Sex and parasites: genomic and transcriptomic analysis of Microbotryum lychnidis-dioicae, the biotrophic and plant-castrating anther smut fungus

Michael H Perlin, Joelle J. Amselem, Eric Fontanillas, Su San Toh, Zehua Chen, Jonathan Goldberg, Sébastien Duplessis, Bernard Henrissat, Sarah Young, Qiandong Zeng, et al.

▶ **To cite this version:**

Michael H Perlin, Joelle J. Amselem, Eric Fontanillas, Su San Toh, Zehua Chen, et al.. Sex and parasites: genomic and transcriptomic analysis of Microbotryum lychnidis-dioicae, the biotrophic and plant-castrating anther smut fungus. BMC Genomics, 2015, 16 (461), 10.1186/s12864-015-1660-8 . hal-01186421

HAL Id: hal-01186421
https://hal.science/hal-01186421v1

Submitted on 24 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Perlin et al. BMC Genomics (2015) 16:461
DOI 10.1186/s12864-015-1660-8

BMC Genomics

RESEARCH ARTICLE Open Access

Sex and parasites: genomic and transcriptomic analysis of *Microbotryum lychnidis-dioicae*, the biotrophic and plant-castrating anther smut fungus

Michael H Perlin[1*], Joelle Amselem[2,3†], Eric Fontanillas[4,5†], Su San Toh[1†], Zehua Chen[6†], Jonathan Goldberg[6], Sebastien Duplessis[7,8], Bernard Henrissat[9,10], Sarah Young[6], Qiandong Zeng[6], Gabriela Aguileta[11], Elsa Petit[4,5,9], Helene Badouin[4,5], Jared Andrews[1], Dominique Razeeq[1], Toni Gabaldón[11,12,13], Hadi Quesneville[2], Tatiana Giraud[4,5], Michael E. Hood[14†], David J. Schultz[1] and Christina A. Cuomo[6*]

Abstract

Background: The genus *Microbotryum* includes plant pathogenic fungi afflicting a wide variety of hosts with anther smut disease. *Microbotryum lychnidis-dioicae* infects *Silene latifolia* and replaces host pollen with fungal spores, exhibiting biotrophy and necrosis associated with altering plant development.

Results: We determined the haploid genome sequence for *M. lychnidis-dioicae* and analyzed whole transcriptome data from plant infections and other stages of the fungal lifecycle, revealing the inventory and expression level of genes that facilitate pathogenic growth. Compared to related fungi, an expanded number of major facilitator superfamily transporters and secretory lipases were detected; lipase gene expression was found to be altered by exposure to lipid compounds, which signaled a switch to dikaryotic, pathogenic growth. In addition, while enzymes to digest cellulose, xylan, xyloglucan, and highly substituted forms of pectin were absent, along with depletion of peroxidases and superoxide dismutases that protect the fungus from oxidative stress, the repertoire of glycosyltransferases and of enzymes that could manipulate host development has expanded. A total of 14 % of the genome was categorized as repetitive sequences. Transposable elements have accumulated in mating-type chromosomal regions and were also associated across the genome with gene clusters of small secreted proteins, which may mediate host interactions.

Conclusions: The unique absence of enzyme classes for plant cell wall degradation and maintenance of enzymes that break down components of pollen tubes and flowers provides a striking example of biotrophic host adaptation.

Keywords: *Microbotryum violaceum*, Anther smuts, CAZymes, Transposable elements, Mating-type chromosomes, Pathogen alteration of host development

* Correspondence: michael.perlin@louisville.edu; cuomo@broadinstitute.org
†Equal contributors
[1]Department of Biology, Program on Disease Evolution, University of Louisville, Louisville, KY 40292, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article

BioMed Central

© 2015 Perlin et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Members of the genus *Microbotryum* are pathogenic fungi that have a global distribution and infect over nine families of host plants, with most replacing floral structures with fungal spores [1]. Within this genus, the anther-smut fungi of plants in the Caryophyllaceae, consist of many recognized and cryptic species, most possessing a very narrow host range, although some are more generalist [2]. Hybrid incompatibilities for the species examined in this complex leads to post-zygotic isolation in the forms of inviability and sterility [3–5]. The high rates of selfing and ecological specialization on different host plants are factors that should promote speciation in *Microbotryum* [6, 7]. The anther smut fungi thus allow examination of the ecology and evolution of host/pathogen interactions in "wild," non-agricultural environments [8, 9], where the genetically variable hosts provides an important contrast to the heavily studied, more monocultural hosts of agricultural systems. *Microbotryum* species also serve as a model for emerging infectious disease through host shifts [9, 10], for studying the evolution of mating systems, non-recombining mating-type chromosomes and sex chromosomes [11, 12], and for examining pathogens that alter the development of the host [13].

As obligate parasites, *Microbotryum* anther smut fungi must complete their life cycle in association with a host plant. Their fungal diploid teliospore masses give the flowers a dark, powdery appearance, thus the name "anther smut." Teliospores of *Microbotryum* are transported from diseased to healthy plants by direct transmission when plants are in close proximity [14] or by pollinating insects [15], where once deposited the diploid fungus germinates and undergoes meiosis to give rise to four haploid cells [6]. Each of these cells can bud off yeast-like sporidia on the plant surface. New infectious dikaryotic hyphae are rapidly produced by conjugation of two cells of opposite mating-type (a_1 and a_2) and enter the host tissue to grow endophytically until they reach the bud meristems and anthers [16]. Here the nuclei fuse (karyogamy) and teliosporogenesis occurs thus completing the life cycle (Fig. 1; [6, 17–19]).

The commandeering of insect pollinators for disease transmission is associated with significant pathogenic alterations of the host's floral morphology, particularly in relation to male and female structures. Although diseased plants are only slightly affected in vegetative morphology and survival [20], infection often results in complete host sterility, as no pollen is produced in the anthers and the ovary becomes rudimentary. Interestingly, in dioecious or gynodioecious species of *Silene* (*e.g.*, *S. latifolia* and *S. vulgaris*, respectively), ovaries of diseased female plants are aborted while a male morphology develops with spore-bearing

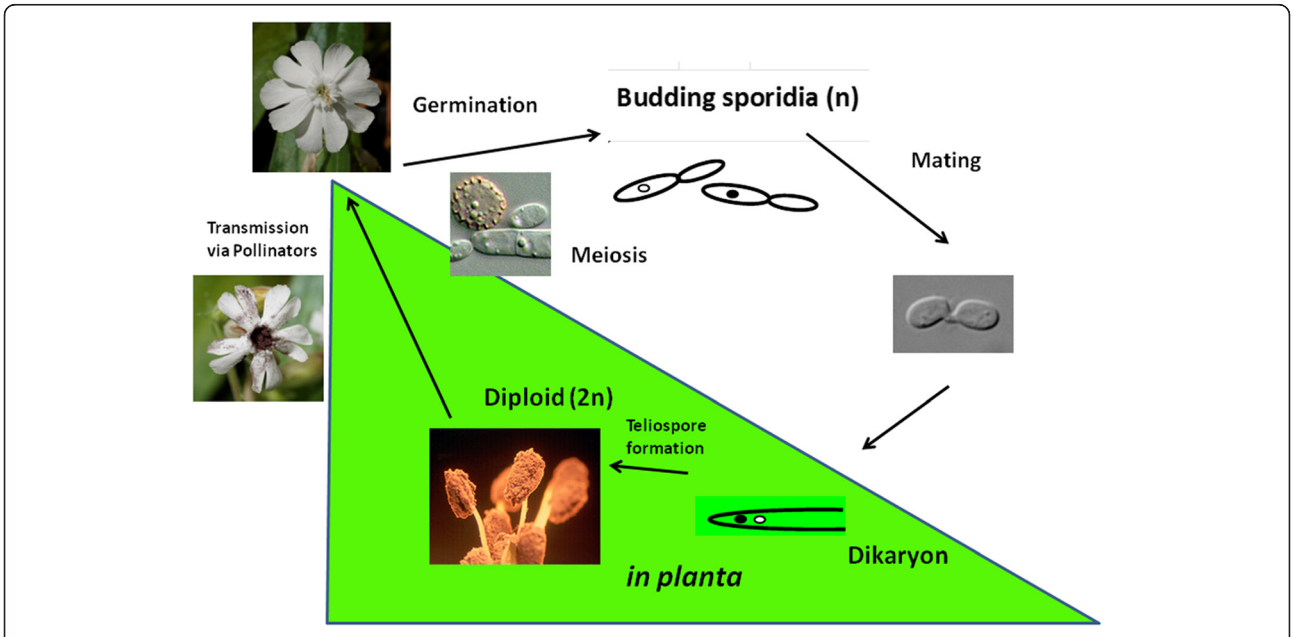

Fig. 1 Lifecycle of *Microbotryum lychnidis-dioicae*. The infection cycle for *M. lychnidis-dioicae*) is shown. Infection begins when diploid teliospores germinate on a suitable plant surface, and after meiosis, produce linear tetrads of haploid basidiospores. These cells may mate with cells of opposite mating-type (*i.e.*, *a1* or *a2*), either directly within the tetrad or after budding to yield free yeast-like sporidia. Mated cells form conjugation bridges and, upon receiving suitable cues from the host plant, develop into dikaryotic hyphae that penetrate the host tissue. The hyphae progress systemically through the plant and migrate to the flower primordia. There nuclear fusion (karyogamy) occurs, as the hyphae break up and develop into diploid teliospores that replace the pollen in the anthers of the developing flowers. Pollinator species are then able to transmit the spores to other uninfected flowers, thus completing the cycle

anthers. The basis for the changes in female plants is not fully understood. Studies have identified host genes expressed during infection of females that are also normally expressed in uninfected males (*MEN* genes [21]; *SLM2* [22]).

The formation of a hyphal dikaryon by mating between haploid cells is a prerequisite for infection in *Microbotryum*, and thus each newly diseased plant represents the completion of a sexual reproductive cycle. In heterothallic fungi, mating can occur only between haploid cells carrying different mating-types, which are controlled by alleles at one or two loci [23, 24]. In a few fungi, recombination suppression has evolved around the mating-type locus/loci [25–27], sometimes leading to structurally dimorphic chromosomes similar to the sex chromosomes in plants and animals [28, 29]. The suppression of recombination on mating-type chromosomes is expected to lead to degeneration; this can be manifested as transposable element accumulation [30], codon usage degeneration [31], accumulation of deleterious mutations [32], and eventually to chromosomal dimorphism, and thus the emergence of allosomes [12]. Such degeneration has in fact been observed on the mating-type chromosomes of some fungi, including the ascomycete *Neurospora tetrasperma* [31] and the basidiomycete *Microbotryum lychnidis-dioicae* [12, 30, 33], the most well-studied representative of the anther-smut fungi.

We sequenced the genome of haploid isolate Lamole p1A1 [11], of the a_1 mating type, to represent the *M. lychnidis-dioicae* species found in association with the perennial, dioecious host, *Silene latifolia.* RNA-Sequencing of distinct life cycle stages was incorporated to validate gene content and measure expression changes during infection. We identified gene family expansions that could play a role in plant infection by comparing *M. lychnidis-dioicae* to other basidiomycetes, taking advantage of increasing genome coverage of the Pucciniomycotina subphylum. The identification of genes that are induced or repressed during infection highlighted carbohydrate active enzymes (CAZymes) that may be involved in host cell degradation or manipulation of host development. Additionally, the *M. lychnidis-dioicae* genome is riddled with a diverse array of transposable elements (TEs), including a higher proportion of Helitron elements than found in the much larger and more highly-repetitive genomes of related rust fungi. Genome regions corresponding to the mating-type chromosomes of *M. lychnidis-dioicae* [11] are enriched for repetitive sequence. Further analysis of the sequence of the entire a_1 mating-type chromosome identified more than 300 genes linked with mating-type. Together, these findings provide an in-depth portrait of genetic architecture and adaptation in a specialized fungal plant pathogen.

Results

Genome sequence and content

The 25.2 Mb haploid genome of *M. lychnidis-dioicae* was sequenced using 454 technology, generating high coverage of three different-sized libraries (Additional file 1), and assembled using Newbler (Table 1). The assembly was comprised of 1,231 scaffolds where the average base was present in a scaffold of 185 kb and a contig of 50 kb (N50 measure, Table 1). Despite the large number of contigs, the assembly was a nearly complete representation of the sequenced genome, comprising 97 % of sequenced bases. The assembly included five scaffold ends with the typical fungal telomere repeat (TTAGGG), though three of these scaffolds were smaller than 1 kb in size.

High coverage strand-specific RNA-Seq, generated from three biological conditions (Additional file 2) assisted with the prediction of 7,364 protein coding genes and identified expression changes potentially important for the pathogenic lifecycle. Sampled conditions included two *in vitro* conditions, haploid cells grown on yeast peptone dextrose media (YPD) agar (referred to as rich) or on 2 % water agar (referred to as nutrient limited). These were compared to a sample from infected male plant tissue during the late stages of fungal development, where teliospores form on partially and fully opened smutted flowers [34, 35] (referred to as "MI-late"). Incorporation of RNA-Seq into predicted gene structures (Methods) defined UTRs for the vast majority (more than 6,100) predicted genes; the average length of 5′ and 3′ UTRs was 183 bases and 253 respectively. Coding sequences average 1,614 bases (median of 1,338 bases) in length and contain 5.6 exons; genes are separated by intergenic regions 502 bases in

Table 1 Genome statistics of nuclear genome and mating-type chromosome regions

	Nuclear genome	NRR[a] regions	PAR[b] regions
Assembly size (Mb)	26.1	1.86	0.38
Scaffolds (count)	1,231	85	2
Scaffold N50 (kb)	185	48	381
Contigs (count)	2,104	229	16
Contig N50 (kb)	50	13	45
GC content (%)	55.4	54.6	53.9
TE content (%)	14	41	13
Protein coding genes	7,364	350	99
Mean coding length	1,614	1,344	1,408
Median coding length	1,338	954	1,302
Mean exons/gene (count)	5.6	4.6	5.1
Mean intercds length (bp)[c]	1,181	2,600	1,861
tRNAs	134	5	2

[a]Non-recombining regions (NRR). [b]Pseudo-autosomal regions (PAR). [c]average length between coding sequence (cds) start and stop

length on average. The *M. lychnidis-dioicae* gene set has high coverage of a core eukaryotic gene set [36], highest in fact than any of the fungal gene sets used in comparative analysis (Additional file 3), suggesting the assembly includes a highly complete gene set.

The mitochondrial genome consisted of a single finished contig of 97 kb and included the canonical set of genes. These were the respiratory-related proteins of the NADH dehydrogenase family (*nad1-6* and *nad4L*), apocytochrome b (*cob*), cytochrome oxidases (*cox1-3*), the proteins related to ATP synthesis (*atp6*, *atp8* and *atp9*); ribosomal RNAs (*rns*, and sequence similar to *rnl*); ribosomal proteins (*rps3*); DNA polymerase (*dpo*) and 25 tRNAs. Homing endonucleases of the LAGLIDADG and GIY-YIG families and a maturase protein were located in mitochondrial intronic regions, including within three introns of *cox1*.

To examine the presence of AT-rich isochores and more generally the genome structure in GC composition in *M. lychnidis*-dioicae, we measured the fluctuation of GC percent along the assembly. Although the genome is not organized in discrete isochores as in some fungal genomes [37], we observe some large-scale variation (>100 kb) in base composition within chromosomes, as well as finer-scale fluctuations (Additional file 4A). The GC content was positively correlated with coding density, with the most significant correlations for 5 and 10 kb windows explaining 16.8 % of the variance (Additional file 4B). This correlation has been explained in other systems by biased codon usage toward GC-rich codons or alternatively biased gene conversion occurring more frequently in coding than in non-coding sequences [38]. In fact, an analysis of the preferred codons (*i.e.,* the most frequently used codons in the predicted genes), showed that 17 out of 18 had a GC base in the third position, which is the most degenerate position and therefore primarily influences the GC composition of genes (Additional file 5).

Shifts in transposable element type, location, and impact of RIP

The genome of *M. lychnidis-dioicae* contained diverse transposable elements (TEs), represented by 286 consensus elements, covering 14 % of the total assembly. The overall TE content was lower than of other species in Pucciniomycota; in *Puccinia graminis* f. sp. *tritici* or *Melampsora larici-populina*, in which TEs account for nearly 45 % of the assembled genomes, contributing to the expanded genome size of these fungi [39]. Among class I retrotransposon elements (36 % of TE sequences), Long Terminal Repeat (LTR) elements were the most common (28 % of TE sequences), in particular *Copia*-like elements (20 % of TE sequences), in agreement with prior studies of genome sampling and expressed sequence profiles [30, 40]. The remainder of class I elements consisted of Long Interspersed Nuclear Element (LINE) (7 % of TE sequences) and *Dictyostelium* Intermediate Repeat Sequence (DIRS) elements (1 % of TE sequences). Among class II DNA transposons (23 % of TE sequences), Terminal Inverted Repeat (TIR) and *Helitron* elements, which transpose by rolling-circle replication [41] account for 12 % and 10 % of TE sequences, respectively (Additional file 6, Table 2). The *Helitron* proportion was an order of magnitude higher than in the more repetitive genomes of other Pucciniomycota fungi *P. graminis* f. sp. *tritici* and *M. larici-populina*, for which a similar analysis characterized only 1 % of TE elements as *Helitrons* [39].

The TE categories varied significantly in their proximity to genes. A chi-squared test of heterogeneity found a significant difference in the TE content of regions nearby genes, comparing regions less than versus greater than 1 kb upstream and downstream of genes (upstream region p-value < 2.2e-16; downstream region p-value < 2.2e-08, Additional file 7). In particular, class II elements (TIR and *Helitrons*) were closer to genes than class I elements (LTR and LINE), with a greater enrichment upstream of genes compared to downstream (Additional file 7). In addition, there appeared to be an association between TE-rich regions in *M. lychnidis-dioicae* and genes for Small Secreted Proteins (SSPs), which can include effector proteins involved in host-pathogen interactions as suggested in some pathogen genomes [42, 43]. SSPs were indeed located nearer to TEs than the set of all other non-SSP genes (Chi-squared p-value < 5e-4; Additional file 8).

Hypermutation in TEs that resembles the genome defense, Repeat-Induced Point mutation (RIP), has previously been observed in the LTR elements (*copia*-like and *gypsy*-like elements) and *Helitron* transposons of *M. lychnidis-dioicae* [44, 45]. Some genes that appeared similar to those necessary for RIP in *Neurospora crassa* [46, 47] were found in the *M. lychnidis-dioicae* genome. These include a cytosine methyltransferase (MVLG_04160), but establishing the orthology with the RIP-essential *rid* gene

Table 2 Genome coverage of TE families (10,283 copies of 286 REPET consensus sequences)

TE copies	Copy number	Coverage relative to TE space		Coverage relative to assembly size	
LTR	1018	27.70	Class I 35.64	3.89	Class I 5.01
DIRS	57	1.34		0.19	
LINE	392	6.61		0.93	
SINE	0	0		0	
TIR	438	12.18	Class II 22.76	1.71	Class II 3.20
MITE	39	0.39		0.06	
Helitron	373	10.18		1.43	
Unknown	1942	41.60	41.60	5.85	5.85
Total	4259	100	100	14.06	14.06

from *N. crassa* versus other cytosine methyltransferases (*e.g.* Dim-2) would require further investigation. *M. lychnidis-dioicae* sequences similar to the Dim-5H3 histone methyltransferase that is essential for marking RIP regions in *N. crassa* (MVLG_02125, MVLG_05378) were also found.

Evaluation of dinucleotide signature at transition mutations in 179 TE families (2,298 genome copies, Additional file 9) revealed that 40 % of these TE copies exhibited elevated substitution rates that were particular to which nucleotide was 3′ to the cytosine (Methods, Additional file 10). Eighty per cent of TE copies with high frequencies of cytosine mutation showed a bias toward CpG dinucleotides, consistent with the "CpG effects" [48] of maintenance methylation known in eukaryotes [49–51], including fungi [45, 52] where CpG methylation of TEs have been shown in ascomycete and basidiomycete fungi [53]. Notably, the rate of CpG mutations varied according to TE order and superfamily examined: TE copies exhibiting a pattern of frequent transition at CpG sites appeared lower for class II DNA transposons (*Helitron*-type and TIR elements) than that of class I Retrotransposons (21 % and 58 % respectively) (Additional file 10). In addition, the GC content in TEs (54.4 %) was very close to GC content in genes (56.0 %) and genome (55.4 %).

In addition, the *M. lychnidis-dioicae* genome was found to contain the core RNAi machinery components, that may act to constrain proliferation of transposable elements, contrasting with some other Basidiomycetes that have lost this pathway [54]. The RNAi pathway components were identified based on similarity to known RNAi genes in fungi and validated by examining predicted functional domains (Methods). The genome of *M. lychnidis-dioicae* was also found to contain one copy of a RNA-dependent RNA polymerase (MVLG_02137), two copies of Argonaute (MVLG_06823 and MVLG_06899), and one copy of Dicer (MVLG_01202). All of these components of RNAi machinery were expressed under each of the conditions examined, suggesting the pathway is active across life cycle states, although one copy of Argonaute (MVLG_06823) was significantly more highly expressed (corrected p-value <0.01) during infection compared to nutrient limited and rich agar media.

Mating-type locus and chromosome

The central proteins involved in mating-type determination in basidiomycetes were found in *M. lychnidis-dioicae*. Orthologs of the two-component homeodomain transcription factor that functions in post-mating compatibility, HD1 (MVLG_07149) and HD2 (MVLG_07150), were assembled in a 14.1 kb region; as previously described HD1 and HD2 are adjacent and divergently transcribed in *M. lychnidis-dioicae* [55], similar to other fungi [56]. Both the mating pheromone receptor [57] and the homeodomain compatibility factor identified above are located at the ends of their respective supercontigs, such that the genomic proximity of these two essential mating-type-determining loci is unclear. The chromosomes bearing these two mating-type loci show suppressed recombination across most of their length, with only two small recombining regions at their ends, *i.e.,* pseudo-autosomal regions (PARs) [11, 12, 33, 55, 58]. The assembly scaffolds corresponding to the non-recombining regions (NRRs) and to the PARs were identified based on alignment to an optical map of the mating-type chromosomes [12, 33] and by performing additional sequencing of gel purified chromosomes (Methods). A total of 449 genes mapped to the a_1 mating-type chromosome, including 350 genes found on the NRRs and 99 genes found on the PARs (Table 1, Additional file 11). Other than the genes for the pheromone receptor, the homeodomain transcription factors, and the STE20 protein kinase, no other genes on the mating-type chromosome have a predicted function linked to mating in other systems.

Consistent with the expectation in regions of suppressed recombination (*e.g.*, [59]), the TE density (Fig. 2a) was several fold higher in the non-recombining region (NRR) of the a_1 mating-type chromosome (41 %) relative to the autosomes (9 %), confirming prior studies of a TE accumulation on the mating-type chromosomes as a whole [11, 12, 33]. The pseudoautosomal regions (PARs, supercontigs 37 and 43) displayed a TE content (~13 %) more similar to the estimate for autosomal regions than to the NRR of the mating-type chromosome. Gene density in the NRR of the mating-type chromosome (23 %), estimated as CDS density, was less than half the gene density of the autosomes (49 %) (Fig. 2b). The number of genes predicted per 10 kb positions also indicated a lower density in the NRR of the mating-type chromosome (1.7 genes) than in the autosomal partition (2.9 genes) (Fig. 2c). As with TE content, the PARs displayed gene density values (~40 % for CDS density and ~2.5 for genes per 10 kb) closer to the autosomal estimates than the NRR of the mating-type chromosome. The NRR of the mating-type chromosomes contained genes of the same distribution, with the exception of 2-fold elevated density of small secreted proteins (SSPs) and, in particular, a 5-fold enrichment of Cys-rich SSPs compared to the autosomes.

With regard to base pair composition, again the NRR exhibited a pattern distinct from the PARs or autosomes in GC content. The GC content in protein coding genes, irrespective of codon position, were similar among the NRR of the mating-type chromosome, PAR, and autosomes (Fig. 2d), as were the contents represented by TEs (Fig. 2e). However, in other sequences, representing

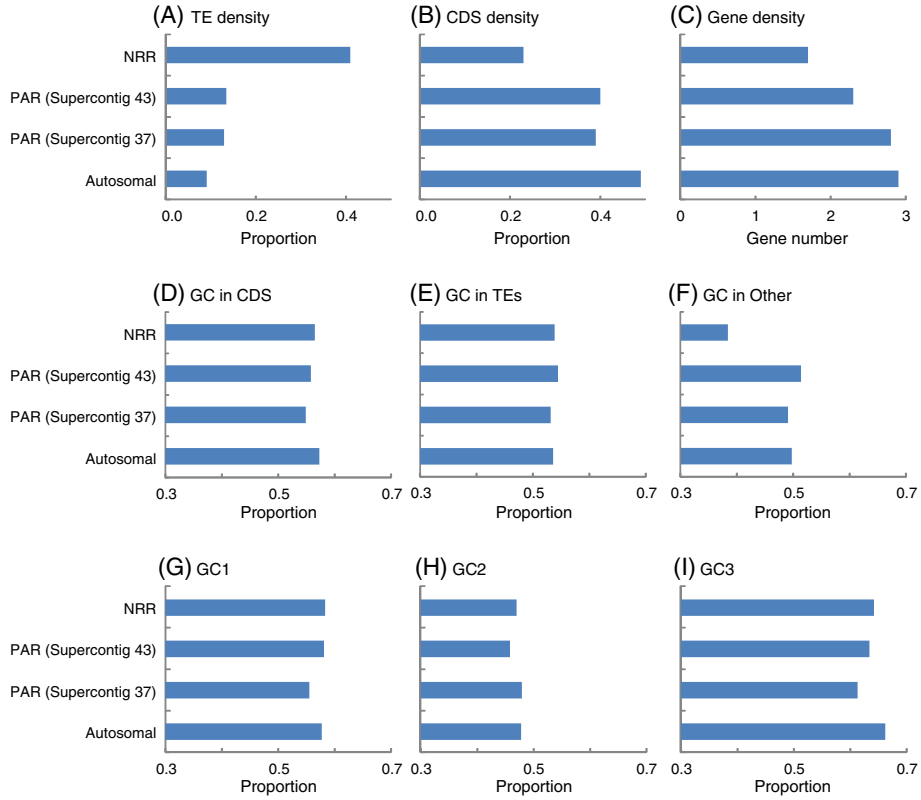

Fig. 2 Comparisons of sequence characteristics in non-recombining regions (NRR) of the mating-type chromosome, pseudoautosomal regions (PAR), and autosomes. The genomic regions are shown, with results for the two supercontigs ("37" and "43") corresponding to the two PARs presented separately. **a** Transposable element (TE) density is shown as the total length of TE sequences over the total length of DNA analyzed. **b, c** Gene density is shown as the proportion of the total length of coding region (CDS) over the total length of DNA analyzed, and number of putative genes identified per 10000 nucleotides, respectively. **d–f** Proportion GC base pair contents are shown for CDS, TEs, and for the remaining, predominantly intergenic regions. **g–i** Proportion GC base pair content for protein-coding genes are shown relative to first-, second-, and third-codon positions ("GC1", "GC2" and "GC3," respectively)

inter-genic regions not consisting of TEs, the NRR displayed markedly reduced GC content (Fig. 2f) while PARs and autosomes had similar GC levels. GC content of codon positions within protein-coding genes did not vary among the NRR, PARs, and autosomes (Fig. 2g–i), notably in that the third-codon position patterns was not reflective of intergenic GC composition variation across regions. This pattern as well as the lower GC content observed for the second-codon position compared to first or third positions is consistent with some prior research [60]. The correspondence analysis of codon usage (Additional file 12) among the non-recombining region of the mating-type chromosome, PAR and autosomal did not indicate obvious differences in codon usage.

Gene conservation and lineage-specific changes

The comparison of 7,364 predicted proteins of *M. lychnidis-dioicae* to those of diverse basidiomycetes revealed gene loss and gain patterns relevant in terms of the growth and pathogenesis of this organism. We included representatives of the three subphyla of basidiomycetes (Pucciniomycotina, Agaricomycotina, and Ustilaginomycotina), as well as three Ascomycota species as outgroups (Fig. 3, Additional file 13). Within the Pucciniomycotina, species compared included *M. lychnidis-dioicae* and two closely related Microbotryomycetes (*Sporobolomyces roseus* and *Rhodotorula glutinis*) and three other more distantly related species; within this group the two other plant pathogens (*P. graminis* f. sp. *tritici* and *M. larici-populina*) are biotrophic, like *M. lychnidis-dioicae*. These comparisons revealed 2,451 *Microbotryum* gene clusters representing 2,613 genes that were broadly conserved in the Basidiomycota (present in at least 13 of the examined 15 other basidiomycete genomes). The gene families specific to Pucciniomycotina, with orthologs present only in *M. lychnidis-dioicae* and/or the other Pucciniomycotina species, (Fig. 3, orange boxes) was composed of a small set of 224 predicted proteins from *M. lychnidis-dioicae*; the two rusts (*P. graminis* and *M. larici-populina*) share a larger number of Pucciniomycotina-specific proteins in part due

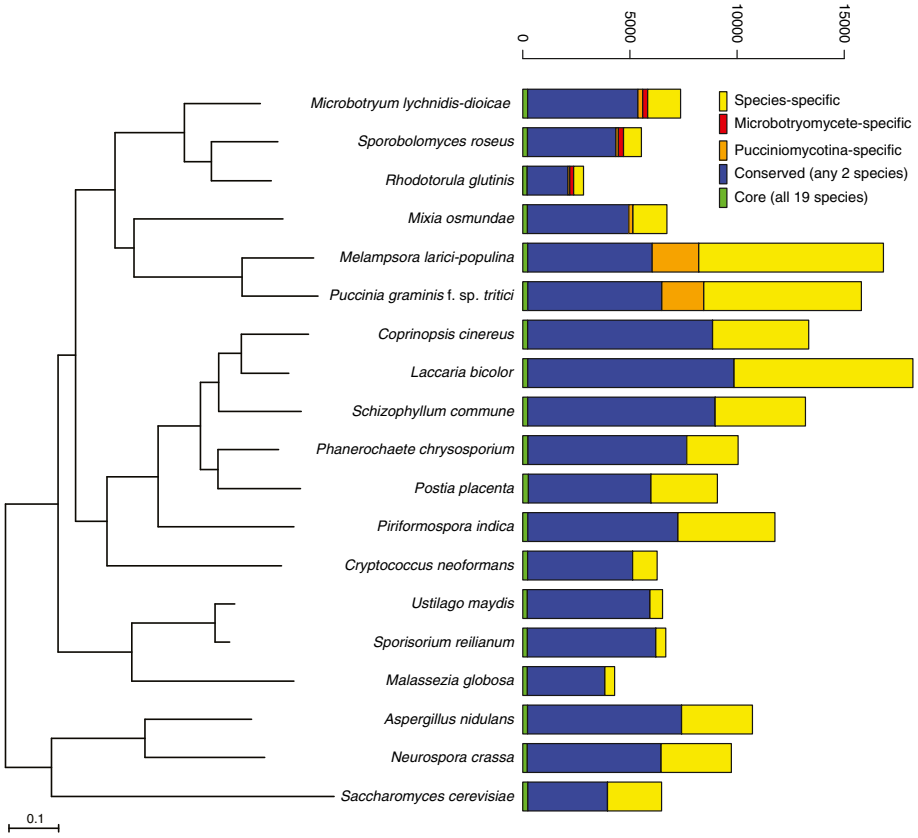

Fig. 3 Phylogenetic relationship and gene conservation of *Microbotryum lychnidis-dioicae* and 18 compared fungi. Phylogeny (left panel) is based on concatenated MUSCLE alignments of 80 single copy genes. Species phylogeny was inferred using RAxML (PROTCATWAG model) with 1,000 bootstrap replicates; all nodes were supported by at least 99 % of replicates. Ortholog conservation (right panel) highlights genes conserved in all species (core, green), genes conserved in at least two species (conserved, blue), genes unique to the Pucciniomycotina or Microbotryomycetes (orange and red, respectively) and genes unique to a given species (species-specific, yellow)

to their expanded genome size. Examining proteins conserved across the Microbotryomycetes, orthologs of 4,844 *M. lychnidis-dioicae* proteins were present in at least one other species, and of these 233 were specific to the Microbotryomycetes. A set of 190 gene duplications occurring specifically along the *M. lychnidis-dioicae* lineage were identified using phylogenetic analysis (Additional file 14, Additional file 15, phylomedb.org); these did not display functional enrichment for gene ontology (GO) term assignments, as GO terms were only assigned for 19 of the 190 genes. While most genes (70 %) were shared with occurrence in at least one other species, the remaining set of 1,534 genes appear specific to *M. lychnidis-dioicae*.

The identification of enriched or depleted PFAM domains for *M. lychnidis-dioicae* compared to other fungi revealed significant differences in functional categories between these genomes. A total of six protein domains were significantly enriched or depleted (q-value < 0.05) in *M. lychnidis-dioicae* compared to the other basidiomycetes examined (Table 3, Additional file 16). Enriched domains include secretory lipases, and two domains of unknown function, DUF23 (glycosyl transferase 92) and DUF1034 (Fn3-like). DUF23 (PF01697) was present in five copies in *M. lychnidis-dioicae* and was only present otherwise in *Mixia osmundae* and *S. roseus* in our comparison; three of these five genes were mapped to the GT2 CAZY family expanded in *M. lychnidis-dioicae* (see below). Both *M. lychnidis-dioicae* and the rusts contain a large number of proteins with the DUF1034 domain; 8 of the 10 *M. lychnidis-dioicae* proteins with a DUF1034 domain also contained a subtilase family protease domain. Phylogenetic analysis of proteins with the DUF1034 domain suggested that independent gene family expansions occurred in different species; the rusts formed a separate clade from *Microbotryum* that is further subdivided, mostly along species lines (Additional file 17). Domains depleted in *M. lychnidis-dioicae* relative to other basidiomycetes were also identified, including Cytochrome p450, NACHT, and F-box domains (Table 3, Additional file 16). A more narrow comparison

Table 3 Expanded or depleted PFAM domains in *Microbotryum lychnidis-dioicae*

PFAM domain	*M. lychnis-dioicae*	*S. roseus*	*R. glutinis*	*M. osmundae*	*M. larici-populina*	*P. graminis-tritici*	*S. reilianum*	*U. maydis*	*M. globosa*	Agaricomycetes[c] (7)	Basidiomycete comparison[a]		Pucciniales comparison[b]	
											p-value	*q*-value	*p*-value	*q*-value
PF00067.15 Cytochrome P450	10	7	5	14	29	17	15	20	6	95	1.43E-11	6.49E-08	1.28E-02	1
PF05729.5 NACHT	1	2	1	1	1	1	1	1	0	44	6.20E-08	1.41E-04	4.51E-01	1
PF01697.20 Glycosyltransferase family 92	5	1	0	1	0	0	0	0	0	0	1.65E-05	2.50E-02	1.99E-02	1
PF06280.5 Fn3-like (DUF1034)	10	0	0	0	5	9	1	1	0	1	2.54E-05	2.88E-02	1.00E + 00	1
PF00646.26 F-box	7	18	12	12	9	11	11	10	3	43	5.20E-05	4.33E-02	2.72E-02	1
PF03583.7 Secretory lipase	7	0	0	0	0	0	3	2	6	0	5.72E-05	4.33E-02	1.76E-03	3.07E-01
PF00734.11 CBM1 Fungal cellulose binding	0	0	3	0	0	0	0	0	0	21	1.21E-04	7.10E-02	6.60E-02	1
PF02816.11 Alpha kinase	0	0	0	0	79	39	0	0	0	5	1.25E-04	7.10E-02	2.80E-27	1.27E-23
PF07690.9 MFS1 Major Facilitator Superfamily	119	110	26	64	90	72	104	98	30	129	4.11E-02	1	2.06E-08	4.67E-05
PF01753.11 zf-MYND finger	7	11	26	3	7	4	2	2	1	41	4.61E-04	2.33E-01	5.16E-08	7.82E-05
PF01083.15 Cutinase	0	0	1	4	21	9	3	4	0	2	7.90E-02	1	6.49E-07	5.89E-04
PF01670.9 Glycosyl hydrolase family 12	0	0	0	10	14	3	0	0	0	1	1.70E-01	1	1.08E-06	6.56E-04
PF11327.1 DUF3129	0	0	0	4	13	10	0	0	0	1	2.62E-01	1	1.08E-06	6.56E-04
PF00097.18 Zinc finger, C3HC4 type	23	11	9	26	22	93	28	24	12	24	6.82E-01	1	1.16E-06	6.56E-04
PF00080.13 Copper/zinc superoxide dismutase	0	0	0	2	6	18	0	0	0	1	2.62E-01	1	1.94E-06	9.79E-04
PF00098.16 Zinc knuckle	10	6	1	6	11	59	8	7	5	11	7.63E-01	1	6.31E-06	2.87E-03
PF06609.6 Fungal trichothecene efflux pump	16	16	5	6	6	3	8	9	2	12	4.10E-02	1	1.18E-05	4.86E-03
PF12013.1 DUF3505	0	0	0	1	4	16	0	10	0	0	2.65E-01	1	2.15E-05	8.14E-03
PF03101.8 FAR1 DNA-binding domain	0	0	0	1	14	4	1	1	0	1	4.09E-01	1	6.65E-05	2.32E-02
PF00083.17 Sugar (and other) transporter	54	56	13	31	46	34	56	46	14	61	3.28E-01	1	1.71E-04	5.54E-02
PF07738.6 Sad1/UNC-like	2	1	2	3	24	7	2	1	2	2	5.87E-01	1	2.45E-04	7.42E-02

[a]M. lychnis-dioicae compared to all other Basidomycetes; [b] Microbotryales (M. lychnis-dioicae, S. roseus, R. glutinis) compared to other Pucciniales (M. larici-populina, P. graminis tritici, M. osmundae); [c]Agaricomycetes represent average of the 7 species in this group; see Additional file 11 for counts per species

of the three Microbotryomycetes to the other species within the Pucciniomycotina identified additional expansions and depletions common to the species in this lineage. A fungal trichothecene efflux pump (PF06609) gene family was enriched in *M. lychnidis-dioicae* and *S. roseus*, with 16 copies in each genome (Table 3). By contrast, the alpha-kinase family that is highly expanded in the rusts [39] was absent in the Microbotryomycetes. The cutinase domain shows a similar conservation pattern; while multiple cutinase genes are found in other biotrophic pathogens, no copies were detected in *M. lychnidis-dioicae*. At lower levels of significance, the CBM1 cellulose binding domain was detected as absent from all species in the Pucciniomycotina with the exception of *R. glutinis* (Table 3, Additional file 16). More specific analysis of these enriched and depleted domains is presented below.

The expansion of the secretory lipases appeared specific to *M. lychnidis-dioicae* within the Pucciniomycotina (Table 3). Among all other basidiomycete genomes compared, secretory lipases are also highly represented in *Malassezia globosa* (Ustilaginomycotina), a skin fungus associated with human dandruff and dependent on its host for lipids. *Malassezia globosa* has an additional gene family expansion associated with lipid acquisition; this species has 6 copies of phospholipase C, whereas *M. lychnidis-dioicae* contains only a single phospholipase C protein. Unlike *M. globosa*, *M. lychnidis-dioicae* does not depend on lipids for growth, and contains a predicted fatty acid synthase (MVLG_04698). The secretory lipase family is also present at lower copy number in the two Ustilaginomycotina corn smuts, *Sporisorium reilianum* and *Ustilago maydis*, of which the latter responds to lipids, including corn oils, as part of a developmental switch [61]. A phylogenetic analysis of the secretory lipases in this comparison revealed that the *M. lychnidis-dioicae* lipases have undergone a lineage specific expansion, as in *M. globosa* (Fig. 3a). Most lipases were predicted to be secreted including three of the seven *M. lychnidis-dioicae* lipases. However four of the seven genes appeared partial based on alignment of the protein sequences; the missing 5′end from two genes deleted the region containing a secretion signal in paralogous copies. Further refinement of the assembly or transcripts is needed to identify the full length version of these genes or confirm if they are perhaps pseudogenes (see below), and establish their relative location in the genome.

Previous work in *M. lychnidis-dioicae* has shown that mating mixtures of haploid cells produce hyphae in response to phytols and to tocopherols [62]. We therefore exposed haploid sporidial cells (p1A1 or p1A2 strains) or mated cells (p1A1 and p1A2 mixed together) to various lipids or oils, including commercially available corn oil, (+/−)-α-tocopherol, and phytol. Each stimulated filamentous growth of mated mixtures (Fig. 4b–e), but had no observable effect on haploid cells (unmated). The observation that mated cells also respond to corn oil supports the hypothesis that lipid response may be important for the development of this species. This is possible, as lipids are likely present on the host meristem [63]. Moreover, we find that at least three lipase genes are differentially expressed when exposed to phytols (a constituent of chlorphyll), compared to similarly treated cells in the absence of phytols (see below; Additional file 18).

An important group of proteins identified as being expanded or depleted in *M. lychnidis-dioicae* are those predicted to be involved in cell wall modifications, both in terms of fungal cell walls, but also as might affect host plants. Families of structurally-related carbohydrate catalytic and carbohydrate-binding modules (or functional domains) are described in the CAZy database (www.cazy.org) [64]. Such enzymes break down, modify, or build glycosidic bonds. The assignment of the predicted proteins derived from a genome to CAZy families helps to shed light on the particular glycobiological features of an organism [65]. A total of 236 *M. lychnidis-dioicae* protein models were mapped to protein families in the CAZy database based on sequence conservation (percentage identity over CAZy domain length) (Additional file 19). The CAZy family profile of *M. lychnidis-dioicae* was then compared to that recently published for 33 basidiomycetes [66], in order to identify expanded and reduced families (Table 4, Additional file 20).

With 98 candidate glycosyltransferases (GTs), *M. lychnidis-dioicae* has more than any of the 33 basidiomycetes used in the comparison (average = 69.5; min = 41; max = 95). Both *M. lychnidis-dioicae* and *Puccinia graminis* contain a candidate fucosyltransferase, which is not present in the other basidiomycetes surveyed. This is compatible with the known presence of fucose in the cell wall of *Microbotryum* [67, 68]. Two other families expanded in *M. lychnidis-dioicae* include alpha-mannosyltransferases (GT32 and GT62), suggesting that the cell wall includes a larger fraction of alpha-mannan than in other species. In fungi, the synthesized cell wall carbohydrates are frequently remodeled by the action of dedicated glycoside hydrolases and transglycosidases that are found in distinct CAZy glycosyl hydrolase (GH) families. A notable feature of *M. lychnidis-dioicae* is that it has a reduced number β-1,3-glucan cleaving or modifying enzymes of families (GH16, GH72, GH81, and GH128).

Examination of the other GH families and of the other categories involved in carbohydrate breakdown (PL, CE, AA and ancillary CBM) revealed that *M. lychnidis-dioicae* completely lacked cellulases (no GH6, GH7, GH8, GH9, GH12, GH44, GH45, nor any GH5 with highest sequence similarity to characterized cellulases) (Table 4, Additional file 20). In addition no cellulose-targeting lytic

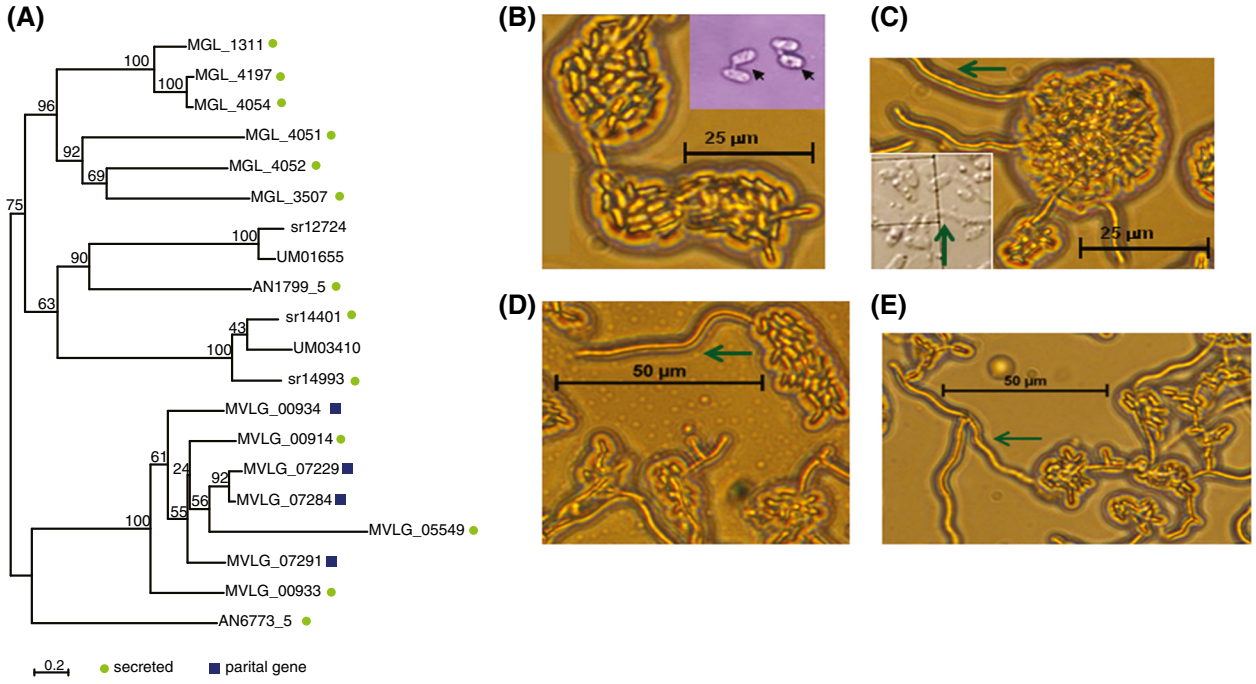

Fig. 4 Expansion of lipase gene family in *Microbotryum lychnidis-dioicae* and *Malassezia globosa*. Both *M. lychnidis-dioicae* and *M. globosa* contain a higher number of proteins with a lipase domain (PF03583.7) relative to the other fungi examined. **a** Phylogenetic tree of proteins with this domain. Protein sequences were aligned with MUSCLE, and a phylogeny was inferred from this alignment using RAxML (PROTCAT model, DAYHOFF matrix). A total of 1000 bootstrap replicates were performed, and the percent of replicates shown is on the tree nodes. All *M. lychnidis-dioicae* proteins are indicated by their respective MVLG designation. Similar lipase genes from other species are included, with the gene prefix denoting the species as follows: AN, *Aspergillus nidulans*; UM, *Ustilago maydis*; sr, *Sporisorium relianum*; MGL, *Malasezzia globosa*. Scale corresponds to substitutions per site. **b–e** Treatment of mated *M. lychnidis-dioicae* cells with various lipids. **b** water-treated mated cell control (Inset: higher magnification showing cells with conjugation bridge, black arrows); (**c**) mated cells treated with commercially-available corn oil (Inset: higher magnification of filamentation); D, mated cells treated with α-tocopherol; (**e**), mated cells treated with phytol. The green arrows show areas of filamentation emanating from mated cells after treatment with each specific type of lipid. **b** and **c** size bars, 25 µm; (**d**) and (**e**), size bars, 50 µm

polysaccharide monooxygenase of family AA9, nor any broad specificity β-glucanase of family GH131, nor any cellulose-binding module of family CBM1 could be found. This clearly shows that *M. lychnidis-dioicae* does not interact with nor digests cellulose during its interaction with plants, a finding confirmed by the failure of *M. lychnidis-dioicae* to grow on cellulose as a sole carbon source (Additional file 21). *Microbotryum lychnidis-dioicae* also completely lacks xylanase (no GH10, GH11, GH30), xyloglucanase (no GH74), and the enzymes for the cleavage of side-chains of xylan, xyloglucan and rhamnogalacturonan (no GH29, GH95, GH51, GH115), indicating that these cell wall polymers are not a carbon source for the fungus (Table 4, Additional file 20). Consistent with this prediction, *M. lychnidis-dioicae* failed to grow on xylan as a sole carbon source (Additional file 21).

Microbotryum lychnidis-dioicae was able to grow on pectin as a sole carbon source (Additional file 21),

Table 4 Selected CAZY expansions and depletions

	GT total	GH total	Beta glucan modification[a]	Cellulose related[b]	Xylan related[c]	GH26 beta-mannanases	Pectin/pectate lyases
Microbotryum lychnidis-dioicae	98	82	4	0	0	4	0
Average basidiomycete[d]	69.5	179.7	28.9	37.3	14.0	0.5	2.9
Piriformospora indica	65	194	31	98	41	1	12
Ustilago maydis	58	100	26	4	8	1	1
Puccinia graminis f. tritici	81	154	11	15	5	4	4
Melampsora laricis-populina	84	169	14	23	9	3	4

[a]GH16,GH72,GH81,GH128. [b]GH6, GH7, GH8, GH9, GH12, GH44, GH45, GH131, AA9, and CBM1. [c]GH10, GH11, GH30, GH29, GH95, GH51, GH115, GH74. [d]Average count for 33 basidiomycete genomes; see Additional file 20

although it does not break down pectin by the action of pectin/pectate lyases, as these enzymes are also absent from the genome (Table 4, Additional file 20). Instead, the genome harbors a suite of six family GH28 enzymes, which cleave polygalacturonic acid after its methylester groups have been removed by the action of six family CE8 pectin methylesterases. This CE8 family is present at high numbers in the two rust fungi and *M. lychnidis-dioicae*; the copy number amplification in *M. lychnidis-dioicae* appears to be due to tandem duplication, with one array of two genes and a second array of four genes. Four of the six CE8 copies have a predicted secretion signal, supporting a potential role in interacting with the host plant. Compared to 33 other basidiomycetes, *M. lychnidis-dioicae* stands out in having a significant expansion of its enzymatic arsenal for the breakdown of β-mannan, a polysaccharide present throughout plants but more abundant in flowers, siliques and stems [69]. *Microbotryum lychnidis-dioicae* encodes four candidate β-mannanases of family GH26 and a comparison with biochemically characterized enzymes shows that 10 out of its 19 GH5 enzymes also target β-mannan (the other *M. lychnidis-dioicae* GH5 enzymes target β-1,3-glucans (5 proteins), glucocerebrosides (3 proteins) and β-1,6-glucan (1 protein)). This β-mannan digestion arsenal is augmented by the presence of a GH2 enzyme, which shows a strong relatedness to characterized β-mannosidases.

Homogalacturonan is a major component (60 %) of plant pectin and the degradation pathway is required in several stages of plant development. One of these stages is anther dehiscence when pollen grains are released; this process requires pectinesterases and polygalacturonases. As indicated above, a total of six CE8 family pectin methylesterases are found in *M. lychnidis-dioicae*, of which four are predicted secreted proteins (MVLG_02682, 04072, 04073, 4074). Part two of the pathway requires polygalacturonase; a total of six *M. lychnidis-dioicae* proteins contain the polygalacturonase GH28 (PF00295) domain, of which MVLG_02498 is highly induced (over 1,000 fold) in MI-late. The homogalactorunan degradation pathway of *M. lychnidis-dioicae* may thus perform a similar role as pollen in anther dehiscence when the flowers bloom, since during teliospore formation of *M. lychnidis-dioicae*, host pollen is no longer available to perform that function.

Multiple classes of transporters are expanded in *M. lychnidis-dioicae* (Table 3), enabling uptake of diverse substrates. Major facilitator transporters, sugar transporters, and the fungal trichothecene efflux pump are present at high copy number relative to other Pucciniomycotina. A fungal trichothecene efflux pump, *TRI12*, was first described in *Fusarium sporotrichioides* as part of the gene cluster involved in trichothecene biosynthesis [70]; trichothecenes are a group of mycotoxins produced by various species of fungi. As the TRI12 domain is present at high copy number in *M. lychnidis-dioicae* and other Basidiomycetes are not known to produce trichothecenes, this suggests that this domain may have a role in transporting other small molecules.

Sugar transporters also play an important role in virulence of biotrophic plant pathogens, such as *Ustilago maydis* and several species of rust fungi. Specifically, a plasma membrane-localized sucrose transporter (Srt1) in *U. maydis* facilitated direct utilization of sucrose, thus eluding the plant defense mechanism [71]. The HeXose Transporter 1 (Hxt1) gene in the rust fungus *Uromyces fabae* is localized to haustoria to take advantage of that structure for sugar uptake [72]. The *M. lychnidis-dioicae* genome contains a total of 26 potential sugar transporters, with multiple high identity matches to Srt1 and Hxt1, which may fulfill similar roles.

One additional contrast between *M. lychnidis-dioicae* and the two plant pathogenic rust fungi examined suggests a difference in the relative importance of response to superoxides. Reactive oxygen species (ROS), including superoxides or H_2O_2 produced by the host plant, are a canonical part of the defense response to pathogens. *Microbotryum lychnidis-dioicae* is depleted in domains for Peroxidase (PF01328) and copper/zinc superoxide dismutase (PF00080); the two other Microbotryomycetes also lack proteins with these domains. Despite the reduced repertoire of such predicted proteins in *M. lychnidis-dioicae* relative to the rust fungi, six proteins (MVLG_00980, MVLG_03089, MVLG_03931, MVLG_02439, MVLG_03568, and MVLG_04684) were identified as containing peroxidase 2 (PF01328), peroxidase (PF00141), redoxin (PF08534), or Glutathione peroxidase (PF00255) domains. Of these predicted proteins, only MVLG_03089 was differentially expressed under the conditions examined and was up-regulated in MI-late relative to growth *in vitro* in rich medium. In addition, four predicted proteins with iron/manganese superoxide dismutase domains (PF02777 and PF00081), glutaredoxin (PF00462) or catalase (PF00199) domains were found (MVLG_00659, MVLG_06630, MVLG_06939, MVLG_04131). Finally, pathway analysis via MetaCyc predictions (http://fungicyc.broadinstitute.org/) suggests that *M. lychnidis-dioicae* contains components of the glutathione-mediated detoxification pathway: Glutathione transferase (EC 2.5.1.18: MVLG_05985, MVLG_04790) and membrane alanyl aminopeptidase (EC 3.4.11.2: MVLG_03673). However, there appears to be a missing component (3.4.19.9) in this pathway to facilitate formation of an intermediate of a glutathione-toxin conjugate. Biochemical and functional analyses will be required to determine the importance of these predicted enzymes in the ability of the pathogen to survive and flourish in its host.

Secreted proteins (SP) and candidate effectors

A total of 279 secreted proteins (SPs) were predicted in *M. lychnidis-dioicae* and their expression and conservation examined to identify candidates for interacting with the host (Table 5, Additional file 22). Among the 71 SPs that were smaller than 250 amino acids (small secreted proteins, SSPs), 46 were species specific in our comparative set and further do not share sequence similarity (e-value <1e-3) with any protein in the NCBI protein database. SSPs indeed often appear species-specific, likely because they co-evolve rapidly in an arms race with their hosts [43]. Notably, 48 of the SSPs were significantly up-regulated during plant infection (MI-late compared to rich media), but were not differentially expressed when comparing expression on rich and nutrient limited agar (Table 5), suggesting that these SSPs may play a specific role during plant infection.

Several cysteine-rich multigene families were identified among predicted secreted proteins. In some cases these families include tandemly duplicated genes; the MVLG_04105 family contains 4 members predicted to be SSPs (MVLG_04105, 04106, 04107, and 04096), three of which are adjacent in the genome on the mating-type chromosome (see below). Although these proteins lack PFAM domains, two of these are induced during infection. An additional family of Cys-rich proteins with nine members has a subset of six clustered in the genome (MVLG_05513, MVLG_05514, MVLG_05515, MVLG_05533, MVLG_05534, MVLG_05538). Seven of the nine proteins in this family were predicted to be secreted, yet their expression was highly variable, with two up-regulated in nutrient limited conditions and two down-regulated during infection. A small subset of Cysteine-rich proteins contains known protein domains. Two proteins (MVLG_02283 and MVLG_02288) contain the Cysteine-rich secretory protein family domain (PF00188). In addition, a total of 9 proteins contain the fungal-specific Cysteine rich CFEM domain (PF05730). All of these CFEM proteins were predicted to be secreted; four of these were significantly induced and three were repressed in MI-late relative to rich and nutrient limited agar.

To identify genes that could provide a mechanism for linking flower development to fungal development, we compared expression of the predicted secreted proteins with the *S. latifolia* EST library produced from flowers [73]. A total of 37 genes share sequence similarity with *S. latifolia* ESTs; ten secreted proteins matched plant ESTs with at least an e-value of e-19. One *S. latifolia* EST (09F02) showed similarity (BlastX, e-value < 7e-24) with two *M. lychnidis-dioicae* proteins (MVLG_02043 and MVLG_02936); the best match, MVLG_02043, encodes a predicted secreted gamma-glutamyltranspeptidase (GGT). Another EST (33C05) shared sequence similarity with two *M. lychnidis-dioicae* proteins (MVLG_00083, MVLG_02276); these in turn share similarity with the expansin family of "ripening related" proteins and were down-regulated during either growth on nutrient limited agar or late in infection (MI-late). The precise function of plant expansins is poorly defined at the molecular level, and the predicted function of the similar fungal proteins is even less well established. However, one such expansin-related protein in *Laccaria bicolor* was recently found to be expressed specifically in the extracellular matrix (ECM) of symbiotic tissues and localized within the fungal cell wall [74].

Table 5 Properties of predicted secreted proteins

Protein length	SP count	Induced in MI late	Repressed in MI late	Induced in water	Repressed in water	FPKM > 1	Highly expressed only in MI late
100	15	4	0	0	0	11	3
150	23	5	0	1	0	20	4
200	15	5	0	2	1	14	
250	18	5	1	1	0	14	3
400	62	8	10	8	3	59	2
500	56	8	7	5	2	55	
600	33	4	2	7	0	33	
700	23	7	1	0	0	23	
800	7	0	2	1	0	7	
900	6	1	0	0	0	6	1
1000	13	2	2	0	1	13	
2000	6	0	1	1	0	6	
3000	2	0	0	0	0	2	
Total	279	49	26	26	7	263	

Two cysteine-rich secreted proteins (MVLG_02288 and MVLG_02283) matched a *S. latifolia* flower EST annotated as having similarity with the plant PR-1 class of pathogenesis related proteins (PRs); these are proteins defined as encoded by the host plant but induced only in pathological or related situations (possibly of non-pathogenic origin). To be included among the PRs, a protein must be induced upon infection but not necessarily in all pathological conditions [75]. Another *S. latifolia* gene of interest, *SLM2,* is expressed in the stamens of smut infected flowers but not uninfected flowers [22]; four *M. lychnidis-dioicae* proteins showed blast similarity with e-value less than e-6. All of the hits were hypothetical proteins that contained the SRF-type transcription factor domain (PF00319), and three of the four were predicted to be targeted to the nucleus (MVLG_04297, MVLG_06278, and MVLG_07052).

Response to oxidative environments

Laccase-like multi-copper oxidase proteins are capable of degrading phenolic compounds like polymeric lignin and humic substances [76] and may be involved in the interacton of fungal pathogens with their host plants. Four proteins in *M. lychnidis-dioicae* contain the three multi-copper oxidase (MCO) domains (PF07731, PF07732, PF00394). Two MCO proteins were predicted to be secreted, and another MCO was predicted to have a GPI-anchor to the membrane. The fourth MCO was a membrane-anchored protein (MVLG_03092), with the N-terminus of the polypeptide outside the cell.

The glyoxal oxidase catalyses the oxidation of aldehydes to carboxylic acid and is an essential component of the extracellular lignin degradation pathway of the root rot fungus, *Phanerochaete chrysosporium*. Of a total of seven *M. lychnidis-dioicae* proteins that contain the glyoxal oxidase N-terminal domain (PF07250), two are predicted to be secreted and four have a predicted GPI-anchor. While two of these glyoxal oxidase genes are adjacent in the genome, they do not form a gene cluster with any MCO as observed at the lignin peroxidase gene cluster in *P. chrysosporium*.

Genes similar to plant hormone synthesis genes

Since *M. lychnidis-dioicae* infection of female *S. latifolia* hosts can alter normal flower development so as to produce pseudomale flowers, we investigated whether the genome might contain genes for pathways that could be associated with such changes. The predicted *M. lychnidis-dioicae* protein database was examined for components of biosynthesis pathways of eight plant hormones (Additional file 23), as well as for other signaling pathways that could have an impact on host gene expression or development. Based on sequence similarity to components for these pathways from plants and microbes, as well as additional confirmation of some complete pathways using MetaCyc predictions (http://fungicyc.broadinstitute.org/), we found evidence that *M. lychnidis-dioicae* encodes enzymes that could participate in hormone biosynthetic pathways, such as polyamine biosynthesis pathways, which produces compounds known to play developmental and stress-response roles in plant physiology [77] (see proposed model in Fig. 5). If these enzymes are, in fact, promoting hormone biosynthesis pathways, one possible explanation is that precursors for these pathways are provided by the plant. Alternatively, the potential components of these pathways we have identified are used by the fungus for functions other than manipulating host development. For example, cytokinin degradation could be mediated by a predicted FAD-oxidase (MVLG_04134), if this enzyme can function as a cytokinin dehydrogenase (EC 1.5.99.12); however, this predicted protein most closely resembles other fungal D-lactate dehydrogenases based on sequence similarity. Similarly, a predicted 2β-dioxygenase (EC1.14.11.13) (MVLG_00840) could be involved in gibberellin inactivation via hydroxylation, although this protein falls more generally into the 2OG-Fe(II) oxygenase superfamily.

Another potential set of pathways for plant signaling involves the production of glycerol lipids, such as diacylglycerol (DAG) or triacylgycerol (TAG). Many studies have demonstrated the importance of compounds like DAG in mammalian signaling [78], and DAG is important in pollen tube elongation in some plant species [79]. Such pathways often involve the action of phospholipase C; a predicted phospholipase C gene (MLVG_07108) and potential phospholipase A (MVLG_03207, MVLG_03384, MVLG_4789) and D (MLVG_01917, MVLG_03610) homologues are found in the genome. Therefore, the organism contains the proteins necessary for synthesizing 1,2-diacylglycerol, 1,2-diacyl-*sn*-glycerol-3-phosphate, 2-lysophosphatidylcholine, and 1-lysophosphatidylcholine. Based on MetaCyc prediction, *M. lychnidis-dioicae* possesses the requisite components of the CDP-diacylglycerol biosynthesis I pathway. In addition, a number of the secretory lipases whose PFAM domain is enriched in *M. lychnidis-dioicae* (PF03583, see above) are implicated in these phospholipase pathways.

Gene expression changes during infection and in response to lipids

To focus on genes potentially important for infection and interaction with the host, we identified genes whose expression was altered in MI-late. A total of 1,254 genes were differentially expressed in MI-late compared to either rich or nutrient limited agar; of these, a common set of 307 genes were induced in MI-late and 126 were repressed in MI-late compared to both other conditions (corrected p-value <0.001, Additional file 24, Methods). Of the 138 genes in both comparisons with a predicted

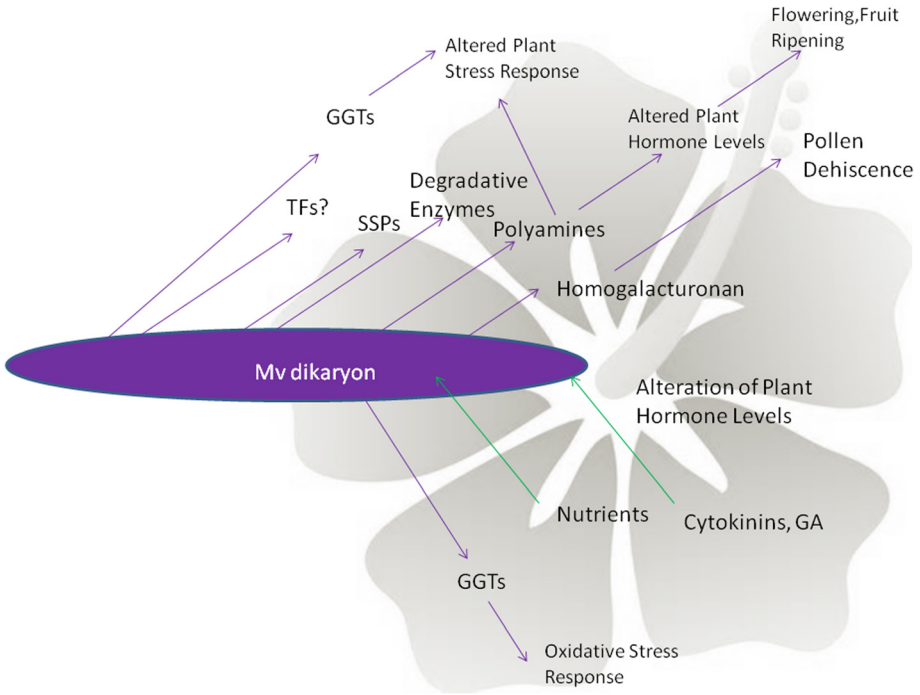

Fig. 5 Model of *Microbotryum lychnidis-dioicae* interactions with its host. The potential pathways identified in *M. lychnidis-dioicae* based on inventory of the genome were used to predict products potentially secreted or taken up by the fungus that could affect host development (see text for more detailed description). GA, gibberellic acid

PFAM domain, transporter domains were most frequently observed; a total of 20 of the 138 functionally assigned proteins corresponded to transporters, with MFS, sugar, OPT, and amino transporters each represented by two or more genes (Additional file 25). Carbohydrate active enzymes, kinases and transcription factors were also highly represented in these MI-late induced genes.

Several different classes of transporters were transcriptionally induced during infection, potentially promoting the uptake of small molecules and nutrients from the host plant. Significant enrichments include domains found in MFS transporters (Fisher's exact test, corrected p-value < 0.001) and sugar transporters ($p < 0.005$), the largest classes of transporters in *M. lychnidis-dioicae* and many fungi. The oligopeptide transporter protein was also enriched at a lower level of significance ($p < 0.1$) and four predicted proteins with this domain (MVLG_03106, MVLG_07217, MVLG_03161, and MVLG_00149) were up-regulated in MI-late. An MFS domain-containing protein of note up-regulated in MI-late was a nitrite transporter (TIGR00886.2; MVLG_00642); this gene is linked to two other genes associated with nitrate assimilation, a nitrite reductase (MVLG_00638) and a nitrate reductase (MVLG_00637). All three genes involved in nitrate assimilation were significantly up-regulated during infection.

In evaluating the expression of the 279 proteins predicted to be secreted, 48 were induced *in planta*. Several cysteine-rich secreted proteins were induced during infection. A pair of linked cysteine rich small secreted proteins (MVLG_04106 and MVLG_04107) was induced during infection. Four other proteins (MVLG_00115, MVLG_00802, MVLG_00815, and MVLG_00859) with the Cys-rich CFEM domain-containing family (PF05730) were significantly induced in infection. These proteins are good candidates for being potential effectors, based on proteins with similar properties that have been shown to be effectors in other systems [80].

Cell wall degrading enzymes may play a role in the infection, particularly during the stage where the fungus causes necrosis of host plant tissue. A total of nine glycoside hydrolases were up-regulated in MI-late; among GH families, GH28 polygalacturonase proteins are mostly highly enriched during infection ($p < 0.08$). Polygalacturonase is required for the second part of the homogalacturonan pathway implicated in pollen dehiscence. MVLG proteins that contain a glyoxal oxidase domain (PF07250) were also significantly enriched ($p < 0.009$) in genes induced during MI-late.

Mated cells of *M. lychnidis-dioicae* respond to corn oil, in addition to the other lipids previously reported [45]. This supports the hypothesis that lipid response may be important for the development of this species.

Although we observed no alteration of phenotype for un-mated haploid cells treated with lipids, three predicted cytoplasmic proteins with a secretory lipase domain (MVLG_07229, 07284, and 07291) were highly induced in haploids grown on nutrient limited agar compared to rich media or MI-late infections. To validate the expression levels from RNA-Seq data, relative levels for lipase genes were measured for mated and unmated cells, grown in nutrient limited, rich media, or treated with phytol by qRT-PCR (Additional file 18). Notably, when mated haploid cells were treated with phytol, after 12 h treatment, two genes (MVLG_00914, MVLG_05549) displayed substantial up-regulation, while the remaining three either increased slightly or decreased (Additional file 18). This could reflect priming of cells that are ready to mate for plant cues that would ultimately lead to stable dikaryon formation after successful mating.

Discussion

This analysis highlights genomic features of *M. lychnidis-dioicae* that reflect its particular lifecycle. *Microbotryum lychnidis-dioicae* grows as a biotroph, similar to the rust fungi, for most of the plant infection cycle. Notably in its capacity as a castrating pathogen, *M. lychnidis-dioicae* also causes necrosis that appears to be limited to developing flowers. Our analysis revealed that gene content contains a different profile than purely biotrophic or necrotrophic plant pathogens. This included a global number of CAZymes that is larger than other biotrophic fungi (Additional file 20); however, at the same time it shows complete loss of many CAZyme families that target the plant cell wall. This is consistent with a primarily intercellular colonization pattern of host apical meristems [81], though it raises questions about how appressorium penetration is accomplished.

Necrotic growth stages may be enabled by the subtilases, laccases, and copper radical oxidases, which are ligninolytic enzymes [82–84]; subtilases, in particular may play a key role in regulating the activities of laccase [76]. By contrast, certain CAZymes present in the biotrophic rust fungi are absent in *M. lychnidis-dioicae*; in particular the absence of cutinases suggests that penetration of the plant surface by *M. lychnidis-dioicae* does not require cutin degradation. This may reflect the fact that the normal portal of entry for infection is via the flower, a tissue that poses less of a barrier to the fungus. Of the plant polysaccharides that constitute a carbon source for many fungi, *M. lychnidis-dioicae* has lost the ability to digest cellulose, xylan, xyloglucan, and the highly substituted forms of pectin (rhamnogalacturonan). Retention of enzymes that breakdown polygalacturonic acid and β-mannan, components of pollen tubes and flowers, respectively, illustrates the high degree of specialization of this fungus.

Whereas smut fungi generally cause necrosis as a space-making process during sporulation [85], *Microbotryum* anther smuts are more aptly characterized as a growth-altering parasite [86]. At the stage of anther development, *M. lychnidis-dioicae* causes abortion of pollen production, and replacement by the diploid teliospores for dispersal by pollinator species *[*87*]*. Moreover, in the female, atrophy of pistils occurs during infection. There are a number of candidate pathways that might participate in pollen tube elongation or in blocking this process. Aspartyl proteases are involved in pollen tube elongation or prevention by *Metschnikowia reukaufii* [88]; seven candidate aspartyl proteases are predicted in *M. lychnidis-dioicae*. Additionally, consistent with transcriptional up-regulation of some component enzymes for the homogalactorunan degradation pathway of *M. lychnidis-dioicae*, this pathway may take over the role of pollen in anther dehiscence when the flowers bloom, since during teliospore formation of *M. lychnidis-dioicae*, host pollen is no longer available to perform that function Fig. 5.

Finally, since *M. lychnidis-dioicae* appears to lack a large repertoire for dealing with host-generated defenses that utilize reactive oxygen species (ROS) one could expect that the interaction with its host normally does not elicit such host responses or the fungus actively down-regulates them. Paradoxically, *M. lychnidis-dioicae* also appears to secrete some proteins that would serve to bring on this plant response by serving as prooxidants (*e.g.*, secreted gamma-glutamyltranspeptidases). Further elucidation of the roles of the predicted peroxidases superoxide dismutases, and glutathione-mediated detoxification pathway components will require functional analyses to evaluate their biological importance, if any, in the interaction of the pathogen with its hosts.

Secretory lipases represent one of the most significantly expanded gene families in *M. lychnidis-dioicae* compared to the other fungi examined. While haploid cells show no outward phenotype alteration in response to lipids like phytol, several secretory lipases of *M. lychnidis-dioicae* in haploid cells are up-regulated on nutrient limited agar. We propose that regulation of secretory lipases primes haploid cells so that, after mating, they can respond to the appropriate plant-derived cues (including lipids) to progress to the next developmental stage, stable dikaryotic hyphae. In fact, most of the secretory lipases we investigated were up-regulated in mated cells and when such cells were exposed to phytol (Additional file 18).

The identification of genes induced during infection (MI-late) suggests their involvement in host invasion and evasion of physical and chemical defense systems of the

plant. A role for CAZymes, including pectin methylesterases and GHs, during plant infection in *M. lychnidis-dioicae* is further supported by increased transcription in the MI-late sample. Analysis of the gene expression profile of both wheat stem and poplar rust (*P. graminis* and *M. larici-populina*) also found that many CAZyme genes related to cell wall degradation were up-regulated during plant infection [39]. In addition to the CAZymes, *M. lychnidis-dioicae* shows significant induction of diverse transporters during plant infection, which may be critical for uptake of small molecules during biotrophic growth.

As in other plant pathogenic fungi, candidate effectors in *M. lychnidis-dioicae* were predicted based on predicted localization, expression during infection, and sequence conservation. Notably, SSPs are located closer to TEs than other protein coding genes, suggesting that this could impact SSP expression or duplication. TEs probably play a role in the expansion of such a family. Indeed they contribute to genome rearrangements and gene duplications [89]. In *Fusarium oxysporum* f. sp. *lycopersici,* effector genes are present on chromosomes or regions enriched for DNA transposons [90]. Some secreted proteins are predicted to act on host cell walls and proteins, either for the remodeling of host development in the flower or for acquisition of additional nutrients by the fungus.

In assembling sequence of the a_1 mating-type chromosome, we characterized how the content of this allosome differs from autosomal regions, contrasting the non-recombining regions of the mating-type chromosome with the pseudo-autosomal regions (PARs) capable of recombination and with autosomes. Both the lower gene density and higher transposable element content in the non-recombining region of the mating-type chromosome relative to autosomal regions are consistent with a reduced efficiency of purifying selection due to the suppression of recombination, as occur on non-recombining sex chromosomes [59]. The two recombining PARs of the mating-type chromosome displayed TE content and gene density more similar to autosomes than the non-recombining part of the mating-type chromosome. By contrast, we observed no difference in codon usage nor in the GC content at third codon positions between autosomes and the mating-type chromosome, though this has been observed in the non-recombining part of the fungal mating-type chromosome of *Neurospora tetrasperma* [31]. Overall the maintenance of homologous meiotic pairing and recombination in PAR regions may render them more similar to autosomes than allosomes with respect to evolutionary forces of selection and drift. However, their physical linkage to the non-recombining region of mating-type chromosomes suggests intermediate modes of evolution [91].

Conclusion

Altogether, this study provides an in-depth genomic portrait of a fungal castrating, biotrophic plant pathogen reflecting its unique life cycle. In particular, the unique absence of enzyme classes for plant cell wall degradation and maintenance of enzymes that break down components of pollen tubes and flowers provides a striking example of biotrophic host adaptation. In addition, while there are fewer enzymes to digest cellulose, xylan, xyloglucan, and highly substituted forms of pectin, as well as proteins that could protect the fungus from oxidative stress, the repertoire of predicted cell wall modifying enzymes and those that could manipulate host development has expanded (see model in Fig. 5). Given the place of *M. lychnidis-dioicae* in a large species-complex with a vast host species pool, the insights from this genomic and transcriptomic analysis combined with comparative approaches with other members of the *Microbotryum* species complex will be most informative on the evolutionary processes involved in a radiation and specialization on a wide array of plant species from different genera.

Methods

Microbotryum lychnidis-dioicae lineage(s) and *Silene latifolia* host(s)

The *focal* lineage of *M. lychnidis-dioicae* for this work is the most studied in the context of disease ecology ("Lamole strain": GenBank I00-15Lamole.1; [9, 11]) and belongs to the recently-refined species designation *M. lychnidis-dioicae*, parasitizing *Silene latifolia*. From the original isolate, haploid sporidial strains were generated via micromanipulation of the meiotic products from a single tetrad, yielding the strains Lamole p1A1 and p1A2 that differ in electrophoretic karyotypes only in the mating-specific chromosome. For the work in this report, the haploid p1A1 strain (mating-type *a1*) was used as the source of the *focal* genome. Additionally, the *focal* genome contains size-heteromorphic sex chromosomes that share many features with sex chromosomes in plant and animal systems [11]. The corresponding a_2 strain, p1A2 was used together with its partner strain p1A1 in plant infections and in RNA-Seq analysis.

High molecular weight DNA preparation

Microbotryum lychnidis-dioicae Lamole p1A1 was grown on yeast peptone dextrose media (YPD; 1 % yeast extract, 10 % dextrose, 2 % peptone, 1.5 % agar) at room temperature for 5 days and ultimately extracted using a phenol chloroform isoamyl extraction method [92]. Harvested fungal cells were ground into fine powder using liquid nitrogen and resuspended in OmniPrep Genomic Lysis Buffer (G-Biosciences, cat no: 786–136) according to manufacturer's recommended tissue to reagent ratio.

The sample was heated in a 55-60 °C water bath for 15 min after extensive vortexing. Chloroform was added to the sample after allowing it to cool to room temperature. Using wide bore tips thereafter, 3–4 extractions using phenol chloroform isoamyl (25:24:1) solution were performed, followed by a final extraction with chloroform isoamyl (24:1) solution. Nucleic acid was then precipitated and the pellet was rinsed twice with ice-cold 70 % ethanol and then air-dried. The pellet was rehydrated using Tris-EDTA buffer (pH 8.0) (100 μl per 100 mg of ground tissue powder used) and treated with RNase (*Longlife* RNase, 5 mg/ml; G-Biosciences); 1 μl of RNase was added for every 100 μl of TE buffer used.

RNA isolation

Haploid cells

Haploid fungal cells of either Lamole p1A1 or p1A2 grown separately under rich conditions for 5 days on yeast peptone dextrose media (YPD; 1 % yeast extract, 10 % dextrose, 2 % peptone, 2 % agar) at room temperature were harvested for RNA extraction. RNAs were checked for quality using a Bioanalyzer (Agilent). The RNAs were then pooled in equal quantity (in terms of mass) based on the Bioanalyzer quantification. The same procedure was also performed for the haploid cells grown separately on 2 % water agar for 2 days, to compare the gene expression when haploid cells were subjected to nutrient free environment without the mating partner. Again, haploid cell samples, p1A1 and p1A2, were extracted as independent samples, and then mixed in equal proportion for RNA sequencing.

Host plant infection for RNA-Seq (MI-late stage)

Silene latifolia seeds (harvested in Summer 2009 from Lamole, Italy) were sterilized and hydrated by soaking them in a sterilizing solution (40 % household bleach, 20 % absolute ethanol and 1 drop of Triton X-100 as surfactant per 50 ml of solution) and washing five times in sterile distilled water, for 2 min per wash with constant agitation. Each seed was then individually planted in closed milk jars on sterile 0.3 % phytagar (Life Technologies), 0.5× MS (Murashige and Skoog) salts (Sigma-Aldrich) and 0.05 % MES (2-(N-morpholino)ethanesulfonic acid) buffer (Brand). Each jar was placed at 4 °C for 5 days to synchronize germination. The jars were then transferred to a 20 °C growth chamber with 13 h of fluorescent light daily. Germination starts within 3 days with the appearance of the radicle. When the seedlings were 15 days old, they were transplanted into 2″ square pots filled with Sunshine MVP Professional Growing Mix (Sun Gro Horticulture Canada Ltd, cat no. 02392868) soil and replaced into the growth chamber. Humidity was kept high initially using dome covers and flood trays. Seedlings were gradually exposed to chamber environment for increasing amounts of time daily in order for the seedling to harden and adapt to the lower humidity. The plants were transplanted to 4″ round pots when they began to bolt at about 30 days old. They were further transplanted into 7″ round pots when they had almost attained maximum height or when the volume of soil was not sufficient to provide hydration requirement for the plant. The plants were watered every other day with 100-ppm fertilizer (Peters Professional® 15-16-17 Peat-Lite Special, Formula no: S12893).

To infect the host plants, mated cells were prepared as follows. Haploid cells grown on rich media were harvested and resuspended in distilled water, adjusted to a concentration of 1×10^9 cells/ml in equal proportion before being spotted onto nutrient-free solid agar media (2 % agar) in 50 μl spots. The plates were allowed to dry and then incubated at 14 °C for about 48 h. Cells were inspected for conjugation tubes under the microscope and then 5 μl of 1×10^6 cells/ml resuspended in distilled water with anionic surfactant was pipetted onto the floral meristem when the cotyledon was fully developed (11–12 days). Infection was determined by the consistent blooming of fully smutted flowers. The floral buds were staged according to previous literature [35] under a dissecting scope (Nikon, Model: SMZ-U) and parts of the floral buds were measured with a glass stage micrometer (Imaging Research, Inc.).

Tissue originating from host plants was collected in RNA-later RNA stabilizing reagent (QIAGEN, cat no: 76106) and left at 4 °C overnight until sufficient tissue had been collected for the RNA extraction. The solution was removed before storing the sample at −80 °C. For infected male plants, we collected floral tissue from buds ranging in size from 4 mm to fully open smutted flowers. These tissue samples were pooled to yield the source for 'MI-late' RNA used in RNA-Seq analysis. Thus, they provide a pooled average picture of gene expression for this size range of infected tissue.

All RNA samples were extracted using the RNeasy Plant Mini Kit (QIAGEN, cat no: 74904) according to the manufacturer's instructions. DNase treatment was performed using Ambion's TURBO DNA-free (Applied Biosystems, cat no: AM1907), also according to manufacturer's instruction. For quality assessment before Illumina sequencing, 5 μg of DNase-treated RNA was reverse transcribed with SuperScript III First Strand Synthesis System for RT-PCR (Life Technologies, cat no: 18080–051). PCR was performed using TaKaRa Ex Taq Hot-Start DNA Polymerase (Takara, cat. no: RR001B) using 25 μl reaction volume. To check for DNA contamination and possible inhibitory substances in the RNA, three sets of housekeeping primers (Eurofins/MWG/Operon) were used. Amplification of a region of the *S. latifolia* partial *wdr1x* gene for a putative WD-repeat protein (GenBank IDs Y18519, Aj310656) was used to assess host cDNA and contaminating genomic DNA; the forward primer, 5′- CTCTG

CTGGAGGTGGAACAT-3′ and reverse primer, 5′- AG CACTGAACACCCCAACTT-3′; in this case a 253 bp fragment would be produced for cDNA, vs. a 335 bp fragment for genomic DNA. Targeting the *M. lychnidis-dioicae mepA* gene, we used as forward primer, 5′- CT TTTGCGTAGGAAGAATGC-3′ and as reverse primer, 5′- AGCACTGAACACCCCAACTT-3′; this combination yielded a 532 bp fragment from cDNA, compared with a 1039 bp fragment from genomic DNA. The other primer combination targeted the *M. lychnidis-dioicae* beta-tubulin gene, with forward primer, 5′- CGGACACCGT TGTCGAGCCT -3′, and reverse primer, 5′- TGAGGT CGCCGTGAGTCGGT-3′, yielding a 150 bp fragment from cDNA compared with a 215 bp fragment from genomic DNA. The PCR program was 30 s at 94 °C, 30 s at 60 °C and 1 min at 72 °C for 35 cycles. RNA quality was also evaluated using an Agilent BioAnalyzer; all samples had RNA integrity number scores of at least 7.8, indicating highly intact RNA.

Treatment of cells with lipids

Haploid fungal cells of Lamole p1A1 and p1A2 were grown separately under rich conditions for 5 days on YPD at room temperature, then harvested into sterile distilled water. The concentration was adjusted and resuspended in equal proportions in each type of medium, to achieve a final concentration of 1×10^9 cells/ml.

To allow better solubility of the lipids, 50 % ethanol was used as the solvent for the lipids. We used 1 % corn oil (Carlini), (±)-α-tocopherol (Sigma-Aldrich, cat no: T3251-5G), and phytol (Sigma-Aldrich, cat no: P3647), dissolved in the solvent and used as the resuspension media for the fungal cells. The mixtures were then spotted in 50 μl spots onto 2 % water agar and allowed to mate for 2 days at 14 °C. The cells were then observed under the microscope for conjugation tubes and filamentous structures. The solvent served as the control media to ensure that changes in phenotype were not due to the ethanol present.

Genome and transcriptome sequencing, assembly, and annotation

For genome sequencing, we constructed three libraries (Additional file 1) with different insert sizes and sequenced each using 454 Technology. The reads were assembled with Newbler (version MapAsmResearch-04/19/2010-patch-08/17/2010). The total assembly size of 26.1 Mb in scaffolds includes 99.6 % of bases of at least Q40 quality; gaps encompass 3.45 % of the total scaffold length.

For RNA-Seq, we purified polyA RNA and constructed a strand-specific library for each sample as previously described [93, 94] and sequenced each with Illumina technology generating 76 base paired reads. Across the three libraries, 96 % of reads met the Illumina Passing Filter (PF) quality threshold. Read alignment rates to the genome varied between three libraries; for the rich and nutrient limited samples, 90 % or 89 % of reads aligned respectively; for the MI-late sample only 23 % of reads aligned. This was expected as these samples also contain the host *Silene* RNAs. To assemble transcripts for use in annotation, RNA-Seq reads were aligned to the assembly with Blat, and then assembled using Inchworm [95] in the genome-guided mode.

To predict genes, we first generated a high confidence training set of 775 transcripts of at least 900 nt using Genemark [96] and the assembled RNA-Seq data. This was used to train Augustus [97] and Glimmerhmm [98]. RNA-Seq data was processed by PASA [99] to generate longer transcripts, and ORFs of at least 600 nt were predicted. Available ESTs from Genbank and Microbase were also utilized. EVM [99] was then used to select a preliminary gene set from the *ab initio* gene calls (Augustus, Genemark, GlimmerHmm, and SNAP), Genewise [100], ESTs, PASA ORFs, and the training set, with the highest weight given to the RNA-Seq based PASA ORFs. The EVM gene set was compared to the PASA ORFs, and non-repeat genes found only in the PASA set were added to the gene set. Finally, PASA was run on the final gene set to all updates of all gene structures with the RNA-Seq and incorporate alternatively spliced transcripts. Genes likely corresponding to repeats were filtered out using TransposonPSI (requiring 1e-10 and 30 % overlap), PFAM domains, Blast similarity to repetitive elements and 7 or more Blast hits to other genes in the set. Genes with flagged features (proteins ≤50 aa, internal exons ≤6 nt, introns ≤20 nt, introns ≥1500 nt, exons spanning gaps in the assembly, internal stop codons, overlapping other coding sequences, overlapping ncRNAs (tRNAs, rRNA, or other)) were manually reviewed and corrected where supported by the evidence. Gene names were assigned with the locus prefix MVLG.

The completeness of the gene set was evaluated by examining the conservation and completeness of core eukaryotic genes (CEGs, [36]). We compared the gene set of *M. lychnidis-dioicae* and of the 18 other fungal genomes used in comparisons to the CEGMA set, and identified Blast hits above and below the recommended 70 % coverage threshold (Additional file 3). A tool for streamlined analysis and visualization of conservation of CEGs is available on SourceForge (http://sourceforge.net/projects/corealyze/).

Differential expression analysis

We used differential expression analysis scripts in the Trinity pipeline [95, 101] to process RNA-Seq data generated from the three conditions (nutrient limited, rich, and MI-late). Briefly, we first extracted protein coding

gene sequences from the *M. lychnidis-dioicae* genome sequence based on coordinates of gene models, and added 100 bases of flanking sequence on each side to approximate UTRs. Then the RNA-Seq reads from each of the three samples were aligned to the extracted coding sequences using bowtie [102]. The alignment files were used to quantify transcript abundances by RSEM [103]. Differential gene expression analysis was conducted using edgeR with TMM normalization [104, 105] using a corrected p-value [106] cutoff of 1e-3. In comparing all pairs of the three conditions, a total of 1,413 genes were differentially expressed across the comparisons (Additional file 24).

TE detection and annotation

Two pipelines from REPET package (http://urgi.versailles.inra.fr/tools/REPET) were run on the *M. lychnidis-dioicae* contigs. The TEdenovo pipeline [107] was used to search for repeats in the genome. The first step uses Blaster with the following parameters [identity > 90 %, HSP (High Scoring segments Pairs) length >100b & <20Kb, *e*-value ≤ 1e-300]. HSPs found were clustered by 3 different methods: Piler [108], Grouper [109] and Recon [110]. Multiple alignments (MAP) of 20 longest members of each cluster (918 clusters) containing at least 3 members were used to derive a consensus. Consensus sequences were then classified based on their structure and similarities against Repbase Update (v15.11) [111] before removing redundancy (Blaster + Matcher). Consensus sequences without any known structure or similarity were classified as "Unknown".

The library of 425 classified consensuses provided by the TEdenovo pipeline was used to annotate TE copies in the whole genome using TEannot pipeline [109]. Annotation is based on 3 methods (Blaster, Censor, RepeatMasker). HSPs provided were filtered and combined. Three methods (TRF, Mreps and RepeatMasker) were also used to annotate SSR. TE duplicates and SSR were then removed. Finally a "long join procedure" [107] was used to address the problem of nested TEs. This procedure finds and connects fragments of TEs interrupted by other TEs inserted more recently to build a TE copy. The nesting patterns of such insertion must respect the three constraints: fragments must be co-linear (both on the genome and the same TE consensus reference), have the same age and separated by younger TE insertion. The identity percentage with the reference consensus is used to estimate the age of a copy. Using results of this first TEannot pipeline, we filtered out 111 consensus sequences without full-length copy in the genome. A copy may be built using one or more fragments joined by the TEannot long join procedure. We ran a second TEannot using the 306 consensus elements remaining after filtering out TE consensus without any full-length copy.

We also used gene prediction and proceeded to manual curation in order to improve TE annotation. We removed TE copies of consensuses that were identified as host genes. Indeed, these consensuses built from family of repeats containing at least 3 members and classified as unknown by the TEdenovo pipeline has been predicted as host genes belonging to multigenic families. We also filtered out TE copies not satisfying the criteria (identity > 0.8 & length > 150 & identity*length > 150) and those corresponding to low complexity region of the consensus Mivi-B-R219-Map20_classI-LINE-incomp very highly represented in the genome included in predicted genes. The few copies just over these thresholds were manually removed, depending on their location in genes and evidence of the gene (PFAM domain not related to TEs).

Search for signature of transition type (C to T) mutation bias

We performed pairwise alignments between each copy and respective consensus to finally provide multiple alignments for each family (consensus) using in-house scripts. TE copies with less than 80 % of identity with consensus and smaller than 400 bp were filtered out. We also filtered out TE families with less than 5 sequences in the multiple alignments. RIPCAL [112] was run on each multiple alignment to count both potential single mutations (transitions and transversions) and di-nucleotide target used in all possible transitions. Results were analysed using in-house R scripts to select most reliable mutated copies (if transition rate > 2 * transversion rate). For 40 % of copies exhibiting a transition mutation bias (of 2298 total copies, 179 consensus families (Additional file 10)), we considered that dinucleotide targets ($CA + TG^1$, $CC + GG^1$, $CG + CG^1$, $CT + AG^1$; ^{1} for reverse complement), were preferentially used if they represent a minimum of 30 % of the addition of the four possible. We expect 25 % of each if they are equiprobable.

Measurement of distance between genes and TEs

We computed the distance from each gene to the closest TE (case 1), or from each TE to the closest gene (case 2) using getDistance.py from S-MART package [113]. Only distances up to 10 kb were considered. For case 1, we also compared the subset of genes encoding predicted secreted proteins with the set of all other genes for different classes of distance intervals. For the case 2, we compared different TE categories in two classes of distance intervals (<1kbp and >1kbp). A chi^2 test of homogeneity (Pearson's chi-squared) was computed to test that the observed difference between the sets did not occur by chance (p-value < 0.05). The graphics and statistical test were performed using in-house R-Scripts.

Identification of the mating-type chromosome supercontigs

Using the same haploid genotype from which the whole genome was sequenced, DNA enriched for the a_1 mating-type chromosome was isolated from agarose gels after pulsed-field electrophoresis. With this technique, the isolated bands could include of small amounts of autosomal fragments that co-migrate with mating-type chromosomes, though these preparations have been shown to be strongly enriched for mating-type chromosome DNA [12]. The isolated DNA was amplified by whole genome amplification (REPLI-g kit, QIAGEN). The DNA was sequenced using 2- and 5 kb-insert size mate-paired libraries and 454 technology version Titanium (www.roche.com). Assembly of non-duplicated reads and excluding autosomal contamination yielded ~20-fold coverage.

The assembly was compared to the a_1 whole haploid genome sequence using NUCmer (http://mummer.sourceforge.net/) to validate assemblies and identify scaffolds corresponding to the mating-type chromosome in the whole genome assembly. The regions corresponding to autosomal contamination were identified by uneven and low read coverage and were excluded from further analyses; mitochondrial DNA was also excluded (GenBank NC_020353). All whole-genome scaffolds with more than 20-fold depth of the enriched mating-type chromosome sequence were confidently assigned to the mating-type chromosomes (Additional file 11). It was not possible to anchor the scaffolds onto the available optical map of the a_1 mating-type chromosome [12] due to the small sizes of the contigs relative to spacing of the restriction enzyme cut sites in the map. Annotation for mapped supercontigs regions was parsed from the genome-level annotation.

TE content was assessed using de novo TE annotation as described above. TE content was compared between the nonrecombining part of the mating-type chromosome (Table 1), the PARs, and the autosomes. GC content was compared between the non-recombining region of the mating-type chromosome, the PARs and the autosomes. Significance of the mean difference was assessed using a *t*-test and a nonparametric Wilcoxon test. In a second step, the GC content at the 3rd codon positions was inspected separately in identified coding regions. Mean GC contents at the 3rd codon positions were compared between the CDS on the non-recombining part of the mating-type chromosome, the PARs and the autosomal coding regions. All GC content analyses were conducted using in-house python and R scripts.

Prediction of the secretome

To predict a high confidence set of secreted proteins, results from several different software tools were integrated. These include TargetP1.1 [114], SignalP3.0, [115], SignalP4.0 (http://www.cbs.dtu.dk/services/SignalP/) [116], TMHMM2.0 [117], PredGPI [118], Phobius [119], NucPred [120], Prosite [121], and WoLF PSORT [122]. A subset of these tools were used to first exclude proteins as not secreted if they had transmembrane domains (two or more, from TMHMM or Phobius), an ER retention signal (0.00014 from Prosite), GPI anchor (specificity of >99.5 % using the general model of PredGPI), or nuclear localization (>0.8 threshold in NucPred). Secreted proteins were then predicted based on passing four of the six thresholds examined (TargetP secreted localization, SignalP3.0 NN Dscore > 0.43, SignalP3.0 HMM Sprob > 0.8, SignalP4.0 D-score > 0.45, WoLFPsort 'Extr' listed as major neighbor, or Phobius secreted localization).

Additional criteria were used for ambiguous predictions. If both TMHMM and Phobius agreed on the existence of 1–2 transmembrane (TM) domains in the protein, the protein was excluded from the probable secretome pool. If a protein was predicted to have a lowly probable GPI linkage (PredGPI specificity >99.0 % and <99.5 %) and a TM predicted by TMHMM and/or Phobius around the same region, this served as corroborating evidence for GPI anchorage to the membrane.

Where evidence conflicted or was insufficient for determining secretome status, BLASTp and Pfam domains were used to establish probable orthologs, followed by referencing the UniProtKB [123, 124] and FunSecKB [125] database for confirmation of localization of the orthologs, where available. Out of 7,360 predicted proteins, 6,899 proteins were excluded from the pool of the secretome based on the criteria described above. Of the remainder, 189 proteins had no contradictory calls in the positive prediction for SP. Another 272 went through further confirmation, of which 182 of these were confirmed to be non-SP and 63 were SP. Of the rest, 27 of them could not be finalized due to lack of ortholog matches in the NCBI database and lack of conserved domain for reference.

Gene clustering and comparative analysis

We compared *M. lychnidis-dioicae* to 18 other fungi (Additional file 13) that sample the three subphyla in Basidiomycota, including 5 other Pucciniomycotina, 7 Agaricomycotina, 3 Ustilaginomycotina, as well as 3 Ascomycota outgroups. For *M. lychnidis-dioicae* and the 18 other fungal genomes, we identified ortholog clusters using OrthoMCL [126] version 1.4 with a Markov inflation index of 1.5 and a maximum e-value of 1×10^{-5}. Two genomes, *R. glutinis* and *P. placenta*, are missing more broadly conserved orthologs than the other genomes; examining the 961 *Microbotryum* gene clusters with an ortholog missing in just one other genome, the number of missing clusters in any one Basidiomycete genome ranged from 1 to 34 with the exception of *R. glutinis* and *P. placenta*, missing 410 and 393 of these

highly conserved clusters, respectively. PFAM domains within each gene were identified using Hmmer3 [127], and gene ontology terms were assigned using BLAST2GO [128].

To examine gene duplication history, the phylome, or complete collection of phylogenetic trees for each gene in a genome, was reconstructed for *Microbotryum lychnidis-dioicae* and 19 other fungi, including those used for OrthoMCL (Additional file 13) and *Serpula lacrymans*. Phylomes were reconstructed using the previously described pipeline [129]. All trees and alignments have been deposited in PhylomeDB [129] and can be browsed online (www.phylomedb.org, phylome code 180). Trees were scanned to detect and date duplication events [130].

RNAi components from other other fungi were used as Blast queries to find homologs in *M. lychnidis-dioicae*; the queries used include *U. hordei* RdRp (CCF48827.1), *C. neoformans* Ago1 (XP_003194007), and *N. crassa* Dcl2 (Q75CC1.3) and Dcl1 (Q758J7.1). The putative function was confirmed by examining protein domains. The identified domains for each protein include: Piwi, PAZ and DUF1785 found in both copies of Argonaute (MVLG_06823, MVLG_06899); DEAD/DEAH helicase, double-stranded RNA binding, and RNAseIII (MVLG_01202). Sugar transporters were identified based on homology to the *Ustillago maydis* Srt1t transporter (Genbank: XP_758521) and the *Uromyces viciae-fabae* Hxt1 (Genbank: CAC41332).

The *M. lychnidis-dioicae* protein models corresponding to carbohydrate-active enzymes were assigned to families of glycoside hydrolases (GH), polysaccharide lyases (PL), carbohydrate esterases (CE), carbohydrate-binding modules (CBM), auxiliary activities (AA) and glycosyltransferases (GT) listed by the CAZy database [64], exactly as previously done for the analyses of dozens of fungal genomes [39, 66, 131, 132].

Data access

The assembly and annotation of *M. lychnidis-dioicae* was submitted to GenBank under accession number AEIJ01000000.

Additional files

Additional file 1: is a table providing Genome sequencing statistics.

Additional file 2: is a table showing RNA-Seq read statistics.

Additional file 3: is a figure showing Conservation of core eukaryotic (CEGMA) genes.

Additional file 4: is a figure Correlation between GC content and gene density.

Additional file 5: is a figure presenting Preferred codons for the different amino-acids.

Additional file 6: is a figure of Frequency of transposable element classes.

Additional file 7: is a figure displaying Proximity of TE copies for main orders (Class I LTR and LINE, Class II TIR and Helitrons) to the closest gene.

Additional file 8: is a figure presenting Comparison of TE proximity for secreted proteins compared to all other genes.

Additional file 9: is a table that shows TE elements used in RIPCAL analysis.

Additional file 10: is a figure that shows Frequency of mutation transition types in different TE classes.

Additional file 11: is a table presenting Genes predicted on non-recombining mating type regions, pseudoautosomal regions ("PAR") and autosomes.

Additional file 12: is a figure presenting Codon usage in autosomal and mating-type-specific genome regions.

Additional file 13: is a table showing Source of genome data used in comparative analysis.

Additional file 14: is a table that shows Gene family duplications from Phylome analysis.

Additional file 15: is a figure displaying Phylome for *M. lychnidis-dioicae*.

Additional file 16: is a figure of the Phylogenetic tree of DUF1034 proteins.

Additional file 17: is a table with Species counts of enriched or depleted protein domains.

Additional file 18: is a table that presents qRT-PCR validation of secretory lipase expression.

Additional file 19: is a table with CAZymes content comparisons.

Additional file 20: is a table that shows Phylogenetic comparison of CAZyme content.

Additional file 21: is a figure showing growth of *M. lychnidis-dioicae* on different sole carbon sources, including dextrose, cellulose, xylan, and pectin [133].

Additional file 22: is a table of the Full set of predicted secreted proteins.

Additional file 23: is a table showing Sequences similar to plant hormone-related genes.

Additional file 24: is a table with Genes differentially expressed between nutrient limited, rich, and MI-late samples.

Additional file 25: is a table of PFAM domains enriched in MI-late induced genes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MHP, DJS and CAC designed the research project. MHP, SY, EF, ZC, and CAC performed assembly and mapping. JG and QZ annotated the genome. MHP, JP, PW, EP, MEH, ZC, SST, and CAC analyzed the genomic sequences and transcriptome data. SST, SD, JA, EF, EP, JG, HB, BH, MHP, GA, TGE and CAC performed data analyses. SST, JMA, and DR conducted experimental analysis of secretory lipases and validation via qRT-PCR of RNA-Seq predictions. CAC, MHP, SST, JA, EF, TG, SD, BH and MEH wrote the paper. All authors approved the final version.

Acknowledgements

We acknowledge the Broad Institute Sequencing Platform for generating all DNA and RNA sequence described here, and Sinéad Chapman for coordinating the sequencing. We thank Mark Lawrence for sharing the *Rhodotorula glutinis* ATCC 204091 genome and annotation methods prior to publication. This project was supported by NSF award #0947963 to MHP, DJS, and CAC, the ANR-09-BLAN-064 and ERC GenomeFun 309403 grants to TG, and BIO2012-37161 and NPRP 5-298-3-086 to TGE. We would also like to thank Vincent Lombard, Elodie Drula and Anthony Levasseur for their help with the day-to-day development of the CAZy database.

Author details
[1]Department of Biology, Program on Disease Evolution, University of Louisville, Louisville, KY 40292, USA. [2]Institut National de la Recherche Agronomique (INRA), Unité de Recherche Génomique Info (URGI), Versailles, France. [3]Institut National de la Recherche Agronomique (INRA), Biologie et gestion des risques en agriculture (BIOGER), Thiverval-Grignon, France. [4]Ecologie, Systématique et Evolution, Bâtiment 360, Université Paris-Sud, F-91405 Orsay, France. [5]CNRS, F-91405 Orsay, France. [6]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [7]INRA, UMR 1136, Interactions Arbres-Microorganismes, Champenoux, France. [8]UMR 1136, Université de Lorraine, Interactions Arbres-Microorganismes, Vandoeuvre-lès-Nancy, France. [9]Centre National de la Recherche Scientifique (CNRS), UMR7257, Université Aix-Marseille, 13288 Marseille, France. [10]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia. [11]Centre for Genomic Regulation (CRG), Barcelona, Spain. [12]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [13]Institució Catalana d'Estudis Avançats (ICREA), Barcelona, Spain. [14]Department of Biology, Amherst College, Amherst, MA 01002, USA.

Received: 18 December 2014 Accepted: 28 May 2015
Published online: 16 June 2015

References

1. Kemler M, Lutz M, Göker M, Oberwinkler F, Begerow D. Hidden diversity in the non-caryophyllaceous plant parasite members of *Microbotryum* (Pucciniomycotina: Microbotryales). Syst Biodivers. 2009;7:297–306.
2. De Vienne DM, Hood ME, Giraud T. Phylogenetic determinants of potential host shifts in fungal pathogens. J Evol Biol. 2009;22:2532–41.
3. Le Gac M, Hood ME, Fournier E, Giraud T. Phylogenetic evidence of host-specific cryptic species in the anther smut fungus. Evol Int J Org Evol. 2007;61:15–26.
4. Le Gac M, Hood ME, Giraud T. Evolution of reproductive isolation within a parasitic fungal species complex. Evol Int J Org Evol. 2007;61:1781–7.
5. De Vienne DM, Refregier G, Hood ME, Guigue A, Devier B, Vercken E, et al. Hybrid sterility and inviability in the parasitic fungal species complex *Microbotryum*. J Evol Biol. 2009;22:683–98.
6. Giraud T, Yockteng R, Lopez-Villavicencio M, Refregier G, Hood ME. Mating system of the anther smut fungus *Microbotryum violaceum*: selfing under heterothallism. Eukaryot Cell. 2008;7:765–75.
7. Gibson AK, Hood ME, Giraud T. Sibling competition arena: selfing and a competition arena can combine to constitute a barrier to gene flow in sympatry. Evol Int J Org Evol. 2012;66:1917–30.
8. Alexander HM. An experimental field study of anther-smut disease of *Silene alba* caused by *Ustilago violacea*: genotypic variation and disease incidence. Evolution. 1989;43:835–47.
9. Antonovics J, Hood ME, Partain J. The ecology and genetics of host shift: *Microbotryum* as a model system. Am Nat. 2002;160:S40–53.
10. Refrégier G, Le Gac M, Jabbour F, Widmer A, Shykoff JA, Yockteng R, et al. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. BMC Evol Biol. 2008;8:100.
11. Hood ME, Antonovics J, Koskella B. Shared forces of sex chromosome evolution in haploid-mating and diploid-mating organisms: *Microbotryum violaceum* and other model organisms. Genetics. 2004;168:141–6.
12. Hood ME, Petit E, Giraud T. Extensive divergence between mating-type chromosomes of the anther-smut fungus. Genetics. 2013;193:309–15.
13. Hughes CF, Perlin MH. Differential expression of mepA, mepC and smtE during growth and development of *Microbotryum violaceum*. Mycologia. 2005;97:605–11.
14. Roche BM, Alexander HM, Maltby AD. Dispersal and disease gradients of anther-smut infection of *Silene Alba* at different life stages. Ecology. 1995;76:1863–71.
15. Jennersten O. Butterfly visitors as vectors of Ustilago violacea spores between caryophyllaceous plants. Oikos. 1983;40:125–30.
16. Akhter S, Antonovics J. Use of internal transcribed spacer primers and fungicide treatments to study the anther-smut disease, *Microbotryum violaceum* (=*Ustilago violacea*), of white campion *Silene alba* (=*Silene latifolia*). Int J Plant Sci. 1999;160:1171–6.
17. Kokontis J, Ruddat M. Promotion of Hyphal growth in *Ustilago-Violacea* by host factors from *Silene-Alba*. Arch Microbiol. 1986;144:302–6.
18. Kokontis JM, Ruddat M. Enzymatic hydrolysis of Hyphal growth factors for *Ustilago violacea* isolated from the host plant *Silene alba*. Bot Gaz. 1989;150:439–44.
19. Scutt CP, Kamisugi Y, Sakai F, Gilmartin PM. Laser isolation of plant sex chromosomes: studies on the DNA composition of the X and Y sex chromosomes of *Silene latifolia*. Genome Natl Res Counc Can Genome Cons Natl Rech Can. 1997;40:705–15.
20. Alexander HM, Antonovics J. Spread of anther-smut disease (*Ustilago violacea*) and character correlations in a genetically variable experimental population of *Silene alba*. J Ecol. 1995;83:783–94.
21. Robertson SE, Li Y, Scutt CP, Willis ME, Gilmartin PM. Spatial expression dynamics of Men-9 delineate the third floral whorl in male and female flowers of dioecious Silene latifolia. Plant J Cell Mol Biol. 1997;12:155–68.
22. Kazama Y, Koizumi A, Uchida W, Ageez A, Kawano S. Expression of the floral B-function gene SLM2 in female flowers of Silene latifolia infected with the smut fungus *Microbotryum violaceum*. Plant Cell Physiol. 2005;46:806–11.
23. Billiard S, López-Villavicencio M, Devier B, Hood ME, Fairhead C, Giraud T. Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. Biol Rev Camb Philos Soc. 2011;86:421–42.
24. Billiard S, López-Villavicencio M, Hood ME, Giraud T. Sex, outcrossing and mating types: unsolved questions in fungi and beyond. J Evol Biol. 2012;25:1020–38.
25. Fraser JA, Heitman J. Evolution of fungal sex chromosomes. Mol Microbiol. 2004;51:299–306.
26. Fraser JA, Hsueh YP, Findley KM, Heitman J. Evolution of the Mating-Type Locus: The Basidiomycetes. In: Heitman J, Kronstad J, Taylor J, Casselton L, editors. Sex in Fungi. Washington, D.C: ASM Press; 2007. p. 19–34.
27. Menkis A, Jacobson DJ, Gustafsson T, Johannesson H. The mating-type chromosome in the filamentous ascomycete *Neurospora tetrasperma* represents a model for early evolution of sex chromosomes. PLoS Genet. 2008;4, e1000030.
28. Hood ME. Dimorphic mating-type chromosomes in the fungus *Microbotryum violaceum*. Genetics. 2002;160:457–61.
29. Hood ME, Antonovics J. Intratetrad mating, heterozygosity, and the maintenance of deleterious alleles in *Microbotryum violaceum* (=*Ustilago violacea*). Heredity. 2000;85(Pt 3):231–41.
30. Hood ME. Repetitive DNA in the automictic fungus *Microbotryum violaceum*. Genetica. 2005;124:1–10.
31. Whittle CA, Sun Y, Johannesson H. Degeneration in codon usage within the region of suppressed recombination in the mating-type chromosomes of *Neurospora tetrasperma*. Eukaryot Cell. 2011;10:594–603.
32. Whittle CA, Johannesson H. Evidence of the accumulation of allele-specific non-synonymous substitutions in the young region of recombination suppression within the mating-type chromosomes of *Neurospora tetrasperma*. Heredity. 2011;107:305–14.
33. Fontanillas E, Hood ME, Badouin H, Petit E, Barbe V, Gouzy J, et al. Degeneration of the non-recombining regions in the mating-type chromosomes of the anther-smut fungi. Mol Biol Evol. 2014;32(4):928–43. doi:10.1093/molbev/msu396. Epub 2014 Dec 21.
34. Farbos I, Oliveira M, Negrutiu I, Mouras A. Sex organ determination and differentiation in the dioecious plant *Melandrium album* (*Silene latifolia*): a cytological and histological analysis. Sex Plant Reprod. 1997;10:155–67.
35. Grant S, Hunkirchen B, Saedler H. Developmental differences between male and female flowers in the dioecious plant *Silene latifolia*. Plant J. 1994;6:471–80.
36. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma Oxf Engl. 2007;23:1061–7.
37. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, et al. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. Nat Commun. 2011;2:202.
38. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.
39. Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, et al. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci. 2011;108:9166–71.
40. Yockteng R, Marthey S, Chiapello H, Gendrault A, Hood ME, Rodolphe F, et al. Expressed sequences tags of the anther smut fungus, *Microbotryum violaceum*, identify mating and pathogenicity genes. BMC Genomics. 2007;8:272.
41. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 2001;98:8714–9.
42. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, Cano LM, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature. 2009;461:393–8.

43. Gladieux P, Ropars J, Badouin H, Branca A, Aguileta G, de Vienne DM, et al. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Mol Ecol. 2014;23:753–73.
44. Hood ME, Katawczik M, Giraud T. Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. Genetics. 2005;170:1081–9.
45. Horns F, Petit E, Yockteng R, Hood ME. Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi. Genome Biol Evol. 2012;4:240–7.
46. Cambareri EB, Jensen BC, Schabtach E, Selker EU. Repeat-induced G-C to A-T mutations in *Neurospora*. Science. 1989;244:1571–5.
47. Selker EU, Cambareri EB, Jensen BC, Haack KR. Rearrangement of duplicated DNA in specialized cells of *Neurospora*. Cell. 1987;51:741–52.
48. Walser J-C, Furano AV. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res. 2010;20(7):875–82. doi:10.1101/gr.103283.109. Epub 2010 May 24.
49. Fryxell KJ, Moon W-J. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 2005;22:650–8.
50. Jiang C, Zhao Z. Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome. BMC Genomics. 2006;7:316.
51. Morton BR, Bi IV, McMullen MD, Gaut BS. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics. 2006;172:569–77.
52. Amselem J, Lebrun M-H, Quesneville H. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics. 2015;16:141.
53. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328:916–9.
54. Nicolás FE, Torres-Martínez S, Ruiz-Vázquez RM. Loss and retention of RNA interference in fungi and parasites. PLoS Pathog. 2013;9, e1003089.
55. Petit E, Giraud T, de Vienne DM, Coelho MA, Aguileta G, Amselem J, et al. Linkage to the mating-type locus across the genus *Microbotryum*: insights into nonrecombining chromosomes. Evol Int J Org Evol. 2012;66:3519–33.
56. Gillissen B, Bergemann J, Sandmann C, Schroeer B, Bölker M, Kahmann R. A two-component regulatory system for self/non-self recognition in Ustilago maydis. Cell. 1992;68:647–57.
57. Devier B, Aguileta G, Hood ME, Giraud T. Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes. Genetics. 2009;181:209–23.
58. Votintseva AA, Filatov DA. Evolutionary strata in a small mating-type-specific region of the smut fungus Microbotryum violaceum. Genetics. 2009;182:1391–6.
59. Bachtrog D. Accumulation of Spock and Worf, two novel non-LTR retrotransposons, on the neo-Y chromosome of Drosophila miranda. Mol Biol Evol. 2003;20:173–81.
60. Elhaik E, Landan G, Braur D. Can GC content at third-codon positions be used as a proxy for isochore composition? Mol Biol Evol. 2009;26:1829–33.
61. Klose J, de Sá MM, Kronstad JW. Lipid-induced filamentous growth in *Ustilago maydis*. Mol Microbiol. 2004;52:823–35.
62. Castle AJ. Isolation and Identification of *α* -Tocopherol as an Inducer of the Parasitic Phase of *Ustilago violacea*. Phytopathology. 1984;74:1194.
63. Suh MC, Samuels AL, Jetter R, Kunst L, Pollard M, Ohlrogge J, et al. Cuticular lipid composition, surface structure, and gene expression in arabidopsis stem epidermis. Plant Physiol. 2005;139:1649–65.
64. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(Database issue):D490–5.
65. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37(Database issue):D233–8.
66. Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, et al. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A. 2014;111:9923–8.
67. Prillinger H, Deml G, Dörfler C, Laaser G, Lockau W. A contribution to the systematics and evolution of higher fungi - yeast-types in the basidiomycetes, part II: microbotryum-type. Bot Acta. 1991;104:5–17.
68. Roeijmans H, Prillinger H, Umile C, Sugiyama J, Nakase T, Boekhout T (1998) Analysis of carbohydrate composition of cell walls and extracellular carbohydrates. In: Kurtzman C, Fell JW, Boekhout T (ed) The Yeasts - A Taxonomic Study. Volume 1. 5th edition. Elsevier, p. 103–105.
69. Liepman AH, Nairn CJ, Willats WGT, Sørensen I, Roberts AW, Keegstra K. Functional genomic analysis supports conservation of function among cellulose synthase-like a gene family members and suggests diverse roles of mannans in plants. Plant Physiol. 2007;143:1881–93.
70. Alexander NJ, McCormick SP, Hohn TM. TRI12, a trichothecene efflux pump from *Fusarium sporotrichioides*: gene isolation and expression in yeast. Mol Gen Genet. 1999;261:977–84.
71. Wahl R, Wippel K, Goos S, Kämper J, Sauer N. A novel high-affinity sucrose transporter is required for virulence of the plant pathogen *Ustilago maydis*. PLoS Biol. 2010;8, e1000303.
72. Voegele RT, Struck C, Hahn M, Mendgen K. The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. Proc Natl Acad Sci U S A. 2001;98:8133–8.
73. Moccia MD, Oger-Desfeux C, Marais GA, Widmer A. A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. BMC Genomics. 2009;10:243.
74. Veneault-Fourrey C, Commun C, Kohler A, Morin E, Balestrini R, Plett J, et al. Genomic and transcriptomic analysis of Laccaria bicolor CAZome reveals insights into polysaccharides remodelling during symbiosis establishment. Fungal Genet Biol FG B. 2014;72:168–81.
75. Antoniw JF, Ritter CE, Pierpoint WS, Loon LCV. Comparison of three pathogenesis-related proteins from plants of two cultivars of tobacco infected with TMV. J Gen Virol. 1980;47:79–87.
76. Baldrian P. Fungal laccases - occurrence and properties. FEMS Microbiol Rev. 2006;30:215–42.
77. Takahashi T, Kakehi J-I. Polyamines: ubiquitous polycations with unique roles in growth and stress responses. Ann Bot. 2010;105:1–6.
78. Brose N, Betz A, Wegmeyer H. Divergent and convergent signaling by the diacylglycerol second messenger pathway in mammals. Curr Opin Neurobiol. 2004;14:328–40.
79. Dong W, Lv H, Xia G, Wang M. Does diacylglycerol serve as a signaling molecule in plants? Plant Signal Behav. 2012;7:472–5.
80. Stergiopoulos I, de Wit PJGM. Fungal effector proteins. Annu Rev Phytopathol. 2009;47:233–63.
81. Bauer R, Oberwinkler F, Vanky K. Ultrastructural markers and systematics in smut fungi and allied taxa. Can J Bot. 1997;75:1273–314.
82. Palmieri G, Bianco C, Cennamo G, Giardina P, Marino G, Monti M, et al. Purification, characterization, and functional role of a novel extracellular protease from *Pleurotus ostreatus*. Appl Environ Microbiol. 2001;67:2754–9.
83. Palmieri G, Cennamo G, Faraco V, Amoresano A, Sannia G, Giardina P. Atypical laccase isoenzymes from copper supplemented *Pleurotus ostreatus* cultures. Enzyme Microb Technol. 2003;33:220–30.
84. Whittaker MM, Kersten PJ, Cullen D, Whittaker JW. Identification of catalytic residues in glyoxal oxidase by targeted mutagenesis. J Biol Chem. 1999;274:36226–32.
85. Luttrell ES. Tissue replacement diseases caused by fungi. Annu Rev Phytopathol. 1981;19:373–89.
86. Cashion NL, Luttrell ES. Host parasite relationships in Karnal bunt of wheat. Phytopathology. 1988;78:75–84.
87. Schäfer AM, Kemler M, Bauer R, Begerow D. The illustrated life cycle of *Microbotryum* on the host plant *Silene latifolia*. Botany. 2010;88:875–85.
88. Eisikowitch D, Lachance MA, Kevan PG, Willis S, Collins-Thompson DL. The effect of the natural assemblage of microorganisms and selected strains of the yeast *Metschnikowia reukaufii* in controlling the germination of pollen of the common milkweed *Asclepias syriaca*. Can J Bot. 1990;68:1163–5.
89. Biémont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics. 2010;186:1085–93.
90. Schmidt SM, Houterman PM, Schreiver I, Ma L, Amyotte S, Chellappan B, et al. MITEs in the promoters of effector genes allow prediction of novel virulence genes in Fusarium oxysporum. BMC Genomics. 2013;14:119.
91. Otto SP, Pannell JR, Peichel CL, Ashman T-L, Charlesworth D, Chippindale AK, et al. About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet TIG. 2011;27:358–67.
92. Luo H, Perlin MH. The gamma-tubulin-encoding gene from the basidiomycete fungus, *Ustilago violacea*, has a long 5′-untranslated region. Gene. 1993;137:187–94.
93. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010;7:709–15.

94. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37, e123.
95. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
96. Borodovsky M, Lomsadze A, Ivanov N, Mills R. Eukaryotic gene prediction using GeneMark.hmm. Curr Protoc Bioinforma. 2003;Chapter 4:Unit4 6.
97. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32(Web Server issue):W309–12.
98. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinforma Oxf Engl. 2004;20:2878–9.
99. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7.
100. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14:988–95.
101. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
102. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
103. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
104. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.
105. Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. Algorithms Mol Biol AMB. 2012;7:5.
106. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
107. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in *de novo* annotation approaches. PLoS ONE. 2011;6, e16526.
108. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21 Suppl 1:i152–8.
109. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1:166–75.
110. Bao Z, Eddy SR. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
111. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
112. Hane JK, Oliver RP. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics. 2008;9:478.
113. Zytnicki M, Quesneville H. S-MART, a software toolbox to aid RNA-seq data analysis. PLoS ONE. 2011;6, e25988.
114. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2:953–71.
115. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004;340:783–95.
116. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.
117. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.
118. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008;9:392.
119. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res. 2007;35(Web Server issue):W429–32.
120. Brameier M, Krings A, MacCallum RM. NucPred–predicting nuclear localization of proteins. Bioinformatics. 2007;23:1159–60.
121. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38(Database issue):D161–6.
122. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585–7.
123. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 2011;39(Database issue):D214–9.
124. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012;40(Database issue):D71–5.
125. Lum G, Min XJ. FunSecKB: the Fungal Secretome KnowledgeBase. Database J Biol Databases Curation. 2011;2011:bar001.
126. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
127. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7, e1002195.
128. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
129. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 2011;39(Database issue):D556–60.
130. Huerta-Cepas J, Gabaldón T. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinforma Oxf Engl. 2011;27:38–45.
131. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, et al. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog. 2012;8, e1003037.
132. Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, et al. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science. 2012;336:1715–9.
133. Holliday R. Ustilago Maydis. In: King RC, editor. Handbook of Genetics. New York: Plenum Press; 1974. p. 575–95.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central