Insight intotrade-off between wood decay and parasitism: Insights from the genome of a fungal forest pathogen

Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen

Ake Olson1, Andrea Aerts2, Fred Asiegbu3, Lassaad Belbahri4, Ourdia Bouzid5, Anders Broberg6, Bjorn Canba¨ck1, Pedro M. Coutinho7, Dan Cullen8, Kerstin Dalman1, Giuliana Deflorio9, Linda T.A. van Diepen10, Christophe Dunand11, Se´bastien Duplessis12, Mikael Durling1, Paolo Gonthier13, Jane Grimwood14, Carl Gunnar Fossdal15, David Hansson6, Bernard Henrissat7, Ari Hietala15, Kajsa Himmelstrand1, Dirk Hoffmeister16, Nils Ho¨gberg1, Timothy Y. James10, Magnus Karlsson1, Annegret Kohler12, Ursula Ku¨es17, Yong-Hwan Lee18, Yao-Cheng Lin19, Marten Lind1, Erika Lindquist2, Vincent Lombard7, Susan Lucas2, Karl Lunde´n1, Emmanuelle Morin12, Claude Murat12, Jongsun Park18, Tommaso Raffaello3, Pierre Rouze´19, Asaf Salamov2, Jeremy Schmutz14, Halvor Solheim15, Jerry Sta°hlberg20, Heriberto Ve´le¨z1, Ronald P. de Vries5,21, Ad Wiebenga21, Steve Woodward9, Igor Yakovlev15, Matteo Garbelotto22*, Francis Martin12*, Igor V. Grigoriev2* and Jan Stenlid1*

1Department of Forest Mycology and Pathology, Swedish University of Agricultural Sciences, Box 7026, Ullsva¨g 26, 750 05 Uppsala, Sweden; 2US DOE Joint Genome Institute, Walnut Creek, CA 94598, USA; 3Department of Forest Ecology, PO Box 27 Latokartanonkaari 7, 00014 University of Helsinki, Helsinki, Finland; 4Laboratory of Soil Biology, University of Neuchaˆtel, Rue Emile Argand 11, CH-2000 Neuchaˆtel, Switzerland; 5Microbiology, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands; 6Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, 750 05 Uppsala, Sweden; 7AFMB UMR 6098 CNRS/UI/UII, Case 932, 163 Avenue de Luminy 13288 Marseille cedex 9, France; 8Forest Products Laboratory, Madison, WI 53726, USA; 9Department of Plant and Soil Science, Institute of Biological and Environmental Sciences, University of Aberdeen, Cruickshank Building, St. Machar Drive, Aberdeen, AB24 3UU, Scotland UK; 10Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA; 11Laboratory of Cell Surfaces and Plant Signalisation 24, University Paul Sabatier (Toulouse III), UMR5546- CNRS, Chemin de Borde-Rouge, BP 42617, Auzeville 31326 Castanet-Tolosan, France; 12UMR INRA-UHP 'Interactions Arbres/Micro-Organismes' IFR 110 'Genomique, Ecophysiologie et Ecologie Fonctionnelles' INRA-Nancy 54280 Champenoux, France; 13Department of Exploitation and Protection of Agricultural and Forest Resources (Di. Va. P. R. A.) – Plant Pathology, University of Torino, Via L. da Vinci 44, I-10095 Grugliasco, Italy; 14HudsonAlpha Institute for Biotechnology, 601 Genome Way Huntsville, AL 35806, USA; 15Norwegian Forest and Landscape Institute, Høgskoleveien 8, NO-1432 A ° s, Norway; 16Pharmaceutical Biology, Friedrich-Schiller-Universita¨t Jena, Winzerlaer Str. 2, 07745 Jena, Germany; 17Bu¨sgen-Institute, Section Molecular Wood Biotechnology and Technical Mycology, University of Go¨ttingen, Bu¨sgenweg 2, D-37077 Go¨ttingen, Germany; 18Department of Agricultural Biotechnology, Seoul National University, Seoul 151-921, Korea; 19VIB Department of Plant Systems Biology, Ghent University, Bioinformatics and Evolutionary Genomics, Technologiepark 927, B-9052 Gent, Belgium; 20Department of Molecular Biology, Swedish University of Agricultural Sciences, Box 590, Husargatan 3, 751 24 Uppsala, Sweden; 21CBS-KNAW Fungal Biodiversity Centre, Uppsalalaan 8, 3584 CT Utrecht, the Netherlands; 22University of California, 338 Hilgard Hall Berkeley CA 94720 USA

Summary

• Parasitism and saprotrophic wood decay are two fungal strategies fundamental for succession and nutrient cycling in forest ecosystems. An opportunity to assess the trade-off between these strategies is provided by the forest pathogen and wood decayer Heterobasidion annosum sensu lato.

• We report the annotated genome sequence and transcript profiling, as well as the quantitative trait loci mapping, of one member of the species complex: H. irregulare. Quantitative trait loci critical for pathogenicity, and rich in transposable elements, orphan and secreted genes, were identified.
• A wide range of cellulose-degrading enzymes are expressed during wood decay. By contrast, pathogenic interaction between H. irregulare and pine engages fewer carbohydrate-active enzymes, but involves an increase in pectinolytic enzymes, transcription modules for oxidative stress and secondary metabolite production.
• Our results show a trade-off in terms of constrained carbohydrate decomposition and membrane transport capacity during interaction with living hosts. Our findings establish that saprotrophic wood decay and necrotrophic parasitism involve two distinct, yet overlapping, processes.

Introduction

Fungi are heterotrophs that play several distinctive roles in ecosystems as saprotrophs, parasites of plants and animals, and mutualistic symbionts of photosynthetic organisms. Generally, each species is specialized in one of these strategies although recent findings indicate that there might be partial physiological capacity overlap among fungi with different primary trophic strategies (Vasiliauskas *et al.*, 2007; Newton *et al.*, 2010). The use of more than one strategy by a single species might convey flexibility towards local changes in environment and competition with other organisms, as well as access to a wider ecological niche, but is likely to result in a trade-off in terms of constrained use of its full genomic capacity under specific environmental conditions.

Heterobasidion annosum sensu lato (*s.l.*) is a cosmopolitan fungal pathogen in coniferous forests. In 1995, economic losses caused by this pathogen were on the order of €600 million annually to forest owners in Europe through tree mortality and wood decay (Woodward *et al.*, 1998). Although the economic consequences for North American forestry are less well documented, they are expected to be of a similar magnitude. The frequency of root rot is increasing, with *c.* 23% per decade in managed forests in northern Europe (Thor *et al.*, 2005). In addition to threatening forest health, this white rot fungus causes massive release of CO_2 from decaying wood, thus representing a major threat to the ability of coniferous forests to serve as a natural carbon sink. The species complex comprises three Eurasian (*Heterobasidion annosum sensu stricto* (*s.s.*), *Heterobasidion parviporum* and *Heterobasidion abietinum*) and two North American (*Heterobasidion occidentale* and *Heterobasidion irregulare*) species, each with a different, but overlapping, host range (Niemelä & Korhonen, 1998; Otrosina & Garbelotto, 2010). Infections by *Heterobasidion* spp. are initiated in fresh wounds or newly created tree stump surfaces, followed by spread via root to root infection through living bark, and subsequent decay of and survival in the root and trunk of standing trees (Woodward *et al.*, 1998). This infection cycle relies on mechanisms for both saprotrophic wood decay and pathogenic interactions with a living host, allowing us to study the potential trade-off between the two trophic strategies. The switch between saprotrophy and parasitism could be associated with an activation of distinct gene sets for the two growth modes, or governed by the differential regulation of common metabolic processes. A major goal of this genomic study was to identify key elements in the molecular repertoire required for balancing between the two trophic strategies.

Materials and Methods

Selection of *H. irregulare* strain and isolation of genomic DNA and RNA

The sequenced *Heterobasidion irregulare* Garbelotto & Otrosina strain TC32-1 (Chase, 1985) has been well characterized and has been used in many studies: for example, the strain constitutes one of the monokaryotic parental isolates in a genetic hybrid AO8 used for the generation of the mapping population for the *H. annosum* (*s.l.*) genetic linkage map (Lind *et al.*, 2005). Already published expressed sequence tags (ESTs) from the TC32-1 strain are available (Karlsson *et al.*, 2003). In addition, a BAC library of 2688 clones was prepared by BIO S&T Inc. (Montreal, Quebec, Canada) according to their standard protocols. DNA was extracted using established methods (Sambrook & Russell, 2001), and extractions of total RNA were performed according to cetyltrimethylammonium bromide (CTAB) and phenol/chloroform methods (Karlsson *et al.*, 2003).

Genome sequencing, assembly and annotation

All sequencing reads for the whole-genome shotgun sequencing were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA. Three different sized libraries were used as templates for the plasmid subclone sequencing process and both ends were sequenced: 214 143 reads from the 2.7-kb library, 192 768 reads from the 6.0-kb library and 63 168 reads from the 39.1-kb fosmid library were sequenced (Supporting Information Table S1).

A total of 406 752 reads was assembled using a modified version of Arachne v.20071016 (Jaffe *et al.*, 2003) with the parameters maxcliq1 = 100, correct1_passes = 0 and BINGE_AND_PURGE=True. This produced 53 scaffold sequences, with L50 of 2.2 Mb (half of the assembled genome (N50) is organized in scaffolds of this size or larger), 19 scaffolds larger than 100 kb and a total scaffold size of 33.9 Mb. Each scaffold was screened against bacterial proteins, organelle sequences and GenBank and removed if found to be a contaminant. Additional scaffolds were removed if the scaffold contained only unanchored rDNA sequences. The final draft whole-genome shotgun assembly contained scaffolds that covered 33.1 Mb of the genome with a contig L50 of 127.0 kb and a scaffold L50 of 2.2 Mb (Table S2).

Genome improvement

All genome improvement reactions were performed at the HudsonAlpha Genome Sequencing Center, Huntsville, AL, USA. In order to improve and finish the genome of *H. irregulare*, the whole-genome shotgun assembly was broken into scaffold-sized pieces and each scaffold piece was reassembled with Phrap (Green, 1999). The scaffold pieces were then finished using a Phred/Phrap/Consed pipeline (Gordon *et al.*, 1998). Initially, all low-quality regions and gaps were targeted with computationally selected sequencing reactions completed with 4 : 1 BigDye terminator:dGTP chemistry (Applied Biosystems, Foster City, CA, USA). These automated rounds included directed primer walking on plasmid subclones using custom primers.

After the completion of the automated rounds, each assembly was manually inspected. Further reactions were manually selected to complete the genome. These reactions included additional custom primer walks on plasmid subclones or fosmids using 4 : 1 BigDye terminator:dGTP chemistry. Smaller repeats in the sequence were resolved by transposon hopping 8-kb plasmid clones. To fill large gaps, resolve larger repeats or resolve chromosome duplications and extend into chromosome telomere regions, shotgun sequencing and finishing of BAC fosmid clones were used. During the course of the improvement project, 5376 BAC ends were sequenced to add additional contiguity. Finally, the sequences were compared with markers on the available genetic map (Lind *et al.*, 2005) and two map joins were made based on map evidence.

Each assembly was validated by independent quality assessment, which included a visual examination of subclone paired ends and visual inspection of discrepancies containing high-quality sequence and all remaining low-quality areas. All available EST resources were also placed on the assembly to ensure completeness.

cDNA library construction and sequencing

Heterobasidion irregulare TC32-1 polyA + RNA was isolated from total RNA for two RNA samples: RNA1 – cells grown in liquid Hagem medium (Stenlid, 1985) and RNA2 – cells grown in liquid high nitrogen MMN medium (Marx, 1969) using the Absolutely mRNA Purification kit and the manufacturer's instructions (Stratagene, La Jolla, CA, USA). cDNA synthesis and cloning involved a modified procedure based on the 'SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning' (Invitrogen, Carlsbad, CA, USA). One to two micrograms of

polyA + RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT-NotI primer (5′-GACTAGTTCTAGATCGCGAGCGGCCGCCCT15VN-3′) were used to synthesize first-strand cDNA. Second-strand synthesis was performed with *Escherichia coli* DNA ligase, polymerase I and RNaseH, followed by end repair using T4 DNA polymerase. The SalI adaptor (5′-TCGACCCACGCGTCCG and 5′-CGGACGCGTGGG) was ligated to the cDNA, digested with NotI (New England Biolabs, Ipswich, MA, USA) and subsequently size selected by gel electrophoresis (1.1% agarose). Size ranges of cDNA were cut out of the gel for the RNA1 sample, yielding two cDNA libraries (JGI library codes CCPA for range 0.6–2 kb and CCOZ for range > 2 kb), and JGI library codes CCPC and CCPB (same respective sizes) for the RNA2 sample. The cDNA inserts were directionally ligated into the SalI and NotI-digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (5′-GTAAAACGACGGCCAGT) and M13-R (5′-AGGAAACAGCTATGACCAT) to determine the fraction of insertless clones. Colonies from each library were plated onto agarose plates (254-mm plates from Teknova, Hollister, CA, USA) at a density of *c.* 1000 colonies per plate. Plates were grown at 37°C for 18 h, after which individual colonies were picked and each used to inoculate a well containing Luria–Bertani (LB) medium with appropriate antibiotics in a 384-well plate (Nunc, Rochester, NY, USA). Clones in 384-well plates were grown at 37°C for 18 h. Contained plasmid DNA for sequencing was produced by rolling circle amplification (Templiphi, GE Healthcare, Piscataway, NJ, USA). Subclone inserts were sequenced from both ends using primers complementary to the flanking vector sequence (Fw, 5′-ATTTAGGTGACACTATAGAA; Rv, 5′-TAATACGACTCACTATAGGG) and BigDye terminator chemistry, and run on an ABI 3730 instrument (Applied Biosystems).

EST sequence processing and assembly

A total of 40 807 ESTs, including 8 840 from CCPA, 8 759 from CCOZ, 13 280 from CCPC, 8 263 from CCPB and 1 665 from external sources, were processed through the JGI EST pipeline (ESTs were generated in pairs, a 5′ and 3′ end read from each cDNA clone). To trim vector and adaptor sequences, common sequence patterns at the ends of ESTs were identified and removed using an internally developed tool. Insertless clones were identified if either of the following criteria were met: > 200 bases of vector sequence at the 5′ end or < 100 bases of nonvector sequence remained. ESTs were then trimmed for quality using a sliding window trimmer (window = 11 bases). Once the average quality score in the window was below the threshold (Q15), the EST was split and the longest remaining sequence segment was retained as the trimmed EST. EST sequences with < 100 bases of high-quality sequence were removed. ESTs were evaluated for the presence of polyA or polyT tails (which, if present, were removed) and the EST was re-evaluated for length, removing ESTs with < 100 bases remaining. ESTs consisting of > 50% low-complexity sequence were also removed from the final set of ‘good ESTs’. In the case of resequencing the same EST, the longest high-quality EST was retained. Sister ESTs (end pair reads) were categorized as follows: if one EST was insertless or a contaminant, then, by default, the second sister was categorized as the same. However, each sister EST was treated separately for complexity and quality scores. Finally, EST sequences were compared with the Genbank nucleotide database in order to identify contaminants; nondesirable ESTs, such as those matching rRNA sequences, were removed.

For clustering, ESTs were evaluated with malign, a kmer-based alignment tool, which clusters ESTs on the basis of sequence overlap (kmer = 16; seed length requirement = 32; alignment ID ≥ 98%). Clusters of ESTs were further merged on the basis of sister ESTs using double linkage. Double linkage requires that two or more matching sister ESTs exist in both clusters to be merged. EST clusters were then each assembled using CAP3 (Huang & Madan, 1999) to form consensus

sequences. For cluster consensus sequence annotation, the consensus sequences were compared with Swissprot using blastx and the hits were reported. Clustering and assembly of all 33 539 ESTs resulted in 10 126 consensus sequences and 1503 singlets.

Whole-genome exon oligoarray

The *Heterobasidion irregulare* custom exon expression array (4 × 72K), manufactured by Roche NimbleGen Systems Limited (Madison, WI, USA) (http://www.nimblegen.com/products/exp/index.html), contained five independent, nonidentical, 60-mer probes per gene model coding sequence. For 12 199 of the 12 299 annotated protein-coding gene models, probes could be designed. For 19 gene models, no probes could be generated, and 81 gene models shared all five probes with other gene models. Included in the array were 916 random 60-mer control probes and labelling controls. For 2 032 probes, technical duplicates were included on the array.

Total RNA was extracted using CTAB/phenol/chloroform and LiCl precipitation. The RNA was DNase I treated and cleaned with a Qiagen RNA Cleanup Kit. Arrays were performed from *H. irregulare* mycelium grown in liquid MMN medium (three biological replicates), from the cambial zone of necrotic bark of pines inoculated with *H. irregulare* (21 d post-inoculation; three biological replicates), from fruit bodies collected in California (four biological replicates) as well as from *H. irregulare* grown on wood shavings from pine (four biological replicates), *H. irregulare* grown in liquid medium amended with lignin (Kraft Pine lignin B 471003-500G; Sigma-Aldrich) (two biological replicates) and *H. irregulare* grown in liquid medium amended with cellulose from spruce (22182-KG Fluka; Sigma-Aldrich) (two biological replicates). Cultures were harvested after 3 wk of incubation at 22°C in darkness.

Total RNA preparations were amplified using the SMART PCR cDNA Synthesis Kit (Clontech, Mountain View, CA, USA) according to the manufacturer's instructions. Single dye labeling of samples, hybridization procedures, data acquisition, background correction and normalization were performed at the NimbleGen facilities (NimbleGen Systems, Reykjavik, Iceland) following their standard protocol. Microarray probe intensities were quantile normalized across all 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 to be nonspecific if it shared > 90% homology with a gene model other than that for which it was made) and renormalized using ARRAYSTAR software (DNASTAR, Inc., Madison, WI, USA). For 621 gene models, no reliable probe was left. A transcript was deemed to be expressed when its signal intensity was three-fold higher than the mean signal-to-noise threshold (cut-off value) of the random oligonucleotide probes present on the array (50–100 arbitrary units). Gene models with an expression value higher than three-fold the cut-off level were considered to be transcribed. The maximum signal intensity values were *c.* 65 000 arbitrary units. Student's *t*-test with false discovery rate (FDR) (Benjamini–Hochberg) multiple testing correction was applied to the data using ARRAYSTAR software (DNASTAR, Inc.). Transcripts with a significant *P* value (< 0.05) were considered to be differentially expressed. The complete expression dataset is available as a series (accession number GSE30230) at the Gene Expression Omnibus at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/geo/).

Genome annotation

The *H. irregulare* genome was annotated using the JGI annotation pipeline, which takes multiple inputs (scaffolds, ESTs and known genes), runs several analytical tools for gene prediction and annotation, and deposits the results in the JGI Genome Portal

(http://www.jgi.doe.gov/Heterobasidion) for further analysis and manual curation. Genomic assembly scaffolds were masked using RepeatMasker (Smit *et al.*, 1996–2010) and the RepBase library of 234 fungal repeats (Jurka *et al.*, 2005). Using the repeat-masked assembly, several gene prediction programs falling into three general categories were employed: *ab initio*– FGENESH (Salamov & Solovyev, 2000), GeneMark (Isono *et al.*, 1994); homology-based – FGENESH+, Genewise (Birney & Durbin, 2000) seeded by BLASTx (Altschul *et al.*, 1990) alignments against GenBank's database of nonredundant proteins (NR: http://www.ncbi.nlm.nih.gov/BLAST/); and EST-based – EST_map (http://www.softberry.com/) seeded by EST contigs. Genewise models were extended, where possible, using scaffold data to find start and stop codons. EST BLAT alignments (Kent, 2002) were used to extend, verify and complete the predicted gene models. The resulting set of models was then filtered for the best models, based on EST and homology support, to produce a nonredundant representative set. This representative set of 11 464 gene models was subjected to further analysis and manual curation. Measures of model quality included the proportions of the models complete with start and stop codons (86% of models), consistent with ESTs (48% of models) and supported by similarity with proteins from the NCBI NR database (70% of models) (Table S3). Approximately 90% of the models showed expression in at least one of the conditions (*H. irregulare* growth on wood shavings from pine, growth in liquid MMN medium, growth in liquid MMN medium amended with lignin or cellulose, *H. irregulare* fruit bodies or cambial zone of necrotic pine bark inoculated with *H. irregulare*) analyzed in the NimbleGen array. The characteristics of the predicted genes are listed in Table S4. Multigene families were predicted with the Markov clustering algorithm (Enright *et al.*, 2002) to cluster the proteins, using BLASTp alignment scores between proteins as a similarity metric. Orthologs with other sequenced basidiomycete genomes were determined on the basis of the best bidirectional blast hits (Table S5).

All predicted gene models were functionally annotated using SignalP (Nielsen *et al.*, 1997), TMHMM (Melen *et al.*, 2003), InterProScan (Zdobnov & Apweiler, 2001) and BLASTp (Altschul *et al.*, 1990) against NR, and hardware-accelerated double-affine Smith–Waterman alignments (deCypherSW; http://www.timelogic.com/decypher_sw.html) against SwissProt (http://www.expasy.org/sprot/), KEGG (Kanehisa *et al.*, 2008) and KOG (Koonin *et al.*, 2004). KEGG hits were used to assign EC numbers (http://www.expasy.org/enzyme/), and Interpro and SwissProt hits were used to map gene ontology (GO) terms (http://www.geneontology.org/). Functional annotations are summarized in Table S6. The top 30 PFAM domains are listed in Table S7. Community-wide manual curation of the automated annotations was performed using the web-based interactive editing tools of the JGI Genome Portal (http://www.jgi.doe.gov/Heterobasidion) to assess predicted gene structures, assign gene functions and report supporting evidence.

Results

Genome structure

Using a whole-genome shotgun approach, the 33.6MB genome of *H. irregulare* (formerly *H. annosum* North American P-type) was sequenced to 8.5 × coverage. Genome improvement, finishing and gap closure resulted in 33 649 967 bp with an estimated error rate of < 1 error in 100 000 base pairs. The genome is represented in 15 scaffolds ranging in size from 3 591 957 to 8 087 bp. Six of the scaffolds represent complete chromosomes with sequence spanning from telomere to telomere. Seven other scaffolds have an identified telomere only at one end (Fig. 1). The final assembly statistics are shown in Table S8.

The published linkage map (Lind *et al.*, 2005) was anchored to the sequenced genome using simple sequence repeat (SSR) markers designed from the genome sequence and evenly distributed across

the scaffolds (Fig. 1). Segregation analysis of 179 sequence and SSR markers supported a genome organized into 14 chromosomes, which is consistent with pulsed-gel electrophoresis data (Anderson *et al.*, 1993). The linkage map was used to locate quantitative trait loci (QTL) for pathogenicity, growth rate and fungal interactions (Olson, 2006; Lind *et al.*, 2007a,b) onto the genome sequence, allowing the identification of the genes in the targeted regions.

Transposable elements (TEs) comprised 16.2% of the *H. irregulare* genome and were not uniformly distributed across the scaffolds ($P < 0.05$) (Figs S1, S2). The Gypsy-like elements were the most frequent TE, corresponding to 9.3% of the assembly. Class II terminal inverted repeats (TIRs) represented the second most frequently categorized element (1.1%), and 3.7% of the genome comprised TEs belonging to unknown families. The insertion age of full-length long terminal repeats (LTRs) shows that *H. irregulare* underwent a recent transposon activity which peaked at an estimated 0.2 million years ago (Mya) and an old activity that occurred at 4–8 Mya (Fig. S3, Notes S1). The genome of *H. irregulare* contains 100 467 SSRs corresponding to, on average, 2 895 SSR Mb^{-1} with a distribution about twice as dense in the intergenic relative to the intragenic region (Fig. S4). High frequencies of tri- and hexarepeats, which are less likely than SSRs to cause frameshift mutations with other repeat units, were found in exonic regions as well as in the 5′ untranslated regions (UTRs) and regions immediately upstream from genes, indicating a possible role of these repeats in overall gene expression (Table S9).

The mitochondrial genome, shown to influence *H. irregulare* virulence (Olson & Stenlid, 2001), spans 114 193 bp and is one of the largest sequenced in fungi (Notes S2). In addition to genes coding for proteins of the oxidative phosphorylation system, we found 14 genes encoded within introns, two genes and two pseudogenes probably derived from a mitochondrial plasmid and six nonconserved hypothetical genes (Table S10). Seven of the 15 protein-encoding genes contain one or more introns (Table S10).

A total of 11 464 gene models have been predicted in *H. irregulare*, with half shared across Basidiomycotina (Notes S3, Tables S5, S11–S15). The transcription factor distribution is comparable with other fungal taxa and the signal transduction pathways are conserved (Notes S3, Figs S5–S7). The largest gene families include transporters and signaling domains (MFS, p450, WD40, protein kinases) (Table S7). Sequenced ESTs and microarray analysis supported 90% of the predicted genes. In the microarray experiment, 1 615 gene models showed differential expression among probes representing a particular gene model in a given growth condition, indicating an alternative gene model structure to those predicted or alternative splice variants present. In genomic regions rich in TEs, such models were more abundant (Fig. 1). A smaller fraction (59%) of the gene models in the TE-rich regions showed similarity to genes from NCBI compared with 70% for the overall genome. However, the existence of most gene models (95%) without homology was verified by ESTs or microarray expression data.

Mating compatibility and heterokaryon formation in tetrapolar Agaricomycetes are controlled by two unlinked mating type loci encoding two different classes of gene: the homeodomain transcription factors controlling nuclear division (*MAT-A*) and the pheromones and their transmembrane receptors controlling nuclear migration and communication (*MAT-B*) (Brown & Casselton, 2001). Bipolar species, such as *H. irregulare*, have only a single mating type locus, and, where known, the mating type loci of bipolar Agaricomycetes only encode homeodomain proteins (James *et al.*, 2006). We found evidence of unlinked regions homologous to both *MAT-A* and *MAT-B* in the *H. irregulare* genome. The *MAT-A* region of *H. irregulare* is much like that of other Agaricomycetes, because it is found on the largest chromosome and displays extensive locally conserved gene order with other species (Fig. S8). However, the *MAT-A* locus appears to also encode either a novel class of *MAT* protein or a highly derived homeodomain transcription factor

that has replaced one of the two typically divergently transcribed homeodomain pairs (Notes S4, Fig. S8). Interestingly, the member of the homedomain pair that has been replaced (HD2) possesses a more typical DNA-binding motif than the nonreplaced HD1 protein (Kües, 2000). The region homologous to *MAT-B* encodes a cluster of five pheromone receptor genes and three putative pheromone genes. We used both evidence on DNA polymorphism and the co-segregation of *MAT*-specific markers to determine whether the *MAT-A* or *MAT-B* region controls mating incompatibility in *H. irregulare*. As predicted under the intense balancing selection observed at *MAT* loci (May *et al.*, 1999), we found that *MAT-A* was highly polymorphic, whereas *MAT-B* genes were not. In a segregation analysis of a progeny array of *H. irregulare*, the putative *MAT-A* genes, but not the putative *MAT-B* genes, co-segregated with mating type, demonstrating that the mating is controlled by the *MAT-A* and not *MAT-B* locus in this bipolar fungus (Fig. S9).

Wood degradation machinery

Heterobasidion irregulare showed a broad spectrum of carbohydrate-active enzymes. The number of glycoside hydrolase (GH), polysaccharide lyase (PL) and carbohydrate esterase (CE) genes was comparable with that of the white rotter *Phanerochaete chrysosporium* and the mycorrhizal symbiont *Laccaria bicolor*, but smaller than that in the saprotrophs *Coprinopsis cinerea* and *Schizophyllum commune*, and in the pathogen *Magnaporthe oryzae* (Notes S5, Table S16). However, *H. irregulare* is almost as well equipped as *S. commune* and *C. cinerea* with regard to the gene families involved specifically in plant cell wall degradation, encoding the enzymes required to digest cellulose, hemicellulose (xyloglucan and its side chains) and pectin and its side chains (Table S17. By contrast, *H. irregulare* has a limited potential for the degradation of a second hemicellulose structure, xylan, with only two xylanases from family GH10 and none from family GH11 (Table S17). *Heterobasidion irregulare* possesses two GH29 fucosidase genes, which may act on living/fresh material, and two GH5 (β-mannanases), which may play a role in softwood hemicellulose degradation, as softwood is known to contain a large proportion of glucomannan (Wiedenhoeft & Miller, 2005).

The growth of *H. irregulare* on various carbohydrate substrates correlated well with its enzymatic repertoire (Notes S5). For example, as suggested by the presence of an invertase (GH32) gene, a feature shared with many phytopathogens (Parrent *et al.*, 2009), *H. irregulare* thrives on sucrose. Enzymes active on cell wall polymers were prominent during transcriptome analysis; of the 282 carbohydrate-active enzymes present in the *H. irregulare* genome, 36 were more than two-fold up-regulated ($P < 0.05$) during early wood degradation (Notes S5). This subset is dominated by putative cellulose-degrading enzymes in the groups GH1, GH5, GH6, GH7, GH10, GH12, GH45 and cellulose oxidoreductase GH61 (Notes S5, Table S18). Pectate lyase and pectin hydrolase (GH28), both up-regulated on early wood decay of pine, are likely to act on the middle lamellae in a softwood-specific manner. Hemicellulose of wood fibres can be degraded by GH5 and GH10, five and two members of which, respectively, were up-regulated. In addition, 14 sugar transporters showed elevated transcript levels during wood degradation when compared with liquid culture growth (Table S18).

Oxidative enzymes implicated in ligninolysis by white rot fungi include lignin peroxidase, manganese peroxidase, glyoxal oxidase and laccase (Hatakka, 1994). With its eight manganese peroxidases and lack of lignin peroxidases, *H. irregulare* has a lower peroxidase potential than *P. chrysosporium* (Fig. S10, Table S19), but a larger number of phenol oxidases and laccases, 18 and five, respectively. To generate H_2O_2, *H. irregulare* encodes 17 quinone oxidoreductases, four glyoxal oxidases, 34 glucose-methanol-choline oxidoreductases and four manganese superoxide dismutases (Notes S5, Table S19).

Generally, genes involved in oxidative lignocellulose degradation showed a lower expression level than carbohydrate hydrolyzing enzymes during wood degradation (Table S18). However, when compared with growth in liquid medium, some of these genes were significantly up-regulated: one of the eight *H. irregulare* manganese peroxidase genes showed three-fold higher expression during wood degradation than in liquid medium. Two of the five glyoxal oxidase genes were significantly up-regulated and one of the glucose-methanol-choline oxidoreductase genes showed 15-fold higher expression levels, whereas only one of the laccase genes was moderately up-regulated. The gene for cellobiose dehydrogenase, responsible for the oxidation of cellobiose, was up-regulated 14-fold (Table S18). Lipids and proteins are minor constituents of softwood but, being easily digestible substrates, they may be of importance in the early stages of wood colonization. During *H. irregulare* degradation of pine sapwood, lipase and protease genes showed significantly higher expression relative to liquid culture conditions.

Pathogenicity

Heterobasidion spp. are recognized as producers of at least 10 different secondary metabolites which are produced in both axenic cultures and during interaction with plants and other fungi (Sonnenbichler *et al.*, 1989). Genome analysis using a secondary metabolite unique regions finder (SMURF) web-tool (http://www.jcvi.org/smurf/index.php) and manual curation identified solitary and clustered putative natural product genes (Tables S20, S21). The genome of *H. irregulare* contained genes for three polyketide synthases (PKSs), 13 nonribosomal peptide synthetase-like (NRPS-like) enzymes, three terpene cyclases and several tailoring enzymes, including one dimethylallyltryptophan synthase (DMATS) predicted to be involved in secondary metabolite production. The phytotoxins fomannosin and fomannoxin were accumulated in culture filtrates of *H. irregulare* (Fig. 2, Notes S6), and terpene cyclase and DMATS genes coding for the enzymes required for fomannosin and fomannoxin biosynthesis, respectively, were identified in the gene repertoire (Notes S6, Table S20).

Three genes encode members of the small (*c.* 150-amino-acid) secreted protein family Cerato-platanin (Cp) (Notes S3) originally identified in *Ceratocystis platani* (Pazzagli *et al.*, 1999) (Fig. S11). During interaction with host tissue, necrotrophic plant pathogens produce reactive oxygen species (ROS) which contribute to the host-mediated oxidative stress and facilitate infection. For the production of ROS, fungi utilize NADPH oxidase homologues (NOx) and ferric reductase (FRe) (Gessler *et al.*, 2007). NOxs are necessary for superoxide generation during developmental processes, whereas FRes are required to acquire iron from the infected host, and are potentially related to pathogenicity, as this gene family of seven members in *H. irregulare* is large relative to that in nonpathogenic basidiomycetes (Notes S3).

Transcriptome analysis showed that 55 of the 250 genes with the highest expression during pathogenic interaction had secretion signal sequences, and 18 of these encoded enzymes putatively active in carbohydrate degradation (Notes S6, Table S22). Sixty-two genes were differentially expressed ($P < 0.05$) in the pathogenic interaction relative to growth on defined liquid medium, with 47 up- and 15 down-regulated (Notes S6, Table S23). Sixteen differentially expressed genes encoded proteins which were predicted to be secreted. Of these, there are five likely to act on carbohydrate substrates (PL1, GH5, GH28, CE16, CBM1), one lipase and two oxidases active on saccharide molecules. The remaining eight with secretion signals showed no similarity to any protein of known function. Two of the differentially expressed genes with secretion signals had four transmembrane conserved domains, which suggests a membrane localization.

By re-mapping virulence data from Lind *et al.*, (2007a), we located three major QTL regions important for pathogenic interactions with Norway spruce and Scots pine, one on scaffold 1 and two

on scaffold 12 (Fig. 1). These QTL regions included 178, 142 and 299 predicted gene models, respectively, one-third of which had an expression distinct from background and showed no cross-hybridization with plant transcripts (Fig. 3). Transcriptional data from mycelia grown in cambium limited the virulence candidates within the QTLs to a handful of significantly ($P < 0.05$) differentially expressed genes. The most highly up-regulated gene was a high-affinity sugar transporter (70-fold) on scaffold 12. This QTL also contained a gene model with no sequence homology, which showed four-fold greater expression in mycelia grown in cambium relative to liquid culture (Fig. 3). The QTL on scaffold 1 contained one gene model with no sequence homology, which showed significantly lower expression in mycelia infecting pine relative to liquid media, and a putative flavin containing Baeyer-Villiger monooxygenase, with a 9.4-fold higher expression during infection (Fig. 3). This type of monooxygenase is needed for one of the biosynthetic steps required for the synthesis of phytotoxin fomannosin, making it a very strong pathogenicity candidate. Two overlapping secondary metabolite clusters harboring, altogether, 43 gene models were located in the QTL region on scaffold 12. The clusters included three NRPSs (3, 4, and 11), several oxidative enzymes and transport proteins (Table S21). Furthermore, the sequence similarities of the QTL regions to other Basidiomycota genomes were low (Notes S6), and the frequency of orphan genes was higher than in other parts of the genome (53% relative to 34%). In addition, the TE density was higher within the pathogenicity QTLs than in other parts of the genome (Fig. 1).

Two-way hierarchical cluster analysis of mean gene expression showed that the biological samples could be separated depending on whether or not mycelia had been grown together with wood components (Fig. 4). Mycelia grown in the cambial zone showed a distinct pattern in the heat map, indicating that interaction with living tissue is very different from the other growth conditions analyzed. Cluster 3 represented gene models specifically expressed in the fruit body (Fig. 4). The gene models more highly expressed during interaction with living tissue were represented in clusters 1 and 5 (Fig. 4). Cluster 13 consisted of gene models with increased expression in contact with wood components relative to mycelia grown in liquid media or fruit body (Fig. 4).

Global transcript profiling demonstrated that genes induced during saprotrophic wood degradation, but not on interaction with living host tissue, represented a trade-off between the two trophic strategies (Notes S7, Tables S24, S25). Gene expression during saprotrophic growth on wood showed highest correlations with gene expression during growth on cellulose and lignin, but much lower correlations with growth in the fruit body. Gene expression during growth in the cambial zone of pine showed an intermediate correlation with the levels in wood (Table 1). The number of genes with a significantly ($P < 0.05$) higher expression was highest during saprotrophic growth on wood, followed by fruit body, and lowest during growth in the cambial zone of pine (Fig. 5). The majority of genes with a higher expression during growth in the cambial zone of pine were also expressed significantly ($P < 0.05$) more strongly during growth on wood, and included pectinolytic enzymes and part of the cellulolytic capacity (Fig. 5). The genes with a significantly ($P < 0.05$) higher expression during growth in the cambial zone of pine and in the fruit body were distinct, whereas some overlap could be found between the genes expressed during growth in the fruit body and during saprotrophic growth on wood.

Discussion

The forest pathogen and wood decayer *H. annosum* (*s.l.*) uses two ecological strategies: parasitism and saprotrophic wood decay. During the life cycle, it infects and lives within standing conifer

trees, but also continues to colonize and degrade the dead tissue. Our analyses of gene expression during these two trophic stages reveal a trade-off in terms of restricted carbon acquisition. Fewer genes encoding carbohydrate-active enzymes and transporters are expressed during pathogenic growth than during saprotrophic wood decay, indicating that the fungus is not using its full capacity for energy acquisition during necrotrophic growth, that is, its full arsenal of wood-degrading enzymes. Instead, the living tissue triggers an expanded metabolic repertoire involving genes associated with, for example, toxin production, protection against plant defenses, handling of low oxygen pressure and other abiotic stresses. We conclude that there is a trade-off between maximal nutritional gain and access to a different ecological niche which must be balanced by the fungus.

Gene expression during saprotrophic growth on wood correlated most strongly with expression during growth on cellulose and lignin, but only moderately with expression during growth in the cambial zone of pine. Presumably, *H. irregulare* detects wood as a source of cellulose-derived energy, whereas living tissue only partly functions in this manner. The genes induced specifically during infectious growth enable *H. irregulare* to access energy sources, such as carbon bound in the macromolecules of living organisms, unavailable to other organisms with which it would otherwise compete.

The re-mapping of two different virulence measurements on two hosts, pine and spruce, revealed that three major regions on two chromosomes are involved, harboring 178, 142 and 299 gene models, respectively. These regions are characterized by a large number of TEs and orphan genes with no homologous genes reported from other species. These orphan models constitute a resource for the exploration of novel enzymatic functions and biological mechanisms. Together, the enrichment in orphan genes and repetitive elements indicates that these are highly dynamic regions with a high evolutionary rate. The characteristics, with a large number of orphan genes and many repetitive elements, of these regions are comparable with the effector regions identified in *Phytophthora infestans* (Haas *et al.*, 2009).

TEs are not equally distributed within and among chromosomes. The major TEs observed are younger than 12 million years and the decrease detected probably reflects element deterioration leading to a loss of ability to detect older elements. As TE proliferation within the pathogenicity QTLs is clearly younger than speciation (Fig. S3) (Dalman *et al.*, 2010), we hypothesize that transposon activity may have contributed to the shaping of the species-specific characteristics of *H. irregulare*. TEs have been implicated to co-locate with important factors for pathogenicity in other pathosystems, for example, *Phytophthora infestans* (Cuomo *et al.*, 2007; Haas *et al.*, 2009).

Transcriptome analyses combined with the QTL approach proved to be a powerful method to reduce the number of candidate virulence genes of the QTL regions. Gene models that are present in the QTL regions for virulence and are significantly up-regulated during pathogenic interaction with pine are strong candidates. Three candidate genes fulfill the criteria: a sugar transporter, a monooxygenase and a gene model without homology to known genes. As the QTLs are based on a mapping population derived from a cross between *H. irregulare* and *H. occidentale*, these genes are the main candidates to explain the difference in virulence between the species. As host specificity probably plays an important role in speciation, these genes could constitute a crucial step towards an understanding of the separation of *H. annosum* (*s.l.*) into separate species.

Hybrids between *H. irregulare* and *H. occidentale* show that the mitochondria have an influence on pathogenicity in these species (Olson & Stenlid, 2001). The analysis of the mitochondrial genome also revealed the largest mitochondrial genome among fungal sequences so far, containing 24 gene models in addition to those involved in oxidative phosphorylation. These genes constitute interesting candidates for the mitochondrial connection to virulence.

The *H. irregulare* genome shows a great potential for both saprotrophic and biotrophic lifestyles. It encodes a wide arsenal of enzymes required to digest cellulose, hemicellulose and pectin, making it almost as well equipped with regard to plant cell wall degradation as obligate saprotrophs, such as *S. commune* and *C. cinerea.* In contrast with *P. chrysosporum*, *H. irregulare* possesses two GH29 fucosidase genes which may act on living/fresh material and two GH5 (β-mannanases) which may play a role in softwood-specific glucomannan degradation. Furthermore, *H. irregulare* growth on sucrose correlates well with the presence of an invertase (GH32) gene, a feature shared with many phytopathogens. Sucrose is one of the major sugars found in fresh pine stump surfaces (Asiegbu, 2000), and the capacity to utilize it during the initial phase of colonization might provide *H. irregulare* with a selective advantage compared with saprotrophs lacking invertase activity, which could help to explain why *H. annosum* spp. are so competitive in industrially managed forests.

The analysis of culture filtrates revealed the presence of fomannosin and fomannoxin, whereas genome analysis identified terpene cyclase and DMATS, possibly involved in the respective synthesis of these known phytotoxins. In addition, genes predicted to be involved in the production of other secondary metabolites were identified. The presence of genes for putative PKSs, NRPS-like enzymes and halogenase, however, implies that the biosynthetic capacity of *H. irregulare* has not been fully explored, as, to date, no polyketides, nonribosomal peptides or halogenated compounds have been identified.

Conclusion

Heterobasidion irregulare is an economically vastly important nonmodel organism with a uniquely strong potential for versatility between pathogenic interaction and wood decay. The key enzymes and pathways of these central processes can be identified by our genomic approach. Comparing growth on dead and living tissue, we reveal a switch in gene expression during living host interaction towards toxin production, protection against plant defenses and the handling of abiotic stress at the expense of carbohydrate decomposition and membrane transport capacity. The combination of QTL and transcriptome analyses is a powerful approach to elucidate the gene sets involved in important phenotypic traits of a species. The virulence QTL regions are characterized by the over-representation of TEs, orphan and secreted genes. We demonstrate that a limited number of genes fulfill the criteria of being located within a QTL and being significantly differentially expressed during host interaction relative to in liquid media. Some of these genes encode proteins that are linked to secondary metabolite production. This approach enables the identification of new candidate pathogenicity factors, but, at the same time, limits the number.

Acknowledgements

The work conducted by the US Department of Energy Joint Genome Institute was supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Financial support from the Swedish Foundation for Strategic Research is gratefully acknowledged. Bioinformatic analyses carried out at INRA-Nancy were supported by Region Lorraine and FABELOR grants to F.M. The assembly and annotations of the *H. irregulare* genome are available from the JGI Genome Portal at http://www.jgi.doe.gov/Heterobasidion and have been deposited at DDBJ/EMBL/GenBank under the accession number AEOJ00000000. The complete expression dataset is available as a series (accession number GSE30230) at the Gene Expression Omnibus at NCBI (http://www.ncbi.nlm.nih.gov/geo/).

Table 1. Linear regression of global gene expression in *H. irregulare* under different growth conditions

Condition	Lignin	Cellulose	Wood	Fruit body	Cambial zone
Lignin	1	0.6934[a]	0.7256	0.2965	0.5328
Cellulose		1	0.7365	0.2244	0.4880
Wood			1	0.3699	0.6158
Fruit body				1	0.2489
Cambial zone					1

1. [a], least squares regression (r^2).
2. n=3590.

Fig. 1. The 14 postulated chromosomes of *Heterobasidion irregulare*. The upper black bar of each chromosome denotes linkage map coverage, with pathogenicity quantitative trait loci (QTLs) marked in green. The wide yellow-to-brown bar describes the gene density (upper half) and gene model quality (lower half) for every 50-kb segment of the sequence. Gene density is calculated as the number of gene models, ranging from > 27 (brown) to < 10 (white). Gene model quality is calculated on the basis of microarray experiments using five probes for each gene model. The color indicates the percentage of models in which all five probes hybridized, ranging from 100% (brown) to < 10% (white). The lowest bar of each chromosome indicates transposon regions, as masked by RepeatMasker. Blue 'T's show an identified telomere region in the corresponding chromosome end.

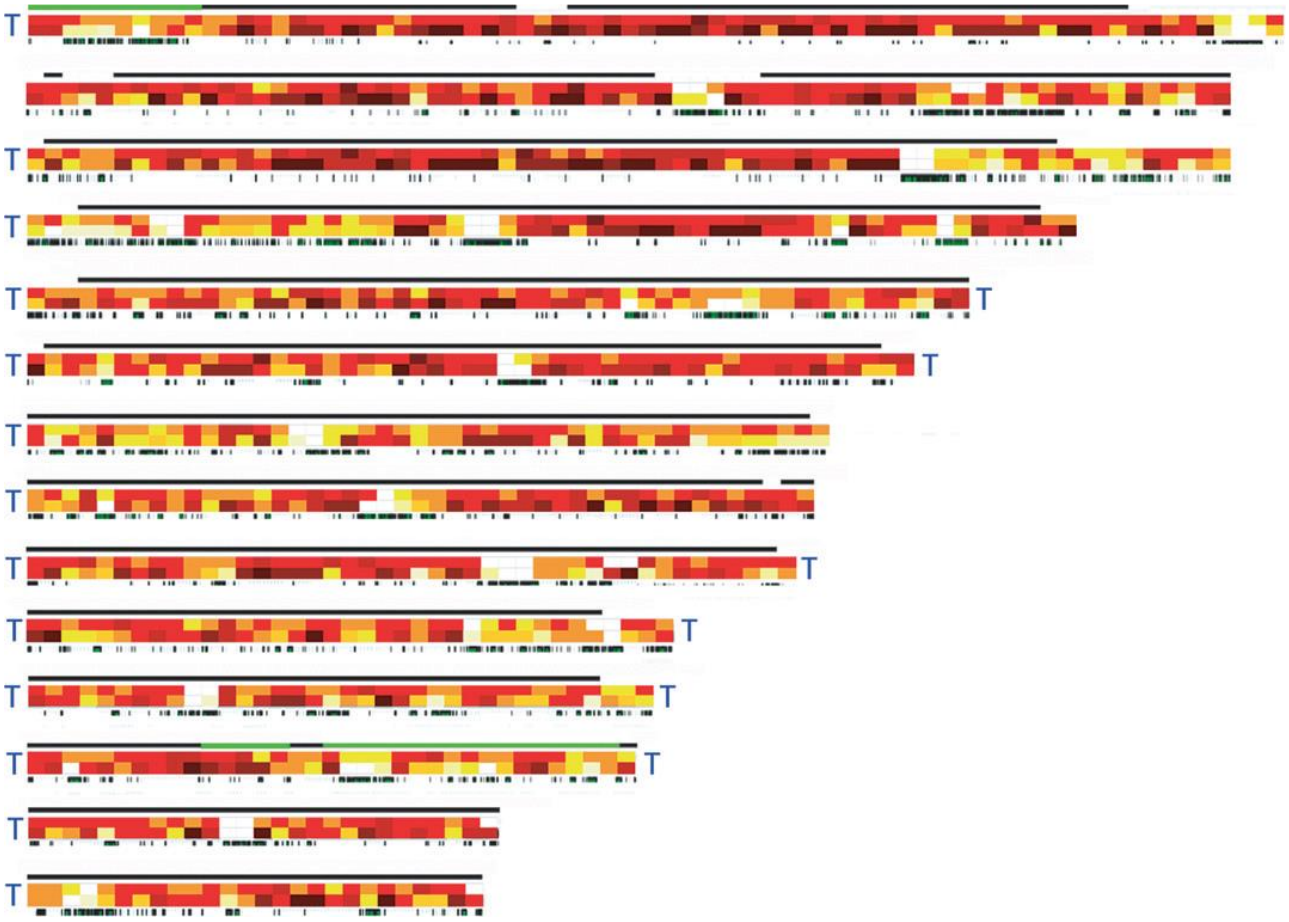

Fig. 2. Preparative gradient reversed-phase HPLC chromatogram of *Heterobasidion irregulare* monoculture filtrate after solid phase extraction. Selected identified compounds are shown: 1, fomannoxin (t_R 39.5); 2, 5-formyl-2-(isopropyl-1′-ol)benzofuran (t_R 26.2); 3, fomannosin (t_R 22.7).

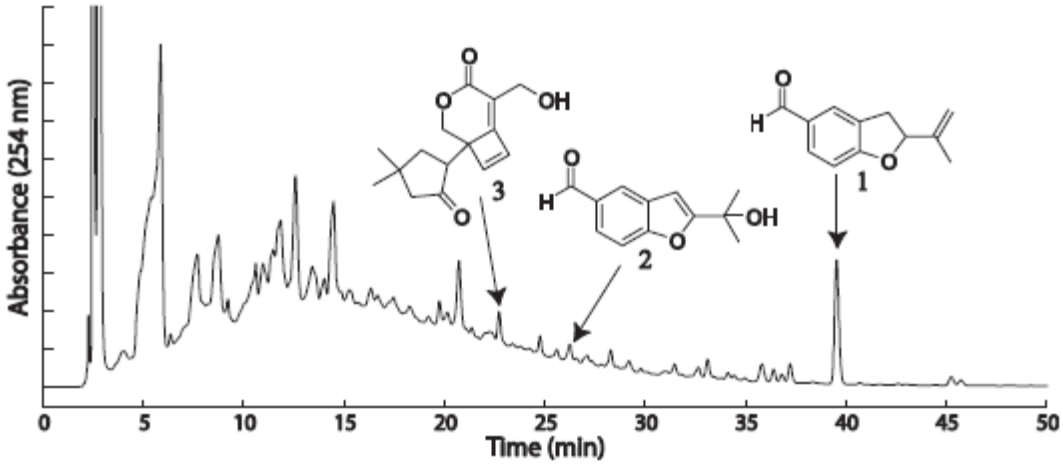

Fig. 3. Quantitative trait locus (QTL) regions on scaffolds 12 (a) and 1 (b), and, within them, the up- and down-regulated gene models during the mycelial growth of *Heterobasidion irregulare* in the cambial zone of pine bark relative to growth in liquid culture. The left *y*-axis denotes the logarithm of odds (LOD) value for the QTL effect, with the horizontal bar (LOD 1.9) indicating the 5% level of significance. The right *y*-axis has a logarithmic scale of fold change for gene models, where '1.00′ indicates no change in expression level above background. Blue vertical bars indicate gene models with fold changes not significantly different from liquid media; red bars indicate models with significant ($P < 0.05$) fold changes. The *x*-axis shows the markers of the QTL regions (below bar) and their scaffold positions (bp) (above bar). The QTL curve is adjusted to fit the physical distance between the markers rather than the genetic distance.

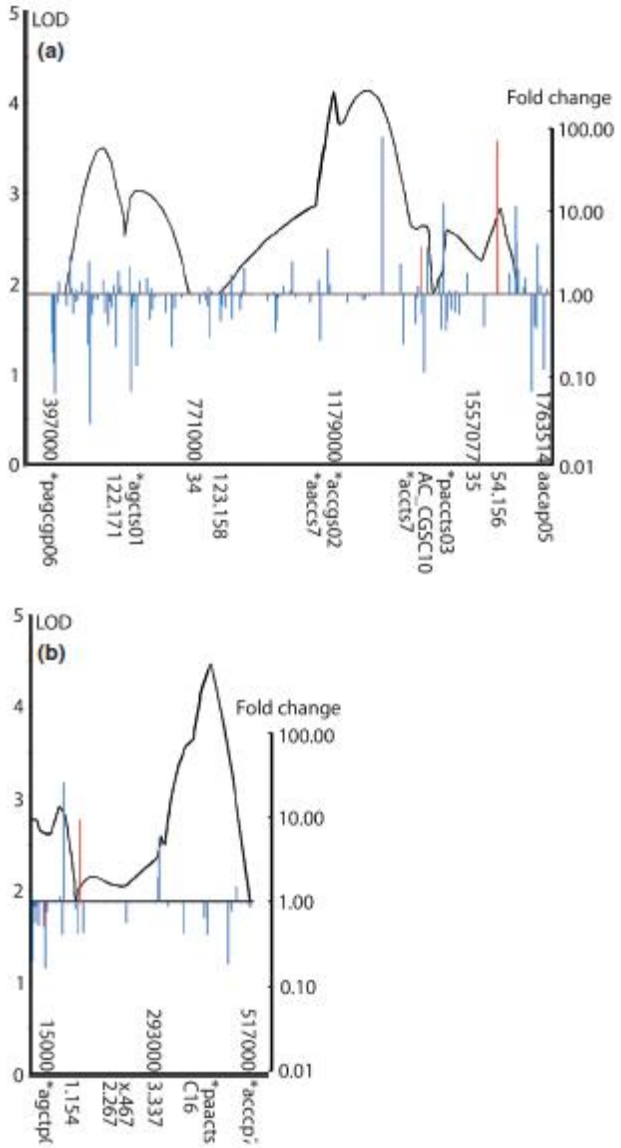

Fig. 4. Dendrogram of two-way Ward's hierarchical clustering of the normalized microarray mean expression of *Heterobasidion irregulare* genes (a) and expression profiles of clusters 1, 3, 5 and 13 (b–e). The heat map shows color-coded expression from blue (low) to red (high). The treatments were liquid MMN (mycelia grown in liquid MNN), cellulose and lignin (mycelia grown in liquid MNN amended with cellulose or lignin), fruit body (mycelia from fruit body), wood (mycelia grown on pine wood) and cambial zone (mycelia grown in the cambial zone of pine bark).

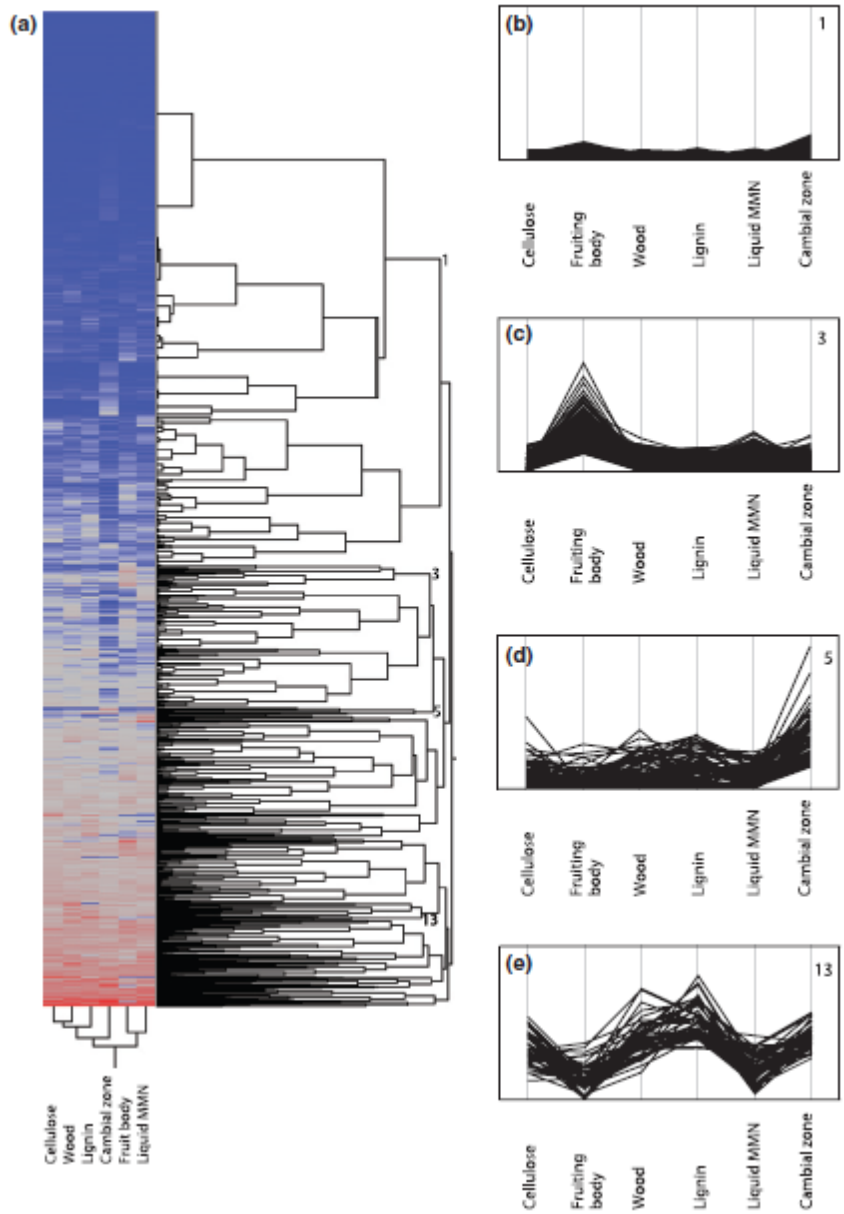

Fig. 5. Venn diagram showing the number of significantly up-regulated (unique and common) genes in *Heterobasidion irregulare* during mycelial growth in wood, growth in the cambial zone of pine or in the fruit body, relative to gene expression from mycelia grown in liquid culture.