Supporting Information

Duplessis et al. 10.1073/pnas.1019315108

SI Text

Background Information. The poplar leaf rust fungus *M. larici-populina* is the most devastating and widespread pathogen of poplars, and has limited the use of poplars for environmental and wood production goals in many parts of the world. Almost all known poplar cultivars are susceptible to *M. larici-populina*, and new virulent strains are continuously developing (1). This disease therefore has a strong potential impact on current and future poplar plantations used for production of forest products (principally pulp and consolidated wood products), carbon sequestration, biofuels production, and bioremediation. *M. larici-populina* belongs to the Basidiomycota (Pucciniomycotina, Pucciniomycetes, Pucciniales, Melampsoraceae). It requires a *Populus* and a *Larix* host to complete its life cycle. The rust overwinters as teliospores on dead *Populus* leaves on the ground. These spores germinate in the spring, producing windborne basidiospores, which results in infection of larch needles. A few days later, masses of yellow-orange aeciospores are produced on needles of the coniferous host. They serve as inoculum for infection of live *Populus* leaves during the spring. Urediniospores (in yellow-orange pustules) are then produced on *Populus* leaves, serving as inoculum for rust epidemics on *Populus* throughout the summer. In late summer, teliospores (the overwintering spores) are again produced on *Populus* leaves, completing the rust's life cycle. The sequenced isolate of *M. larici-populina* was strain 98AG31 (virulence 3–4-7). This isolate was collected in 1998 in Moÿ-de-l'Aisne (France) on *Populus trichocarpa x Populus deltoides* cv. Beaupré leaves and urediniospores were maintained in a cryotheque at Institut National de la Recherche Agronomique Nancy (France). For DNA production, dikaryotic urediniospores of strain 98AG31 were multiplied on detached leaves of *P. deltoides x Populus nigra* cv. Robusta as previously described (2).

P. graminis, the causal agent of stem rust (black rust), infects cereal crops (wheat, barley, rye and oat), as well as many native and cultivated grasses (3). Stem rust has plagued wheat production worldwide and is the most feared pathogen of wheat because of its ability to devastate a healthy field of wheat in less than 1 month. A new race of the wheat stem rust pathogen (*P. graminis* f. sp. *tritici*), Ug99, was first identified from Uganda in 1999. Ug99 is a highly virulent strain that is able to overcome resistance in approximately 80% of all of the wheat and barley currently grown. *P. graminis* belongs to the Basidiomycota (Pucciniomycotina, Pucciniomycetes, Pucciniales, Pucciniaceae) and is a typical macrocyclic, heteroecious rust fungus with five distinct spore stages and two hosts. The asexual, uredinial stage is found on cereals and grasses, and, under optimal conditions, produces a new generation every 8 to 12 d. Urediniospores (dikaryotic) are distributed by wind and can travel long distances in the upper atmosphere. The sexual stage begins with the formation of telia, typically in late summer or early fall. Teliospores are thick-walled and allow the fungus to overwinter. In the spring, germinating teliospores produce haploid basidiospores that infect the alternate host (*Berberis* spp.), resulting in the production of pycnia. Sexual mating results in the formation of aecia and the infection of cereal/grass host with aeciospores completes the life cycle. The sequenced isolate of *P. graminis* f. sp. *tritici* was strain CDL 75–36-700–3, race SCCL. This isolate was collected in 1975 in Pennsylvania from wheat, and pure urediniospores are maintained at the US Department of Agriculture Agricultural Research Service's Cereal Disease Laboratory.

Genome Sequencing and Assembly. ***Shotgun sequencing strategy and results.*** For *M. larici-populina*, genomic DNA was isolated from dikaryotic urediniospores of strain 98AG31. Three plasmid libraries (4, 6.3, and 8.5 kb) and a fosmid library (36.9 kb) were end-sequenced by using Sanger technology. Whole-genome shotgun sequence was assembled with Arachne (4), yielding a haploid genome assembly of 101.1 Mb comprising 462 scaffolds (N50 of 27 scaffolds; L50 of 1.1 Mb) at a 6.9-fold coverage (Dataset S1, Tables S1 and S2). Gaps accounted for 3.4% of the main genome scaffolds in the assembly. This is the largest basidiomycete genome sequenced so far (5). A total of approximately 650 kb across 102 scaffolds of the assembly shared more than 95% homology to larger scaffolds for at least half of their coverage; these were classified as tentative alternate haplotypes (Dataset S1, Table S2). The mitochondrial genome was comprised of four scaffolds for a total sequence length of 79 kb (Dataset S1, Table S2). Available EST reads generated by the JGI (*cDNA Libraries, EST Clustering, and EST Annotation*) were clustered, and a representative for each cluster was selected to assess completeness of the assembly. At 90% identity and 90% coverage, 2,420 of 2,494 cluster representatives (97.03%) align to the assembly. An additional 29 sequences (1.16%) showed partial alignments, and 45 (1.80%) were not found in this assembly. The assembly was deposited at GenBank under the project accession AECX00000000. The sequence traces were deposited in the National Center for Biotechnology Information (NCBI) trace repository. All data are also available from the JGI Web site (http://jgi.doe.gov/Melampsora).

For *P. graminis* f. sp. *tritici*, strain CDL 75–36-700–3, race SCCL was sequenced. Genomic DNA was prepared from urediniospores, a dikaryotic spore stage, as described (6). Two plasmid libraries (4.5 kb or 8.6 kb inserts) and a fosmid library (40.1 kb inserts) were end-sequenced by using Sanger technology. Reads were trimmed for quality and to remove vector sequence, and trimmed reads total 908 Mb (Dataset S1, Table S1*B*). The sequence was assembled using Arachne (4). The assembly totals 81.52 Mb in a total of 4,557 sequence contigs (L50 of 39.5 kb), which are linked into 392 scaffolds (N50 of 30 scaffolds; L50 of 965 kb) totaling 88.6 Mb (Table 1). Gaps account for 8% of the total scaffold size. The assembled sequence has a high consensus quality, with 96.3% of bases at Q40 (one error every 10,000 bases) or greater. The assembly was deposited at GenBank under the project accession AAWC01000000. The sequence traces were deposited in the NCBI trace repository. All data are also available from the Broad Institute Web site (http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html).

For *P. graminis* f. sp. *tritici*, a fingerprint map was constructed from the end-sequenced clones in the fosmid library. The clones were restriction-fingerprinted (7, 8) by using a double digest of EcoRI and PstI. The resulting fingerprints were assembled into a map using the FingerPrinted Contig (FPC) software (9, 10) and automated in-house scripts. The map contains a total of 21,520 clones in 1,969 contigs, and 729 singleton clones. A total of 1,817 contigs (92%) could be mapped back to the assembly, covering 332 assembly scaffolds and 99% of the assembly (total of 87.6 Mb). For these mapped fingerprint contigs, 99% aligned without any conflicting restriction sites. In cases of conflict, either the fingerprint contig matched to multiple places in the assembly, suggesting uncertain placement as a result of repetitive content, or the fingerprint contig suggested that two scaffolds be joined in a way not supported by sequence. The fingerprint map also suggested that six pairs of scaffolds could be joined, but by examining the sequence, it was not possible to make these joins without

removing large regions of sequence. The 92% of mapped fingerprint contigs contain 98% of the fosmids in the fingerprint map. The 152 contigs which did not map back to the assembly contain a total of 467 fosmids, so are on average small in size. Both Internet Contig Explorer (11) and FPC versions of the map are openly available at the Genome Sciences Centre Web site, the Internet Contig Explorer software can be downloaded from http://www.bcgsc.ca/platform/bioinfo/software/ice, and the FPC version of the map is available on the data downloads page, http://www.bcgsc.ca/data/data.

Although a dikaryotic spore stage was sequenced for *P. graminis* f. sp. *tritici*, only a small fraction of the assembly was identified as potentially representing two haplotypes or large duplications. By aligning the assembly against itself using Arachne, we identified scaffolds that align internally to other scaffolds, where the alignment spans the smaller scaffold. A total of 13 scaffolds ranging in size from 5 kb to 174 kb fit this criteria; the largest is scaffold 110 which aligns to scaffold 40. In total, these regions contain 326 kb, which covers only 0.4% of the assembly.

SNPs. SNPs were identified for *M. larici-populina* by mapping the sequencing reads back to the assembled genome. Only sequencing reads with unique placement on the genome assembly were used for the SNP detection. Each base is covered by at least four reads (two from the consensus reads and two from the SNP) but no more than 25 reads. In total, 88,083 SNPs were detected in the dikaryotic genome. There was no SNPs density difference observed between the coding (0.84 SNPs per 1 kb) and noncoding region (0.87 SNPs per 1 kb). More than 70% of 1-kb genome sequence bins contained less than one SNP and a total of 254 1-kb genome sequence bins contained more than 10 SNPs.

For *P. graminis* f. sp. *tritici*, SNPs were called from 147 million Illumina 76b paired reads, which were aligned to the genome assembly using BWA (12). The resulting alignments of 113 million reads covered 99.8% of the assembled bases at an average of 78-fold depth. Filtering for unique alignments and mapping quality of 30 or greater resulted in 49.7 million read alignments which covered 85.99% of the assembly at 41-fold depth. To identify SNPs, consensus genotypes were called from these alignments using SAMtools pileup command. Variants were then filtered using the samtools.pl varFilter, to require a depth of four or more, and a maximum of one SNP in a 10-base window, and a quality of at least 20 to remove uncertain calls. Positions with alternate (nonreference) allele frequency between 20% and 80% were classified as heterozygous; positions with less than 20% of an alternate allele were classified as homozygous reference calls and removed from the set of SNPs. Last, the coverage distribution was examined by using a boxplot, and positions with more than 1.5 times the interquartile range above the 75th percentile were classified as outliers and removed. A total of 129,172 were identified; based on normalization for potential SNP positions (positions with sufficient uniquely aligned reads), the rate of variation was calculated at 2.28 SNPs/kb of coding sequence and 1.72 SNPs/kb of intergenic sequence, higher rates than that found in *M. larici-populina*.

cDNA Libraries, EST Clustering, and EST Annotation. For *M. larici-populina*, RNA was isolated from resting and germinating (4 h growth on 1% agar medium) urediniospores and mixed in equal volumes (50%/50%). Two cDNA libraries (CBWO and CBWP) were constructed from the same RNA mix. The libraries were made from oligo(dT)-primed cDNA and cloned into the pCMV-SPORT6 vector (Invitrogen). PolyA RNA was primed with an oligo dT primer (5′-GACTAGTTCTAGATCGCGAGCGGCCGCCCTTTTTTTTTTTTTT-3′), ligated to a SalI adapter (5′-TCGACCCACGCGTCCG and 5′- CGGACGCGTGGG), and digested with NotI. cDNA was size selected from 2 to 8 kb by using 1.1% agarose gel electrophoresis, then ligated into NotI and SalI-digested pCMV-SPORT6 vector. The library was constructed and sequencing was performed at the US Depatment of Energy JGI (Walnut Creek, CA). A total of 25,792 and 23,225 ESTs were sequenced from the CBWO and CBWP libraries, respectively. Sequences were filtered for quality and assembled following standard processing procedures in the JGI pipeline. In total, 11,535 contigs were generated and used to assist in gene annotation.

For *P. graminis* f. sp. *tritici*, four cDNA libraries were constructed and sequenced. RNA was isolated from urediniospores, germinated (24 h) urediniospores, teliospores, and isolated haustoria. Haustoria were isolated from wheat leaves infected with rust strain 21–0 (accession no. 540129) as described (13). Samples were reverse-transcribed, and cDNAs were cloned in pDONR221 for the spore samples and in pTriplEx2 for the haustorial sample. Sequences were filtered for quality and aligned to the *P. graminis* f. sp. *tritici* assembly using Blat to assist in annotation.

Repeat Analysis. ***TEs of M. larici-populina and P. graminis f. sp. tritici.*** The REPET pipeline (http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET) (14) was run on the *M. larici-populina* and *P. graminis* f. sp. *tritici* genome contigs. A de novo repeat search was performed using Blaster [ID > 90%, high scoring pairs (HSPs) length > 100 b and < 20 Kb, E-value 1e-300). The cumulative length of de novo repeats correspond to 29% and 36% of *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes, respectively. HSPs identified in the first step were grouped into clusters (14–16). Multiple alignments of the 20 longest members of each cluster containing at least three members (5,141 clusters for *M. larici-populina* and 6,967 clusters for *P. graminis* f. sp. *tritici*) were used to derive a consensus for each. Consensus sequences were classified using TEclassifier and by removing redundancy with Blaster and Matcher. Complete TEs have a structure compatible with a full TE and similarity with known TEs from Repbase Update (v. 14.05) (17). Incomplete TEs have evidence of TE structure or similarity but not both. Consensus sequences without any known structure or similarity were classified as "NoCat." Three methods (Blaster, Censor, RepeatMasker) were used to annotate TE copies in the whole genome based on the TE consensus from the TE de novo pipeline. The adjacent or overlapping HSPs from the same TE categories were filtered and combined. To annotate simple sequence repeats (SSRs), three methods (TRF, Mreps, and RepeatMasker) were used. TE/SSR doublons included in the TE annotation were then removed. Finally a "long join procedure" was used to address the problem of nested TEs. This procedure finds and connects the split segments of one TE interrupted by several other TEs as a result of recent insertion. Consensus sequences of *M. larici-populina* and *P. graminis* f. sp. *tritici* TEs (2,020 and 2,171, respectively) were used to annotate cluster members in these genomes. A larger proportion of class II TIR elements (~14%) was found in the *M. larici-populina* genome compared with class I LTR retrotransposons (~8%). Interestingly, this proportion differs with the TE content in the *P. graminis* f. sp. *tritici* genome, in which 12.4% and 9.8% of LTR and TIR elements, respectively, are found (Dataset S1, Tables S3 and S4). Fig. S1 details the distribution of different TE types on scaffolds 1, 2, and 3 of *M. larici-populina*. The distribution of predicted genes is also shown. Fig. S2 details the overall distribution of TEs and predicted genes on *P. graminis* f. sp. *tritici* genome scaffolds. The distribution of genes conserved between basidiomycete genomes is also detailed.

TE insertion age. To identify full-length LTR retrotransposons in *M. larici-populina* and *P. graminis* f. sp. *tritici* genome assemblies, a de novo search was performed with LTR_STRUC (18). This program yielded 131 full-length candidate LTR retrotransposon sequences for *M. larici-populina* and 266 for *P. graminis* f. sp. *tritici*, which were checked for homology using the BLASTN algorithm (19) against the sequences from the RepBase database. Among the 131 putative full-length LTRs of *M. larici-populina*, 72 were attributed to *Gypsy*/*Ty3*-like elements and 45 to *Copia*/*Ty1*-like. Fourteen other elements did not exhibit a significant homology

with known TE families or have homologies with non-LTR retrotransposons. For *P. graminis* f. sp. *tritici*, 95 were attributed to *Gypsy*/*Ty3*-like elements, 62 to *Copia*/*Ty1*-like, and 109 other elements did not exhibit a significant homology with known TE families or have homologies with non-LTR retrotransposons. To determine the insertion age of *Gypsy*/*Ty3*-like elements and *Copia*/*Ty1*-like, the divergence of LTRs were examined. The insertion age was determined from the evolutionary distance between 5′- and 3′-solo LTR of full length elements derived from a ClustalW (20) alignment of the two solo LTR sequences using the Kimura correction. To convert the sequence distance to a putative insertion age, a substitution rate of 1.3×10^{-8} mutations per site per year was used (21). Most full-length *Gypsy*/*Ty3*-like and *Copia*/*Ty1*-like elements were inserted in the *M. larici-populina* and *P. graminis* f. sp. *tritici* genome recently (<1 Mya).

SSRs. MISA (http://pgrc.ipk-gatersleben.de/misa/download/misa.pl) was used to identify mono- to hexanucleotide SSR motifs using default parameters, i.e., a minimum repeat number of 10, 6, 5, 5, 5, and 5, for mono, di-, tri, tetra-, penta-, and hexanucleotide motifs, respectively. A total of 9,972 SSRs were identified in the *M. larici-populina* genome corresponding to 5,224 mono-, 1,612 di-, 2,580 tri-, 220 tetra-, 178 penta-, and 158 hexanucleotide (Dataset S1, Table S5). In the *P. graminis* f. sp. *tritici* genome, 31,173 SSRs were identified corresponding to 22,667 mono-, 3,576 di-, 4,185 tri-, 344 tetra-, 301 penta-, and 100 hexanucleotide motifs (Dataset S1, Table S5). To date, the *P. graminis* f. sp. *tritici* genome is the most SSR-rich Basidiomycota genome (22). The relative abundance was calculated as the number of SSRs per Mb. For all the SSRs, the relative abundance was 98 and 254 SSR per Mb in *M. larici-populina* and *P. graminis* f. sp. *tritici* genome, respectively (Dataset S1, Table S5).

Gene Prediction and Annotation. For *M. larici-populina*, a combination of various gene callers was used for gene prediction and included ab initio and homology based Fgenesh (23), Genewise (24), and EST based estExt (25), as well as EuGene (26). Based on the 49,017 urediniopsores ESTs, a subset of 300 genes was carefully annotated by the *Melampsora* Genome Consortium to determine their parameters (e.g., donor and acceptor splice sites, intron mean length, stop codons) and to train gene callers for gene prediction. Different sets of gene calls were found by the distinct algorithms. Gene models were then combined to produce a non redundant set of genes using a heuristic approach implemented in the JGI pipeline to conserve a single best gene model per locus. An initial set of 16,694 predicted gene models was made available on the *Melampsora* genome Web site (January 2008). Manual curation of genes was performed by the *Melampsora* Genome Consortium for selected gene categories (*Targeted Annotation and Analysis of Specific Gene Families*), and new gene models were found by using *M. larici-populina* ESTs sequenced from urediniopsores and infected plant tissues (present study and ref. 27) and recursive BLAST searches against the genome. Dubious genes corresponding to probable TE were filtered from the gene catalog based on REPET results, TE-related PFAM domains, and searches in TE databases (http://www.girinst.org/). Genes showing an obvious TE-related PFAM hit were excluded from the predicted *M. larici-populina* genes. Genes of unknown function identified by the REPET analysis and showing homology to TE in databases were also excluded. Other genes which overlapped repeat elements in the REPET analysis but were not obviously TEs, that could represent new genes belonging to multigene families, were not excluded. Finally, a total of 16,399 gene models are predicted in the *M. larici-populina* catalog (April 2011). The distribution of coding sequence compared with TE is detailed for the three largest scaffolds in Fig. S1.

For *P. graminis* f. sp. *tritici*, gene structures were predicted by using a combination of manual annotation, automated gene callers, and EST-based transcript identification. More than 87,000 ESTs sequenced as part of this project were aligned to the genome using Blat, and alignments were clustered to construct reference transcripts. We also predicted potential genes using Fgenesh (23), GeneID (28), and Augustus (29), which were trained on the subset of EST-based transcripts that covered entire ORFs without splicing or frame conflicts. The gene model with the best alignment with BLAST hits and agreement with splice sites inferred from ESTs was selected for each locus. Gene models with potential problems were manually reviewed and edited where possible. The resulting gene set of 20,567 genes was then examined for potential false-positive calls (as detailed above). A total of 2,794 genes were either similar to repetitive elements or low-confidence gene models and were flagged as dubious. Subtracting these from the gene set resulted in a total of 17,773 predicted proteins. The distribution of coding sequence compared with TE is detailed in Fig. S2. Distribution of 4,640 conserved genes, which contained a potential orthoMCL orthologue in at least five of the six other basidiomycetes, all of which are publicly available [*M. larici-populina* (JGI), *U. maydis* (Broad Institute), *Sporobolomyces roseus* (JGI), *C. cinerea* (Broad Institute), *C. neoformans* (Broad Institute), and *L. bicolor* (JGI)] is also shown.

Comparisons of the *P. graminis* f. sp. *tritici* or *M. larici-populina* proteins to the nonredundant protein database at GenBank identified homologues for only 35% or 41% of the predicted genes, respectively (BLASTP, E-value $\leq 10^{-5}$). However, comparison between the two rust gene catalogs identified matches for as many as 57% of the proteins (unidirectional BLASTP, E-value $\leq 10^{-5}$), indicating that the two rust species share genes specific to the Pucciniales lineage. Ongoing genome sequencing of rust species of the *Puccinia* and the *Cronartium* genera will assist in refining this set. Overall, however, rust gene homologues exhibit low levels of identity supporting the divergence of rust sequences at the nucleotide level (Fig. S3).

Orthology, Synteny, Tandem Repeats, and Multigene Families.

Multigene families and evolutionary analysis of multigene families. To examine patterns of gene loss and gain in the rust genomes, we collected proteins sets from 12 publicly available fungal genomes [10 Basidiomycota: *M. larici-populina* (JGI), *P. graminis f. sp. tritici* (Broad Institute), *C. cinerea* (Broad Institute), *C. neoformans* (Broad Institute), *Postia placenta* (JGI), *L. bicolor* (JGI, Frozen gene catalog), *Malassezia globosa*, *Phanerochaete chrysosporium* (JGI), *S. roseus* (JGI, v1) *and U. maydis* (Broad Institute); and two Ascomycota: *N. crassa* (Broad Institute) *and M. oryzae* (Broad Institute)]. In total, this analysis included 16,399 genes for *M. larici-populina* and 17,773 genes for *P. graminis* f. sp. *tritici*. Gene families were then constructed based on sequence similarity and grouped by TribeMCL (30). This resulted in a dataset of 15,012 gene families and 18,138 orphans (i.e., genes without homology to other sequence in the dataset; Dataset S1, Table S6). Excluding the orphan genes, *M. larici-populina* has 5,304 gene families with average family size 2.71 genes per family, whereas *P. graminis* f. sp. *tritici* has 5,413 gene families and the average gene family size is 2.67 genes per family; both are slightly more than the average 2.55 genes per family observed in *L. bicolor* (31). The rust genomes both experienced large gene family expansions. There are 19 gene families (1,597 genes) expanded to more than 50 gene copies in *M. larici-populina* and 14 gene families (1,183 genes) in *P. graminis* f. sp. *tritici*. On the other hand, *L. bicolor* had only 13 gene families (868 genes) with such large expansion.

To infer phylogenetic relationships, protein alignments were generated using MUSCLE (32) for each of 105 single-copy gene families. Unconserved regions in each multiple alignment were removed by using an in-house script. The conserved region of each single copy gene family was concatenated into one sequence and the phylogenetic relationship of fungal species was inferred by using PhyML (33) with default parameters. The phylogenetic profiles of each gene family were constructed to

reflect the absence or presence of a particular gene family in a given species. We combined the phylogenic profiles and the species tree to reconstruct the parsimonious series of gene gain and loss events (34) for these fungal genomes. The DOLLOP program from the PHYLIP (35) package was used to define the minimum gene set for ancestral nodes of the phylogenetic tree. The DOLLOP program is based on the Dollo parsimony principle, which assumes that gene(s) have arisen exactly once on the evolutionary tree and can be lost independently in different evolutionary lineages.

The protein sequences in each fungal genome were searched against the NCBI nr protein database with threshold E-value of less than 1e-5 and were stored in XML format. By using these blast hits, the Gene Ontology (GO) vocabulary for each protein sequence was predicted using the Blast2GO pipeline (36). To further enrich the mapping of proteins with GO annotation, the InterProScan result of each protein sequence was combined with the Blast2GO result. Due to the stage of each genome annotation and the manual curation of GO terms for each genome project, the number of predicted GO terms for each genome is highly variable. For example, the *M. oryzae* genome has the largest number of homology-based GO assignments; it was published in 2005, with more than six revisions to the genome annotation, and more importantly, a comprehensive manual GO annotation curation. By contrast, the *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes contain many lineage-specific gene families that are not homologous to proteins in the current nr database, and both contain fewer predicted GO annotations.

Based on the Dollop analysis, we identified large numbers of lineage-specific gene families in *M. larici-populina* and *P. graminis* f. sp. *tritici* (909 and 1,241 families with 5,798 and 6,139 genes, respectively). Among these families, five GO terms are overrepresented [false discovery rate (FDR) < 0.01] in *M. larici-populina* and four GO terms in *P. graminis* f. sp. *tritici* (Dataset S1, Tables S7 and S8), corresponding to zinc ion binding and nucleic acid binding activities in both rust fungi, as well as distinct redox-related functions (*Redox Control and Oxidative Stress*). Trehalose biosynthetic process GO is overrepresented in the wheat stem rust, whereas heat shock protein binding and ribonuclease activities are detected in the poplar rust. By contrast, 39 and 78 GO terms are underrepresented in *M. larici-populina* and in *P. graminis* f. sp. *tritici*, respectively.

To obtain a clear picture of gene families expanded in *M. larici-populina* and *P. graminis* f. sp. *tritici*, the SD and the mean gene family size were calculated for all gene families (excluding orphans and lineage-specific families). The counts by species for each family were transformed into a matrix of *z*-scores to center and normalize the data. The 100 families with the greatest *z*-scores in *M. larici-populina* and/or *P. graminis* f. sp. *tritici* were selected. These profiles were hierarchically clustered (complete linkage clustering) by using the Pearson correlation as a distance measure; clustering and visualization was performed using MeV (37). The biological function of each family was predicted based on the sequence similarity from the Interpro protein domain database and the UniProt–SwissProt protein database (Fig. S4*A*). To illustrate clade-specific expanded families in rust fungi, the *z*-scores for *M. larici-populina* and *P. graminis* f. sp. *tritici* were summed and the 100 greatest values were selected and similarly clustered (Fig. S4*B*). Although the total number of transporters detected in the two rust genomes were lower compared with those reported for other fungi (*Transporters*), several transporters families are clearly expanded, particularly oligopeptide transporters and amino acid transporters (Fig. S4*A*), indicating unique adaptations of the biotrophy-related transport machinery of rust fungi (*Transporters*). Additional expanded families include transcription factors, copper/zinc SODs, and α-kinase families (*Redox Control and Oxidative Stress* and *Signal Transduction Pathways*). Overrepresented GO terms observed in the two rust fungi genomes (as detailed earlier) are consistent with expanded families highlighted in Fig. S4*B*, such as specific expanding families of zinc-finger proteins encoding genes in *M. larici-populina* or *P. graminis* f. sp. *tritici*. Among other expanding gene families are peptidases, lipases, and carbohydrate active enzymes [*Annotation of Putative CAZymes*, *Proteases*, and *Lipases (Triacylglycerol Hydrolases)*]. Other than these, most gene expansions detected in one or the other rust genome are related to unknown functions and encompass families of genes encoding SSPs (Fig. S4 *A* and *B*). To determine whether the proliferation of TE observed in rust fungi genomes could be related with the expansion of specific gene families, GO-enrichment tests were performed for 50 kb genomic windows containing more than 70% TE. No significant enrichment could be detected by such approach. Similarly, no significant localization of genes families encoding lineage- or clade-specific SSPs was detected in TE-rich regions of the rust fungi genomes.

Lack of genome duplication and synteny between M. larici-populina and P. graminis f. sp. tritici. The identification of tandem genes, duplicated blocks, and syntenic regions between the two genomes was measured by i-ADHoRe 2.0 (Automatic Detection of Homologous Regions) (38). Gene pairs were regarded as homologous if they belong to the same gene family from *Multigene Families and Evolutionary Analysis of Multigene Families*. Tandemly duplicated genes were defined as two homologous genes separated by less than 10 nonhomologous gene on the same scaffold, independent of orientation. In the *M. larici-populina* genome, we identified 117 duplication blocks with three to eight paralogous gene pairs (1,467 genes in total) ranging from 5 kb to 285 kb in size. Furthermore, 1,495 genes are tandemly duplicated in 664 tandem arrays. The *P. graminis* has 90 duplication blocks with three to 26 paralogous gene pairs (1,955 genes in total) ranging from 4 kb to 467 kb in size and 1,282 tandemly duplicated genes are arranged in 561 tandem duplicated arrays. No significant gene functional enrichment could be detected in the duplicated regions. There are 39 synteny blocks between *M. larici-populina* and *P. graminis* f. sp. *tritici.* The largest synteny block has six orthologous gene pairs with 51 predicted genes spanning on 281 kb of genomic sequence (Fig. S5).

The rate of sequence evolution between gene pairs was estimated by calculating the rate of synonymous substitution (Ks) by using a method described previously (39) (Fig. S3) using PAML (40). Duplicated blocks between the two genomes showed higher Ks values, suggesting an older date of duplication, compared with duplicated blocks in each rust genome.

Whole-Genome Exon Oligoarrays. The *M. larici-populina* custom-exon expression oligoarray (4 × 72K) manufactured by Roche NimbleGen contained four independent, nonidentical, 60-*mer* probes per gene model. Included in the oligoarray were 17,556 predicted coding sequences and 1,063 random 60-*mer* control probes and labeling controls (41). For 440 gene models, technical duplicates were included on the array for internal quality control. The 17,556 selected sequences corresponded to an early version of the predicted gene catalog before manual annotation and included some ESTs and TEs. Of the 16,399 gene models now included in the *M. larici-populina* gene catalog, only 14,232 were represented on the array. Urediniospores of *M. larici-populina* (isolate 98AG31) were propagated on detached leaves of susceptible *Populus deltoides* × *P. nigra* "Robusta" as previously reported (2). Germinated urediniospores were obtained from 1 mg of urediniospores grown on water agar medium (2%) in Petri dishes for 3 h at 19 ± 1 °C. For poplar leaf infection, the susceptible *P. trichocarpa x P. deltoides* "Beaupré" was grown in a greenhouse from dormant cuttings and leaves were spray-inoculated and incubated for 96 h postinoculation (hpi) as described (2). Tissues were snap-frozen in liquid nitrogen, and RNA extraction was carried out using the RNeasy Plant Mini Kit including a DNase treatment (Qiagen). RNA quality and integrity were checked before cDNA synthesis using the Bio-Rad

Experion analyzer and Experion RNA StdSens analysis kit (Bio-Rad). Total RNA preparations (three biological replicates) were amplified by using the MessageAmp II aRNA amplification kit (Ambion) according to the manufacturer's instructions, and double-stranded cDNA was synthesized according to NimbleGen specifications. cDNA were synthesized from 2.5 μg of aRNA using the Invitrogen SuperScript Double-Stranded cDNA Synthesis Kit according to the NimbleGen user protocol. Single dye labeling of samples, hybridization procedures, and data acquisition were performed at the NimbleGen facilities, following their standard protocol. Microarray probe intensities were quantile-normalized across chips. Average expression levels were calculated for each gene from the independent probes on the array and were used for further analysis. Raw array data were filtered for nonspecific probes (a probe was considered as nonspecific if it shared more than 90% homology with a gene model other than the gene model for which it was made) and renormalized using the ArrayStar software (DNASTAR). For 859 gene models, no reliable probe was left. A transcript was deemed expressed when its signal intensity was threefold higher than the mean signal-to-noise threshold (cutoff value) of 1,063 random oligonucleotide probes present on the array. Gene models with an expression value higher than threefold the cutoff level (105, 97, and 98 for 96 hpi, resting urediniospores, and germinating urediniospores, respectively, in arbitrary fluorescence units) were considered as transcribed. A total of 281 transcripts showed evidence of cross-hybridization with mock-inoculated (i.e., water) poplar leaf transcripts and were not further considered in the analysis. A Student *t* test with FDR (Benjamini–Hochberg) multiple testing correction was applied to the data using the ArrayStar software (DNASTAR). Transcripts with a significant *P* value (<0.05) and more than a threefold change in transcript level were considered as differentially expressed. The complete expression datasets are available at the GEO database (NCBI) as series GSE23097.

In total, 13,093 gene models from the *M. larici-populina* gene catalog (~80%) were assayed for gene expression. Of the 13,093 genes tested, no expression was detected for 3,845 genes (29%) *in planta* at 96 hpi or in resting and germinating urediniospores; by contrast, 6,466 genes were expressed in all conditions. In total, 71% of the genes tested were expressed in at least one biological stage. Strikingly, 1,116 genes were uniquely expressed *in planta* and not in spores, whereas only 92 and 207 genes were expressed only in resting and germinating urediniospores, respectively. However, a total of 983 genes were expressed both in resting and germinating spores and not *in planta*, representing putative spore-related genes. Specific expression of selected manually annotated genes families are presented and/or discussed in the following sections. The top 100 most highly expressed genes detected during poplar leaf infection at 96 hpi are presented in Dataset S1, Table S9. The top 100 most highly induced genes at 96 hpi in poplar or in germinating urediniospores compared with resting urediniospores are presented in Dataset S1, Tables S10 and S11, respectively. Finally, a selection of significantly regulated genes ≥ Ten-fold change *in planta* compared with urediniospores and presenting homology with known functions are listed in Dataset S1, Table S12.

A *P. graminis* f. sp. *tritici* custom expression oligoarray (385,000 features) was manufactured by Roche NimbleGen. Probe design was attempted for the initial set of 20,567 *P. graminis* f. sp. *tritici* genes, 578 EST clusters which aligned to the assembly but were at least 200 nucleotides from a gene call, and 41 wheat and barley sequences. The final probe set covered 20,228 genes, 558 of the EST clusters, and 41 plant sequences; all probes were replicated in triplicate on the array. Nearly all genes (99.7%) were represented by five independent, nonidentical, 60-*mer* probes. The oligoarray also included 77,436 random 60-*mer* control probes.

RNA was prepared from four different conditions, with three biological replicates per condition. The conditions sampled were urediniospores, germinated urediniospores, infected wheat, and infected barley. *P. graminis* f. sp. *tritici* urediniospores were collected from wheat seedling leaves (42) and stored at 4 °C and 30% relative humidity for 1 to 3 d before RNA extraction. Fresh urediniospores were germinated for 24 h as described (42). Seven-day-old wheat seedlings (cultivar McNair 701) were inoculated as described (42) and grown in a Conviron growth chamber for 8 d with 16 h light periods at 20 °C and 8 h dark periods at 16 °C. Seven-day-old barley seedlings (cultivar Hypana) were inoculated as described (42) and grown in a Conviron growth chamber for 8 d with 16 h light periods at 20 C and 8 h dark periods at 16 °C. Wheat and barley seedlings were placed in a random block array within the same growth chamber. Infected wheat and barley leaves were harvested after 8 d dpi, when macroscopic flecking was visible and non infected leaf tissue was removed. Total RNA was extracted using a hot (60 °C) TRIzol (Invitrogen) method followed by Rneasy Midi kit (Qiagen), a modification of the method described by (43). RNA was labeled and hybridized to oligoarrays by Roche NimbleGen, all data were RMA-normalized, and the hybridization intensity to each gene calculated by using Nimblescan. The intensity values for each gene between biological replicates were highly correlated; the average pairwise correlation coefficients between replicates was 0.99 for urediniospores, 0.98 for germinated urediniospores, 0.99 for infected wheat, and 0.95 for infected barley. Correlation coefficients between similar conditions was also high, particularly for infected wheat and infected barley samples (0.96), and but also for the resting and germinated urediniospores (0.82).

To identify differentially expressed genes, *P* values were determined using paired two sample *t* test (using mattes in Matlab), and FDR and q-values were calculated (using mafdr in Matlab). By plotting a histogram of the $\log_2$ expression values of the control and gene probes, we observed clear separation of the signal for most genes. Therefore, we estimated a background level of hybridization to include 95% of the values of control probes; genes which fell below this threshold for a given experiment were considered to be nonspecific background levels of hybridization. Across all four conditions, a hybridization value greater than background was detected for 9,818 genes; for 6,570 genes, expression was detected in all four conditions. The top 100 most highly expressed genes detected during wheat infection at 8 d dpi are presented in Dataset S1, Table S13. The top 100 most highly induced genes at 8 dpi in wheat or in germinating urediniospores compared with resting urediniospores are presented in Dataset S1, Tables S14 and S15, respectively. Finally, a selection of significantly regulated genes showing a 10-fold change *in planta* compared with urediniospores and presenting homology with known functions are listed in Dataset S1, Table S16. The complete *P. graminis* f. sp. *tritici* expression datasets are available at the GEO database (NCBI) as series GSE25020.

Targeted Annotation and Analysis of Specific Gene Families. Gene categories corresponding to proteins playing a role in fungal development, virulence, and biotrophy, such as effector-like proteins, carbohydrate degrading enzymes, and transporters, were analyzed in greater detail, and some annotations were updated upon review.

Effector/secretome. From several studies, it has become apparent that SSPs can play important and decisive roles in the manipulation of the plant immune system. Given the importance of SSP for virulence/avirulence, careful annotation of fungal genomes is required to accurately identify this commonly under annotated class of genes (44). The SignalP, TargetP, and TMHMM algorithms (45) allowed us to identify *M. larici-populina* proteins predicted to carry a signal peptide and no additional transmembrane domains, of which 1,184 had a protein length less than 300 aa following manual curation. To identify potential SSP missed by the gene callers, we used ESTs from poplar rust haustoria (27) and poplar rust-infected leaves for de novo gene discovery. Recursive TBLASTN searches of these candidates against the *M. larici-populina* genome helped in identifying additional

paralogous sequences. We identified 170 unpredicted SSP genes (>10% of the initial set of predicted SSPs). Interestingly, most of these corresponded to small cysteine-rich proteins (mean length, 111.7 aa; mean number of cysteine residues, 6.9). The small size and sequence divergence of these gene families have probably contributed to their underrepresentation in the gene predictions, as observed for small cysteine-rich peptides in plants (46). TribeMCL analyses identified 199 SSP families, but more relationships were unraveled using recursive BLAST analyses. SSP genes are organized in 169 families of two to 111 genes (Dataset S1, Table S17 lists families with more than three genes). In total, 814 of the 1,184 SSPs had no identifiable homologue in international databases or the wheat stem rust genome (BLASTP, E-value > 1e-6) and represent putative *Melampsora* specific SSP genes. Apart from the presence of the signal peptide, the only recognizable feature of SSPs is often a high percentage of cysteine residues. Of the 1,184 SSPs present in the *M. larici-populina* genome, 63% had more than four cysteine residues; for small SSPs (length of 101–150 aa), genes with more than eight cysteine residues are overrepresented, mostly because of the largest SSP family encompassing 111 members (Fig. S6*A*). These cysteines are presumed to play an important role in the stability of secreted proteins, and are typical features of some fungal and oomycete effectors (47, 48). Despite low sequence identity even for a given class, most of these proteins shared a common structure of five exons, with the first full codon of exons 2, 4, and 5 being a cysteine. The recent description of a conserved [Y/F/W]xC motif in the N-terminal region of secreted candidate effectors of the powdery mildew *B. graminis* and *Puccinia* spp. rust fungi (49, 50) following secretion signal cleavage site raised the question of existence of domains that could be involved to translocation of those effectors into the host cytoplasm similar to the RxLR domain in oomycete effectors (51). Following the procedure described by Godfrey and collaborators (49), [Y/F/W]xC motifs were also found abundant in poplar rust SSP genes compared with larger secreted proteins and other genes such as ribosomal proteins or conserved core fungal genes described in *Multigene Families and Evolutionary Analysis of Multigene Families*. However, in the case of *M. larici-populina*, the motifs are predominantly found after 30 aa following the secretion signal cleavage site. Also, the motif is not restricted to the N-terminal region and is also found in the C-terminal region of several SSPs as illustrated in the SSP family 1 (Fig. S6*A*). However, this YxC motif was also found in many other genes encoding nonsecreted proteins, and particularly proteins with overrepresented GO terms related to zinc binding and nucleic acid binding, suggesting that these [Y/F/W]xC motifs might not be involved in translocation mechanism but could have a role related to zinc binding or nucleic acid binding, like zinc finger transcription factors. Of the 22 known or putative effector proteins described in rusts, 20 were present in the *M. larici-populina* genome, and some exist in multigene families. These included the rust transferred protein (RTP1) from bean rust, as well as 19 HESPs and avirulence proteins AvrM and AvrP4 from flax rust. By using a lower stringency for identifying gene clusters (defined as groups of at least three SSPs with no more than four intervening genes), 106 clusters with three to six SSPs were found in *M. larici-populina*. These clusters were scattered across the genome assembly. Different studies have reported the presence of avirulence genes or candidate effectors in regions enriched in TEs and repeats (52). Localization in such genomic environments may have helped in faster evolution of these gene families to adapt to the host defense response. SSP genes families were not significantly localized in TE-rich regions (*Multigene Families and Evolutionary Analysis of Multigene Families*). To determine whether the amplification and diversification of the large families of paralogous SSPs in the *M. larici-populina* genome could be related to the expansion of specific transposons, repeats and TEs (defined in *Transposable Elements of* M. larici-populina *and* P. graminis *f. sp.* tritici), locations were identified in vicinity of SSP belonging to multigene families in the MCL analysis (*Multigene Families and Evolutionary Analysis of Multigene Families*) for 15 kb-windows (7.5 kb in 5′; 7.5 kb in 3′). A total of 36 elements (eight class II TIR-types, four class I LTR-types, and 23 NoCat types consensus) were linked to SSP genes falling in 11 different families (19 elements with SSP genes in family 1; Fig. S6*A*). Two families in which all genes were systematically associated with the same TIR-type TE were identified. SSP family-36 members (ProteinIDs 54662, 54664, 123264, 123266, 123267) are associated with the class II TIR Mela-B-R1199-MAP7 and SSP family-44 members (ProteinID 84257, 91014, 94957, 123215, 123905) are associated with the class II TIR Mela-B-G2809-MAP3. The NoCat type Mela-B-R386-MAP5 was found in the vicinity of 25 genes of the SSP family 1; however, we were not able to demonstrate a related expansion of the SSP genes and the NoCat consensus elements identified. *M. larici-populina* SSPs expression levels measured for urediniospores, germinated urediniospores, as well as during poplar leaf infection at 96 hpi revealed a large proportion of SSPs (49%) expressed above background for at least one biological condition. Among them, only 22% are expressed in urediniospores, 25% in germinated urediniospores, and 32% at 96 hpi. Interestingly, many transcripts encoding expansin related protein (ProteinIDs 34090, 78206, 71932, and 105838) and degradative enzymes (ProteinIDs 49700 and 109181) were highly expressed during germination, suggesting their preferential role during first steps of fungal development. In contrast, several SSPs transcripts peaked during biotrophic growth *in planta*. At 96 hpi, when haustorial structures are established *in planta*, massive inductions of transcripts displaying similarities with the *M. lini* HESPs or *U. fabae* Rust Transferred Protein 1 were observed (53–55). Despite the detection of known SSPs during parasitic growth, particularly after haustoria formation, the majority of transcripts accumulated *in planta* encodes unknown proteins, specifically identified in *M. larici-populina* genome, and could represent a large reservoir of new rust fungal effectors.

For *P. graminis* f. sp. *tritici*, similar methods were used to identify secreted proteins. All predicted proteins were first screened for the presence of a potential secretion signal using SignalP, requiring a HMM signal probability (Sprob) of at least 0.9; a total of 1,934 proteins fit this criteria. Proteins were then analyzed with TargetP to filter out potential mitochondrially targeted proteins (RC1 or 2). Next, proteins with potential transmembrane domains predicted by TMHMM were removed, requiring an helix of at least 18 aa not in the first 60 aa, to avoid overlap with the secretion signal, and at least one predicted helix. Last, proteins with predicted GPI-anchor sites predicted by the big-PI fungal predictor were flagged (total of 136 proteins). The final set of predicted secreted proteins total 1,386. To identify families of SSPs, the set of 1,106 predicted secreted proteins that were at most 300 aa in size were clustered. This set was compared with itself using BLAST, and the resulting hits were clustered based on the expect value into families using the mcl algorithm (version MCL-09-308) with an inflation value of 1.1. The resulting 164 clusters varied in size from two to 44 proteins. The largest cluster of 44 proteins contains a conserved set of eight cysteines, but otherwise weak overall similarity. Although most members of this family are not highly expressed in wheat, four proteins (PGTG_07275, PGTG_03101, PGTG_00970, and PGTG_00967) were highly represented in the haustorial EST set (by 57, 22, 14, and seven ESTs, respectively). The 44 proteins in this largest cluster were aligned using muscle, and a phylogeny was estimated using RAxML (56) (with the PROTMIXBLOSUM62 model); the highest likelihood tree was selected from 10 different starting trees, and the agreement with 1,000 bootstrap replicates were summarize (Fig. S6*B*). The alignment was converted into a sequence logo showing the location of the eight conserved cysteines (Fig. S6*B*); the conserved cysteines are distributed across the 221 average length of these proteins. Of the largest 10 clusters, only the ninth contained proteins with the previously described N-terminal [Y/F/W]xC motif (49).

The *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes contain all protein complexes needed for the classical eukaryotic secretion pathway. These proteins are well conserved among fungi, except for the signal recognition particle. Whole-genome expression array of *M. larici-populina* genes indicate that 96.5% of these gene models are constitutively and highly expressed, suggesting that the identified secretion pathway is functional and active both in urediniospores and *in planta*.

Transporters. *M. larici-populina* and *P. graminis* f. sp. *tritici* are obligate parasite and consequently must derive their nutrients from their plant hosts. As a consequence, a fine-tuned association exists between the two partners, whereby carbon and nitrogen nutrients must first enter fungal cells through efficient transport systems. Genes encoding transporters were annotated using Blastp and tBlastx algorithms with transporter encoding genes retrieved from NCBI GenBank, Transport Classification database (57), TransportDB (58), and the Saccharomyces Genome Database as queries. Lineage-specific gene gain and loss were estimated using CAFÉ (59).

Comparative analysis with other transportomes of basidiomycetes (Dataset S1, Table S18) highlights a significant reduction of the MFS transporters in both *M. larici-populina* and *P. graminis* f. sp. *tritici* lineages. However, five *M. larici-populina* genes encoding putative sugar porters are highly expressed *in planta*, whereas two others, which are closely related to *U. fabae* HXT1 (60), are specifically expressed during plant infection. *M. larici-populina* and *P. graminis* f. sp. *tritici* do not contain an orthologue of the *U. maydis Srt1* gene, recently demonstrated as a sucrose transporter (61). This is consistent with *in planta* expression of *M. larici-populina* and *P. graminis* f. sp. *tritici* invertases. Taken together these results suggest that hexose sugars are the main sugar supply for the two rust fungi *M. larici-populina* and *P. graminis* f. sp. *tritici*.

Interestingly, four genes classified as MFS anion:cation symporters (TC. 2.A.1.14) are expressed at 96 hpi in *M. larici-populina*. These four genes belong to the TNA1 clade, a high affinity transporter for nicotinic acid (vitamin B3) or structurally related compounds (62, 63), catalyzing uptake of nicotinic acid even at low concentrations. These results support a role of rust haustoria for vitamin B3 uptake and suggest that *M. larici-populina* might be dependent from its host for vitamin B3 supply. Another notable adaptation was found in the Amino acid–Polyamine–organoCation superfamily. Phylogenetic analyses reveal an expansion in *M. larici-populina* lineage for the Yeast Amino Acid Transporter family (Dataset S1, Table S18). This expansion is likely a result of duplication of *M. larici-populina* genes, orthologous to a broad specificity amino acid transporter *AAT3* from *U. fabae* (64). In addition, three genes are specifically expressed during plant infection and display homology with PIG2 (AAT1p, specific for lysine and histidine) and PIG27 (AAT2p, unknown substrate specificity) from *U. fabae* (65, 66). These results support the findings that haustoria contain amino acid transporters allowing uptake of plant amino acids.

In contrast to MFS superfamily, the *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes display an increased genetic potential for peptide uptake through oligopeptide transporters. Based on phylogenetic analysis, this expansion is specific to *M. larici-populina* and *P. graminis* f. sp. *tritici* lineages (Dataset S1, Table S18) and could be explained by recent and extensive duplication. Among this rust fungi specific clade, three *M. larici-populina* genes are expressed only *in planta* and one only in urediniospores, suggesting functional divergence. Third of the *M. larici-populina* OPT encoding genes are expressed *in planta* specifically or with a high level of expression. After uptake, the internalized peptides can be rapidly hydrolyzed by peptidases and used as source of amino acids, nitrogen, or carbon. This is consistent with the increased genetic potential of secreted proteases. Amplification of OPT gene family in both obligate fungi *P. graminis* f. sp. *tritici* and *M. larici-populina* might reveal a genomic adaptation to peptides available from the plant host.

Comparative analysis also highlights significant expansion of the Metal Ion (Mn^{2+}-iron) Transporter (Nramp) Family (TC.2.A.55) and the AE Auxin efflux family transporters (TC.2.A.69) in both *M. larici-populina* and *P. graminis* f. sp. *tritici* lineages. The former family contains transporters catalyzing uptake of divalent metal cations such as Fe^{2+}, Mn^{2+}, Zn^{2+}. *M. larici-populina* Nramp-encoding genes are expressed *in planta*, suggesting that these metals are important for plant infection. A striking expansion of auxin efflux family transporters is observed in both rust fungi. Furthermore, four genes of seven in *M. larici-populina* are up-regulated at 96 hpi compared with urediniospores (ratio from nine to 236), and two of them display an expression specific to 96 hpi. In *P. graminis* f. sp. *tritici*, three of six genes are expressed during wheat infection, one being specifically and highly expressed. In plants, such transporters are well characterized and control polar transport of auxins. Recent work highlights the crosstalk between plant defense and auxin signaling pathways (reviewed in ref. 67), and some suggest that auxin promotes virulence during biotrophic interactions (68, 69). However, it is not clear yet whether *M. larici-populina* or *P. graminis* f. sp. *tritici* are able to synthesize auxin. Both rust fungi clearly have homologues of potential aromatic amino acid aminotransferase genes (*tam1* and *tam2*) and indole-3 acetaldehyde dehydrogenase genes (*iad1* and *iad2*) involved in auxin synthesis in *U. maydis* (70); however, other intermediate genes likely required in the auxin synthesis pathway remain to be identified.

Annotation of putative CAZymes. Genes encoding carbohydrate-active enzymes in *M. larici-populina* and *P. graminis* f. sp. *tritici* were annotated by using the CAZY database (http://www.cazy.org) (71), organizing enzymes into families of GHs, GTs, PLs, CEs, and ancillary carbohydrate-binding modules (CBMs; Dataset S1, Table S19). *M. larici-populina* and *P. graminis* f. sp. *tritici* CAZyme repertoires were compared with those of other sequenced fungi, including plant parasites, and saprobic fungi.

The smallest number of GHs was found in organisms that have no significant interaction with plants (*M. globosa*, *S. cerevisiae*, *S. pombe*, and *C. neoformans*). The numbers of GHs encoded by the 21 fungal genomes examined (Fig. S8) revealed that, with 173 and 158 GHs, *M. larici-populina* and *P. graminis* f. sp. *tritici* have a relatively small number of GH-encoding genes. Although this number is higher than that found in the biotroph *U. maydis* and the ectomycorrhizal symbiont *T. melanosporum*, it is much lower than that found in phytopathogens such as *M. oryzae* ($n = 233$) or *Gibberella zeae* ($n = 248$) or saprotrophs (*Podospora anserina*, $n = 230$).

These two rust fungi clearly lack or present reduced sets of secreted hydrolytic enzymes active on PCW, indicating that rust fungi might use a limited set of CAZy to achieve colonization of plant tissue (three bottom clades in Fig. S8). Significantly, the genomes of the two rust fungi do not encode any protein with a cellulose-binding module of family CBM1. The absence or highly reduced number of proteins appended to a CBM1 domain is also observed in *U. maydis* and with the symbiotic fungi *L. bicolor* and *T. melanosporum.* The pectinolytic enzyme profile of *M. larici-populina* and *P. graminis* f. sp. *tritici* is reduced in accordance with their localization (i.e., leaf and stem as opposed to fruit).

Compared with other basidiomycetes and ascomycetes, the overall decrease in PCW digestion enzymes is accompanied by the conservation and even the expansion of several unexpected families such as family GH47 (α-mannosidases) and the multifunctional family GH5, which groups together enzymes that depolymerize the PCW (cellulose, mannan) or the fungal cell wall (β-1,3-glucan, β-1,6-glucan). Compared with the biotroph *U. maydis* or the hemibiotroph *M. oryzae*, the two rust fungi even show moderate expansions of a few GH families involved in cellulose and hemicellulose breakdown (e.g., GH7, GH10, GH12, GH26, and GH27). Several GH47, GH5, GH7, GH10,

GH12, and GH26 encoding genes are expressed during plant infection at 96 hpi compared with urediniospores. Finally, proteins distantly related to plant expansins are highly expressed during the plant infection. A major feature of rust fungal biotrophy is the formation of haustoria which are fungal invaginations into the plant plasmalemma. To build such a structure, the fungus needs to locally breach the PCW and then to derive a specific extrahaustorial matrix at the interface with the plant host. The specific PCW-targeting CAZymes identified in the poplar rust genome might reflect such requirements.

M. larici-populina and *P. graminis* f. sp. *tritici* genomes are enriched with chitin deacetylase CE4 encoding genes, as also shown in the symbiont *L. bicolor*. Conversion of exposed chitin to chitosan mediated by chitin deacetylase is thought to play an important role for the fungus to avoid recognition by their host defense system (72). *M. larici-populina* and *P. graminis* f. sp. *tritici* CE4 transcripts expression detected *in planta* support the role of these enzymes for fungal progression in the plant. The large number of CE4 found both in pathogenic and symbiotic biotrophs indicates a possible convergent adaptation to achieve colonization of the plant tissue by evading the plant defense system.

P. graminis f. sp. *tritici* and *M. larici-populina* differ widely in the number of cutinase genes (family CE5), with 12 genes in the latter and none in the former, which might reflect specialization to their host.

Proteases. Protease encoding genes were catalogued in the predicted proteome of *M. larici-populina*, *P. graminis f. sp. tritici*, *L. bicolor*, *P. chrysosporium*, *S. roseus*, *Mycosphaerella graminicola*, *C. cinerea*, *U. maydis*, *C. neoformans var. grubii*, and *M. oryzae*, according to the MEROPS database. The genome-wide analysis of gene families encoding putative secreted peptidases revealed that rust fungi possess the full endo- and exoprotease arsenal expected (belonging mainly to MEROPS subfamilies A01, S08, S09, and S10) (73) to cooperate efficiently in protein digestion (Dataset S1, Table S20).

The most noteworthy expansions among endopeptidase gene families [comparatively to their most common ancestor, *S. roseus*, and other fungal pathogens (*U. maydis* and *C. neoformans* var. *grubii*)] are of secreted aspartic proteases (MEROPS family A01) and subtilisin serine-proteases (family S08; Dataset S1, Table S20). Comparative genomics of members of the secreted aspartic protease A01 family revealed strong expansion of gene models encoding homologs to secreted CnAP1 peptidases (eight and five gene models in *M. larici-populina* and *P. graminis* f. sp. *tritici*, respectively) and polyporopepsin (seven models only in *M. larici-populina*), the latest expansion being *M. larici-populina* lineage-specific. Expansion of the subtilisin family appears specific to the *M. larici-populina* and *P. graminis* f. sp. *tritici* lineages. Among the eight genes, two display a strong expression at 96 hpi in *M. larici-populina*. Interestingly, trypsin and subtilisin serine-proteases may also be involved in direct cell wall degradation, as hydroxyproline-rich glycoproteins such as extensins are possible targets of these enzymes (74–76). We did not identify trypsin-like proteins in *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes, but found an expanded family of subtilisins in both genomes. This may indicate that members of subtilisins may play a role in cell wall degradation, required for fungal establishment within plant tissues.

Distribution of the exopeptidase families (including amino- and carboxypeptidases) found in *M. larici-populina* and *P. graminis* f. sp. *tritici* contrasted strongly with those found in saprotrophic (*C. cinerea*) and hemibiotrophic (*M. graminicola* and *M. oryzae*) fungi. Globally, genes encoding both nonsecreted and secreted metallopeptidases are poorly represented in rust fungi (Dataset S1, Table S20), with the consequence that families of secreted carboxypeptidase A (MEROPS family M14) and aminopeptidase Y (family M28) are drastically reduced in these fungi. On the contrary, *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes display an expansion of dipeptidyl-peptidases (family S9) and serine-carboxypeptidases (family S10) with three or one S9 peptidases and eight or nine S10 peptidases in *M. larici-populina* and *P. graminis* f. sp. *tritici*, respectively. In addition, expression profiling indicated that eight of these *M. larici-populina* genes are significantly up-regulated *in planta* and one is strongly expressed during spore germination.

Selection for the obligate biotrophic lifestyle may have favored the expansion (and/or retention) of highly specialized gene families of proteases comparatively to saprotrophic fungi and opportunistic pathogens, which are known to produce the broadest spectrum of protein and polysaccharide degrading enzymes, indicative of their less specialized nutritional status (77). Selection of one specific gene family of exopeptidases over another in rust fungi suggests that differences in the properties of the enzymes could have provided advantages during the process of evolution toward obligate pathogenesis.

One additional striking result concerns the diversification of the bleomycin hydrolases subfamily (C01B; papain family), exclusively in *M. larici-populina*. We identified a single cluster (37.3 kb) of four bleomycin hydrolase gene models (501–522 aa) interrupted by one TE on scaffold 26. They seem to have derived from each other by recent gene duplication event(s) and rearrangement, which likely occurred through repeated element(s). Bleomycin hydrolases are nonsecreted proteases being involved in cell stress and detoxification processes (inactivates bleomycin B2, a cytotoxic glycometallopeptide) and/or homocysteine-thiolactonase activities (protecting the cell against homocysteine toxicity). The incidence of such proteins putatively involved in "stress response" could be considered as a direct result of a stressful environment induced by the defense responses mounted by the host, even in compatible interactions (78).

Lipases (triacylglycerol hydrolases). The genome-wide analysis of lipase genes revealed that the rust fungi *M. larici-populina* and *P. graminis* f. sp. *tritici* possess genes representative of the different enzyme classes from the lipase extended family. Cutinases and carboxylesterases are the two superfamilies with the most noticeable expansions (Dataset S1, Table S21).

M. larici-populina genome contains a number of putative cutinases as high as *M. oryzae* genome. The expansion of cutinase genes in *M. larici-populina* and *P. graminis* f. sp. *tritici* is monophyletic, with a more recent diversification in *P. graminis* f. sp. *tritici* than in *M. larici-populina*. Eight of the 14 cutinases are organized in tandem that mostly originated from gene duplication. Interestingly, eight of these genes are expressed neither *in planta* nor in urediniospores, which raises the question of maintaining high number of cutinase encoding genes within the genome. These genes might have a role during infection of the alternate host, which was not included in the study. However, two genes, expressed only in spores and germinated spores, are good candidates for a role in the formation of adhesion pads in *M. larici-populina* (79, 80). Two other cutinase encoding genes are expressed only *in planta* at 96 hpi, a time point at which the vast majority of the fungus lives inside plant tissues. Functions of such genes remain to be clarified.

The superfamily abH3 (*Candida rugosa* lipase-like) containing carboxylesterase displays a significant expansion compared with other basidiomycetes and ascomycetes genomes (Dataset S1, Table S21). A phylogenetic analysis revealed that *M. larici-populina* proteins of the carboxylesterase family belong to two phylogenetic clades, with one of the clades encompassing 22 of the 23 members. Two different carboxylesterases expressed at the surface of *Uromyces viciae-fabae* urediniospores have been proposed to play a role in the adhesion of spores to the plant surface (75). In addition, different carboxylesterase isoforms were detected within symbiotic tissues in the ectomycorrhiza, whereas almost no activity of carboxylesterase was observed in pure culture of the mycorrhizal basidiomycete *Suillus bovinus* (81), but there is no hypothesis to explain the role of such carboxylesterase.

Cytochrome P450. Cytochrome P450 enzymes are a superfamily of heme-containing monooxygenases that show extraordinary diversity. In fungi, their role in compound detoxification has been found to be important in establishment in specialized niches (82). They have also been implicated in pathogenesis, in particular in detoxification of plant defense (83, 84).

In total, 29 and 17 P450s were found in *M. larici-populina* and *P. graminis* f. sp. *tritici*, respectively, which is far less than the large numbers and diversity of P450s found in wood-decomposing fungi such as 353 in *Postia placenta*, and 140 in *P. chrysosporium* (85) or 84 in the mycorrhizal fungus *L. bicolor* (31). Interestingly, no P450s were found in the genomes of two intracellular obligate microsporidia (*Encephalitozoon cuniculi* and *Antonospora locustae*) (86). It can be hypothesized that this low number of P450s in rust fungi is linked to the nonintrusive nature of the infection process of rust fungi and the establishment of haustoria within host cells without cell death. Twenty-two *M. larici-populina* P450s belonged to the CYP4/CYP19/CYP26 subfamily associated with lipid transport and metabolism, whereas three P450s belonged to the CYP2 subfamily that has been associated with drug and toxin metabolism. An additional four were unspecified monooxygenases.

Redox control and oxidative stress. Genes coding for thiol oxidoreductases, ROS scavenging systems, and GSTs have been identified in the genomes of rust fungi. After manual curation, the *M. larici-populina* and *P. graminis* f. sp. *tritici* genes were compared with a set of 11 sequenced fungal genomes, including five basidiomycetes, five ascomycetes and one zygomycete (Dataset S1, Tables S22–S24).

For the thiol oxidoreductase enzymes, the number of thioredoxin and glutaredoxin (GRX) genes vary between species. Compared with most fungi, which possess two or three genes, *M. larici-populina*, as well as two other basidiomycetes (*L. bicolor* and *P. chrysosporium*) with completely different lifestyles, possesses four thioredoxin genes (Dataset S1, Table S22). This expansion could be explained by duplication events in specific organisms, but does not reflect adaptation to common physiology. For the GRX family, the gene number ranges from four to seven, with no clear relationship to the fungal phylum or physiology.

For ROS-detoxifying enzymes, *M. larici-populina* and *P. graminis* f. sp. *tritici* have more SOD and catalase (CAT) than any other fungal genome examined. This is especially remarkable for the *P. graminis* f. sp. *tritici* SOD family, which is composed of 20 members (Dataset S1, Table S23). For the CAT family, there are three and eight members, respectively, in *M. larici-populina* and *P. graminis* f. sp. *tritici*, whereas other fungi possess one to five CAT encoding genes. Whether all of the *M. larici-populina* genes are expressed is uncertain, as only one gene model is supported by ESTs and, from microarrays, four genes are not expressed in the conditions tested. The impact of the CAT family expansion might be attenuated by the presence of only four TPX genes in both fungi, whereas other fungi possess five to nine TPXs.

A recent phylogenomic analysis updated the classification of fungal GSTs into seven groups (87). Some relationships between the occurrence or expansion of specific GST classes and the fungal physiology have been established. Overall, the number of GST genes ranges from five to 32, with mycorrhizal and saprobic fungi having the largest family (Dataset S1, Table S24). With 5 and 10 genes, respectively, *P. graminis* f. sp. *tritici* and *M. larici-populina* are among the less well equipped organisms, and the absence of GTE is consistent with their obligate biotrophic nature.

Examining expression profiles of all these defense-related genes during a compatible reaction, eight *M. larici-populina* genes are strongly (>10-fold) and significantly up-regulated after 96 hpi. In particular, the two most induced genes encode a GRX and a sulfhydryl oxidase, presumably located in the endoplasmic reticulum. By contrast, of the 15 *M. larici-populina* genes encoding SOD and CAT, only three are significantly up-regulated at this time point. This might be correlated to the absence of ROS production at infection sites in the rust-poplar model both during compatible or incompatible interactions (2).

Signal transduction pathways. Genome-wide analysis of signaling gene families was carried out for *M. larici-populina* and *P. graminis* f. sp. *tritici* based on previous catalogs of annotated fungal signaling proteins (88–91). Gene families targeted include G protein-coupled receptors (GPCR), sensory transduction histidine kinases, G proteins (small monomeric G protein Ras, α, β and γ subunits of heterotrimeric G proteins, regulators of G protein), phospholipases C, cyclic-AMP (cAMP)-related signaling proteins (adenylate cyclase AC, AC-associated protein CAP, cyclic nucleotide phosphodiesterase, and catalytic and regulatory subunits of cAMP-dependent protein kinases PKA), and MAP kinases that mediate input signals through cascade phosphorylation (i.e., MAP kinases, MAPK kinases, and MAKK kinases).

Our survey indicates that both *M. larici-populina* and *P. graminis* f. sp. *tritici* possess proteins involved in pathways controlling processes such as stress response, filamentous growth, mating, and virulence (92, 93) (Dataset S1, Table S25). A total of four pheromone receptors genes were identified (*Mating Type Genes in Rust Fungal Pathogens*) and new types of GPCR-like proteins with typical features of this type of membrane protein were also identified, as in the genomes of other basidiomycetes (31). A number of MAP kinases encoding genes similar to those reported for pathogenic fungi were identified in the poplar rust genome, notably homologues of the MAPKs of the Slt2 pathway (Bck, Mkk, Mpk) related to virulence in *U. maydis* (93). In general, most gene families were highly expressed in resting and germinating urediniospores of both rust fungi, suggesting that this spore type is equipped with a complete set of transduction proteins at the time first contacts are made with the host. Only a few signaling genes presented noticeable transcripts levels in infected plant tissues; these include the G proteins Ras1 and three GPa subunits as well as Gpb and Gpg subunits, and components related to the cAMP-pathways, such as cAMP-related GPCR, CAP, and PKA catalytic subunits (although AC expression was not detected). *P. graminis* f. sp. *tritici* genome lacks a gene coding a low-affinity phosphodiesterase that is present in *M. larici-populina*.

M. larici-populina and *P. graminis* f. sp. *tritici* contain large numbers of sequences with regions matching the α-kinase hidden Markov model (PF02816). Alignment of these sequences and known α-kinases (from the Kinase Database; http://www.kinase.com) indicates that *M. larici-populina* and *P. graminis* f. sp. *tritici* have 79 and 32 lineage-specific α-kinases, respectively, 12 and seven of which contain a full complement of structural and catalytic residues (94). The expansion of the active α-kinases in *P. graminis* f. sp. *tritici* and *M. larici-populina* is a result of lineage-specific paralogous expansion (Fig. S7); the predicted inactive α-kinases largely fall into two lineage-specific clusters based on phylogenetic estimation for both active and inactive α-kinases. *M. larici-populina* and *P. graminis* f. sp. *tritici* also contain, respectively, five and 34 α-kinase-like sequences that match only the C-terminal region of the domain; these are missing all predicted catalytic residues.

Mating type genes in rust fungal pathogens. Consistent with a heterothallic nature, possibly mediated by two unlinked mating type loci, the rust fungi *M. larici-populina* and *P. graminis* f. sp. *tritici* have HD1 and HD2 homeodomain transcription factor genes as well as pheromone and pheromone receptor genes, as found in mating type loci of other basidiomycetes (95). The *M. larici-populina* and *P. graminis* f. sp. *tritici* loci encoding homeodomain transcription factors display a simple structure with just one divergently transcribed gene pair for proteins with an HD1 and an HD2 homeodomain motif, respectively. Based on this structure shared with the *b* mating type locus of the Ustilaginomycotina (96), we also call this putative mating type locus *b* and the enclosed genes *bE* and *bW*. Sequences of the homeodomains in the *M. larici-populina* and *P. graminis* f. sp. *tritici bE1* and *bW1* transcription factors are similar to the ones of other basidiomycetes (95).

M. larici-populina and *P. graminis* f. sp. *tritici* genomes contain 11 and 3 putative pheromone precursor genes, respectively. *MlpPh1* to *MlpPh11* are found on five different scaffolds, three of which (scaffolds 83, 113, and 172) also carry a putative pheromone receptor gene (*MlpSTE3*). *MlpPh1* and *MlpSTE3.4* are directly adjacent to each other on scaffold 72, whereas 43 kb and 80 kb separate *MlpPh2* and *MlpSTE3.2* (scaffold 83) and *MlpPh3* and *MlpSTE3.3* (scaffold 113), respectively. A fourth receptor gene, *MlpSTE3.1*, is present on scaffold 3. Interestingly, there are several TE insertions in the scaffolds with pheromone precursor genes and pheromone receptor genes, which hinders assembling the scaffolds into larger units. It is therefore unclear whether some or all of them might form a very large single mating type locus with multiple pheromone precursor and pheromone receptor genes possibly presented on a transposon-rich sex chromosome, as known from the anther smut *Microbotryum violaceum* (97, 98).

Devier et al. (99) recently detected an ancient transspecific polymorphism at pheromone receptor genes in basidiomycetes with a 370-Mya split into two distinct lineages within the Pucciniomycotina that defines the alternate mating types of *M. violaceum*. The three known pheromone receptors from the wheat rust *P. graminis* f. sp. *tritici* and the *M. larici-populina* receptors cluster with the receptor pr-MATA2 of the A2 mating type of *M. violaceum*.

Noncoding RNAs. ***tRNA gene abundance, anticodon/codon use, and translational selection.*** The tRNAScan-SE (100) algorithm, applied with default parameters and eukaryotic model to the assembly, resulted in the identification of 253 putative tRNAs in the *M. larici-populina* genome, 24 being found in the mitochondrion scaffolds. A total of 194 of these contain an intron. Forty-nine of 61 possible anticodon tRNA, as well as two selenocysteine tRNAs, were detected and no suppressor tRNAs (i.e., anticodon that binds stop codons) were found. In addition, 10 tRNA pseudogenes were detected by tRNAScan-SE. For the *P. graminis* f. sp. *tritici* genome, tRNAScan-SE predicted a total of 428 tRNAs, of which 366 contain an intron. A total of 47 of the possible 61 anticodon tRNAs were identified. Whereas *M. larici-populina* contains no more than 11 copies of any tRNA, *P. graminis* f. sp. *tritici* contains five tRNAs amplified to 22 to 55 copies. Fifty-two pseudo- and 21 undetermined or questionable tRNAs were also predicted. For comparison, the numbers of tRNA genes detected in the *C. neoformans*, *P. chrysosporium*, and *L. bicolor* genomes were 141, 200, and 279, respectively.

Spliceosomal RNAs (snRNA). The cmsearch program of the INFERNAL package (101) was used to scan the *M. larici-populina* genome for spliceosomal RNA genes, by using appropriate covariance model from Rfam (102). For each covariance model, the window size and trusted cutoff score indicated in the RFAM database were used. Three spliceosomal snRNAs were discovered in the genome, with seven copies of U2 snRNA, two copies of U4 snRNA, and six copies of U11 snRNA. No U1, U5, and U11 snRNA candidate genes were identified in the genome. No U12 snRNA were detected, in accordance with the lack of U12-type introns in this genome as in the genomes of all other fungi analyzed until now. The snRNAs detected in the *M. larici-populina* genome were not verified experimentally.

rRNA. For *M. larici-populina*, *M. larici-populina* 5.8 S ITS and *P. graminis* f. sp. *tritici* 28S and 18S rRNA sequences retrieved from NCBI were used as a query against all *M. larici-populina* scaffolds (main genome, alternate haplotypes, repetitive elements) with the BLASTN algorithm. A complete rDNA tandem unit of 10.3 kb was reconstructed and then used as a query to identify all rDNA units present in the genome sequence. A total of 22 complete rDNA copies were identified and 11 extra 18S and 20S rDNA sequences were found elsewhere in the genome. A total of 16 rDNA copies were present in the main genome scaffolds, nine found in tandem on scaffold 4 (three copies), scaffold 20 (two copies), scaffold 51 (two copies) and scaffold 72 (two copies), one copy was found in the alternate haplotype scaffolds and five in five different scaffolds of the repetitive-elements scaffolds.

For *P. graminis* f. sp. *tritici*, 18S, 28S, and other available rRNA sequences from GenBank were used to search the *P. graminis* f. sp. *tritici* assembly using BLAST. Within the assembly, these sequences match only the start of scaffold 33 and scaffold 392, which is less than 3 kb in size. This suggests that scaffold 392 may be part of the rDNA array at the end of scaffold 33, and that there may be only one tandem rDNA locus in *P. graminis* f. sp. *tritici*. To estimate the copy number of the rDNA repeat, the 28S gene was used to BLAST search all the sequence reads. The total coverage across the 28S was normalized for the average assembly coverage; these data suggests that there are at least 16 copies of the rDNA unit. Any sequencing or cloning bias against this region would underestimate the rDNA copy number.

Mitochondrion Genome. After Arachne assembly, four *M. larici-populina* scaffolds (47.6, 18.7, 8.8, and 4.4 kb) were identified as putative mitochondrial sequence based on high AT content. The mitochondrial genome was estimated at 79.5 kb; the scaffolds include 6% gaps and a GC content of 31.3%. This assembly was not subjected to experimental validation.

The mitochondrial sequence for *P. graminis* f. sp. *tritici* is similar in size, totaling 79.2 kb in six contigs, and will be further described separately.

1. Frey P, Gérard P, Feau N, Husson C, Pinon J (2005) Variability and population biology of *Melampsora* rusts on poplar. *Rust Diseases of Willow and Poplar*, eds Pei MH, McCracken AR (CABI Publishing, Cambridge), pp 63–72.
2. Rinaldi C, et al. (2007) Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina*. *Plant Physiol* 144:347–366.
3. Leonard KJ, Szabo LJ (2005) Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol Plant Pathol* 6:99–111.
4. Jaffe DB, et al. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13:91–96.
5. Cuomo CA, Birren BW (2010) The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol* 470:833–855.
6. Barnes CW, Szabo LJ (2008) A rapid method for detecting and quantifying bacterial DNA in rust fungal DNA samples. *Phytopathology* 98:115–119.
7. Mathewson CA, Schein JE, Marra MA (2007) Large-scale BAC clone restriction digest fingerprinting. *Curr Protoc Hum Genet Apr* Chapter 5:Unit 5.19.
8. Schein JE, et al. (2004) High-throughput BAC fingerprinting. *Methods Mol Biol* 255:143–156.
9. Soderlund C, Longden I, Mott R (1997) FPC: A system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* 13:523–535.
10. Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res* 10:1772–1787.
11. Fjell CD, Bosdet I, Schein JE, Jones SJ, Marra MA (2003) Internet Contig Explorer (iCE)—a tool for visualizing clone fingerprint maps. *Genome Res* 13(6A):1244–1249.
12. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
13. Catanzariti A-M, Mago R, Ellis J, Dodds P (2011) Constructing haustorium-specific cDNA libraries from rust fungi. *Methods in Molecular Biology: Plant Immunity*. ed McDowell JM (Humana Press, Springer Science+Business Media, New York).
14. Quesneville H, et al. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLOS Comput Biol* 1:166–175.
15. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* 12:1269–1276.
16. Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 21:152–159.
17. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.
18. McCarthy EM, McDonald JF (2003) LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* 19:362–367.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
20. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
21. Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* 101:12404–12410.
22. Murat C, et al. (2010) Distribution and localization of microsatellites in the Perigord black truffle genome and identification of new molecular markers. *Fungal Genet Biol*, 10.1016/j.fgb.2010.10.007.
23. Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10:516–522.

24. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14: 988–995.
25. Martinez D, Grigoriev I, Salamov AA (2010) Annotation of protein-coding genes in fungal genomes. *Appl Comput Math* 9:56–65.
26. Foissac S, et al. (2008) Genome annotation in plants and fungi: EuGene as a model platform. *Curr Bioinformatics* 3:87–97.
27. Joly DL, Feau N, Tanguay P, Hamelin RC (2010) Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11:422.
28. Parra G, Blanco E, Guigó R (2000) GeneID in *Drosophila*. *Genome Res* 10:511–515.
29. Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: A Web server for gene finding in eukaryotes. *Nucleic Acids Res* 32(Web Server Issue): W309–W312.
30. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
31. Martin F, et al. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452:88–92.
32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
33. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
34. Martens C, Vandepoele K, Van de Peer Y (2008) Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc Natl Acad Sci USA* 105:3427–3432.
35. Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 266:418–427.
36. Conesa A, et al. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676.
37. Saeed AI, et al. (2006) TM4 microarray software suite. *Methods Enzymol* 411: 134–193.
38. Simillion C, Janssens K, Sterck L, Van de Peer Y (2008) i-ADHoRe 2.0: An improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24: 127–128.
39. Rensing SA, et al. (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol Biol* 7:130.
40. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
41. Hacquard S, et al. (2010) Laser capture microdissection of uredinia formed by *Melampsora larici-populina* revealed a transcriptional switch between biotrophy and sporulation. *Mol Plant Microbe Interact* 23:1275–1286.
42. Liu Z, Szabo LJ, Bushnell WR (1993) Molecular cloning and analysis of abundant and stage-specific mRNAs from Puccinia graminis. *Mol Plant Microbe Interact* 6: 84–91.
43. Caldo RA, Nettleton D, Wise RP (2004) Interaction-dependent gene expression in Mla-specified response to barley powdery mildew. *Plant Cell* 16:2514–2528.
44. Ellis JG, Dodds PN, Lawrence GJ (2007) The role of secreted proteins in diseases of plants caused by rust, powdery mildew and smut fungi. *Curr Opin Microbiol* 10: 326–331.
45. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.
46. Silverstein KAT, et al. (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J* 51:262–280.
47. Kamoun S (2006) A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu Rev Phytopathol* 44:41–60.
48. Stergiopoulos I, de Wit PJGM (2009) Fungal effector proteins. *Annu Rev Phytopathol* 47:233–263.
49. Godfrey D, et al. (2010) Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics* 11:317.
50. Spanu PD, et al. (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330:1543–1546.
51. Kale SD, et al. (2010) External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell* 142:284–295.
52. Oliva R, et al. (2010) Recent developments in effector biology of filamentous plant pathogens. *Cell Microbiol* 12:705–715.
53. Dodds PN, Lawrence GJ, Catanzariti AM, Ayliffe MA, Ellis JG (2004) The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16:755–768.
54. Catanzariti A-M, Dodds PN, Lawrence GJ, Ayliffe MA, Ellis JG (2006) Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18:243–256.
55. Kemen E, et al. (2005) Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol Plant Microbe Interact* 18:1130–1139.
56. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
57. Saier MH, Jr., Tran CV, Barabote RD (2006) TCDB: The Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 34(Database issue):D181–D186.
58. Ren QH, Chen KX, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 35(Database issue):D274–D279.
59. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–1271.
60. Voegele RT, Struck C, Hahn M, Mendgen K (2001) The role of haustoria in sugar supply during infection of broad bean by the rust fungus Uromyces fabae. *Proc Natl Acad Sci USA* 98:8133–8138.
61. Wahl R, Wippel K, Goos S, Kämper J, Sauer N (2010) A novel high-affinity sucrose transporter is required for virulence of the plant pathogen *Ustilago maydis*. *PLoS Biol* 8:e1000303.
62. Llorente B, Dujon B (2000) Transcriptional regulation of the *Saccharomyces cerevisiae* DAL5 gene family and identification of the high affinity nicotinic acid permease TNA1 (YGR260w). *FEBS Lett* 475:237–241.
63. Klebl F, Zillig M, Sauer NF (2000) Transcription of the yeast TNA1 gene is not only regulated by nicotinate but also by p-aminobenzoate. *FEBS Lett* 481:86–87.
64. Struck C, Mueller E, Martin H, Lohaus G (2004) The Uromyces fabae UfAAT3 gene encodes a general amino acid permease that prefers uptake of *in planta* scarce amino acids. *Mol Plant Pathol* 5:183–189.
65. Struck C, Ernst M, Hahn M (2002) Characterization of a developmentally regulated amino acid transporter (AAT1p) of the rust fungus Uromyces fabae. *Mol Plant Pathol* 3:23–30.
66. Hahn M, Neef U, Struck C, Göttfert M, Mendgen K (1997) A putative amino acid transporter is specifically expressed in haustoria of the rust fungus *Uromyces fabae*. *Mol Plant Microbe Interact* 10:438–445.
67. Grant MR, Jones JDG (2009) Hormone (dis)harmony moulds plant health and disease. *Science* 324:750–752.
68. Navarro L, et al. (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* 312:436–439.
69. Ribot C, et al. (2008) Susceptibility of rice to the blast fungus, *Magnaporthe grisea*. *J Plant Physiol* 165:114–124.
70. Reineke G, et al. (2008) Indole-3-acetic acid (IAA) biosynthesis in the smut fungus *Ustilago maydis* and its relevance for increased IAA levels in infected tissue and host tumour formation. *Mol Plant Pathol* 9:339–355.
71. Cantarel BL, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res* 37(Database issue): D233–D238.
72. El Gueddari NE, et al. (2002) Developmentally regulated conversion of surface-exposed chitin to chitosan in cell walls of plant pathogenic fungi. *New Phytol* 156: 103–112.
73. Monod M, et al. (2002) Secreted proteases from pathogenic fungi. *Int J Med Microbiol* 292:405–419.
74. Dow JM, Davies HA, Daniels MJ (1998) A metalloprotease from *Xanthomonas campestris* that specifically degrades proline/hydroxyproline-rich glycoproteins of the plant extracellular matrix. *Mol Plant Microbe Interact* 11:1085–1093.
75. Carlile AJ, et al. (2000) Characterization of SNP1, a cell wall-degrading trypsin, produced during infection by *Stagonospora nodorum*. *Mol Plant Microbe Interact* 13: 538–550.
76. Di Pietro A, et al. (2001) Molecular characterization of a subtilase from the vascular wilt fungus *Fusarium oxysporum*. *Mol Plant Microbe Interact* 14:653–662.
77. St Leger RJ, Joshi L, Roberts DW (1997) Adaptation of proteases and carbohydrates of saprophytic, phytopathogenic and entomopathogenic fungi to the requirements of their ecological niches. *Microbiology* 143:1983–1992.
78. Bindschedler LV, et al. (2009) *In planta* proteomics and proteogenomics of the biotrophic barley fungal pathogen Blumeria graminis f. sp. *hordei*. *Mol Cell Proteomics* 8:2368–2381.
79. Deising H, Nicholson RL, Haug M, Howard RJ, Mendgen K (1992) Adhesion pad formation and the involvement of cutinase and esterases in the attachment of uredospores to the host cuticle. *Plant Cell* 4:1101–1111.
80. Laurans F, Pilate G (1999) Histological aspects of a hypersensitive response in poplar to *Melampsora larici-populina*. *Phytopathology* 89:233–238.
81. Timonen S, Sen R (1998) Heterogeneity of fungal and plant enzyme expression in intact Scots pine-*Suillus bovinus* and-*Paxillus involutus* mycorrhizospheres developed in natural forest humus. *New Phytol* 138:355–366.
82. Stajich JE, et al. (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc Natl Acad Sci USA* 107:11889–11894.
83. Vanetten HD, Matthews DE, Matthews PS (1989) Phytoalexin detoxification: importance for pathogenicity and practical implications. *Annu Rev Phytopathol* 27:143–164.
84. Coleman JJ, et al. (2009) The genome of Nectria haematococca: contribution of supernumerary chromosomes to gene expansion. *PLoS Genet* 5:e1000618.
85. Yadav JS, Doddapaneni H, Subramanian V (2006) P450ome of the white rot fungus Phanerochaete chrysosporium: Structure, evolution and regulation of expression of genomic P450 clusters. *Biochem Soc Trans* 34:1165–1169.
86. Park J, et al. (2008) Fungal cytochrome P450 database. *BMC Genomics* 9:402.
87. Morel M, Ngadin AA, Droux M, Jacquot JP, Gelhaye E (2009) The fungal glutathione S-transferase system. Evidence of new classes in the wood-degrading basidiomycete *Phanerochaete chrysosporium*. *Cell Mol Life Sci* 66:3711–3725.
88. Borkovich KA, et al. (2004) Lessons from the genome sequence of Neurospora crassa: tracing the path from genomic blueprint to multicellular organism. *Microbiol Mol Biol Rev* 68:1–108.
89. Li L, Wright SJ, Krystofova S, Park G, Borkovich KA (2007) Heterotrimeric G protein signaling in filamentous fungi. *Annu Rev Microbiol* 61:423–452.
90. Lafon A, Han K-H, Seo J-A, Yu J-H, d'Enfert C (2006) G-protein and cAMP-mediated signaling in aspergilli: A genomic perspective. *Fungal Genet Biol* 43:490–502.
91. Xue C, Hsueh YP, Heitman J (2008) Magnificent seven: Roles of G protein-coupled receptors in extracellular sensing in fungi. *FEMS Microbiol Rev* 32:1010–1032.
92. Román E, Arana DM, Nombela C, Alonso-Monge R, Pla J (2007) MAP kinase pathways as regulators of fungal virulence. *Trends Microbiol* 15:181–190.
93. Zhao X, Mehrabi R, Xu J-R (2007) Mitogen-activated protein kinase pathways and fungal pathogenesis. *Eukaryot Cell* 6:1701–1714.
94. Drennan D, Ryazanov AG (2004) Alpha-kinases: Analysis of the family and comparison with conventional protein kinases. *Prog Biophys Mol Biol* 85:1–32.

95. Casselton LA, Olesnicky NS (1998) Molecular genetics of mating recognition in basidiomycete fungi. *Microbiol Mol Biol Rev* 62:55–70.
96. Kahmann R, Schirawski J (2007) Mating in the smut fungi: From a to b to the downstream cascades. *Sex in Fungi: Molecular Determination and Evolutionary Implications*, eds Heitman J, Kronstad J, Taylor J, Casselton L (ASM, Washington, DC), pp 377–387.
97. Hood ME (2002) Dimorphic mating-type chromosomes in the fungus *Microbotryum violaceum*. *Genetics* 160:457–461.
98. Hood ME (2005) Repetitive DNA in the automictic fungus *Microbotryum violaceum*. *Genetica* 124:1–10.
99. Devier B, Aguileta G, Hood ME, Giraud T (2009) Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes. *Genetics* 181:209–223.
100. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.
101. Eddy SR (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* 3:18.
102. Griffiths-Jones S, et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33(Database issue):D121–D124.

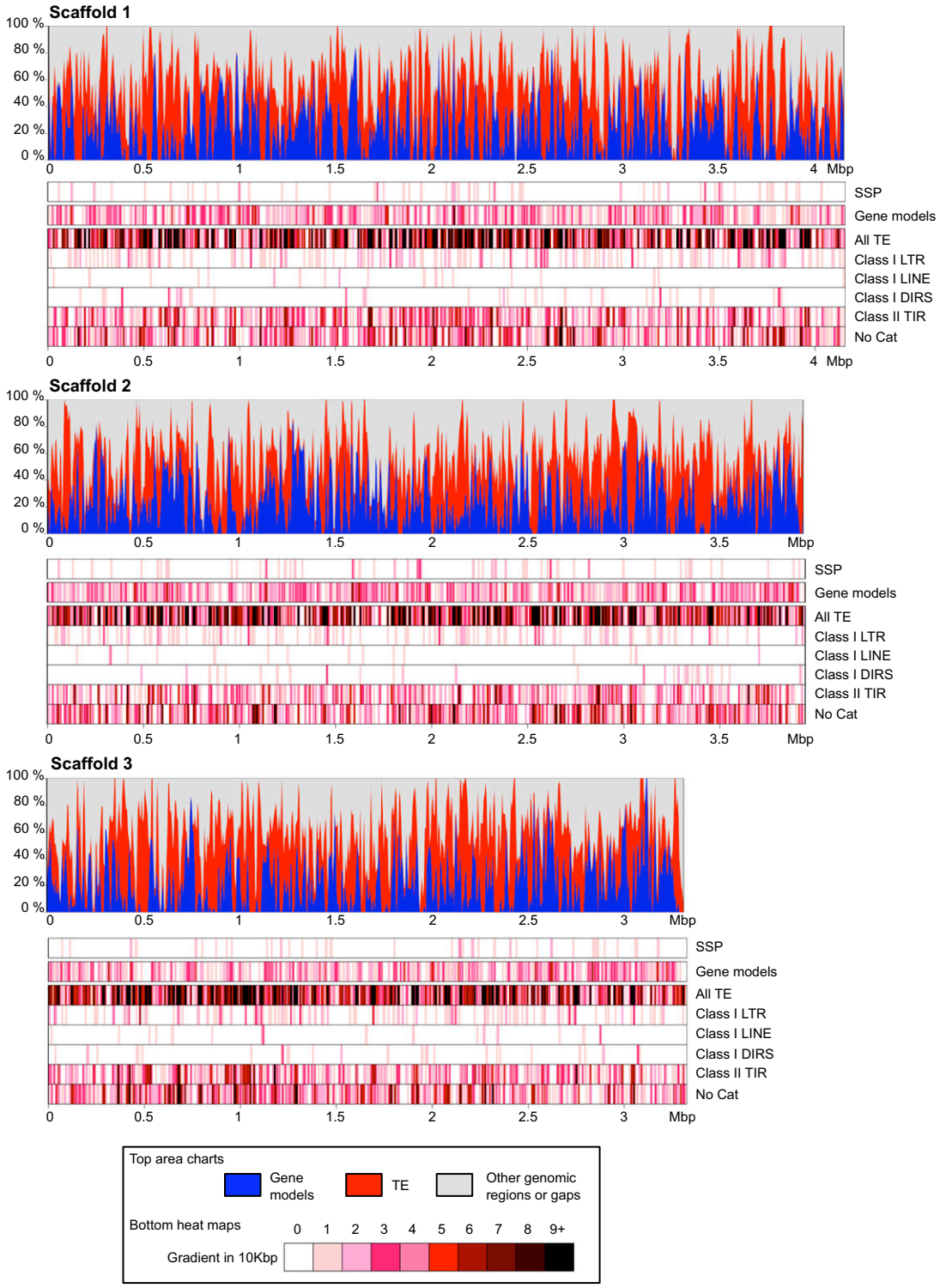

Fig. S1. Genomic landscape of *M. larici-populina*. The top area chart quantifies the distribution of TEs and protein-coding genes along *M. larici-populina* genome scaffolds 1, 2, and 3 (in Mb). The *y* axis represents the percentage of bases corresponding to TE (red), gene models (blue), and other genomic regions or gaps (gray) in 10-kb sliding windows. The heat maps display the distribution of selected elements including SSP-coding gene models, all protein-coding gene models, all TEs, LTR retrotransposons (class I LTR), long interspersed elements (class I LINE), *Dictyostelium* intermediate repeat sequence (class I DIRS), TIRs (class II TIR), and TE of unknown classes (TE No Cat). The heat maps were constructed by counting the number of elements in 10-kb sliding windows along the scaffold sequences. The frequency of the various elements ranged from white (0 element/10 kb) to black (≥9 elements/10 kb) colors.

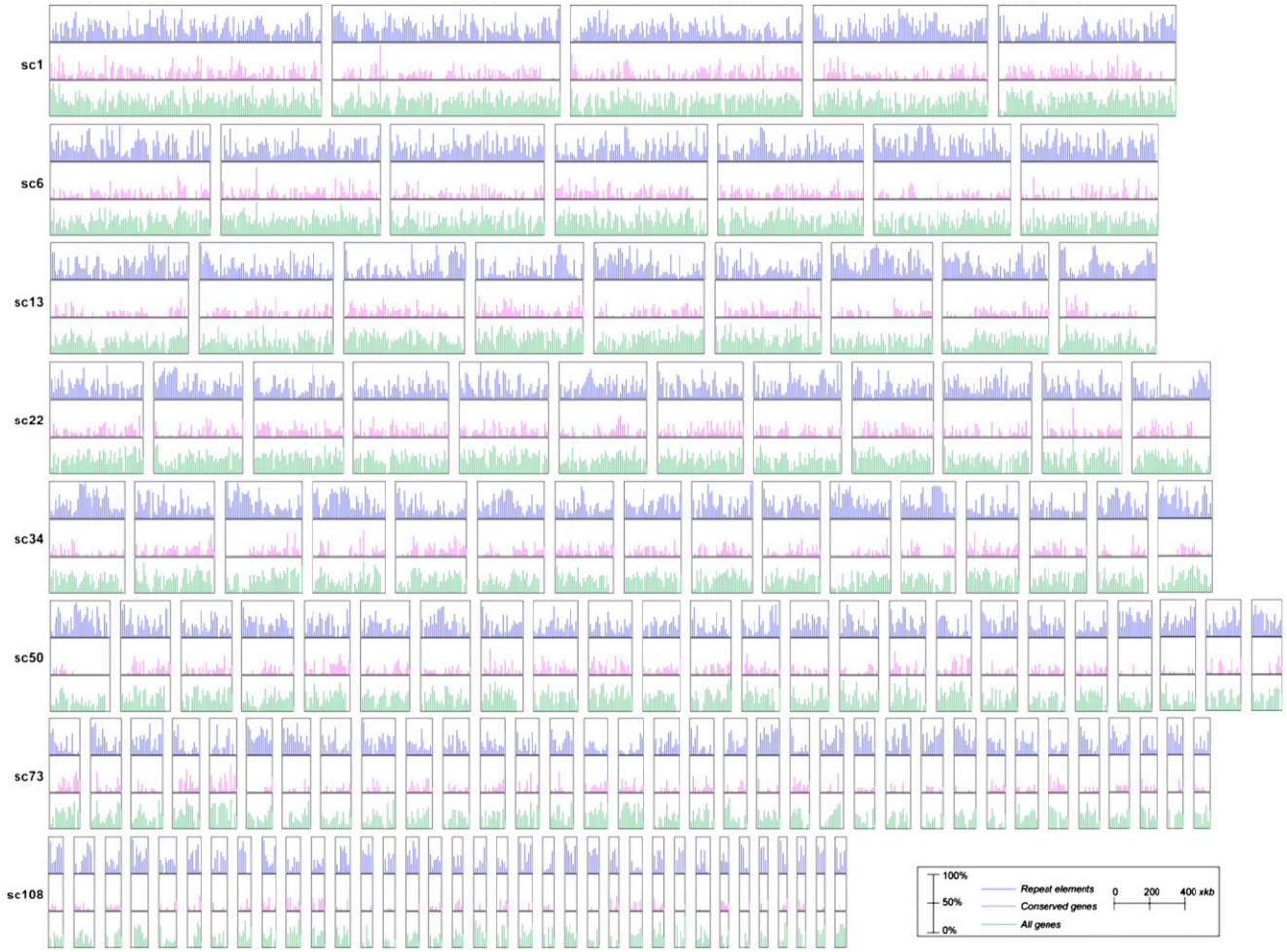

Fig. S2. Distribution of repetitive elements and genes across the *P. graminis* f. sp. *tritici* assembly. The location of repetitive elements (blue) and genes (purple, green) in the *P. graminis* f. sp. *tritici* genome are shown for 20-kb windows as the percentage covered for each window for all scaffolds greater than 100 kb.

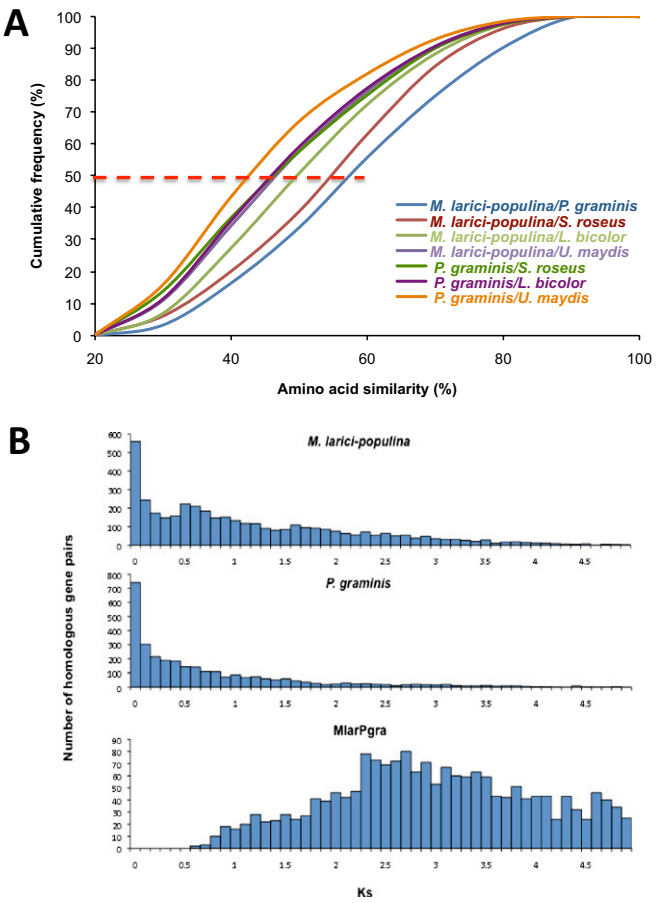

Fig. S3. Molecular divergence between Pucciniomycotina and other Basidiomycota and between Pucciniomycotina paralogous and orthologous gene pairs. (*A*) *M. larici-populina* and *P. graminis* have large protein sequence divergence with other Basidiomycota genomes. More than half of *M. larici-populina* and *P. graminis* orthologous genes have protein sequence similarity of at least 60%. By contrast, half of the orthologous genes between *P. graminis and U. maydis* share less than 40% sequence similarity. Orthologous genes were identified based on reciprocal best hits. Protein sequence similarity was calculated from Smith–Waterman alignments. (*B*) Age distribution of paralogous and orthologous gene pairs. The vertical axis indicates the number of gene pairs and the horizontal axis measures Ks, the synonymous substitution rate, for homologous gene pairs, estimated with codeml in PAML (40). Gene pairs were regarded as homologous if the aligned region was longer than 150 aa and if the sequences shared more than 30% similarity. (MlarPgra: *M. larici-populina* and *P. graminis* orthologous gene pairs)

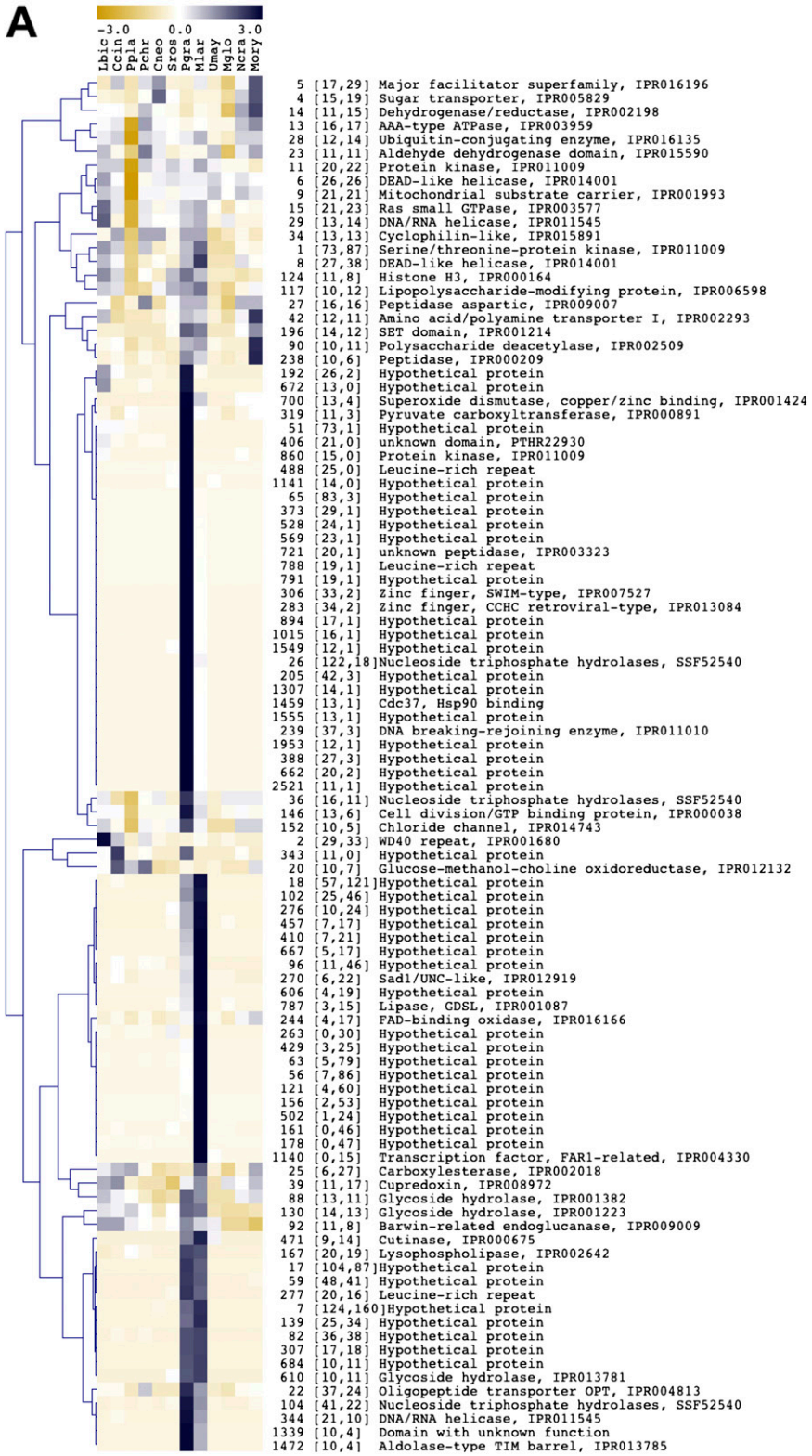

Fig. S4. (Continued)

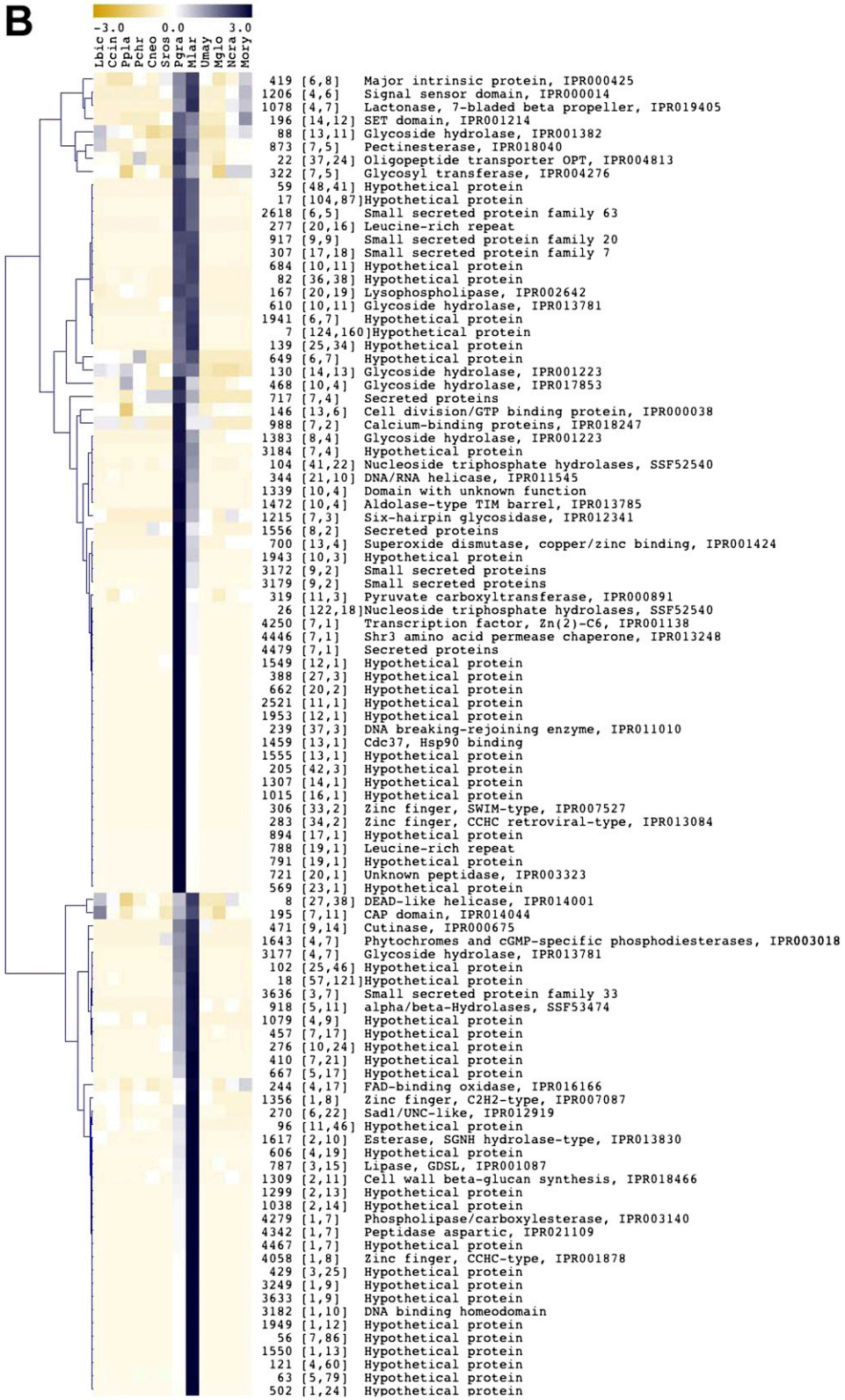

Fig. S4. Hierarchical clustering of *M. larici-populina* (Mlar) and *P. graminis* f. sp. *tritici* (Pgra)-specific gene family sizes. *Left*: Gene family size in each species. The *z*-scores scale indicates that, given a certain gene family and organism, the gene family size is substantially smaller (yellow) or larger (blue) than the mean gene family size (note scale, *Top*). Hence, blue blocks reflect gene family expansions. *Right*: Gene family ID followed by the number of genes in the family in *P. graminis* f. sp. *tritici* and *M. larici-populina* (between brackets) and by the gene description (Pfam families). (*A*) Clustering of the top 100 *z*-scores for *M. larici-populina* and *P. graminis* f. sp. *tritici*. (*B*) Clustering of the top 100 summed *M. larici-populina* and *P. graminis* f. sp. *tritici z*-scores. Ccin, *C. cinerea*; Cneo, *C. neoformans*; Lbic, *L. bicolor*; Mory, *M. oryzae*; Mglo, *Malassezia globosa*; Ncra, *N. crassa*; and Pchr, *Phanerochaete chrysosporium*.

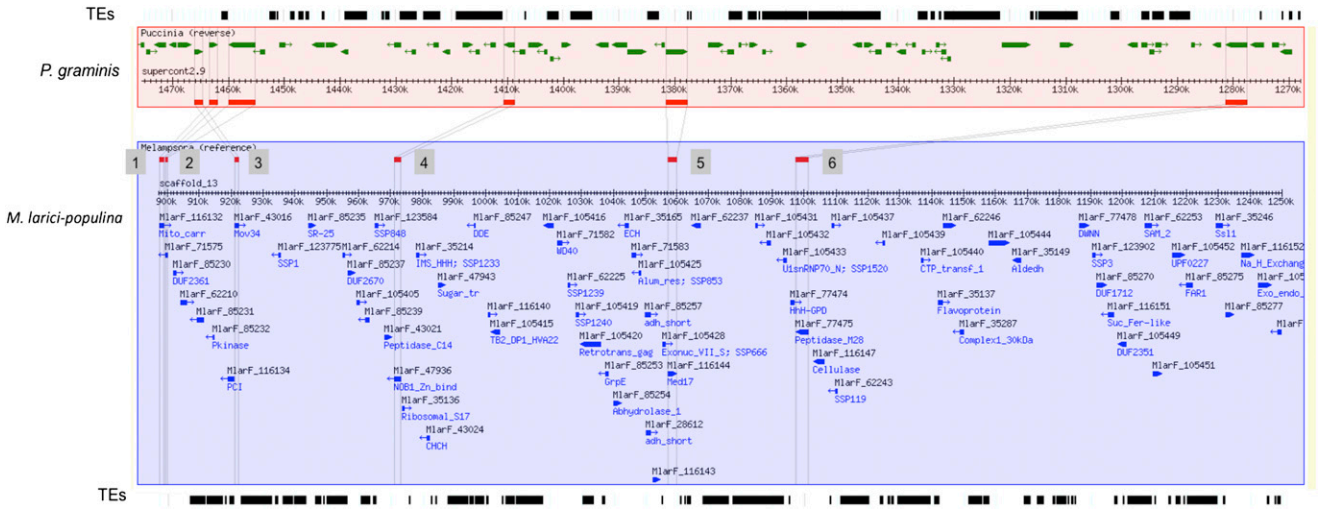

Fig. S5. The longest syntenic region shared between *M. larici-populina* and *P. graminis* f. sp. *tritici*. This segment was obtained from iADHoRe, and includes 200 kb with six anchor genes. TEs from the corresponding region are displayed on the upper or lower panel of the genome sequence. Protein function of these six anchor genes are (*1*) peroxisomal adenine nucleotide transporter, (*2*) hypothetical protein, (*3*) eukaryotic translation initiation factor, (*4*) RNA-binding protein NOB1, (*5*) mediator of RNA poly II transcription subunit 17, and (*6*) uncharacterized zinc metalloprotease.

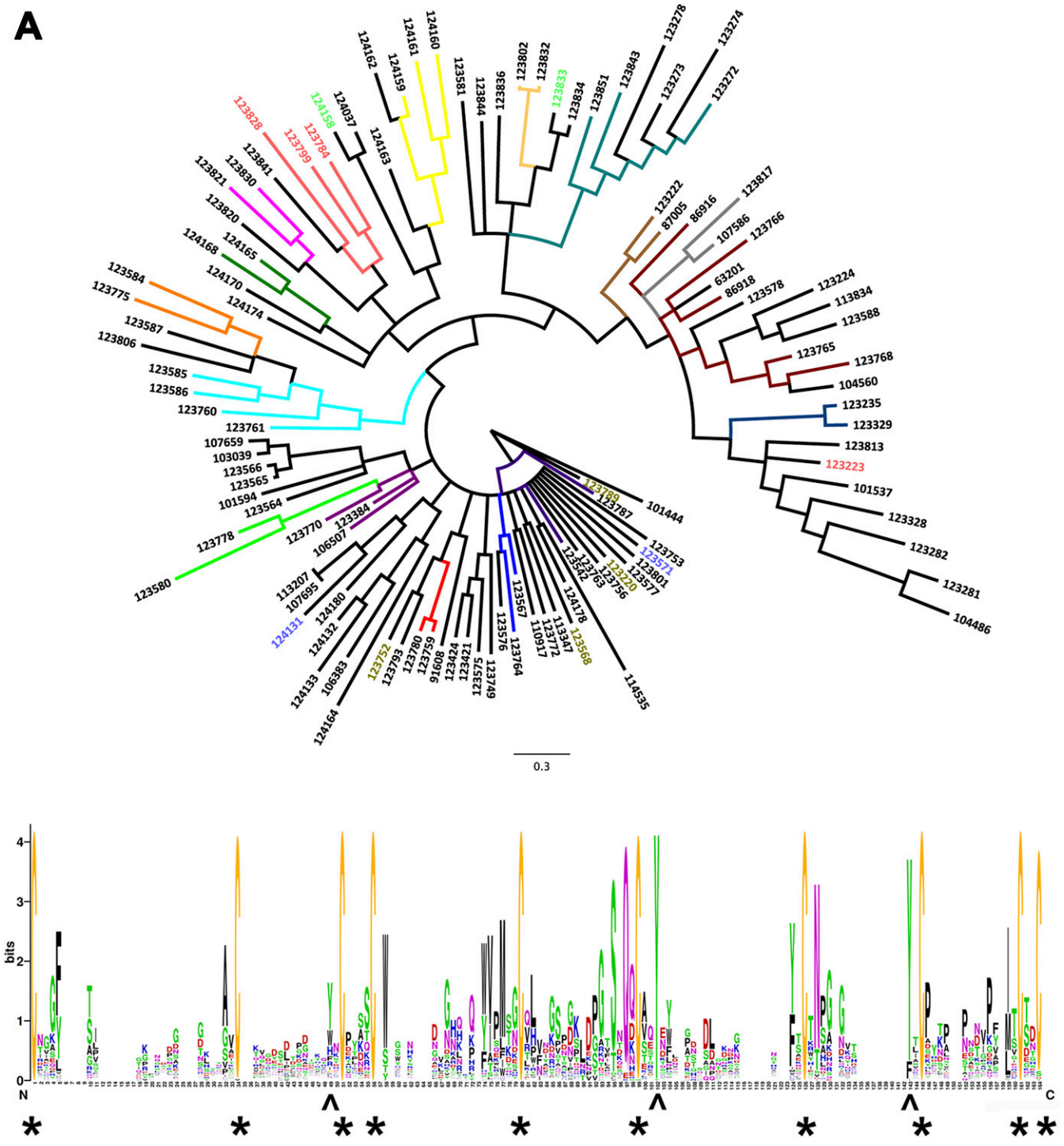

Fig. S6. (Continued)

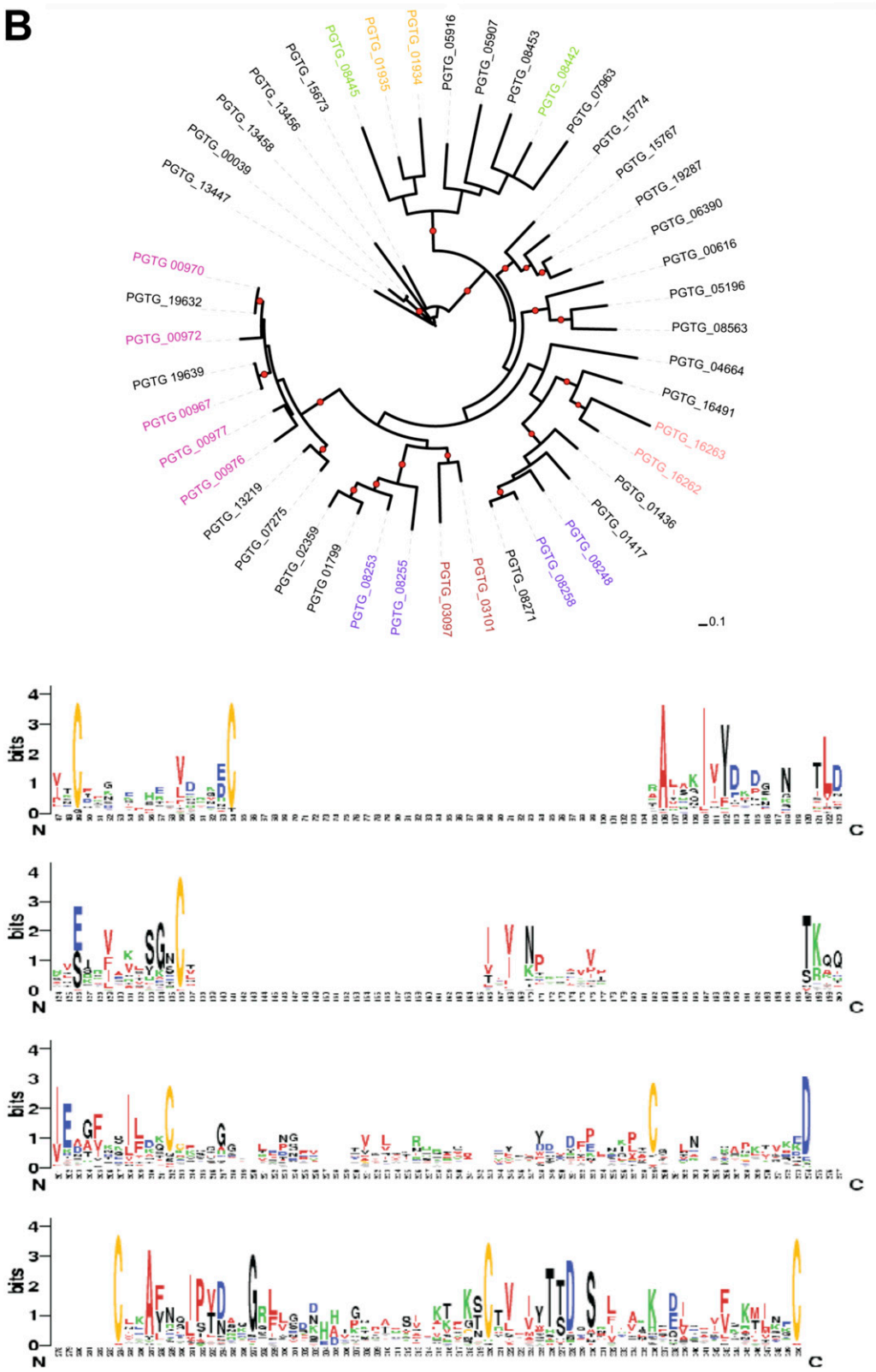

Fig. S6. Bayesian tree and conservation profile for the largest *M. larici-populina* cysteine-rich SSP class (*A*). Different colors are used to indicate genes that reside in close vicinity (≤ 5 genes apart). For clades, the tree branches are colored; for unrelated genes, Protein IDs are colored. Scale bar represents the number of substitutions per site. Cysteine and phenylalanine/tyrosine residues conserved in other SSP cysteine-rich classes are indicated below the conservation profile by asterisks and arrows, respectively. Maximum likelihood tree and conservation profile for the largest *P. graminis* f. sp. *tritici* SSP family (*B*). Different colors are used to indicate genes that reside in close vicinity (≤5 genes apart). Nodes with at least 80% bootstrap support (1,000 replicates) are indicated with a red circle. The best tree was visualized using the interactive tree of life (1). Scale bar represents the number of substitutions per site.

1. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.

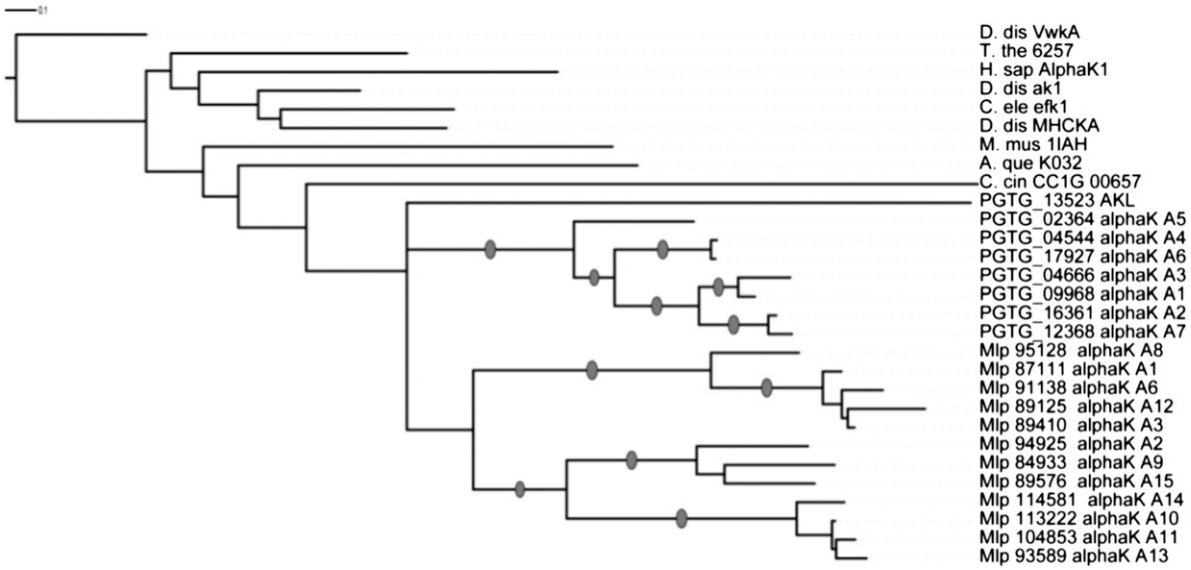

Fig. S7. Independent expansion of α-kinase family in *P. graminis* and *M. larici-populina*. Predicted active α-kinases from *P. graminis* and *M. larici populina* were selected along with one member of the α-kinase–like (AKL) family from *P. graminis* and typical α-kinases for other eukaryotes (http://www.kinase.com). Kinase domains were identified and aligned with the hmmer3 version of hmmalign (1), selecting only the kinase domain. Phylogeny was estimated using RAxML (2). The best tree was visualized using the interactive tree of life (3). Shaded circles indicate branches with more than 80% bootstrap support (1,000 replicates). Species abbreviations: PGTG, *P. graminis*; Mlp, *M. larici-populina*; D. dis, *Dictyostelium discoideum*; T. the, *Tetrahymena thermophila*; C. cin, *C. cinerea*; A. que, *Amphimedon queenslandica*; M. mus, *Mus musculus*; and H. sap, *Homo sapiens*.

1. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23:205–211.
2. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
3. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.

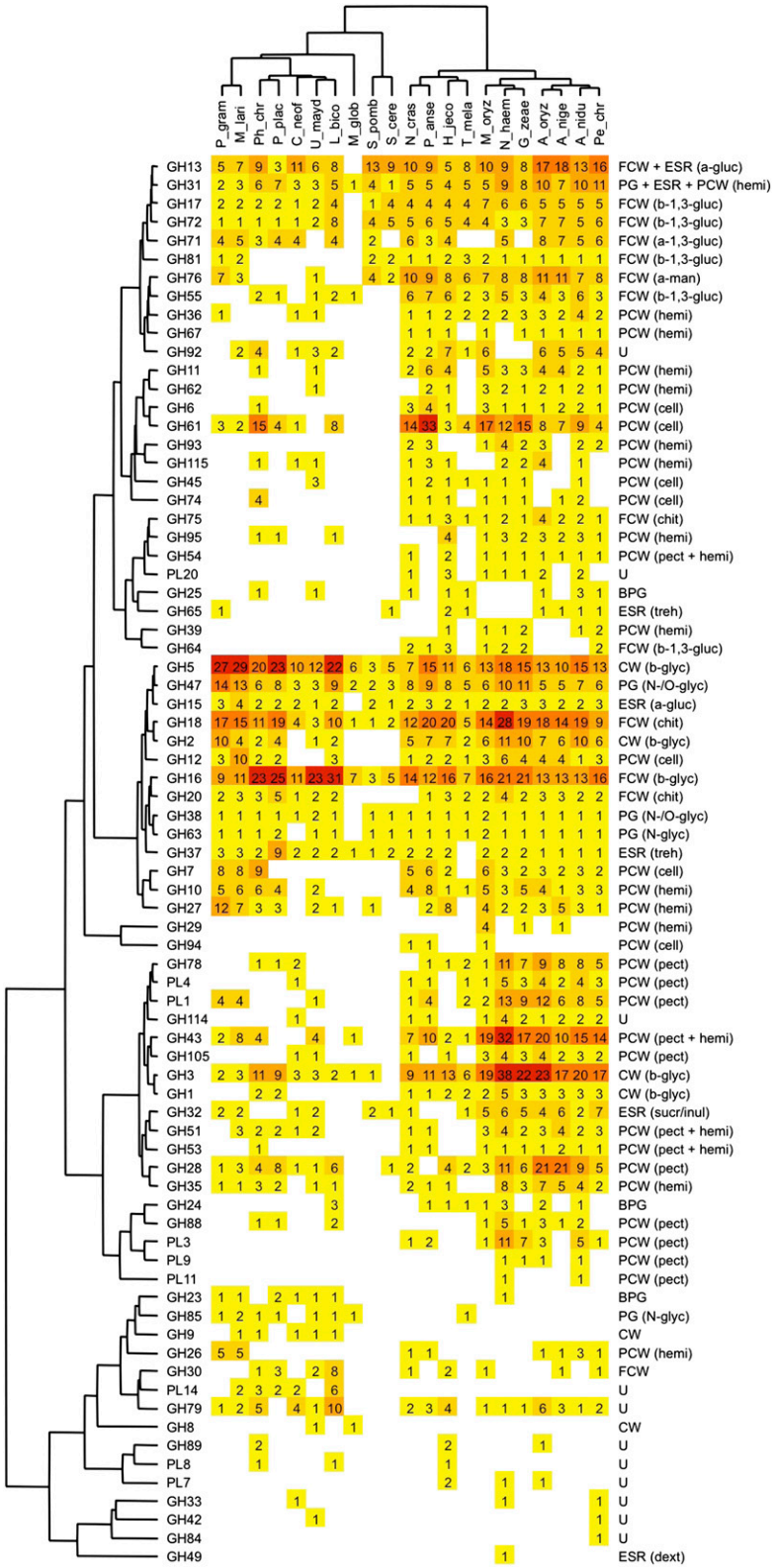

Fig. S8. Biclustering of the carbohydrate-cleaving families (19) from representative fungal genomes. Top tree: The fungi named are *Aspergillus nidulans* (A_nidu), *Aspergillus niger* (A_nige), *Aspergillus oryzae* (A_oryz), *Cryptococcus neoformans* (C_neof), *Gibberella zeae* (G_zeae), *Hypocrea jecorina* (H_jeco), *L. bicolor* (L_bico), *M. oryzae* (M_oryz), *Malassezia globosa* (M_glob), *M. larici-populina* (M_lari), *Nectria hematococca* (N_haem), *N. crassa* (N_cras), *Penicillium chrysogenum* (Pe_chr), *Phanerochaete chrysosporium* (Ph_chr), *Podospora anserina* (P_anse), *Postia placenta* (P_plac), *P. graminis* f. sp. *tritici* (P_gram), *Saccharomyces cerevisiae* (S_cere), *Schizosaccharomyces pombe* (S_pomb), *Tuber melanosporum* (T_mela), and *U. maydis* (U_mayd). Left tree: Enzyme families are represented by their class and family number according to the carbohydrate-active enzyme database (26). Right side: Known substrate of CAZy families (most

Legend continued on following page.

common forms in brackets). BPG, bacterial peptidoglycan; CW, cell wall; ESR, energy storage and recovery; FCW, fungal cell wall; PG,protein glycosylation; U, undetermined; α-gluc, α-glucans (including starch/glycogen); α-man, α-mannan; β-glyc, β-glycans; β-1,3-gluc, β-1,3-glucan; cell, cellulose; chit, chitin/chitosan; dext, dextran; hemi, hemicelluloses; inul, inulin; *N*-glyc, *N*-glycans; *N*-/O-glyc, *N*-/O-glycans; pect, pectin; sucr, sucrose; and treh, trehalose. Abundance of the different enzymes within a family is represented by a color scale from 0 (white) to 33 occurrences (red) per species.

Other Supporting Information Files

Dataset S1 (XLS)