MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 292–306 doi: 10.1111/j.1365-294X.2009.04454.x

# Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens

G. AGUILETA,*†[1] J. LENGELLE,†[1] S. MARTHEY,† H. CHIAPELLO,† F. RODOLPHE,† A. GENDRAULT,† R. YOCKTENG,‡ E. VERCKEN,* B. DEVIER,* M. C. FONTAINE,* P. WINCKER,§ C. DOSSAT,§ C. CRUAUD,§ A. COULOUX§ and T. GIRAUD*

*Ecologie, Systématique et Evolution, Université Paris-Sud, F-91405 Orsay cedex; and CNRS F-91405 Orsay cedex, France, †MIG - INRA UR1077 Domaine de Vilvert 78352 Jouy en Josas Cedex France, ‡UMR 7205, CNRS-MNHN, Origine, Structure et Evolution de la Biodiversité, Département Systématique et Evolution, 16 rue Buffon CP 39, 75005, Paris, France, §Genoscope, Centre National de Sequençage, UMR CNRS 8030, 2 Gaston Crémieux, CP 5706, 91507 Evry, France*

## Abstract

**Numerous genes in diverse organisms have been shown to be under positive selection, especially genes involved in reproduction, adaptation to contrasting environments, hybrid inviability, and host-pathogen interactions. Looking for genes under positive selection in pathogens has been a priority in efforts to investigate coevolution dynamics and to develop vaccines or drugs. To elucidate the functions involved in host specialization, here we aimed at identifying candidate sequences that could have evolved under positive selection among closely related pathogens specialized on different hosts. For this goal, we sequenced *c.* 17 000–32 000 ESTs from each of four *Microbotryum* species, which are fungal pathogens responsible for anther smut disease on host plants in the Caryophyllaceae. Forty-two of the 372 predicted orthologous genes showed significant signal of positive selection, which represents a good number of candidate genes for further investigation. Sequencing 16 of these genes in 9 additional *Microbotryum* species confirmed that they have indeed been rapidly evolving in the pathogen species specialized on different hosts. The genes showing significant signals of positive selection were putatively involved in nutrient uptake from the host, secondary metabolite synthesis and secretion, respiration under stressful conditions and stress response, hyphal growth and differentiation, and regulation of expression by other genes. Many of these genes had transmembrane domains and may therefore also be involved in pathogen recognition by the host. Our approach thus revealed fruitful and should be feasible for many non-model organisms for which candidate genes for diversifying selection are needed.**

*Keywords*: adaptation, coevolution, *Dianthus*, genomics, Phylogenetic analysis by maximum likelihood, pathogenic fungi, *Silene*, *Ustilago*

*Received 15 July 2009; revision received 4 November 2009; accepted 8 November 2009*

## Introduction

Understanding the processes of speciation represents a central challenge in evolutionary biology, and it has immediate and direct implications to broader societal interests concerning biodiversity. The mechanisms of speciation are particularly relevant to studies of pathogens, where the diversification frequently involves adaptation to new hosts and there is great importance in the context of disease emergence in humans, livestock and crops. The genetic architecture of species differences has just begun to re-emerge with great promise

Correspondence: Tatiana Giraud, Fax: 00 33 16915 4697; E-mail: Tatiana.Giraud@u-psud.fr

[1] These authors contributed equally to this work.

from the era of comparative genomics (Wu & Ting 2004; Noor & Feder 2006). Morphological and physiological differences between species that are the result of adaptation leave signatures of positive selection upon the genes responsible, and thus allow evolutionary analyses of divergence for a combination of genetic and phenotypic traits. The genes involved in specialization to different hosts, which likely also contribute to pathogen speciation, are expected to be fast-evolving and therefore to show signs of positive selection (reviewed in Aguileta *et al.* 2009). However, other types of genes are expected to be under positive selection and have an important role in the formation of pathogen species. 'Speciation genes' (i.e. genes contributing directly to hybrid inviability or sterility) have been detected in model organisms, such as Drosophila, and indeed show the signatures of positive selection (Wu & Ting 2004; Noor & Feder 2006). These are genes that have evolved rapidly between geographically separated incipient species, for whatever reason, and then cause genetic incompatibilities in hybrids upon secondary contact. Such 'speciation genes' would likely also occur in pathogens, as in any organism. Also, genes involved in mate recognition and contributing to prezygotic isolation have been shown to be under positive selection in some organisms (e.g. Lee *et al.* 1995; Swanson *et al.* 2001; Galindo *et al.* 2003).

Powerful statistical methods for describing the nature of selection on various types of genes have been recently developed (Yang & Bielawski 2000a; Nielsen 2005; Aguileta *et al.* 2009). The most widely used indicator of positive selection between species is an excess of nonsynonymous nucleotide substitutions, that alter amino-acid sequence, relative to the number of synonymous fixed differences and scaled to the number of such available sites. By this criterion, numerous genes in very diverse organisms have been shown to be under a positive selection pressure (e.g. Ting *et al.* 2000; Swanson *et al.* 2001; Presgraves *et al.* 2003; Nielsen *et al.* 2005; Aguileta *et al.* 2009). Such an approach, looking for signatures of selection on genes to find ecologically important functions, has been recently coined 'reverse ecology' (Fuga Li *et al.* 2008; McKay & Stinchcombe 2008).

Most studies focusing on positive selection in pathogens have targeted specific genes that were likely candidates due to their functional relevance, such as those coding for antigenic proteins or genes involved in drug resistance (for a review see Aguileta *et al.* 2009). More recently, studies have taken advantage of large sequence datasets to identify genes under positive selection without a priori candidates (e.g. Barrier *et al.* 2003; Nielsen *et al.* 2005), in particular in pathogens (Chen *et al.* 2006; Anisimova *et al.* 2007; Ge *et al.* 2008; King *et al.* 2008). These genomic scans have targeted closely related pathogens to detect genes involved in coevolution. Such blind approaches have, however, never been performed among closely related species that are specialized on different types of hosts, which should identify the functions of genes most frequently involved in specialization and speciation. The genes identified in this manner may not necessarily be those expected a priori, and could allow the formulation of new hypotheses about the role of certain molecular processes in virulence and specialization.

Here, we undertook comparative genomics to identify candidates for rapidly evolving genes among closely related species of *Microbotryum*, a fungal species complex responsible for the anther-smut disease on more than one hundred of Caryophyllaceae species (Thrall *et al.* 1993). In diseased plants, the anthers become filled with fungal spores rather than pollen, sterilizing the plant. The spores are spread to healthy plants by insect pollinators, thus representing a case of sexually transmitted disease in plants. *Microbotryum violaceum sensu lato* represents a species complex: isolates from different hosts species have recently been shown to constitute different sibling species (Le Gac *et al.* 2007a), highly specialized on specific host plants. *Microbotryum* species are isolated by strong post-mating barriers, such as hybrid inviability and sterility (Le Gac *et al.* 2007b; Sloan *et al.* 2008; de Vienne *et al.* 2009). A phylogenetic study has shown that speciation events have occurred mostly by host-shifts (Refrégier *et al.* 2008), and field studies have in fact revealed several cases of incipient host shifts in nature (Antonovics *et al.* 2002; Hood *et al.* 2003; Lopez-Villavicencio *et al.* 2005). *Microbotryum* has been widely recognized as an important model for host-pathogen dynamics and fungal genetics (Antonovics *et al.* 2002; Garber & Ruddat 2002; Hedges 2002; Martínez-Espinoza *et al.* 2002; Bernasconi *et al.* 2009). The fungus has no impact on human activities, but is related to major crop pathogens, including the rust and smut fungi (Devier *et al.* 2009).

Because the numerous sibling species in the *Microbotryum* complex are highly host-specific, and because hybrids show inviability and sterility, genes evolving under positive selection should be involved in coevolution, specialization and speciation on different host plants, but also adaptation to the abiotic conditions the host lives in, and adaptation to the microfloral community associated with the particular host. In order to identify candidate genes potentially involved in these processes in the *Microbotryum* complex, we sequenced *c.* 17 000–32 000 ESTs from each of four *Microbotryum* species, specialized on different host plants*: M. lychnidis-dioicae* on *Silene latifolia*, *M. lagerheimii* on *S. vulgaris*, *M. dianthorum* on *Dianthus carthusianorum*, and *M. violaceum*

s. str. on *S. nutans*. RNA was extracted in conditions of mating and of production of the infection stage (Yockteng *et al.* 2007). These conditions were thus designed to identify the genes that are likely involved in speciation and host specialization processes. We then assembled ESTs into contigs and built clusters of orthologs shared by at least three of the four species. Analyses aiming at detecting footprints of selection were performed on the alignments of the ortholog clusters using statistical models implemented in PAML (Yang 1997, 2007). To validate the *in silico* approach, we then analysed 16 of the genes with signals of positive selection in additional *Microbotryum* species, specialized on yet other plant species.

Our specific aims were thus: (i) to detect candidate genes with signals of positive selection among four sibling species of the pathogen *Microbotryum* complex, for further validation and investigation, (ii) if possible, annotate these candidate genes to gain insights into their function and assess whether they are putatively involved in virulence, specialization, and speciation.

## Material and methods

### *Strains, culture conditions, RNA isolation, cDNA library construction and sequencing*

Four strains were used: 100.02 of *M. lychnidis-dioicae,* collected from the host *S. latifolia* in 2001 in the Alps, near Tirano in Italy; 300.27 of *M. lagerheimi,* collected from the host *S. vulgaris* in 2003 near the Pic du Midi de Bigorre in the Pyrénées mountains in France; 309.05 of *M. dianthorum,* collected from the host *D. carthusianorum* in 2003 in the val d'Esquierry in the Pyrénées mountains in France; LB of *M. violaceum* s. str, collected from the host *S. nutans* in 2003 in the Jura mountains, in France. Teliospores from these strains were plated on GMB1 medium (Thomas *et al.* 2003). On such nutritive media, diploid teliospores germinate and produce haploid sporidia of the two mating type A1 and A2. A1 and A2 sporidia lines were identified by pairing with existing stocks of known mating type.

A mixed suspension of A1 and A2 sporidia (250 μL of each) was plated on water agar supplemented with α-tocopherol (10 IU/g) and incubated at 4 °C for 1 week. These conditions of low nutrients with α-tocopherol are thought to mimic the host plant surface for the fungus, because sporidia conjugate and produce hyphae of a few cells (Day & Garber 1988). This was checked using a light microscope (400X). RNA extraction and sequencing were performed as previously reported (Yockteng *et al.* 2007; Giraud *et al.* 2008), except that only the library built from *M. lychnidis-dioicae* was normalized.

### *Sequence cleaning, assembly and annotation*

Sequences are available in Genbank (accessions numbers CU368079-CU457702, FP567975-FP578296). Raw sequence data were cleaned from vector and adaptor sequences. Contaminating plasmid sequences, such as *E. coli*, were removed from the analyses. The SURF (SeqUence Repository and Feature detection) package (Lannuccelli 2005) was used for sequence base-calling, cleaning, and for detection of any contamination in putative inserts. This analysis involved three steps including the use of PHRED (Ewing & Green 1998; Ewing *et al.* 1998), which detected bad quality regions, of RepeatMasker, which masked low complexity regions, and of Crossmatch, which found putative contaminated sequences of the UNIVEC database and single nucleotide repetitions. Only sequences with a SURF score over 20 on at least 100 bp were released in the EST division of the EMBL-EBI Nucleotide Sequence Database.

Expressed sequence tags were aligned and assembled into contigs (Huang & Madan 1999) when the criterion of a minimum identity of 95% over 50 bp was met. When an EST could not be assembled with others in a contig, it was retained as a 'singlet'. The contigs and the singlets should thus correspond to sequences of unique genes, and will be called hereafter 'unisequences'.

The consensus sequences of the contigs and the sequences of the singlets were compared to the sequences in the GenBank database and in the Uniprot database (Consortium 2007) using the tBLASTx and the BLASTx algorithms (Altschul *et al.* 1997). Unisequences showing significant similarity ($E$-value ≤ $10^{-4}$) to database entries were annotated using their most significant match. Unisequences were also classified into Gene Ontology functional categories (http://www.geneontology.org) based on BLAST similarities to known genes of the NCBI nr (non-redundant) protein database and using the Blast2GO annotation tool (Conesa *et al.* 2005).

Finally, a modified version of the ESTIMA tool (Kumar *et al.* 2004), was used to develop a public database named MICROBASE, dedicated to *Microbotryum violaceum* EST management and analysis. This database includes information on EST sequences, contigs, annotations, gene ontology functional categories and search programs to compare similarities of any sequence against the database. MICROBASE is accessible freely through a web interface at the URL http://genome.jouy.inra.fr/microbase. We previously reported the EST library built from *M. lychnidis-dioicae* collected from the host *S. latifolia* (Yockteng *et al.* 2007). The libraries from the three other *Microbotryum* species are now available via MICROBASE.

### *Unisequence CDS predictions and clustering*

The pipeline of the Prot4EST software (Wasmuth & Blaxter 2004) was used to predict unisequence CDS positions and to translate coding regions into protein sequences. In a first step, all putative ribosomal sequences in the dataset were identified through a BLASTn search against the rRNA sequence database (Ribosomal Database II) and the sequences whose BLASTn *E*-value ≥1e-65 were discarded. The second and third steps use the BLASTx algorithm to detect any similarity between unisequences and sequences from both the mitochondrial protein database (NCBI ftp site) and the Uniprot database (Consortium 2007). Unisequences showing a significant BLAST result (i.e. cutoff of e-08) against the mitochondrial database were annotated as mitochondrial genes to be translated subsequently with the relevant genetic code. We removed all sequences corresponding to transposable elements, as well as those containing internal stop codons. Sequences that matched the Uniprot-Swissprot database significantly (i.e. cutoff of e-08) were considered as CDS and a HSP tile path was constructed. This means that Prot4EST then considers that the nascent translation of these sequences can be extended at either end in the same reading frame. Only sequences that yielded no sequence similarity were then submitted to the fourth step of the pipeline which aims at identifying coding regions using hidden Markov models implemented in the ESTscan software (Iseli *et al*. 1999). For this step, a transition matrix was created from the genome sequence of *Ustilago maydis*, a related Basidiomycete fungus, as well as the CDS EMBL-EBI file that is also available for the same species. Predicted polypeptides satisfying a given length threshold criteria (CDS of at least 30 codons in length and covering at least 10% of the input sequence) then underwent the extension process (like for HSP tiling). In a fifth step, the DECODER program (Fukunishi & Hayashizaki 2001) is used to predict the CDS and polypeptide translations for the remaining sequences. DECODER exploits the quality scores of the sequences produced from base calling software (such as PHRED used in the SURF package) and additional text base information (such as optimal codon usage). DECODER computes a likelihood score for each possible CDS, and the one with the lowest score is chosen as the correct CDS. Finally, a last attempt is performed to provide a putative polypeptide translation based on the longest string of amino acids uninterrupted by stop codons from a six-frame translation of the sequence. In spite of all the caution exerted at the previously described stages for predicting CDSs, potential problems could arise from the incorporation of an intron that does not destroy the reading frame. In this case, the predicted CDS would contain a region of neutral sequence that could bias analyses of selective pressure. This phenomenon however should not have been a problem here as gaps were removed from alignments.

Clustering of unisequence CDSs into groups of orthologs included three main steps for which we employed custom-made Perl scripts. The first step detected the single-copy unisequences from each CDS library. To do this, each library was aligned against itself by using a BLASTn algorithm. All CDS sequences having exactly one significant hit (e-value <1e-10) were considered as single-copy unisequences (thus avoiding hidden paralogy) and were kept for the detection of orthologs among the four libraries, using a derivative of the Best Bidirectional Hit for n sequences. The single-copy unisequences from all four libraries were combined in a single file and aligned against themselves using the BLASTn algorithm. All single-copy unisequences with a hit (e-value <0.1) with a unisequence of another library were considered to have an ortholog in the corresponding species and were therefore kept for the last step. The last script compiled alignment results and built clusters of putative orthologs, including either 3 or 4 sequences, each of them belonging to a different species. Although very unlikely, it is not impossible that chimeric sequences were present among the single-copy unisequences, as CAP3 may merge into a single contig ESTs derived form two recently duplicated paralogs that differ at only a few sites.

### *Ortholog gene alignment, filtering and sorting by alignment length*

The predicted protein unisequences of orthologs were aligned using T-coffee (Notredame *et al.* 2000) with default settings. The corresponding nucleotide alignments were performed by using the protein alignments as guide, as implemented in the tranalign program of the EMBOSS package (http://embossgui.sourceforge.net/demo/manual/tranalign.html). In order to keep only reliable alignments, which are crucial for the subsequent detection of selection, the alignments were then filtered using different criteria. First, we required a level of protein sequence identity of at least 70% for all alignments of putative orthologs. Second, the alignments were post-processed to remove gaps and keep only unambiguously aligned blocks of sequence. This step was performed using Gblocks (Castresana 2000) with the maximum number of non-conserved positions set to 8, and the minimal block size set to 5 (for all other parameters, default settings were used). Finally, we used the length of the final alignments to classify the resulting ortholog clusters for subsequent analysis: clusters with at least 300 nucleotides were analyzed individually, and those whose length was less than 300 nucleotides were concatenated, as they were too short to be analyzed individually without risking stochastic sampling errors.

### *Detection of positive selection*

Positive selection was tested using the codeml program of the PAML4 package (Yang 2007). Selective pressure was measured by using the nonsynonymous/synonymous substitution rate ratio ($d_N/d_S$), also referred to as ω. An $\omega < 1$ suggests purifying selection, $\omega = 1$ is consistent with neutral evolution, and $\omega > 1$ is indicative of positive selection (Yang & Bielawski 2000b). Nested codon models implementing the ω ratio can be compared by means of a likelihood ratio test (LRT) (Anisimova *et al.* 2001). We used the null model M1a, which assumes two site classes with $1 > \omega_0 > 0$, and $\omega_1 = 1$, which therefore implicitly supposed that no site is under positive selection, and compared it with the alternative model M2a, which adds an extra class of sites that allows ω to take values >1. We also compared the null model M7, which assumes a beta distribution of ω across sites, with the alternative model M8, which adds an extra class of sites to M7 where ω can take values >1. Thereby positive selection can be detected if a model allowing for positive selection is significantly more likely (as estimated by the LRT) than a null model without positive selection. Because we analyzed different partitions of the same data we corrected for multiple testing using the Bonferroni correction method (Anisimova & Yang 2007). These corrections were done taking into account the number of tests conducted to analyze the same data.

### *Functional annotation*

In order to assign functional annotation to cluster orthologs exhibiting evidence for positive selection, we first used the available MICROBASE annotation of individual unisequences contained in each cluster. As described previously, this annotation was obtained from the BLASTp best hit and the corresponding GO terms.

We then performed three complementary analyses in order to collect maximal functional information contained in the clusters of interest. In a first step, we looked for all possible motifs, signal and domains in the individual sequences using the Interproscan software (assuming default settings) of the Interpro database (Mulder *et al.* 2007). In the second step and third step, we tried to identify distant homologs for each individual sequences using two complementary methods: (i) the PSI-BLAST (Position-Specific Iterated BLAST) algorithm (Altschul *et al.* 1997), which is an efficient method to detect weak but biologically relevant sequence similarities against the Uniprot database, and (ii) the FROST algorithm (Marin *et al.* 2002), a recent threading method allowing to detect three dimensional structure similarities with protein structures described in the PDB (Berman *et al.* 2000). Fold recognition methods such as FROST are often used when sequence or profile alignment methods such as PSI-BLAST are unable to find similar sequences.

### *Sequences of 16 candidate genes in other* Microbotryum *species*

We chose 16 genes among those with signals of positive selection, based on their function or their elevated ω ratio, and we sequenced them in several other *Microbotryum* species, specialized on yet other host species. We designed primers based on the cluster alignments of the EST libraries (Clusters chosen, species analysed and primers used are given in Table S9). Alignments were analyzed for positive selection as described above. We used in addition codeml to compare a model allowing a single $d_N/d_S$ ratio across all branches of the tree with a model that allows each branch to have a different $d_N/d_S$ ratio. This test, known as free-ratios test is not a formal test for ratio variation among branches; however, it gave an approximate idea about possible lineages where the $d_N/d_S$ ratio has increased relative to the rest of the tree. We also conducted a branch-site analysis, which is a more powerful test that evaluates whether there have been positive selection events at specific sites in particular lineages.

## Results

### *Sequence analysis: gene finding, assembly and clustering*

Figure 1 shows a diagram of the pipeline we implemented for the extraction, identification and clustering of ESTs. Four EST libraries were processed, one for each *Microbotryum* species. Following the sequence cleaning and assembly steps, unisequences (i.e. contigs or singlets corresponding to sequences of unique genes) were retrieved and assembled for each species (Table 1). Coding sequences (CDS) were predicted to obtain the coding frame required for detecting synonymous and nonsynonymous substitutions (Table 1). Comparisons between the four libraries yielded 372 clusters of orthologs, including sequences from three or four species, 53 of which being at least 300 nucleotides long. This fairly low number of clusters of orthologs stems from our stringent parameters to avoid hidden paralogy and wrong coding frames. Many sequences had to be

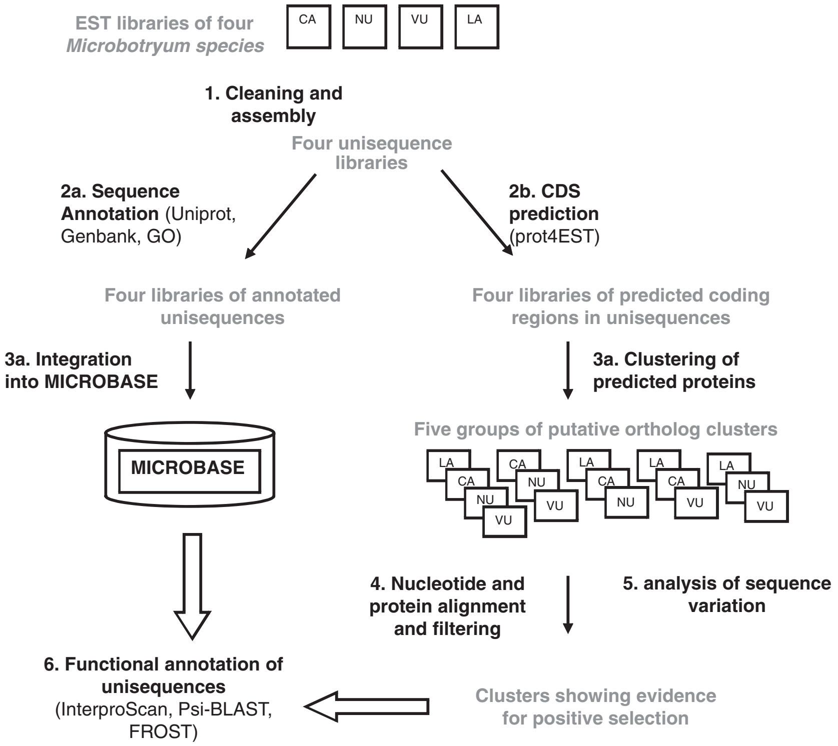
EST libraries of four *Microbotryum species*
CA
NU
VU
LA
1. Cleaning and assembly
Four unisequence libraries
2a. Sequence Annotation (Uniprot, Genbank, GO)
2b. CDS prediction (prot4EST)
Four libraries of annotated unisequences
Four libraries of predicted coding regions in unisequences
3a. Integration into MICROBASE
3a. Clustering of predicted proteins
MICROBASE
Five groups of putative ortholog clusters
LA CA NU VU
4. Nucleotide and protein alignment and filtering
5. analysis of sequence variation
6. Functional annotation of unisequences (InterproScan, Psi-BLAST, FROST)
Clusters showing evidence for positive selection

**Fig. 1** Pipeline used for sequence analysis of EST libraries and identification of genes under positive selection. Abbreviations for species name are *CA: for EST collected from the *Microbotryum* species infecting the host *Dianthus carthusianorum*, *LA: for EST collected from the *Microbotryum* species infecting the host *Silene latifolia*, *NU: for EST collected from the *Microbotryum* species infecting the host *Silene nutans*, *VU: for EST collected from the *Microbotryum* species infecting the host *Silene vulgaris*. Abbreviations for cluster names are concatenations of the abbreviations above.

**Table 1** Results obtained from the sequence analysis of the four *Microbotryum* EST libraries

| EST library | *Microbotryum* on *D. carthusianorum* (CA) | *Microbotryum* on *S. latifolia* (LA) | *Microbotryum* on *S. nutans* (NU) | *Microbotryum* on *S. vulgaris* (VU) |
|---|---|---|---|---|
| Number of EST | 32 233 | 24 126 | 28 446 | 16 896 |
| Number of singlets | 2319 | 4178 | 3316 | 2015 |
| Number of contigs | 2874 | 3587 | 3762 | 3095 |
| Number of unisequences | 5193 | 7765 | 7078 | 5110 |
| Number of CDSs | 4905 | 7409 | 6595 | 4844 |
| Mean CDS length (bp) | 249 | 265 | 429 | 321 |
| Total length (bp) | 1 221 002 | 1 960 417 | 2 830 111 | 1 553 075 |

dropped to avoid misalignment. Clusters of orthologs including sequences from at least three of the four species provides five species groups. These species groups were named according to the combination of host-specific species of *Microbotryum* as follows: LA stands for *Microbotryum* from *Silene latifolia*, CA from *Dianthus carthusianorum*, NU from *Silene nutans* and VU from *Silene vulgaris*. Thus there was one species group called LACANUVU, which contained clusters with orthologous sequences for all four species, and four groups contained sequences from three species, respectively called LACANU, CANUVU, LACAVU and LANUVU (Table 2).

### *Detection of positive selection*

Positive selection is detected when a model of evolution allowing for positive selection appears significantly more likely than a null model without positive selection, as indicated by likelihood ratio tests (LRTs). For the detection of positive selection, the clusters were analyzed in different data partitions (Fig. 2), depending on

**Table 2** Number of clusters of putative orthologs according to the five groups of species (same nomenclature as in Fig. 1)

| Species group | All the clusters | Small clusters | Long clusters |
|---|---|---|---|
| LACANUVU | 45 | 40 | 5 |
| CANUVU | 60 | 49 | 11 |
| LACANU | 103 | 83 | 20 |
| LACAVU | 67 | 62 | 5 |
| LANUVU | 97 | 85 | 12 |
| Total | 372 | 319 | 53 |

the length of the aligned cluster sequences: (i) The first partition thus included the individual 53 clusters whose alignment was at least 300 bp; (ii) the second partition consisted of the concatenation of the 53 long clusters; (iii) the third partition included the 319 clusters whose individual alignments were shorter than 300 bp; (iv) and a final partition included the concatenation of the complete dataset of 372 alignments (Table 2). The clusters whose alignment was shorter than 300 bp had to be concatenated to avoid stochastic sampling errors (i.e. sample size being too small; Anisimova *et al.* 2001).

The individual analysis of the 53 ortholog clusters that were long enough to be analyzed individually indicated that seven clusters showed significant signal of positive selection, distributed among the species groups LACANU, LACAVU and LANUVU (see Tables S1A and B for codeml parameter estimates and LRTs). Most of the 53 longest clusters had very low $d_N$ to $d_S$ ratios, and were thus highly conserved among species (Fig. 3): the ω ratio showed high variation among the 53 clusters, but most of them had ω values between 0 and 0.30, well below 1.

The second data partition included the concatenations by species group of the 53 alignments analyzed above individually. This was done to investigate the effect of concatenation on the detection of positive selection in order to proceed with analysis based on concatenation of sequences shorter than 300 bp. Only two species groups, LACAVU and LANUVU, showed evidence for a significantly accelerated nonsynonymous rate relative to the rate of synonymous substitutions (see Tables S2A and B for codeml parameter estimates and LRTs). Interestingly, the analysis of the concatenated alignment of the species group LACANU did not detect any sites under positive selection, while the individual analysis

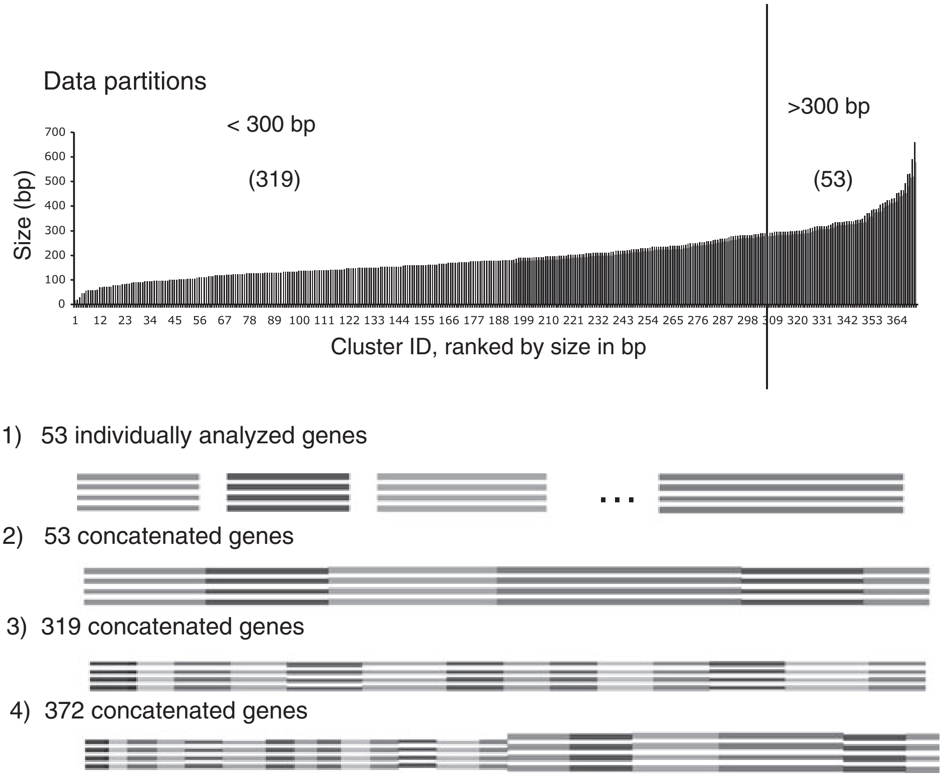

**Fig. 2** Gene cluster distribution, ranked by alignment size in bp, and illustration of the four data partitions analyzed. At the left part of the histogram are the 319 genes whose alignments are shorter than 300 bp, at the right part are the 53 genes whose alignments are longer than 300 bp.

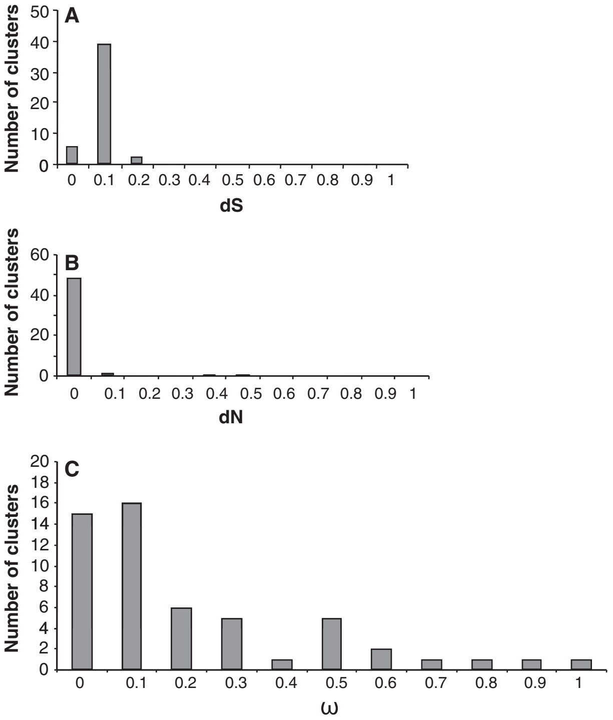

**Fig. 3** $d_N$, $d_S$, ω value distribution for the 53 longest clusters (individual analysis); $d_N$ (A) denotes the number of non synonymous substitutions over the total number of non synonymous substitutions and $d_S$ (B) the number of synonymous substitutions over the total number of synonymous substitutions; ω (C) is defined as the $d_N/d_S$ ratio.

of the same data showed two different clusters with evidence of positive selection. Nevertheless, the sites found to be under positive selection in the concatenated analysis were included in the significant sites from the individual analysis. In both cases, LRTs remained statistically significant even after correction for multiple testing in the concatenated analysis.

The third data partition involved the concatenation of the 319 clusters whose alignments were less than 300 nucleotides long, which were therefore not suitable for individual analysis. A single concatenated alignment was built for each of the five species groups. We found evidence for positive selection in all five species groups (Table S3). A total of 27 genes were found to exhibit sites under positive selection among the 319 concatenated.

Finally, in order to further test the effect of concatenation, we analysed a fourth data partition, which consisted in the concatenation of all 372 clusters. In this case, positive selection was detected in only four of the five species groups, the signal being lost in the LACANU species group (Table S4). A total of 21 genes were found to exhibit sites under positive selection among the 372 tested. Table S12 indicates the overlapping of clusters with significant signal of positive selection between the different data partitions analyzed.

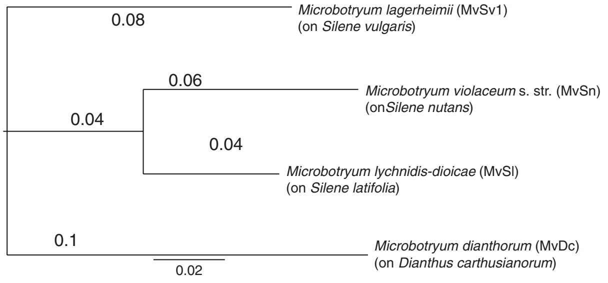
0.08
*Microbotryum lagerheimii* (MvSv1)
(on *Silene vulgaris*)
0.06
*Microbotryum violaceum* s. str. (MvSn)
(on*Silene nutans*)
0.04
0.04
*Microbotryum lychnidis-dioicae* (MvSl)
(on *Silene latifolia*)
0.1
*Microbotryum dianthorum* (MvDc)
(on *Dianthus carthusianorum*)
0.02

**Fig. 4** Phylogenetic tree of the four species analyzed with estimates of sysnonymous substitution rate (d*S*). The tree was built from the concatenation of the alignments of the 41 orthologs detected in all the four species. Branch lengths are estimated based on d*S* substitution rates, which are an indication of among-species divergence. Names of *Microbotyum* species names are in indicated, as in Lutz *et al.* (2005) and in brackets as in Refrégier *et al.* (2008), as well as names of host plants.

In general, we observed a low divergence level among the four species analyzed, the average d*S* values ranging between 0.1 and 0.04. We built a phylogenetic tree from the concatenation of the 43 genes shared by all four species, estimated the rate of synonymous substitutions observed along each branch on the tree and used this measure as an approximate measure of divergence between species (Fig. 4). Given the few changes that have accumulated among species, it is unlikely that our estimates for positive selection are affected by saturated synonymous substitution rates. In spite of the low divergence level, we obtained significant evidence for positive selection after LRTs in 42 clusters of orthologs out of the 372 putative orthologs present in at least three *Microbotryum* species. However, it is important to note that the high number of genes detected as being under positive selection should be interpreted with caution, as it may be inflated by false positives mainly due to the small size of the sample analyzed (e.g. four closely related species). Because our aim was to produce a list of candidate genes, we decided to keep all the 42 genes detected initially for experimental validation in future studies. Table 3 shows a summary of the evidence of positive selection and the level of sequence divergence in each of the concatenated alignments. In general, these comparisons suggest that codeml is less powerful in detecting sites under positive selection with concatenated datasets, as has been previously suggested (Anisimova *et al.* 2001).

We then tested whether codeml was more powerful if the concatenation was done only between genes having similar evolutionary rates. We tested this idea by grouping the 53 longest clusters in three classes according to the branch lengths inferred by codeml. Clusters

**Table 3** Summary of the results for the detection of positive selection in the different data partitions

| Species group | Number of orthologs under positive selection in the 53 long clusters analysed individually | Detection of positive selection and mean sequence identity in the 53 clusters analysed by concatenation | Detection of positive selection and mean sequence identity in the 319 short clusters concatenated | Detection of positive selection and mean sequence identity in the 372 clusters concatenated |
|---|---|---|---|---|
| LACANUVU | 0 | No/98% | Yes/96% | Yes/96% |
| CANUVU | 0 | No/97% | Yes/94% | Yes/95% |
| LACANU | 2 | No/96% | Yes/95% | No/95% |
| LACAVU | 2 | Yes/86% | Yes/94% | Yes/92% |
| LANUVU | 3 | Yes/96% | Yes/95% | Yes/95% |
| Total | 7 | | | |

were then concatenated according to their group species and class of branch length. Interestingly, codeml recovered a significant signal in the LACANU group, as in the individual alignment analysis, whereas it was lost in the total concatenation of the 53 clusters.

### *Functional annotation of the 42 putative genes detected as being under positive selection*

We performed the functional annotation of the 42 genes exhibiting evidence of positive selection. In a first step, we assigned functional categories to genes using both the BLASTp best hit obtained for each of the coding sequences included in the clusters (Table S5). The 42 genes were classified into 6 GO (Gene Ontology) categories according to their function: (i) regulation of gene expression (7 clusters), (ii) respiratory and energy metabolism (5 clusters), (iii) protein degradation and (4) protein secretion (1 cluster), (5) cellular development (6 clusters) and (6) unknown function (23 clusters) (Fig. 5). We compared the proportions of genes in the different GO classes in the whole set of orthologs and in the 42 genes under positive selection. Some GO classes appeared to include a higher proportion of genes under positive selection, as compared to the whole set of orthologs (especially the 'molecular function class'). However, the differences between the two distributions were not significant (Chi-square test, data not shown).

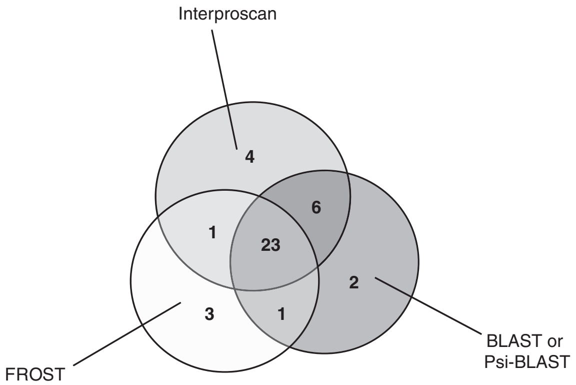

**Fig. 5** Number of clusters with annotation using three different methods. The number of clusters with annotation refers to those among the 42 clusters detected under positive selection. Annotations were obtained using the three different methods (BLAST or Psi-BLAST, INTERPROSCAN and FROST).

Because the first step using BLASTp did not yield annotations for all genes (Table S5, 23 clusters falling in the 'unknown function' GO class), we performed a second step of annotation, by using three complementary methods on each sequence of the 42 genes: (i) motif and domain prediction using the Interpro database (Table S6), (ii) PSI-BLAST comparisons (Table S7), allowing detecting weak but biologically relevant sequence similarities with the Uniprot database and (3) FROST, a recent threading method allowing to search a compatible fold in the PDB structure database (Table S8). Interestingly, these methods provided information on putative functions for genes that obtained no annotation using BLASTp (Fig. 5). Overall, owing to the use of multiple annotation methods, 40 of the total 42 clusters of putative interesting orthologs could be annotated. Among the putative functions of the genes showing significant signal of positive selection, many could be related to virulence and specialization (see discussion).

### *Validation of the blind* in silico *approach by sequencing candidate genes in additional* Microbotryum *species*

In order to validate our blind *in silico* approach for detecting genes under positive selection, we chose 16 genes among the 42 showing signs of positive selection, based on their function or their elevated ω ratio. We chose those that looked interesting based on these criteria even if the associated LRTs were not always significant. We sequenced the 16 genes in nine additional *Microbotryum* species, specialized on other host species (see Table S9 and Fig. 6). To have an independent sample, we looked for signals of positive selection in the alignments containing only these additional *Microbotryum* species. No mixed signals were detected in the chromatograms, which would have indicated hidden paralogs. The analysis using codeml detected sites under positive selection in 12 of these genes. LRTs were significant for only three genes (C1686, P237, P336), but this test is known to be conservative (Anisimova *et al.* 2001). We also investigated variation of $\omega$ among branches for these 16 genes using these data sets with all 13 *Microbotryum* species (the 4 initial species plus the 9 additional species). In 14 of the 16 genes, very high $\omega$ values were inferred in branches for at least one of the new *Microbotryum* species that we added, and often several species had elevated $d_N/d_S$ ratios (see Tables S9, S10 and S11 and an example in Fig. 6). Finally, the branch-site analysis confirmed that, although there were no branches or sites with $d_N/d_S > 1$, there was evidence of accelerated substitution rates along specific lineages (Fig. 6). These results yield strong support for the validity of our prediction of genes under positive selection, showing that genes detected to be under selection based on a blind *in silico* approach also show footprints of positive selection in other closely related pathogenic species.

## Discussion

We have sequenced a large number of ESTs obtained under conditions allowing the detection of candidate genes that are likely involved in the speciation processes of a complex of fungal species that have specialized in infecting a range of related plant hosts (Le Gac *et al.* 2007a,b). The analysis pipeline we have developed makes an efficient use of the information available for non-model species, without complete genomes available, and by using powerful standard methods for the detection of positive selection we have detected a number of proteins with potentially important functions that may have contributed to host specialization and promoted speciation.

Detection of positive selection using a site-based approach has some power limitations, especially when analyzing a few closely related species, as in our case. It is known that LRTs do not have much power when the analyzed sequences are short or have low divergence (Anisimova *et al.* 2001; Suzuki & Nei 2002; Mundy & Cook 2003). We have tried to be conservative in order to reduce the number of false positives in our site-based analyses. We have conducted two LRTs (M1a vs. M2a and M7 vs. M8) and have kept only those genes yielding significant LRTs in both tests. We have analyzed different partitions of the same data to check the convergence of results and made sure to correct for multiple tests of the same data. Overall, a total of 42 genes were detected as candidate genes to be evolving under positive selection. This number includes the short genes that were analyzed only in concatenated analyses. Aligning data from genes that evolve at different rates may however alter the positive selection signal and result in false positives. Nevertheless, these 42 genes represent good candidates, that will warrant and require further validation and investigation. Furthermore, the seven genes detected as being under positive selection by the individual analysis are very strong candidates for being functionally important.

Many of the genes identified as evolving under positive selection among the closely related *Microbotryum* species specialized on different hosts had putative functions that could be related to virulence and specialization. The most striking example is the cluster P350,

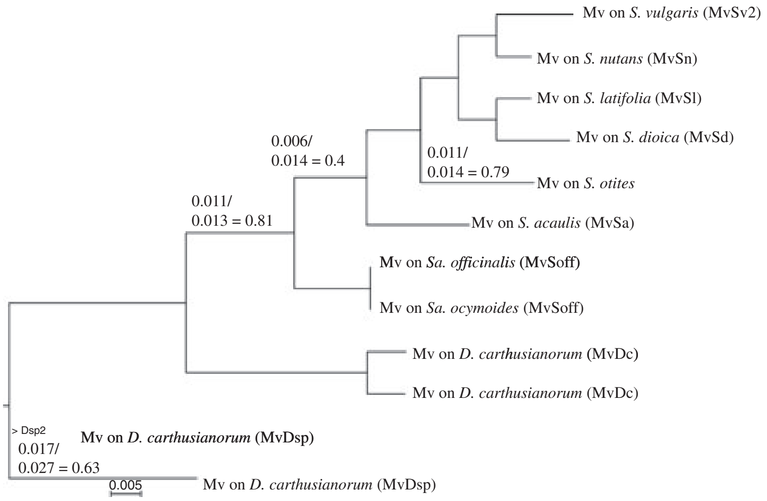
Mv on *S. vulgaris* (MvSv2)
Mv on *S. nutans* (MvSn)
Mv on *S. latifolia* (MvSl)
Mv on *S. dioica* (MvSd)
0.006/
0.014 = 0.4
0.011/
0.014 = 0.79
Mv on *S. otites*
0.011/
0.013 = 0.81
Mv on *S. acaulis* (MvSa)
Mv on *Sa. officinalis* (MvSoff)
Mv on *Sa. ocymoides* (MvSoff)
Mv on *D. carthusianorum* (MvDc)
Mv on *D. carthusianorum* (MvDc)
> Dsp2
Mv on *D. carthusianorum* (MvDsp)
0.017/
0.027 = 0.63
0.005
Mv on *D. carthusianorum* (MvDsp)

**Fig. 6** Example of a gene tree used to conduct the free-ratios test (P336). The free ration test allows each branch to have a different $d_N/d_S$ ratio (estimates with an accelerated $d_N/d_S$ ratio are indicated on the branches). Names of host plants are indicated, as well as *Microbotyum* species names after those in Refs. (Le Gac *et al.* 2007a; Refrégier *et al.* 2008).

similar to the CAP10 gene from the human pathogen fungus *Cryptococcus neoformans* that is required to produce its thick extracellular polysaccharide capsules, which are well recognized as virulence factors (Chang & Kwon-Chung 1999). P133 was similar to plasma membrane vacuolar type H+-ATPases (V-ATPases), which are proton pumps playing a key role in the physiology of fungi, controlling essential functions such as nutrient uptake (Portillo 2000). P306 was annotated as a Proteasome subunit, which are large protein complexes whose main function is to degrade by proteolysis damaged proteins or toxins (Lucyshyn *et al.* 2008), such as those that can be produced by host plants against pathogens. Several genes identified as evolving under positive selection (P1741, P146, P376, P1709, P1711) were putatively involved in respiration under stressful conditions, such as those that can be found within the host plant. The NADH-ubiquinone oxidoreductase for instance enables respiration to continue in the presence of inhibitors. A cAMP binding domain was also detected in a gene under positive selection, which is usually involved in stress response (Berman *et al.* 2005). Transporters were also well represented (sulphate transporters, ABC transporter, equilibrative nucleoside transporter) (P424, P244, P195, P177). These transmembrane proteins have an important role in the uptake of essential elements from the host plants (Lee & Cooksey 2000). Other genes appeared to be involved in the synthesis of secondary metabolites that play a role in pathogenicity. Monooxygenase (P146) for instance has been implicated in the production of secondary metabolites involved in plant infection (Hohn *et al.* 1993). GTP-binding proteins (P246, P1702, P119) are involved in the dimorphic switch from budding to hyphal growth in *U. maydis*, which are an essential step of infection in *Ustilago* and *Microbotryum* (Mahlert *et al.* 2006; Bohmer *et al.* 2007). P20, P821, P1137, P147, P19 are involved in transcription and DNA binding, meaning that the regulation of expression, in addition to the evolution of the proteins themselves, can be important for specialization, as suggested by transcriptomics and other approaches that look at the evolution of regulatory regions (Wong & Nielsen 2004; Hahn 2007; Egea *et al.* 2008).

In summary, many of the genes putatively evolving under positive selection among close species of fungal pathogens specialized on different host plants are thus putatively involved in regulation of other genes, nutriment uptake from the host plants, secondary metabolite synthesis and secretion, respiration under stressful conditions and stress response, and hyphal growth and differentiation. Many of these genes have transmembrane domains and may therefore contribute to the recognition of the pathogen by the host. These functions are indeed likely to be involved in coevolution with the host and especially specialization onto new hosts.

Codeml seemed to lose power to detect the signal of positive selection on concatenated alignments relative to the analysis of individual genes. This may stem from the differences in rates of evolution among genes. Indeed, codeml estimates the rates of synonymous and nonsynonymous substitutions to estimate branch lengths. After proper scaling, codeml compares, site by site, the two rates of substitutions in a ratio. Because it is based on a ratio, positive selection is relative. Therefore, to be detected as being under positive selection, a site must exhibit a higher number of nonsynonymous substitutions if the gene evolves globally faster (i.e. if branches are longer). When sequences from different genes are concatenated into a single alignment, branch lengths will be estimated as if the concatenated sequences were a single gene evolving under the same evolutionary rate. If evolutionary rates are heterogeneous in the concatenation, positive selection will likely be overestimated in fast evolving genes, whereas it will likely be underestimated in slow evolving ones, thus producing both false positives and false negatives.

Another important aspect in the detection of positive selection is the comparison of the models implemented in codeml by means of a LRT. As discussed in Anisimova *et al.* (2001), the LRT can be conservative depending on the sequence divergence level. The LRT compares two nested models (e.g. M1a vs. M2a, and M7 vs. M8), with a $\chi^2$ distribution, but under certain circumstances the assumption that the LRT statistic follows the $\chi^2$ distribution can make it a conservative test (Anisimova *et al.* 2001). In particular, the test can be too conservative if the sequences are short and highly similar, as is the case with the 319 cluster alignments that are less than 300 nucleotides long. All alignments showed high sequence identity among sequences, with values ranging between 92–98%. This high sequence similarity is expected to reduce the power of the LRT to find significant evidence for positive selection (Anisimova *et al.* 2001). Increasing the number of sequences under comparison can increase the power of the LRT, but in our case, at the time of the present study there were only four EST libraries available for *Microbotryum*. It is remarkable that even under conservative conditions (i.e. many of our sequences were short and some were highly similar), we were able to detect positive selection at several sites, as significant LRTs, precisely because they are stringent.

Our approach to detect genes under positive selection without a priori candidates among closely related pathogen species thus allowed getting good candidates for genes involved in specialization on different hosts. Such an approach, especially using only four species, can yield false positives, but we show that it yields a good number of candidate genes that will be used for further testing and validation. We performed a first validation by sequencing 16 of the genes detected to be evolving under positive selection in nine additional species. Signals of positive selection were detected in these genes for several other species than the four initially analyzed, confirming that the genes are rapidly evolving among the close *Microbotryum* species specialized on different hosts. The interests of our approach using ESTs include the lack of introns and pseudogenes in the sequences, both of which can create false signals in whole genome scans aiming at detecting positive selection. The drawbacks are that the sequences do not encompass the complete genes or all the genes from the genomes, but we showed that it nevertheless identified a number of good candidate genes to be under positive selection. Such an approach could therefore be very useful for non-model organisms, without complete genomes available. Also, because our approach used ESTs and not complete genomes, some of our clusters may have contained paralogs. Paralogs do not seem however not be a serious concern as we amplified 16 of the genes detected to be under positive selection in nine additional *Microbotryum* species, and found no evidence of paralog amplification. Indeed, no mixed signals were observed. Even if the positive selection detected in some clusters was due to recent diversification of multigene families instead of rapid evolution between orthologs, this would still be an interesting signal showing that these genes are rapidly evolving. Furthermore, paralogs are not likely to be widespread in *Microbotryum* as the RIP (Repeat-induced Point Mutation) mechanism seems very active (Hood *et al.* 2005). This defence mechanism against transposable elements mutates repeated sequences, which has been shown to impair duplication in other fungi (Dean *et al.* 2005).

In conclusion, our approach to detect genes likely evolving under positive selection without a priori candidates among closely related pathogen species provided valuable insights on possible genes involved in coevolution or specialization on different hosts for biotroph pathogens, some of which were not necessarily expected. It was for instance highly interesting that most of the genes putatively involved in host specialization had functions in hyphal growth, nutrient uptake, respiration under stressful conditions and regulation of other genes. The genes detected here will be subjected to experimental validation, (i) using additional sequencing, in other species and at the intraspecific level, (ii) using expression studies to assess whether they are differentially expressed in conditions of saprophytic growth vs. plant infection, (iii) using transformation studies, to explore the phenotypes of mutants having the candidate genes disrupted, (iv) by investigating whether the candidate genes experience a higher degree of divergence between close *Microbotryum* species than the rest of the genome, then suggesting that they have been involved in early differentiation, (v) by investigating if they colocalize with regions identified as involved in hybrid inviability and sterility using QTL approaches. The ESTs libraries from the four *Microbotryum* species will furthermore be an invaluable resource and for all scientists working on these species, and also for comparative genomic studies in fungi.

## Acknowledgements

We thank Michael E. Hood for comments on a previous version of the manuscript and for help all along the project. We also thank Jüergen Kroymann for comments on a previous version of the manuscript. We thank Annie Tellier, Manuela Lopez-Villavicencio and Alexis Matamoro for help with the sequences. We thank Yves Brygoo and Elisabeth Fournier for discussions when initiating the project. We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing help and computational resources. This work was funded by the grants ANR-06-BLAN-0201 and ANR 07-BDIV-003, by a post-doctoral grant from the French Ile-de-France Region, and by the 'Consortium National de Recherche en Génomique' for sequencing the cDNA libraries. The sequencing of genes under positive selection in other species was also supported by the 'Consortium National de Recherche en Génomique', and the 'service de systématique moléculaire' of the Muséum National d'Histoire Naturelle (CNRS IFR 101). It is part of the agreement no. 2005/67 between the Genoscope and the Muséum National d'Histoire Naturelle on the project 'Macrophylogeny of life' directed by Guillaume Lecointre.

## References

Aguileta G, Refrégier G, Yockteng Y, Fournier E, Giraud T (2009) Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protozoa. *Infection Genetics and Evolution*, **49**, 151–193.

Altschul S, Madden T, Schäffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Anisimova M, Yang ZH (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution*, **24**, 1219–1228.

Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, **18**, 1585–1592.

Anisimova M, Bielawski J, Dunn K, Yang Z (2007) Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evolutionary Biology*, **7**, 154.

Antonovics J, Hood M, Partain J, Heuhsen AM (2002) The ecology and genetics of a host-shift: *Microbotryum* as a model system. *American Naturalist*, **160**, S40–S53.

Barrier M, Bustamante CD, Yu J, Purugganan MD (2003) Selection on rapidly evolving proteins in the *Arabidopsis* genome. *Genetics*, **163**, 723–733.

Berman HM, Westbrook J, Feng Z *et al.* (2000) The protein data bank. *Nucleic Acids Research*, **28**, 235–242.

Berman HM, Ten Eyck LF, Goodsell DS *et al.* (2005) The cAMP binding domain: an ancient signaling module. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 45–50.

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

Bohmer M, Colby T, Bohmer C *et al.* (2007) Proteomic analysis of dimorphic transition in the phytopathogenic fungus *Ustilago maydis*. *Proteomics*, **7**, 675–685.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chang YC, Kwon-Chung KJ (1999) Isolation, characterization, and localization of a capsule-associated gene, CAP10, of *Cryptococcus neoformans*. *Journal of Bacteriology*, **181**, 5636–5643.

Chen SL, Hung CS, Xu JA *et al.* (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli:* a comparative genomics approach. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5977–5982.

Conesa A, Gotz S, Garcia-Gomez J *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Consortium TU (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **35**, D193–197.

Day AW, Garber ED (1988) *Ustilago violacea,* anther smut of the caryophyllaceae. *Advances in Plant Pathology*, **6**, 457–482.

Dean RA, Talbot NJ, Ebbole DJ *et al.* (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.

Devier B, Aguileta G, Hood M, Giraud T (2009) Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes. *Genetics*, **181**, 209–223.

Egea R, Casillas S, Barbadilla A (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, **36**, W157–162.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl M, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Fuga Li Y, Costello JC, Holloway AK, Hahn MW (2008) ''Reverse Ecology'' and the power of population genomics. *Evolution*, **62**, 2984–2994.

Fukunishi Y, Hayashizaki Y (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiological Genomics*, **5**, 81–87.

Galindo BE, Vacquier VD, Swanson WJ (2003) Positive selection in the egg receptor for abalone sperm lysin. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4639–4643.

Garber ED, Ruddat M (2002) Transmission genetics of *Microbotryum violaceum (Ustilago violacea)*: a case study. *Advances in Applied Microbiology*, **51**, 107–127.

Ge G, Cowen L, Feng X, Widmer G (2008) Protein coding gene nucleotide substitution pattern in the apicomplexan Protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Comparative and Functional Genomics*, **2008**, 879023.

Giraud T, Yockteng R, Marthey S *et al.* (2008) 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.

Hahn MW (2007) Detecting natural selection on *cis*-regulatory DNA. *Genetica*, **129**, 7–18.

Hedges SB (2002) The origin and evolution of model organisms. *Nature Reviews Genetics*, **3**, 838–849.

Hohn TM, McCormick SP, Desjardins AE (1993) Evidence for a gene cluster involving trichothecene-pathway biosynthetic genes in *Fusarium sporotrichoides*. *Current Genetics*, **24**, 291–295.

Hood ME, Antonovics J, Heishman H (2003) Karyotypic similarity identifies multiple host-shifts of a pathogenic fungus in natural populations. *Infection Genetics and Evolution*, **2**, 167–172.

Hood ME, Katawezik M, Giraud T (2005) Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. *Genetics*, **170**, 1081–1089.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly Program. *Genome Research*, **9**, 868–877.

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In: *Proceedings International Conference Intelligent Systems for Molecular Biology* (eds Lengaver T, Schneider R, Bork P, Brutlag DL, Glasgow JI, Mewes H-W, Zimmer R), pp. 138–148. AAAI, California.

King C-C, Chao D-Y, Chien L-J *et al.* (2008) Comparative analysis of full genomic sequences among different genotypes of dengue virus type 3. *Virology Journal*, **5**, 63.

Kumar C, LeDuc R, Gong G *et al.* (2004) ESTIMA, a tool for EST management in a multi-project environment. *BMC Bioinformatics*, **5**, 176–187.

Lannuccelli E (2005) *SIGENAE Team*. INRA, Toulouse.

Le Gac M, Hood ME, Fournier E, Giraud T (2007a) Phylogenetic evidence of host-specific cryptic species in the anther smut fungus. *Evolution*, **61**, 15–26.

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

Lee SW, Cooksey DA (2000) Genes expressed in *Pseudomonas putida* during colonization of a plant-pathogenic fungus. *Applied and Environmental Microbiology*, **66**, 2764–2772.

Lee Y, Ota T, Vacquier V (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution*, **12**, 231–238.

Lopez-Villavicencio M, Enjalbert J, Hood ME *et al.* (2005) The anther smut disease on *Gypsophila repens*: a case of parasite sub-optimal performance following a recent host shift? *Journal of Evolutionary Biology*, **18**, 1293–1303.

Lucyshyn D, Abolmaali S, Weindorfer H *et al.* (2008) Ubiquitin and fusarium resistance: lessons from wheat cDNAs conferring deoxynivalenol resistance in yeast, 437–441.

Lutz M, Goker M, Piatek M *et al.* (2005) Anther smuts of Caryophyllaceae: molecular characters indicate hostdependent species delimitation. *Mycological Progress*, **4**, 225–238.

Mahlert M, Leveleki L, Hlubek A, Sandrock B, Bolker M (2006) Rac1 and Cdc42 regulate hyphal growth and cytokinesis in the dimorphic fungus *Ustilago maydis*. *Molecular Microbiology*, **59**, 567–578.

Marin A, Pothier J, Zimmermann K, Gibrat JF (2002) FROST: a filter-based fold recognition method. *Proteins*, **49**, 493–509.

Martínez-Espinoza AD, García-Pedrajas MD, Gold SE (2002) The Ustilaginales as plant pests and model systems. *Fungal Genetics and Biology*, **35**, 1–20.

McKay JK, Stinchcombe JR (2008) Ecological genomics of model eukaryotes. *Evolution*, **62**, 2953–2957.

Mulder NJ, Apweiler R, Attwood TK *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Research*, **35**, D224–228.

Mundy NI, Cook S (2003) Positive selection during the diversification of class I vomeronasal receptor-like (V1RL) genes, putative pheromone receptor genes, in human and primate evolution. *Molecular Biology and Evolution*, **20**, 1805–1810.

Nielsen R (2005) Molecular signature of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nielsen R, Bustamante C, Clark AG *et al.* (2005) A scan of positively selected genes in the genome of humans and chimpanzees. *PLoS Biology*, **3**, e170.

Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Review Genetics*, **7**, 851–861.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–217.

Portillo F (2000) Regulation of plasma membrane H+-ATPase in fungi and plants. *Biochimica Et Biophysica Acta-Reviews on Biomembranes*, **1469**, 31–42.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. *Nature*, **423**, 715–719.

Refrégier G, Le Gac M, Jabbour F *et al.* (2008) Cophylogeny of the anther smut fungi and their Caryophyllaceous hosts: prevalence of host shifts and Importance of delimiting parasite species. *BMC Evolutionary Biology*, **8**, 100.

Sloan D, Giraud T, Hood M (2008) Maximized virulence in a sterilizing pathogen: the anther-smut fungus and its co-evolved hosts. *Journal of Evolutionary Biology*, **21**, 1544–1554.

Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **19**, 1865–1869.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7375–7379.

Thomas A, Shykoff J, Jonot O, Giraud T (2003) Mating-type ratio bias in populations of the phytopathogenic fungus *Microbotryum violaceum* from several host species. *International Journal of Plant Sciences*, **164**, 641–647.

Thrall PH, Biere A, Antonovics J (1993) Plant-life history and disease susceptibility - the occurrence of *Ustilago violacea* on different species within the caryophyllaceae. *Journal of Ecology*, **81**, 489–498.

Ting C-T, Tsaur S-C, Wu C-I (2000) The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5313–5316.

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

Wasmuth JD, Blaxter ML (2004) prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.

Wong WSW, Nielsen R (2004) Detecting selection in noncoding regions of nucleotide sequences. *Genetics*, **167**, 949–958.

Wu C, Ting C (2004) Genes and speciation. *Nature Review Genetics*, **5**, 114–122.

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Bioscience*, **13**, 555–556.

Yang ZH (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang Z, Bielawski JP (2000a) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.

Yang ZH, Bielawski JP (2000b) Statistical methods for detecting molecular adaptation. *Trends in Ecology Evolution*, **15**, 496–503.

Yockteng R, Marthey S, Chiapello H *et al.* (2007) Expressed sequence tags of the anther smut fungus, *Microbotryum violaceum,* identify mating and pathogenicity genes. *BMC Genomics*, **8**, 272.

GA, RY, EV, BD, MCF and TG have been studying the co-evolution between the model fungi *Microbotryum* and their Caryophyllaceous host plants for several years, using population genetics, phylogenetics, genomics and experiments, addressing speciation, specialization, population structure and mating systems. JL, SM, HC, FR and AG belong to the MIG lab, dedicated to the functional analyses of genomes, and contribute to the development of genomics and post-genomics, by the development of methods and tools for in silico genome analysis and information systems for genomics. PW, CD, CC and AC belong to the Genoscope, dedicated to the production and analyses of genomes.

## Supporting information

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

**Table S1** A. Codeml parameter estimates for the seven clusters detected to be under positive selection among the 53 longest clusters (individual analysis). B. LRT results for the seven clusters detected to be under positive selection among the 53 longest clusters (individual analysis)

**Table S2** A. Codeml parameter estimates for the two cluster groups detected to be under positive selection among the 53 longest clusters (concatenated analysis). B. LRT results

for the two cluster groups detected to be under positive selection among the 53 longest clusters (concatenated analysis)

**Table S3** A. Codeml parameter estimates for all five species groups including the 319 shorter clusters (concatenated analysis). B. LRT results for all five species groups including the 319 shorter clusters (concatenated analysis)

**Table S4** A. Codeml parameter estimates for four species groups including all 372 clusters (concatenated analysis). B. LRT results estimates for four species groups including all 372 clusters (concatenated analysis)

**Table S5** Functional annotation of the clusters predicted to be under positive selection, for the 29 clusters for which Blast results were available. NA, not available

**Table S6** Functional annotation of the clusters predicted to be under positive selection, for the 34 clusters for which Interpro results were available. NA, not available

**Table S7** Functional annotation of the clusters predicted to be under positive selection, for the 32 clusters for which Psi-Blast results were available. NA, not available

**Table S8** Functional annotation of the clusters predicted to be under positive selection, for the 28 clusters for which FROST results were available. NA, not available

**Table S9** For each of the 16 genes analysed in additional *Microbotryum* species, species with branches having a $d_N/d_S$ ratio higher than two using the free-ratios test (the ratio is indicated in brackets). Species names are those coined in Refs. (Le Gac *et al.* 2007a; Refrégier *et al.* 2008)

**Table S10** Codeml parameter estimates for the three clusters detected to be under significant positive selection among the 16 clusters including the added fungal species

**Table S11** Likelihood ratio test results for the three clusters detected to be under significant positive selection among the 16 clusters including the added fungal species

**Table S12** Genes found to be under positive selection from the site-based analysis with codeml (LRTs: M1a vs. M2a and M7 vs. M8). Overlapping predictions are indicated in bold

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.