in planta in MvSl were often found at the centre of the sweeps (Fig. 5). In MvSd, another gene with a cysteine-rich CFEM domain was found in a centre of a sweep.

Genes of pathogens involved in the co-evolutionary arms race with the host, such as those encoding effectors interacting with host defence mechanisms, would be expected to display high rates of nonsynonymous substitutions (Stukenbrock & McDonald 2009). We therefore also compared numbers of synonymous (D_S) and nonsynonymous substitutions (D_N) for genes with features or functions commonly associated with pathogenicity, between clusters and between species. Our candidate pathogenicity genes included genes upregulated in planta relative to in vitro conditions previously identified in MvSl in *Silene latifolia* (Perlin *et al.* 2015), genes encoding secreted proteins, major facilitators and sugar transporters. We compared these candidate genes with the other genes in terms of their neutrality index, defined as $NI = (D_N/P_N)/(D_S/P_S)$, P_N and P_S being the numbers of synonymous and nonsynonymous polymorphisms, respectively. NI measures the strength and direction of departure from neutrality, with the inclusion of polymorphism level adding power for detecting selection between closely related species. The NI index suggested that the genes upregulated in planta were more often under positive selection than other genes. We found a twofold, significant enrichment of genes upregulated in planta in the left tail of the distribution of NI index in comparisons of MvSd and MvSl clusters (Table S14, Supporting information). The distribution of NI was also shifted towards lower values for genes upregulated in planta relative to other genes (Fig. S8, Supporting information). This enrichment in the 0.05 quantile of the NI distribution (i.e. low NI, or $D_N/D_S > P_N/P_S$) indicated that the proportion of substitutions resulting from positive selection was larger in genes upregulated in planta than in other genes, consistent with at least some of these genes being involved in

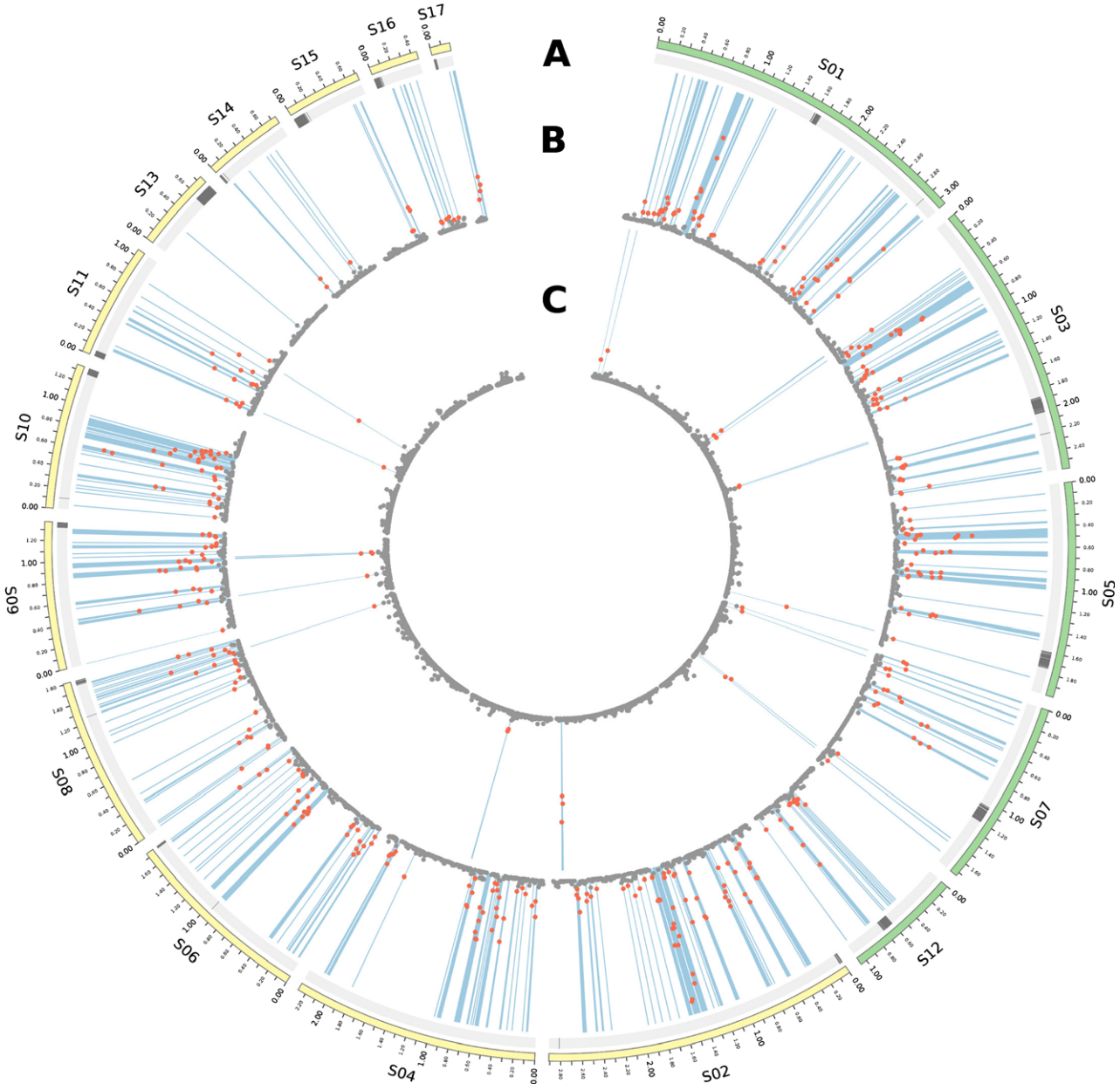

Fig. 2 Composite likelihood ratio (CLR) along the genomes in the anther-smut fungi *Microbotryum lychnidis-dioicae* (MvSl) and *Microbotryum silenes-dioicae* (MvSd). Chromosomes with finished assembly are indicated in green, chromosome arms in yellow and centromeric repeats as grey traits perpendicular to chromosomes. (A) Location of centromeric repeats. (B) and (C) Composite likelihood ratio in northwestern MvSl and MvSd, respectively, with outlier values in red and inferred selective sweeps in blue. The significance thresholds of the CLR were determined with demographic simulations (see 'Materials and methods'). [Colour figure can be viewed at wileyonlinelibrary.com]

an arms race with the host plant. We investigated the putative functions of the genes upregulated in planta with the lowest NI values, and found that the encoded proteins in fact had several putative functions and functional domains known to be implicated in pathogenicity in some fungi (Table S15, Supporting information). They included two major facilitators, two proteins with cysteine-rich CFEM domains (Perlin *et al.* 2015), one secreted and the other anchored to the membrane, a secreted aspartic peptidase, possibly required for anther dehiscence in flowers infected with *Microbotryum* (Perlin *et al.* 2015), a sugar transporter, a glyoxal oxidase (Leuthner *et al.* 2005), a secreted multi-copper oxidase and a ferritin, these last two proteins protecting against host-induced oxidative stress (Perlin *et al.* 2015). Eight of the genes with low NI values were located within

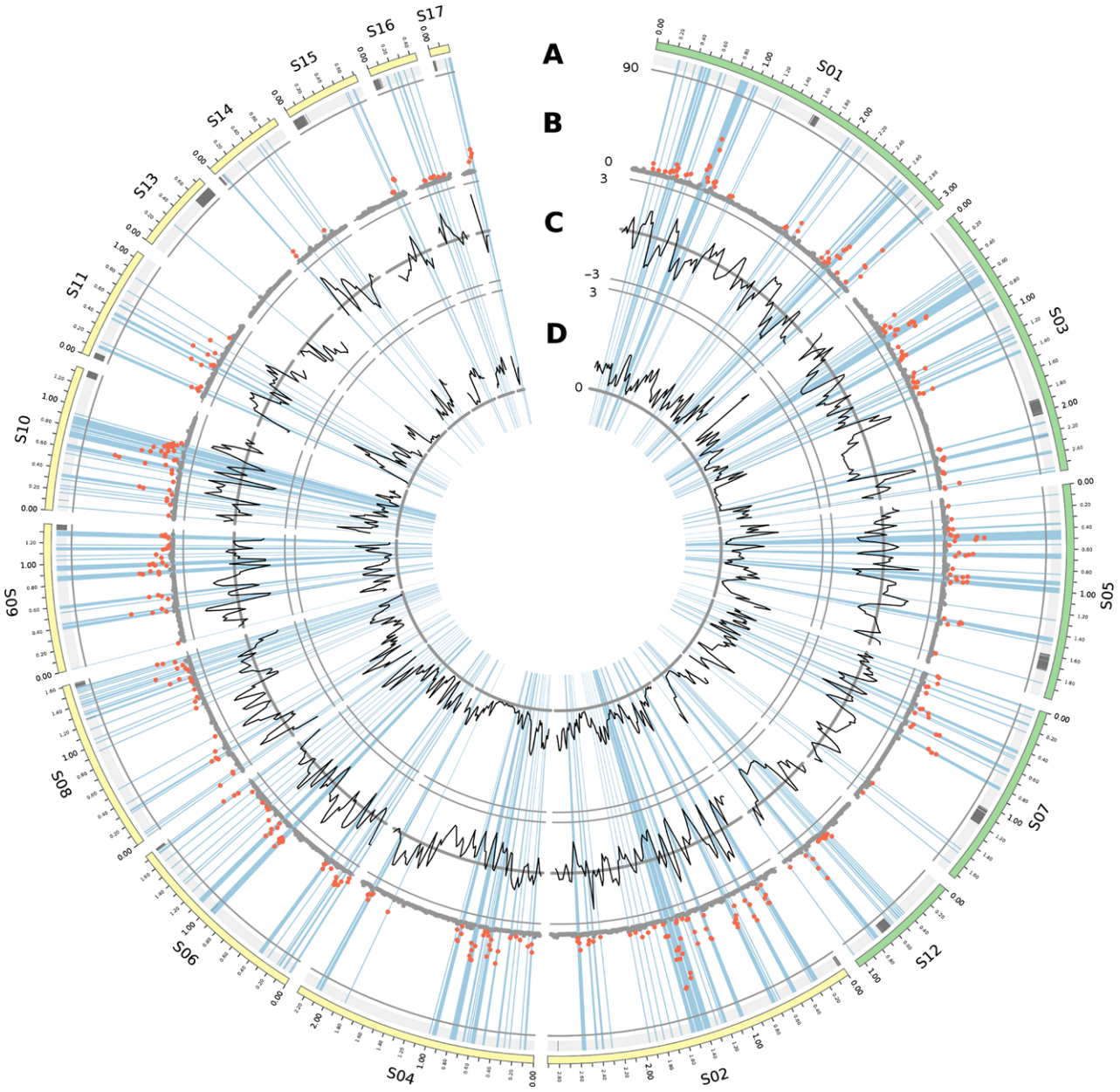

Fig. 3 Composite likelihood ratio (CLR), Tajima's *D* and Fay and Wu's *H* standardized according to Zeng *et al*. (2006), calculated along the genome of the northwestern cluster of the anther-smut fungus *Microbotryum lychnidis-dioicae* (MvSl). Chromosomes with finished assembly are indicated in green, chromosome arms in yellow and centromeric repeats as grey traits perpendicular to chromosomes. (A) Location of centromeric repeats. (B) CLR in northwestern MvSl with outlier values in red and inferred selective sweeps in blue. (C) Tajima's *D* computed using overlapping 50-kb windows. (*D*) θ_π per kb computed using overlapping 50-kb windows. [Colour figure can be viewed at wileyonlinelibrary.com]

putative selective sweeps in MvSl, making them particularly good candidates for being effectors (Table S15, Supporting information).

We also searched for signatures of positive selection by analysing the proportions of nonsynonymous over synonymous polymorphisms and differences between species and clusters (McDonald and Kreitman test and d*N*/d*S*), both genome-wide and focusing on gene categories that were thought a priori to possibly include genes involved in host–pathogen interaction. These analyses did not reveal candidates carrying signatures of positive selection after corrections for false discovery rates and multiple tests (Appendix S1, Tables S16–S18, Supporting information).

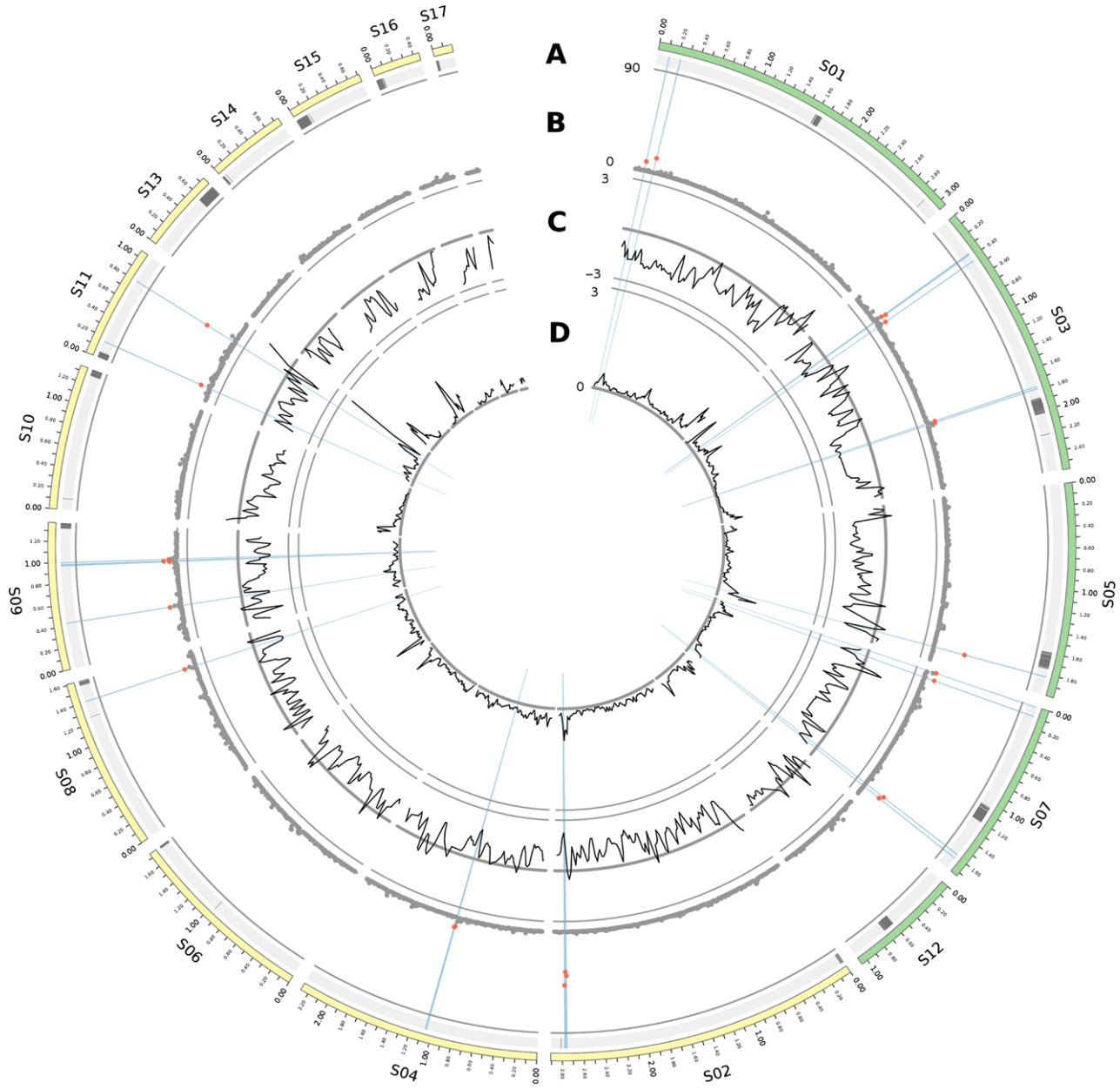

Fig. 4 Composite likelihood ratio (CLR), Tajima's *D* and Fay and Wu's *H* standardized according to Zeng *et al.* (2006) along the genome of the anther-smut fungus *Microbotryum silenes-dioicae* (MvSd). Chromosomes with finished assembly are indicated in green, chromosome arms in yellow and centromeric repeats as grey traits perpendicular to chromosomes. (A) Location of centromeric repeats. (B) Composite likelihood ratio in the northwestern cluster with outlier values in red and inferred selective sweeps in blue. (C) Tajima's *D* computed using overlapping 50-kb windows. (D) θ_π per kb computed using overlapping 50-kb windows. [Colour figure can be viewed at wileyonlinelibrary.com]

Genome-wide impact of selection

We tested whether polymorphism was negatively correlated with levels of LD along the genome, as expected in cases of recurrent selective sweeps or background selection (Cutter & Payseur 2003; Stephan 2010). We used LD as a proxy for the recombination rate. We also tested the influence of other genomic features, such as gene density and GC content, on the distribution of genomic variability along genomes. Previous work showed that GC content can also be positively correlated with levels of diversity along genomes (Alföldi *et al.* 2011), because local enrichment in GC often occurs in regions where recombination or gene conversion occurs frequently (Duret *et al.* 2006). Conversely, a negative correlation between GC content and

Table 1 Number and percentage of the genome in selective sweeps detected using composite likelihood ratios (CLRs) above a threshold determined based on demographic simulations, computing CLRs every 10-, 50- or 100-kb windows

	Northwestern MvSl		MvSd	
	Percentage	Number of sweeps	Percentage	Number of sweeps
10 kb	16.90	208	1.00	19
50 kb	15.57	65	0.93	5
100 kb	17.68	45	1.03	3

MvSl, *Microbotryum lychnidis-dioicae*; MvSd, *Microbotryum silenes-dioicae*.

polymorphism is expected when GC-biased gene conversion occurs (Marais 2003; Glémin *et al.* 2014).

We estimated LD in MvSd and the largest MvSl cluster (i.e. northwestern) using r^2, that is the coefficient of correlation between pairs of SNPs (Hill & Robertson 1968). The rate of LD decay was higher in MvSd than in northwestern MvSl, with r^2 decreasing below 0.2 after 67 kb in MvSd and 96 kb in northwestern MvSl (Fig. S9, Supporting information). LD was stronger around centromeres, as expected as these regions usually show low rates of recombination.

Variations in diversity within each of the four genetic groups, as well as mean LD within the MvSl northwestern cluster, are shown in Fig. S1 (Supporting information). Within-cluster diversity levels along chromosomes were significantly correlated between MvSl clusters (Table S19, Supporting information, e.g. Spearman's $r = -0.46$, $P < 0.0001$ using 100-kb windows between the southern and northwestern clusters), and between MvSl and MvSd (Spearman's $r = -0.41$, $P < 0.0001$ using 100-kb windows between MvSd and the MvSl southern cluster). This suggests that recombination hotspots and coldspots are located in the same regions in the different MvSl clusters and in MvSd.

Positive selection reduces diversity at linked sites and the size of the genomic fragments impacted depends on the local recombination rate, leading to a positive correlation between polymorphism and recombination rate (Cutter & Payseur 2003) and, thus, a negative correlation between local diversity and LD. A positive correlation between polymorphism and recombination rate is thus a further indication of the occurrence of recurrent selective sweeps genome-wide. We therefore looked for the variables explaining within-species diversity across the genome in MvSl and MvSd [using either total (π) or synonymous (πs) diversity], including recombination rates, using LD as a proxy. Genetic hitchhiking by sites linked to sites under selection would be expected to occur more frequently when nonsynonymous substitutions (which may be deleterious or adaptive) are frequent (Tiley & Burleigh 2015), which would lead to a negative correlation between polymorphism and the density of coding sites, the principal targets of selection (Payseur & Nachman 2002). We therefore also examined the effect of dN/dS on polymorphism level. GC content was also used as an explanatory variable for levels of diversity, as GC enrichment has been observed in regions of frequent recombination (Duret *et al.* 2006) and, conversely, GC-biased gene conversion can lead to lower variability associated with a high GC content (Marais 2003). We thus tested for the effects of the following variables, for each cluster and window size: mean LD, GC content, CDS (CDS) density and pairwise dN/dS. Mean LD explained most of the variance in nucleotide diversity throughout the genome (Tables S20 and S21, Supporting information). LD had a significant negative effect on diversity in all MvSl clusters (Fig. 6), but not in MvSd, consistent with frequent selective sweeps in MvSl (Stephan 2010). GC content had a significant negative effect in most models (Tables S15 and S20, Supporting information), as expected at equilibrium with GC-biased gene conversion (Marais 2003), which maintains lower levels of diversity (Marais 2003) (see Appendix S1, Supporting information). GC-biased conversion gene has been shown to occur at recombination hotspots in fungi (Lesecque *et al.* 2013).

We also found a strong negative correlation between polymorphism and per-base number of fixed differences between each MvSl cluster and MvSd, showing that the less polymorphic regions tended to accumulate more fixed differences (Table S22, Supporting information, e.g. Spearman's $r = -0.40$, $P < 0.001$ at 100 kb in the MvSl Italian cluster), as expected under recurrent selective sweeps. We also looked at the distribution of dN/dS across genes (ratio of number of fixed nonsynonymous differences between MvSl and MvSd per nonsynonymous site over number of fixed synonymous differences per synonymous site, Fig. S10, Supporting information). Most of the genes had a dN/dS ratio close to 0, indicating that they evolve under purifying selection; however, a second, smaller peak was observed around $dN/dS = 1$, suggesting that a fraction of the genes may evolve under relaxed selection.

Discussion

Low linkage disequilibrium levels indicate effective recombination despite high selfing rates

Linkage disequilibrium was found to decay relatively slowly, as expected for selfing species. Indeed, r^2 decreased below 0.2 within 67 kb in MvSd and 96 kb in northwestern MvSl. The decay was even slower than in

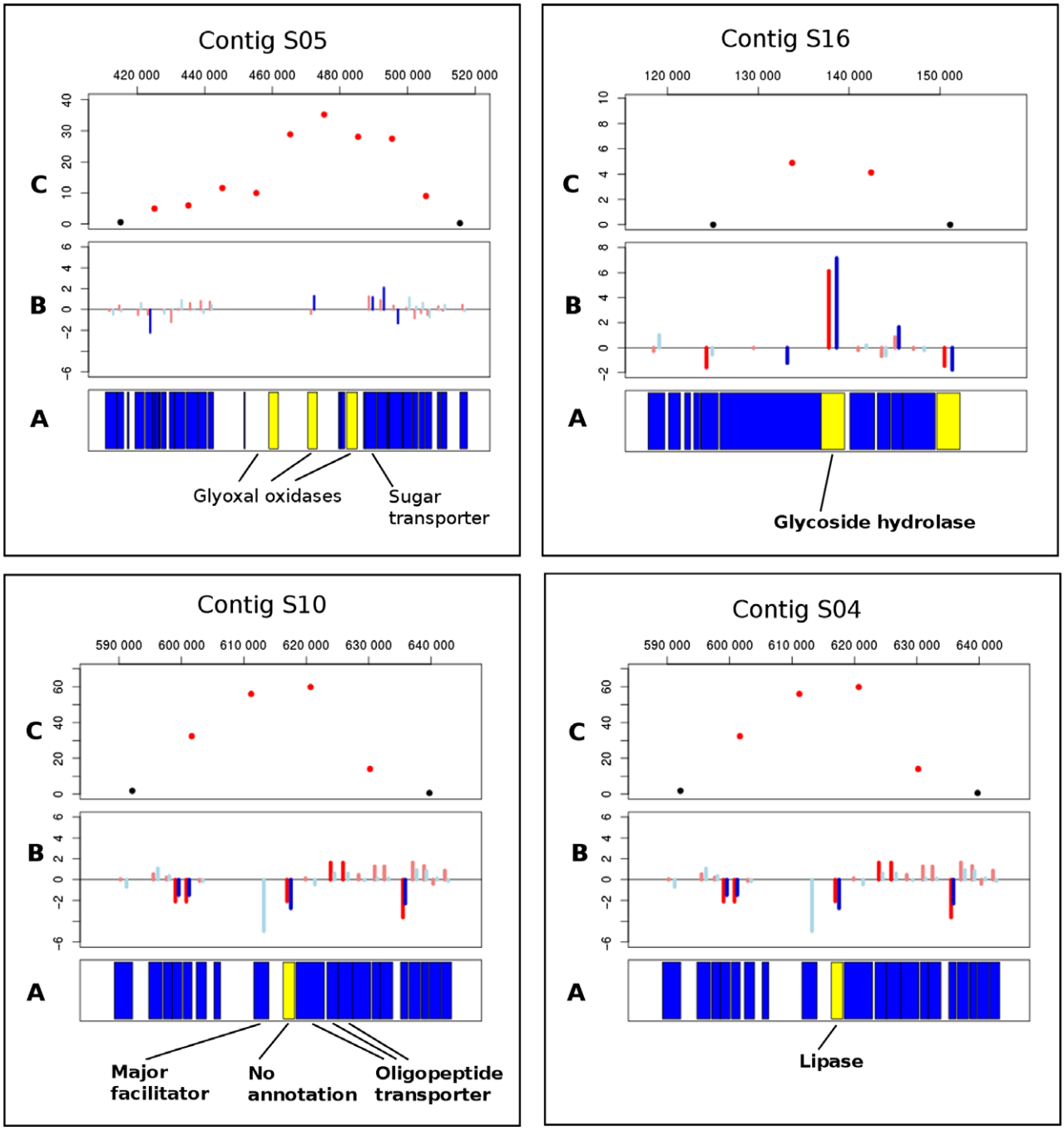

Fig. 5 Zoom on four selective sweeps in *Microbotryum lychnidis-dioicae*. (A) Location of predicted genes in blue, with indication of interesting putative functions or secretion signals (in yellow); lipases, oligopeptide transporter and major facilitator domains are expanded in the *M. lychnidis-dioicae* genome compared to other fungi. (B) Differential expression of genes, indicated as the difference of $\log_2$(FPKM + 1) values in planta versus in vitro rich media (blue) or in planta vs. nutrient-limited conditions (red). Positive values mean higher expression in planta than in vitro. Significant differential expression is indicated by a brighter shade. (C) Composite likelihood ratio (CLR). For each selective sweeps represented, the contig ID and genomic coordinates (in bp) are indicated. [Colour figure can be viewed at wileyonlinelibrary.com]

the selfing plant *Arabidopsis thaliana*, which was found within 50 kb (Nordborg *et al.* 2005). However, the fact that LD does not extend beyond 100 kb indicates that effective recombination regularly occurs in nature despite the high rates of intratetrad mating. Selfing rates were in fact previously found to be below 100%, with estimates based on microsatellite markers ranging between 0.90 and 0.95 in *Microbotryum lychnidis-dioicae*

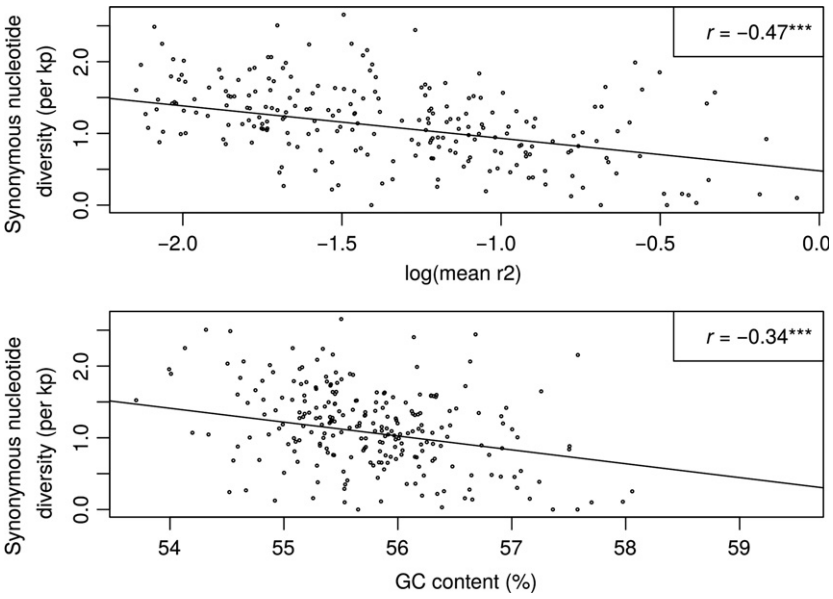

Fig. 6 Correlations between synonymous nucleotide diversity and other parameters in the southern cluster of *Microbotryum lychnidis-dioicae* (MvSl), using 100-kb windows. The nucleotide diversity is plotted in the top panel against the log of mean linkage disequilibrium (r^2) and the bottom panel against the GC content. The regression lines are drawn and Spearman's *r* coefficients are indicated.

and *Microbotryum silenes-dioicae* (Vercken *et al.* 2010). Therefore, while this level of selfing is high enough to greatly depress individual heterozygosity and to induce relatively high LD in genomes, it still allows the reshuffling of alleles, making it possible to perform analyses of population structure and natural selection.

Widespread genomic divergence between species and lack of introgression

We detected no gene flow between the species *M. lychnidis-dioicae* and *M. silenes-dioicae* in whole-genome sequences, while previous studies using microsatellite markers had detected some interspecific hybrids, albeit at a low rate (Gladieux *et al.* 2011). Different habitats of the plants and different pollinator guilds can be invoked to account for the limited level of hybridization despite widespread coexistence in sympatry (Goulson & Jerrim 1997; van Putten *et al.* 2007). However, previous approaches based on a dozen of microsatellite markers could have missed later-generation hybrids, such as backcrossed individuals. In the present study, we deliberately chose to sequence microsatellite genotypes that appeared of pure ancestry in order to be able to assess whether long-term gene flow between the two sympatric sister species parasitizing sister plant species occurred beyond first-generation hybrids. We found no evidence for admixture, indicating that introgression does not persist beyond one or two generations. Experimental crosses had previously evidenced a lack of premating isolation among *Microbotryum* species, even for individuals of different species coexisting at the same sites (Le Gac *et al.* 2007; Refrégier *et al.* 2010), but postmating barriers such as hybrid inviability and sterility appeared to be positively correlated with genetic distance (Le Gac *et al.* 2007; de Vienne *et al.* 2009). The performance of F1 hybrids between the two sister species under study, *M. lychnidis-dioicae* and *M. silenes-dioicae*, was not significantly different from the performance of their parental species in the laboratory (Van Putten *et al.* 2003; Le Gac *et al.* 2007; Refrégier *et al.* 2010), but F2 hybrids produced by selfing F1s had mostly returned to homozygosity, as estimated using three dozens of microsatellite markers, suggesting that genomic content derived from one of the two parental species had already been purged (de Vienne *et al.* 2009). The findings presented here suggest that introgression does not persist in nature, as it is not detected at any genomic regions, further arguing for a very strong genome-wide selection by the host plant, likely promoted by the scattering of genes involved in interactions with the host across the genome. This hypothesis is supported by the finding of selective sweeps and genes upregulated in planta all along the genomes and not clustered in particular locations. Selfing also acts as a strong barrier to interspecific gene flow in *Microbotryum* because of the systematic competition between hybrids and selfed progeny during plant infection (Gibson *et al.* 2012).

Diversity and genomic signatures of selection at linked sites

The overall genetic diversity levels were consistent with the limited variability at microsatellite loci previously reported for these pathogens (Vercken *et al.* 2010), and the smaller effective population size of MvSd than of MvSl (Gladieux *et al.* 2011). We found a strong negative effect of LD on diversity along the genomes in MvSl,

which suggests that recurrent linked selection due to selective sweeps plays a role in shaping diversity (Cutter & Payseur 2003; Stephan 2010). Background selection can also lead to a negative correlation between polymorphism and recombination rates and it has been difficult so far to elucidate whether background selection or positive selection was the main factor driving this pattern (Stephan 2010). The correlation between diversity levels along chromosomes between MvSl clusters is thus consistent with the action of background selection and recombination. Our finding that LD was negatively correlated with genetic diversity in MvSl, but not in MvSd, in which much fewer selective sweeps were found, however, suggests that positive selection plays an important role. The strong negative correlation between polymorphism and per-base number of fixed differences between each MvSl cluster and MvSd, showing that the less polymorphic regions tended to accumulate more fixed differences, is also in agreement with recurrent selective sweeps (Manthey *et al.* 2015). Indeed, lower recombination rates increase the probability of mutation fixation within species on larger regions around sites affected by positive selection (Manthey *et al.* 2015). These findings on patterns of diversity along genomes being correlated with LD and differences between species altogether thus bring independent evidence for widespread and frequent selective sweeps throughout the genome of MvSl.

We found a negative correlation between polymorphism level and GC content, which is expected at equilibrium if GC-biased gene conversion occurs, as GC-biased conversion acts as selection in maintaining lower levels of diversity (Marais 2003). GC-biased conversion has been shown to occur in fungi at recombination hotspots (Lesecque *et al.* 2013). A negative correlation between polymorphism level and GC content can also be due to selection for GC-rich codons. In *M. lychnidis-dioicae*, GC content was found positively correlated with coding density (Perlin *et al.* 2015), which can be due to biased codon usage towards GC-rich codons or biased gene conversion occurring more frequently in coding than in noncoding sequences (Duret & Galtier 2009). An analysis of the preferred codons (i.e. the most frequently used codons in the predicted genes) showed that 17 of 18 had a GC base at the third position (Perlin *et al.* 2015), which is the most degenerate position and therefore primarily influences the GC composition of genes.

Detection of genes under selection

We detected particularly high genetic diversity in the genes upregulated in planta, indicating that these accumulated more nonsynonymous substitutions than other genes, which supports the view that they are likely involved in interaction with the host plant. However, the neutrality index of the genes upregulated in planta was only slightly smaller than the genomic background, which may be due to the fact that only a fraction of these genes is involved in the host–pathogen arms race. We therefore focused on the genes upregulated in planta with a low value of the neutrality index, that is those having fixed the highest proportion of nonsynonymous substitutions, and found several interesting putative functions. Among them, CFEM proteins are cysteine rich with extracellular domains and particularly good candidates for being fungal effectors (Kulkarni *et al.* 2005; Perlin *et al.* 2015); a glyoxal oxidase has been found to be required for pathogenicity in *Ustilago maydis* (Leuthner *et al.* 2005). Multi-copper oxidases and ferritin can protect against host-induced oxidative stress (Festa & Thiele 2012). Aspartic peptidase is necessary for pollen tube elongation and may be necessary for anther dehiscence in flowers infected by *Microbotryum* (Perlin *et al.* 2015). Pectin lyases are also known to be involved in pathogenicity (Ma *et al.* 2015), as well as glycoside hydrolases (Ma *et al.* 2015).

We found that selective sweeps were abundant in *M. lychnidis-dioicae*, affecting nearly 17% of the genome. The sister species *M. silenes-dioicae* displayed fewer selective sweeps, affecting only 1% of its genome. The number of selective sweeps inferred increased when smaller windows (10- and 50-kb windows) were used, which was not surprising given that the LD extends over 50 kb in these highly selfing fungi. Windows larger than the extent of LD should therefore be used, and 100-kb windows inferred 45 selective sweeps in MvSl and three in MvSd. The selective sweeps involved different genomic regions between the two *Microbotryum* species and were scattered along the genome. Moreover, the genes upregulated in planta, and those coding putatively secreted proteins, did not seem to be clustered. Thus, there does not seem to be 'genomic islands' of genes involved in pathogenicity in *Microbotryum*, contrary to what has been observed in other pathogenic fungi (Rouxel *et al.* 2011). Furthermore, we detected fewer selective sweeps in MvSd than in MvSl. This was unlikely due to the mapping of MvSd reads on an MvSl reference as the proportion of pairs of reads properly mapped on the MvSl reference genome and the number of sites passing quality filter were similar in MvSd and MvSl. The fewer selective sweeps in MvSd than in MvSl may be due to a lower level of genetic diversity in MvSd, which both reduces the variation on which selection can act and lower the power for detecting selective sweeps. In addition, the LD extended across larger genomic regions in MvSl, which allows detecting more and older selective sweeps. Alternatively, such differences may reflect a true difference in the number of

selective sweeps between species, reflecting a slower co-evolution of MvSd with its host, or a lower adaptive potential, possibly due to a smaller effective population size in MvSd and/or lower levels of dispersal between populations, as previously reported (Vercken *et al.* 2010). In fact, evidence of co-evolution has been detected in MvSl in cross-inoculation experiments where different rates of infections were observed between sympatric and allopatric populations of plant–fungal pairs (Kaltz *et al.* 1999; Feurtey *et al.* 2016). Future similar experiments in MvSd could test whether smaller differences are detected between allopatric and sympatric populations compared to MvSl, which would support the view of a less intense co-evolution in the MvSd–*S. dioica* than in the MvSl–*Silene latifolia* interactions.

Within the selective sweeps, we did not find any enrichment of genes upregulated in planta, encoding putatively secreted proteins nor having particular gene ontology. However, it is unlikely that all the genes present in those regions represent targets of positive selection given the relatively large LD along the genome. Some interesting putative functions were found for genes at the centre of the regions harbouring signatures of selective sweeps, including again CFEM domain genes upregulated in planta, putatively extracellular and membrane-anchored, a cluster of OPT and major facilitator transporters and a secreted lipase. OPT and major facilitator domains are found in membrane transporters, are expanded in *M. lychnidis-dioicae* and could be involved in nutrient uptake (Perlin *et al.* 2015). Secreted lipases are also expanded in *M. lychnidis-dioicae* and could be involved in penetration of the plant by the fungi through the wax of the cuticle. Some genes located at the centre of sweeps were upregulated in planta, but did not have any annotation. Fungal putative effectors, such as small secreted proteins upregulated during infection, often lack annotations, which would be related to the fact that they evolve rapidly and that sequence homology is difficult to establish, even between close species (Stergiopoulos & de Wit 2009). All these candidates will also be used for future functional studies for investigating their roles in pathogenicity.

Conclusions

We identified numerous selective sweeps throughout the genome of a fungal plant pathogen. The impact of selection in the genome of *M. lychnidis-dioicae* may even be greater, as our approach using demographic models for assessing significant deviations from neutrality of the allelic frequency spectrum has been shown to be conservative (Nielsen *et al.* 2005). Such widespread sweeps along the genome of a natural plant pathogenic fungus may result from selection on polygenic traits, positive epistasis and/or rapid co-evolution with its host plant (Bonhomme *et al.* 2015). The regions with footprints of selective sweeps in fact included genes with putative functions and patterns of upregulation in planta consistent with a role in pathogenicity. This study thus provides clues to the genes involved in pathogen–plant interaction, which have long eluded identification in this well-studied system. The identification of these putative effectors will foster future functional and evolutionary studies, in both the host plant and anther-smut pathogens (Bernasconi *et al.* 2009; Perlin *et al.* 2015).

Selective sweeps have been found abundant in some organisms previously (Sattath *et al.* 2011; Long *et al.* 2013; Bonhomme *et al.* 2015), while not in others (Gossmann *et al.* 2010; Hernandez *et al.* 2011), leading to the hypothesis that organisms with smaller effective population sizes have fewer selective sweeps (Slotte *et al.* 2010; Hernandez *et al.* 2011). Here, we found, in contrast to this hypothesis, widespread selective sweeps in *Microbotryum lychnidis-dioicae*, largely impacting the diversity along its whole genome, despite small effective size. This may be because *M. lychnidis-dioicae* is a natural pathogen experiencing dynamic co-evolutionary arms race with its host. In addition, our findings of a significant correlation between polymorphism and LD only in the species displaying footprints of widespread selective sweeps contribute to the body of evidence that is needed for resolving the long-standing controversy about the relative importance of positive selection and background selection in driving this pattern (Stephan 2010). Our findings therefore broaden knowledge about the occurrence and frequency of selection in natural populations and provide hypotheses to test in further studies. Our results also reinforce the view that GC content and linked selection can have a pervasive impact on the genetic diversity in genomes (Cutter & Payseur 2003; Burri *et al.* 2015; Manthey *et al.* 2015), especially in selfing species with high levels of LD (Cutter & Payseur 2003; Nordborg *et al.* 2005; Khan *et al.* 2009; Andersen *et al.* 2012) and in organisms involved in co-evolution with symbionts (Bonhomme *et al.* 2015).

Acknowledgements

We thank the Genotoul platform for sequencing. T. Giraud acknowledges the ANR Grant ANR-12-ADAP-0009 and the ERC Starting Grant GenomeFun 309403. We thank all the strain collectors (Table S1, Supporting information). We thank Michael Hood for strains and help all along the project. We thank Gilles Deparis and Laetitia Giraud for technical help and Nicolas Bierne, Christophe Lemaire and Christelle Fraïsse for help with analyses.

References

Alföldi J, Di Palma F, Grabherr M *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**, 587–591.

Andersen EC, Gerke JP, Shapiro JA *et al.* (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics*, **44**, 285–290.

Anderson PK, Cunningham AA, Patel NG *et al.* (2004) Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution*, **19**, 535–544.

Ashby B, Gupta S (2014) Parasitic castration promotes coevolutionary cycling but also imposes a cost on sex. *Evolution; International Journal of Organic Evolution*, **68**, 2234–2244.

Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. *Genome Research*, **17**, 1219–1227.

Badouin H, Hood ME, Gouzy J *et al.* (2015) Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus *Microbotryum lychnidis-dioicae*. *Genetics*, **200**, 1275–1284.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692.

Bernasconi G, Antonovics J, Biere A *et al.* (2009) *Silene* as a model system in ecology and evolution. *Heredity*, **103**, 5–14.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bonhomme M, Boitard S, Clemente HS *et al.* (2015) Genomic signature of selective sweeps illuminates adaptation of *Medicago truncatula* to root-associated microorganisms. *Molecular Biology and Evolution*, **32**, 2097–2110.

Branco S, Bi K, Liao HL *et al.* (2017) Continental-level population differentiation and environmental adaptation in the mushroom *Suillus brevipes*. *Molecular Ecology*, doi: 10.1111/mec.13892.

Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. *Genome Research*, **25**, 1656–1665.

Carlsson-Granér U (1997) Anther-smut disease in *Silene dioica:* variation in susceptibility among genotypes and populations, and patterns of disease within populations. *Evolution*, **51**, 1416–1426.

Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 700–707.

Croll D, Lendenmann MH, Stewart E, McDonald BA (2015) The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics*, **201**, 1213–1228.

Cutter AD, Payseur BA (2003) Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Molecular Biology and Evolution*, **20**, 665–673.

Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, **10**, 285–311.

Duret L, Eyre-Walker A, Galtier N (2006) A new perspective on isochore evolution. *Gene*, **385**, 71–74.

Ellison CE, Hall C, Kowbel D *et al.* (2011) Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2831–2836.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.

Festa RA, Thiele DJ (2012) Copper at the front line of the host-pathogen battle. *PLoS Pathogens*, **8**, e1002887.

Feurtey A, Gladieux P, Hood ME *et al.* (2016) Strong phylogeographic co-structure between the anther-smut fungus and its white campion host. *The New Phytologist*, **212**, 668–679.

Fontanillas E, Hood ME, Badouin H *et al.* (2015) Degeneration of the nonrecombining regions in the mating-type chromosomes of the anther-smut fungi. *Molecular Biology and Evolution*, **32**, 928–943.

Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 2977–2982.

Gibson AK, Hood ME, Giraud T (2012) Sibling competition arena: selfing and a competition arena can combine to constitute a barrier to gene flow in sympatry. *Evolution; International Journal of Organic Evolution*, **66**, 1917–1930.

Gibson AK, Refrégier G, Hood Michael E, Giraud T (2014) Performance of a hybrid fungal pathogen on pure-species and hybrid host plants. *International Journal of Plant Sciences*, **175**, 724–730.

Giraud T, Odile J, Shykoff JA (2005) Selfing propensity under choice conditions in a parasitic fungus, *Microbotryum violaceum,* and parameters influencing infection success in artificial inoculations. *International Journal of Plant Sciences*, **166**, 649–657.

Giraud T, Yockteng R, López-Villavicencio M, Refrégier G, Hood ME (2008a) Mating system of the anther smut fungus *Microbotryum violaceum:* selfing under heterothallism. *Eukaryotic Cell*, **7**, 765–775.

Giraud T, Yockteng R, Marthey S *et al.* (2008b) Isolation of 60 polymorphic microsatellite loci in EST libraries of four sibling species of the phytopathogenic fungal complex *Microbotryum*. *Molecular Ecology Resources*, **8**, 387–392.

Giraud T, Gladieux P, Gavrilets S (2010) Linking the emergence of fungal plant diseases with ecological speciation. *Trends in Ecology & Evolution*, **25**, 387–395.

Gladieux P, Vercken E, Fontaine MC *et al.* (2011) Maintenance of fungal pathogen species that are specialized to different hosts: allopatric divergence and introgression through secondary contact. *Molecular Biology and Evolution*, **28**, 459–471.

Gladieux P, Ropars J, Badouin H *et al.* (2014) Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular Ecology*, **23**, 753–773.

Glémin S, Clément Y, David J, Ressayre A (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics*, **30**, 263–270.

Gossmann TI, Song B-H, Windsor AJ *et al.* (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, **27**, 1822–1832.

Goulson D, Jerrim K (1997) Maintenance of the species boundary between *Silene dioica* and *S. latifolia* (red and white Campion). *Oikos*, **79**, 115–126.

Granka JM, Henn BM, Gignoux CR *et al.* (2012) Limited evidence for classic selective sweeps in African populations. *Genetics*, **192**, 1049–1064.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

Haasl RJ, Payseur BA (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, **25**, 5–23.

Hancock AM, Brachi B, Faure N *et al.* (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Hernandez RD, Kelley JL, Elyashiv E *et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science*, **331**, 920–924.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, **33**, 54–78.

Hood ME (2002) Dimorphic mating-type chromosomes in the fungus *Microbotryum violaceum*. *Genetics*, **160**, 457–461.

Hood ME, Antonovics J (2004) Mating within the meiotic tetrad and the maintenance of genomic heterozygosity. *Genetics*, **166**, 1751–1759.

Hood ME, Mena-Alí JI, Gibson AK *et al.* (2010) Distribution of the anther-smut pathogen *Microbotryum* on species of the Caryophyllaceae. *The New Phytologist*, **187**, 217–229.

Hood ME, Petit E, Giraud T (2013) Extensive divergence between mating-type chromosomes of the anther-smut fungus. *Genetics*, **193**, 309–315.

Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **31**, 3026–3039.

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jones BL, Raga TO, Liebert A *et al.* (2013) Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. *American Journal of Human Genetics*, **93**, 538–544.

Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, **338**, 1027–1036.

Kaltz O, Gandon S, Michalakis Y, Shykoff JA (1999) Local maladaptation in the anther-smut fungus *Microbotryum violaceum* to its host plant *Silene latifolia:* evidence from a cross-inoculation experiment. *Evolution*, **53**, 395–407.

Khan A, Taylor S, Ajioka JW, Rosenthal BM, Sibley LD (2009) Selection at a single locus leads to widespread expansion of *Toxoplasma gondii* lineages that are virulent in mice. *PLoS Genetics*, **5**, e1000404.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568–576.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

Krzywinski MI, Schein JE, Birol I *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, **19**, 1639–1645.

Kulkarni RD, Thon MR, Pan H, Dean RA (2005) Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus *Magnaporthe grisea*. *Genome Biology*, **6**, R24.

Lamason RL, Mohideen M-APK, Mest JR *et al.* (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **310**, 1782–1786.

Le Gac M, Hood ME, Giraud T (2007) Evolution of reproductive isolation within a parasitic fungal species complex. *Evolution*, **61**, 1781–1787.

Lesecque Y, Mouchiroud D, Duret L (2013) GC-Biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, **30**, 1409–1419.

Leuthner B, Aichinger C, Oehmen E *et al.* (2005) A H_2O_2-producing glyoxal oxidase is required for filamentous growth and pathogenicity in *Ustilago maydis*. *Molecular Genetics and Genomics*, **272**, 639–650.

Long Q, Rabanal FA, Meng D *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, **45**, 884–890.

López-Villavicencio M, Jonot O, Coantic A *et al.* (2007) Multiple infections by the anther smut pathogen are frequent and involve related strains. *PLoS Pathogens*, **3**, e176.

Ma Z, Song T, Zhu L *et al.* (2015) A *Phytophthora sojae* glycoside hydrolase 12 protein is a major virulence factor during soybean infection and is recognized as a PAMP. *The Plant Cell*, **27**, 2057–2072.

Manthey JD, Klicka J, Spellman GM (2015) Chromosomal patterns of diversity and differentiation in creepers: a next-gen phylogeographic investigation of *Certhia americana*. *Heredity*, **115**, 165–172.

Marais G (2003) Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, **19**, 330–338.

Mita SD, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, **13**, 27.

Morris GP, Ramu P, Deshpande SP *et al.* (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 453–458.

Nair S, Williams JT, Brockman A *et al.* (2003) A selective sweep driven by pyrimethamine treatment in Southeast Asian malaria parasites. *Molecular Biology and Evolution*, **20**, 1526–1536.

Neafsey DE, Barker BM, Sharpton TJ *et al.* (2010) Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. *Genome Research*, **20**, 938–946.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps in

thousands of genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.

Payseur BA, Nachman MW (2002) Gene density and human nucleotide polymorphism. *Molecular Biology and Evolution*, **19**, 336–340.

Perlin MH, Amselem J, Fontanillas E *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.

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

Presti LL, Lanver D, Schweizer G *et al.* (2015) Fungal effectors and plant susceptibility. *Annual Review of Plant Biology*, **66**, 513–545.

van Putten WF, Elzinga Jelmer A, Biere A (2007) Host fidelity of the pollinator guilds of *Silene dioica* and *Silene latifolia:* possible consequences for sympatric host race differentiation of a vectored plant disease. *International Journal of Plant Sciences*, **168**, 421–434.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Refrégier G, Hood ME, Giraud T (2010) No evidence of reproductive character displacement between two sister fungal species causing anther smut disease in *Silene*. *International Journal of Plant Sciences*, **171**, 847–859.

Rödelsperger C, Neher RA, Weller AM *et al.* (2014) Characterization of genetic diversity in the nematode *Pristionchus pacificus* from population-scale resequencing data. *Genetics*, **196**, 1153–1165.

Rouxel T, Grandaubert J, Hane JK *et al.* (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature Communications*, **2**, 202.

Rubin C-J, Zody MC, Eriksson J *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, **464**, 587–591.

Rubin C-J, Megens H-J, Barrio AM *et al.* (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19529–19536.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genetics*, **7**, e1001302.

Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution*, **27**, 1813–1821.

Stajich JE, Berbee ML, Blackwell M *et al.* (2009) The fungi. *Current Biology*, **19**, R840–R845.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1245–1253.

Stergiopoulos I, de Wit PJGM (2009) Fungal effector proteins. *Annual Review of Phytopathology*, **47**, 233–263.

Stukenbrock EH, Croll D (2014) The evolving fungal genome. *Fungal Biology Reviews*, **28**, 1–12.

Stukenbrock EH, McDonald BA (2009) Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Molecular Plant-Microbe Interactions*, **22**, 371–380.

Svetec N, Pavlidis P, Stephan W (2009) Recent strong positive selection on *Drosophila melanogaster* HDAC6, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. *Molecular Biology and Evolution*, **26**, 1549–1556.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tellier A, Moreno-Gámez S, Stephan W (2014) Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. *Evolution*, **68**, 2211–2224.

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9979–9986.

Tiley GP, Burleigh G (2015) The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology*, **15**, 194.

Udpa N, Ronen R, Zhou D *et al.* (2014) Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biology*, **15**, R36.

Van Putten WF, Biere A, Van Damme JMM, May G (2003) Intraspecific competition and mating between fungal strains of the anther smut *Microbotryum violaceum* from the host plants *Silene latifolia* and *S. dioica*. *Evolution*, **57**, 766–776.

Vercken E, Fontaine MC, Gladieux P *et al.* (2010) Glacial refugia in pathogens: European genetic structure of anther smut pathogens on *Silene latifolia* and *Silene dioica*. *PLoS Pathogens*, **6**, e1001229.

de Vienne DM, Refrégier G, Hood ME *et al.* (2009) Hybrid sterility and inviability in the parasitic fungal species complex *Microbotryum*. *Journal of Evolutionary Biology*, **22**, 683–698.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Win J, Chaparro-Garcia A, Belhaj K *et al.* (2012) Effector biology of plant-associated organisms: concepts and perspectives. *Cold Spring Harbor Symposia on Quantitative Biology*, **77**, 235–247.

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

Zeng K, Fu Y-X, Shi S, Wu C-I (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.

T.G. designed the study and supervised the study; H.B., A.S. and S.L.P. performed the experiments; H.B., J.G., S.S., G.A., A.B. and P.G. analysed the data; H.B. and T.G. wrote the manuscript with contributions by all other authors. T.G. and J.G. contributed to the funding of the study.

Data accessibility

Numbers for public database accession on NCBI Sequence Read Archive: SRR2428492 to SRR2428557, Bioproject PRJNA295022.

The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to tatiana.giraud@upsud.fr.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Nucleotide diversity (π) and F_{ST} along the genome in *Microbotryum lychnidis-dioicae* (MvSl) clusters and in *Microbotryum silenes-dioicae* (MvSd).

Fig. S2 Discriminant analysis of principal components (DAPC) in the anther-smut fungi *Microbotryum lychnidis-dioicae* (MvSl) and *Microbotryum silenes-dioicae* (MvSd).

Fig. S3 Population genetic structure in *Microbotryum lychnidis-dioicae* (MvSl) for $K = 3$.

Fig. S4 Best demographic models fitted to (a) *Microbotryum silenes-dioicae* (MvSd), and (b) the three lineages within *Microbotryum lychnidis-dioicae* (MvSl).

Fig. S5 Comparison of the observed joint allele frequency spectrum with that expected under the demographic model fitted using dadi, and corresponding residuals.

Fig. S6 Derived site-frequency spectrum.

Fig. S7 Per-gene distributions of π (normalized per kb) and Tajima's *D* in each of the two species (*Microbotryum silenes-dioicae*, MvSd, and *Microbotryum lychnidis-dioicae*, MvSl) and in each of the three MvSl clusters (northwestern, southern and eastern clusters).

Fig. S8 Distribution of the neutrality index (NI).

Fig. S9 Decay of linkage disequilibrium (r^2) with physical distance in the northwestern cluster of *Microbotryum lychnidis-dioicae* (MvSl) and in *M. silenes-dioicae* (MvSd).

Fig. S10 Distribution of dN/dS across genes (ratio of number of fixed non-synonymous differences between MvSl and MvSd per non-synonymous site over number of fixed synonymous differences per synonymous site).

Table S1 Samples used in the study, with ID, fungal species name, host species name, location of collection, sequence accession number and names of collectors.

Table S2 Statistics of mapping: total number of reads, number of reads mapped as proper pairs and proportion of reads mapped as proper pairs.

Table S3 Number of single nucleotide polymorphisms (SNPs) per strain, percentage of heterozygous SNPs, ploidy, mean and median coverage of heterozygous and homozygous SNPs.

Table S4 Number of shared, exclusive or fixed single nucleotide polymorphisms (SNPs) between clusters of *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S5 Orientation of SNPs.

Table S6 Comparison of demographic models for *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S7 Recombination rates obtained using the software LDHat for each scaffold in *M. silenes-dioicae* (MvSd) and in the three clusters of *Microbotryum lychnidis-dioicae* (MvSl).

Table S8 Pairwise F_{ST} and number of fixed differences per kb between clusters of *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S9 Composition and polymorphism of clusters of *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S10 Parameter estimates and their standard deviation (SD) of demographic models fitted to the data using DADI for *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S11 Summary statistics of genomic regions significant for the composite likelihood ratio (CLR) test in *Microbotryum lychnidis-dioicae* (MvSl) and *M. silenes-dioicae* (MvSd).

Table S12 Coordinates of genomic regions significant for the composite likelihood ratio (CLR) test in *Microbotryum lychnidis-dioicae* (MvSl), and number of genes totally included in each region.

Table S13 Coordinates of genomic regions significant for the composite likelihood ratio (CLR) test in *Microbotryum silenes-dioicae* (MvSd), and number of genes totally included in each region.

Table S14 False discovery rate for one-sided Fisher's exact tests of enrichment of specific gene categories in the 0.05 quantile of the neutrality index (NI).

Table S15 Genes up-regulated *in planta* with neutrality index (NI) values within the 0.05 quantile between *Microbotryum silenes-dioicae* (MvSd) and in *M. lychnidis-dioicae* (MvSl) clusters.

Table S16 Summary statistics on CDS in *Microbotryum lychnidis-dioicae* (MvSl) clusters and in *M. silenes-dioicae* (MvSd).

Table S17 Comparison of numbers of non-synonymous over synonymous substitutions (mean d*N*/d*S*) in candidate gene sets vs. the remaining gene sets.

Table S18 Comparison of mean of the proportions of non-synonymous over synonymous polymorphisms (p*N*/p*S*) in candidate gene sets vs. the remaining gene sets.

Table S19 Correlation between nucleotidic diversity in *Microbotryum lychnidis-dioicae* (MvSl) clusters and in *M. silenes-dioicae* (MvSd).

Table S20 Multiple linear regressions models for explaining total diversity in the clusters of *Microbotryum lychnidis-dioicae* (MvSl) and in *M. silenes-dioicae* (MvSd).

Table S21 Multiple linear regressions models for explaining synonymous diversity in the clusters of *Microbotryum lychnidis-dioicae* (MvSl) and in *M. silenes-dioicae* (MvSd)

Table S22 Spearman's rho and significance level for correlation of nucleotidic diversity (π/kp) and the number of fixed differences (per kp) in 100 kb non-overlapping sliding windows.

Appendix S1 Supplementary text.