Available online at www.sciencedirect.com

SCIENCE DIRECT®

Fungal Genetics and Biology 43 (2006) 343–356

FUNGAL GENETICS AND BIOLOGY

www.elsevier.com/locate/yfgbi

Computational analysis of the *Phanerochaete chrysosporium* v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins

Amber Vanden Wymelenberg [a], Patrick Minges [a], Grzegorz Sabat [b], Diego Martinez [c], Andrea Aerts [d], Asaf Salamov [d], Igor Grigoriev [d], Harris Shapiro [d], Nik Putnam [d], Paula Belinky [e], Carlos Dosoretz [f], Jill Gaskell [g], Phil Kersten [g], Dan Cullen [g,*]

[a] *Department of Bacteriology, University of Wisconsin, Madison, WI 53706, USA*
[b] *Genetics and Biotechnology Center, University of Wisconsin, Madison, WI 53706, USA*
[c] *Joint Genome Institute, Los Alamos National Laboratories, Los Alamos, NM 87545, USA*
[d] *Joint Genome Institute, Walnut Creek, CA 94598, USA*
[e] *Environmental Biotechnology Laboratory, Migal-Galilee Technology Center, P.O. Box 831 Kiryat-Shmona, Israel*
[f] *Civil and Environmental Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel*
[g] *USDA Forest Products Laboratory, One Gifford Pinchot Dr., Madison, WI 53726, USA*

Received 21 November 2005; accepted 11 January 2006
Available online 9 March 2006

Abstract

The white-rot basidiomycete *Phanerochaete chrysosporium* employs extracellular enzymes to completely degrade the major polymers of wood: cellulose, hemicellulose, and lignin. Analysis of a total of 10,048 v2.1 gene models predicts 769 secreted proteins, a substantial increase over the 268 models identified in the earlier database (v1.0). Within the v2.1 'computational secretome,' 43% showed no significant similarity to known proteins, but were structurally related to other hypothetical protein sequences. In contrast, 53% showed significant similarity to known protein sequences including 87 models assigned to 33 glycoside hydrolase families and 52 sequences distributed among 13 peptidase families. When grown under standard ligninolytic conditions, peptides corresponding to 11 peptidase genes were identified in culture filtrates by mass spectrometry (LS–MS/MS). Five peptidases were members of a large family of aspartyl proteases, many of which were localized to gene clusters. Consistent with a role in dephosphorylation of lignin peroxidase, a mannose-6-phosphatase (M6Pase) was also identified in carbon-starved cultures. Beyond proteases and M6Pase, 28 specific gene products were identified including several representatives of gene families. These included 4 lignin peroxidases, 3 lipases, 2 carboxylesterases, and 8 glycosyl hydrolases. The results underscore the rich genetic diversity and complexity of *P. chrysosporium*'s extracellular enzyme systems.
Published by Elsevier Inc.

Keywords: *Phanerochaete chrysosporium*; Secretion; Secretome; Proteome; Gene cluster

1. Introduction

Lignin, the component of plant cell walls that gives strength to wood, is the second most abundant natural polymer on earth. This amorphous and insoluble aromatic material lacks stereoregularity, and unlike hemicellulose and cellulose is not susceptible to hydrolytic attack. A relatively small group of microbes, collectively referred to as 'white-rot' fungi, are uniquely able to completely degrade lignin to gain access to the carbohydrate polymers of plant cell walls, which they use as carbon and energy sources. The white-rot basidiomycete *Phanerochaete chrysosporium* has become the model system for studying the physiology and genetics of lignin degradation (for

[*] Corresponding author. Fax: +1 608 231 9262.
E-mail address: dcullen@facstaff.wisc.edu (D. Cullen).

1087-1845/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.fgb.2006.01.003

review, see Cullen and Kersten, 2004). Initially released in 2002 (Martinez et al., 2004), the *P. chrysosporium* genome assembly and gene models were recently updated (www.jgi.doe.gov/whiterot).

Major components of the *P. chrysosporium* lignin depolymerization systems include lignin peroxidase (LiP), manganese peroxidase (MnP), and a peroxide-generating enzyme, glyoxal oxidase (GLOX). Defined media with limiting amounts of carbon or nitrogen are routinely employed for enzyme production, and under these conditions *P. chrysosporium* secretes multiple LiP and MnP isozymes. Lignin peroxidase gene *lipD* encodes the dominant isozyme in carbon-limited medium, and evidence suggests this protein is dephosphorylated by a mannose-6-phosphatase (M6Pase) (Kuan and Tien, 1989; Rothschild et al., 1997; Rothschild et al., 1999). The LiPs and MnPs are encoded by families of 10 and 5 structurally related genes, respectively. Eight *lips* are closely linked within a 100 kb region (Stewart and Cullen, 1999), and *mnp1* lies 5.7 kb from *mnp4* (Martinez et al., 2004). The *lips* and *mnps* exhibit dramatic differential regulation in response to media composition, but no clear relationship has been observed between transcriptional regulation and genomic organization (reviewed in Cullen and Kersten, 2004).

Beyond the oxidative enzymes and M6Pase, relatively little is known of extracellular proteins present in ligninolytic cultures. Several proteases have been partially characterized from submerged cultures, but it remains uncertain whether extracellular peroxidases are substantially degraded under ligninolytic conditions (Bonnarme et al., 1993; Dass et al., 1995; Dosoretz et al., 1990a,b). The relationship between the proteases produced under such nutrient limitation and those produced in colonized wood pulp (Datta, 1992) or in cellulolytic cultures (Eriksson and Pettersson, 1982) is also unclear. The latter enzymes have been implicated in the activation of cellulase activity (Eriksson and Pettersson, 1982) and in the cleavage of cellobiose dehydrogenase functional domains (Eggert et al., 1996; Habu et al., 1993). With regard to *P. chrysosporium* protease genetics, a cDNA encoding a serine protease was recently isolated from non-ligninolytic cultures (Faraco et al., 2005), and a family of clustered glutamic proteases has been observed within the genome (Sims et al., 2004). LC–MS/MS peptide identification demonstrated the expression of three aspartyl protease genes in a medium containing cellulose as sole carbon source (Vanden Wymelenberg et al., 2005).

Herein, we describe the computational identification and analysis of v2.1 protein models with predicted secretion signals. Shotgun LC–MS/MS on filtrates from carbon- and nitrogen-limited cultures identified the expected peroxidases and glyoxal oxidases in addition to an impressive array of previously unknown extracellular proteins.

2. Methods

2.1. Fungal strains and culture conditions

P. chrysosporium strain RP78 ((Stewart et al., 2000) FGSC strain 9002), a homokaryotic derivative of BKM-F-1767 (Center for Forest Mycology Research, Forest Products Laboratory, Madison, WI), was used throughout. Standard B3 salts media with limiting carbon or nitrogen were grown statically at 39 °C as previously described (Brown et al., 1988; Kirk et al., 1978) and harvested on days 4 and 5, respectively. LiP activities as measured by veratryl alcohol oxidation (Tien and Kirk, 1984) were 8.41 and 18.6 nmol min^{-1} ml^{-1} in C-limited and N-limited cultures, respectively.

2.2. Protein analysis

Carbon- and nitrogen-limited cultures were harvested by filtration through Miracloth (Calbiochem, La Jolla, CA) and the filtrates were stored at −20 °C. One hundred seventy-five milliliters of each filtrate was concentrated 100-fold in an Amicon 8400 stirred ultrafiltration cell with a 5000 MWCO polyethersulfone membrane (Millipore Corp., Bedford, MA). Five hundred microliters Amicon-concentrated protein was further concentrated in a 10,000 MWCO Nanosep centrifugal device (Pall Life Sciences, Ann Arbor, MI) to a final volume of 50 μl. Twenty-five microliters were mixed with 20 μl Laemmli buffer (Bio-Rad Laboratories, Inc, Hercules, CA) and loaded onto a 12.5% Criterion Tris–HCL Ready Gel (Bio-Rad Laboratories) for SDS–PAGE. Electrophoresis was performed in a Bio-Rad Criterion Cell, 200 V, 50 min, 23 °C. Gels were stained with Coomassie Blue R-250 (Bio-Rad Laboratories) to estimate protein abundance and MW distribution. Total protein resolved on the gel was manually fractioned with a surgical blade into 10, 2 mm long and 5 mm wide strips. These gel strips were further cut into ~1 mm pieces and placed in individual siliconized 1.5 ml microcentrifuge tubes (Fisher Scientific) for subsequent enzymatic digestion.

'In Gel' digestion and mass spectrometric analysis were performed as described (www.biotech.wisc.edu/ServicesResearch/MassSpec/ingel.htm.) In short, gel pieces were destained completely in MeOH/H_2O/NH_4HCO_3 (50%:50%:100 mM), dehydrated for 10 min in acetonitrile/H_2O/NH_4HCO_3 (50%:50%:25 mM) and then once more for 1 min in 100% acetonitrile. The samples were dried in a Speed-Vac for 5 min, reduced in 25 mM DTT (dithiothreitol in 25 mM NH_4HCO_3) for 30 min at 56 °C, alkylated with 55 mM IAA (iodoacetamide in 25 mM NH_4HCO_3) in darkness at room temperature for 30 min, washed in H_2O for 20 min, equilibrated in 25 mM NH_4HCO_3 for 10 min, dehydrated for 10 min in acetonitrile/H_2O/NH_4HCO_3 (50%: 50%:25 mM) and then once more for 1 min in 100% acetonitrile. Following drying, samples were rehydrated with 25 μl of trypsin solution (20 ng/μl trypsin (Sequence Grade Modified, Promega Inc., Madison, WI) in 25 mM NH_4HCO_3) or with 25 μl Asp-N solution (8 ng/μl endoproteinase Asp-N, Roche Biochemicals) in 50 mM Na_2HPO_4, 5 mM Tris–HCl, pH 8.0. Additional buffer overlay (~15 μl) was provided to keep gel fragments immersed. The digestions were conducted overnight (18 h) at 37 °C, then terminated by acidification with 0.1% TFA (trifluoroacetic acid). Peptides generated from digestions were extracted in two subsequent steps, first with an equal volume of 0.1% TFA (~50 μl) and vigorous vortexing for 15 min, then with the same volume of acetonitrile/H_2O/TFA (70%:25%:5%) and vortexing. The collected peptide solution was dried completely in a Speed-Vac, re-suspended in 50 μl of 0.1% TFA, and solid phase extracted (C18 SPEC-PLUS™-PT pipette tips Varian, Inc., Lake Forest, CA). Peptides were eluted off the C18 column with acetonitrile/H_2O/TFA (70%:25%:0.2%), dried in a Speed-Vac and finally reconstituted with 45 μl of 0.1% formic acid.

Peptide fractions were individually analyzed by nanoLC–MS/MS using 1100 series LC/MSD Trap SL spectrometer (Agilent, Palo Alto, CA). Chromatography of peptides prior to mass spectral analysis was accomplished using C18 reverse phase HPLC trap column (Zorbax 300SB-C18, 5 μM, 5 × 0.3 mm, Agilent) and separation column (Zorbax 300SB-C18, 3.5 μm, 0.075 × 150 mm, Agilent) onto which 40 μl of each extracted peptide fraction was automatically loaded. An Agilent 1100 series HPLC delivered solvents A: 0.1% (v/v) formic acid in water, and B: 95% (v/v) acetonitrile, 0.1% (v/v) formic acid, at either 10 μl/min to load sample, or at 0.28 μl/min to elute peptides directly into the nano-electrospray. The elution was for 80 min in a gradient from 20 to 60% (v/v) solvent B. Peptides eluting from the HPLC column/electrospray source were trapped in an ion cell and sequential MS/MS spectra spanning from 300 to 2200 *m*/*z* were generated for the four most abundant ions present at each switching event. Redundancy was limited by dynamic exclusion. MS/MS data were converted to matrix generic format (mfg) files using Data Analysis Software (Agilent). Spectrum Mill MS Proteomics Workbench (Agilent) and an in-house licensed Mascot search engine (Matrix Science, London, UK) were used to identify peptides using a dataset of 10,048 gene models described below. Throughout, protein similarity scores are based on the Smith–Waterman algorithm (Smith and Waterman, 1981) using the BLOSUM62 matrix.

2.3. Genome assembly and automated annotation

The v.2.0 assembly and v2.1 gene models are considerably improved relative to earlier versions (Table 1). The v1.0 assembly was based on a 9.75-fold redundant whole genome shotgun (WGS) dataset in paired-end reads from three 3.1 ± 0.2 Kb genomic DNA libraries (Martinez et al., 2004). To supplement this data set and improve the assembly, an additional 116 Mb of high quality shotgun sequence was generated in the form of paired-end sequences from 6.3 Kb plasmid and 35 Kb fosmid clones. All sequence reads are available from the NCBI trace archive (http://www.ncbi.nih.gov/Traces/). The WGS reads were assembled with version 1.0.3 of the JAZZ Assembler. The assembly contains a total of 32.5 Mb of sequence (excluding gaps) and 1252 contigs. Half of this assembled contig sequence (N50) is contained in the largest 44 contigs, the smallest of which is 228 kb in length. There are a total of 232 scaffolds in the assembly, and 95% of the assembled sequence is contained in the longest 24 scaffolds, which all have a net length (excluding internal gaps) of at least 210 Kb.

Table 1
General features of *P. chrysosporium* genome

Property[a]	Release version	
	V1.0	V2.1
Assembly, Mbp	29.8	35.1
N90, scaffolds	161	21
N50, scaffolds	46	8
Total number of genes	11,777	10,048
Gene length, bp	1,164.4	1,667.0
Transcript length, bp	855.8	1,365.7
Protein length, aa	282.2	455.2
Exons per gene	3.6	5.9
Exon length, bp	234.6	233.6
Intron length, bp	118.6	64.2
Genes annotated with:		
KOG	3,578	7,220
EC	1,673	2,252
GO	4,035	4,923

[a] Abbreviations: N90 and N50, number of scaffolds containing 90% and 50% of assembly, respectively. Proteins classified by KOG, eukaryotic orthologous groups (Koonin et al., 2004); EC, enzyme commission numbers assigned by KEGG; GO, gene ontology consortium (http://geneontology.org/).

A total of 10,048 gene models were predicted and annotated in the v2.1 *P. chrysosporium* genome assembly using JGI Annotation pipeline. Predicted genes, supporting evidence, annotations, and analyses are available through interactive visualization and analysis tools from JGI Genome Portal (www.jgi.doe.gov/whiterot).

Gene prediction methods used for annotation of the new assembly include ab initio methods Fgenesh (Salamov and Solovyev, 2000), homology-based methods, Fgenesh+ (www.softberry.com) and Genewise (Birney and Durbin, 2000). Fgenesh was trained on a set of available mRNAs, ESTs, and reliable homology gene models and showed 78.3% sensitivity (fraction of correctly detected true exons) and 77.6% specificity (fraction of true exons among all predicted exons). GeneWise models were extended when possible to include start and stop codons. When multiple models were predicted at the same locus, the model with best homology, including coverage in both model and hit sequences, was selected for the non-redundant set of genes called 'BestModels, v2.1.' This set includes 66.5% Fgenesh(+) and 33.5% Genewise gene models. Only 6% are supported by both methods.

Approximately 75% of genes in the non-redundant set have known functional domains or show homology to known proteins in other genomes. Genes have been annotated and classified according to GO (gene ontology consortium http://www.geneontology.org/), eukaryotic orthologous groups (KOGs (Koonin et al., 2004)), and KEGG metabolic pathways (Kanehisa et al., 2004). Following KEGG annotation, E.C. numbers have been assigned to 2252 genes. 7220 and 4923 genes have KOG and GO assignments, respectively.

Comparison with 11,777 gene models from the earlier annotation release v1.0 (Martinez et al., 2004), where Genewise (Birney and Durbin, 2000) and GrailEXP (Xu and Uberbacher, 1997) were used for gene prediction, shows considerable improvement. Contaminants and transposons have been removed; gene fragments were combined into more complete and dense models. Gene structure statistics summarized in Table 1 indicate smaller introns, higher number of exons, and larger transcripts and proteins. Manual validation of the models also supports these observations.

V1.0 gene models were mapped on the new assembly (displayed as 'BestModels v1.0' track) and compared with v2.1 model set. Seventy-seven percent of gene loci of v1.0 are present in v2.1. While the majority of release-specific gene models are ab initio predictions, 8% of missed models from the old set contain PFAM domain and half of them are related to repeats. Fourteen percent of the new models contain additional functional domains.

Based on analysis of domain composition in both gene sets, the repertoire of molecular functions of the new release became richer. According to analysis of PFAM domains, out of 1637 PFAM domains, 589 occur more frequently in new set with 309 not present in the earlier release. One hundred and eighty-nine lost domains from the previous release include large number of transposon-related domains like integrase and reverse transcriptase that were removed from the current set of genes.

3. Results

3.1. Secretome computational analysis

A secretome dataset of *P. chrysosporium* was generated using the most recent assembly (www.jgi.doe.gov/whiterot). The 10,048 v2.1 protein models were submitted to PHOBIUS (http://phobius.cgb.ki.se/index.html), predictive software with improved discrimination between transmembrane helices and signal peptides (Kall et al., 2004). A total of 874 potentially secreted proteins were predicted, and then further reduced to 769 by filtering out proteins with putative mitochondrial targeting signals (http://www.cbs.dtu.dk/services/TargetP/). Thus, at least 7.6% of the *P. chrysosporium* gene models are predicted to encode secreted proteins. For comparison, 4.5% of the 6165 ORFs of *Candida albicans* were predicted using a similar computational approach (Lee et al., 2003b). It is important to note that some of these proteins are likely not extracellular, including those that may be cell wall bound, residing within vacuoles, or ER-related.

BlastP analysis against the NCBI non-redundant database categorized all but 38 sequences by some similarity (>50 Smith–Waterman score) to current accessions. Nearly half were similar only to hypothetical proteins, many of which were conceptual translations from other fungal genome projects (Fig. 1). ClustalW analysis of the 359 hypothetical proteins revealed few closely related sequences, i.e., a total of five pairs with >80% amino acid identity. Hypothetical proteins sharing >35% sequence identity tended to cluster within the genome. More specifically, we observed 15 separate clusters each containing 2–7 members. Five pairs of structurally related hypothetical genes showed no apparent linkage (Supplemental material). A genome-wide examination of all genes encoding secreted proteins revealed a decidedly non-random distribution. Among the 12 longest scaffolds, the percentage of genes encoding extracellular enzymes ranged from 9.5 (scaffold 1) to 1.6 (scaffold 6). Closer inspection of scaffolds showed clusters unevenly distributed along their length, as is clear from the examples shown in Fig. 2.

The 407 models with similarity to known proteins were broadly categorized as glycoside hydrolases (87), oxidoreductases (84), peptidases (52), and esterases–lipases (21), while others could be assigned to more specific grouping such as hydrophobins (14), lignin peroxidases (10), and manganese peroxidases (5). The 134 models not easily assigned to large families or broad functional categories (Fig. 1A, Misc. proteins), included 103 proteins with significant similarity (Smith–Waterman scores >100) to known proteins. Glycosyl hydrolases and peptidases were allocated to specific families or clans (Figs. 1B and C) according to systems of Coutinho and Henrissat (http://afmb.cnrs-mrs.fr/CAZY/; (Henrissat, 1991)) and Rawlings et al. (http://merops.sanger.ac.uk/; (Rawlings et al., 2004)), respectively.

3.2. Protein identifications in ligninolytic media

Total soluble protein from carbon- and nitrogen-limited cultures was concentrated by ultrafiltration, size fractionated by SDS–PAGE, and subjected to LC–MS/MS analysis. Searches of the 10,048 v2.1 protein database using conservative cut-off scores (>13, SpectrumMill; >40, Mascot), allowed unambiguous assignment of 77 unique peptide sequences to 40 specific gene models (Tables 2 and 3). Expression of 11 of these genes was previously observed in cellulolytic cultures (Vanden Wymelenberg et al., 2005). Of the 29 new extracellular proteins, 15 were detected only in carbon-limited filtrates, while 5 were expressed in both carbon and nitrogen cultures. Analysis of previously acquired spectra using the v2.1 model database demonstrated the expression of 5 additional genes in cellulolytic medium (Table 4). A complete listing of the computational secretome, all peptide sequences, scores, and previously published results are available online (Supplemental material). Results for particular protein families follow.

3.2.1. Peptidases

A total of 31 peptidase sequences were detected in carbon- and nitrogen-limited cultures. Twenty unique sequences were assigned to 11 specific gene models, representing MEROPS peptidase families A1, S10, and S53. No representatives of other gene families, including the large family of glutamic proteases (Fig. 1C; (Sims et al., 2004)), were detected in these studies or in earlier investigations of

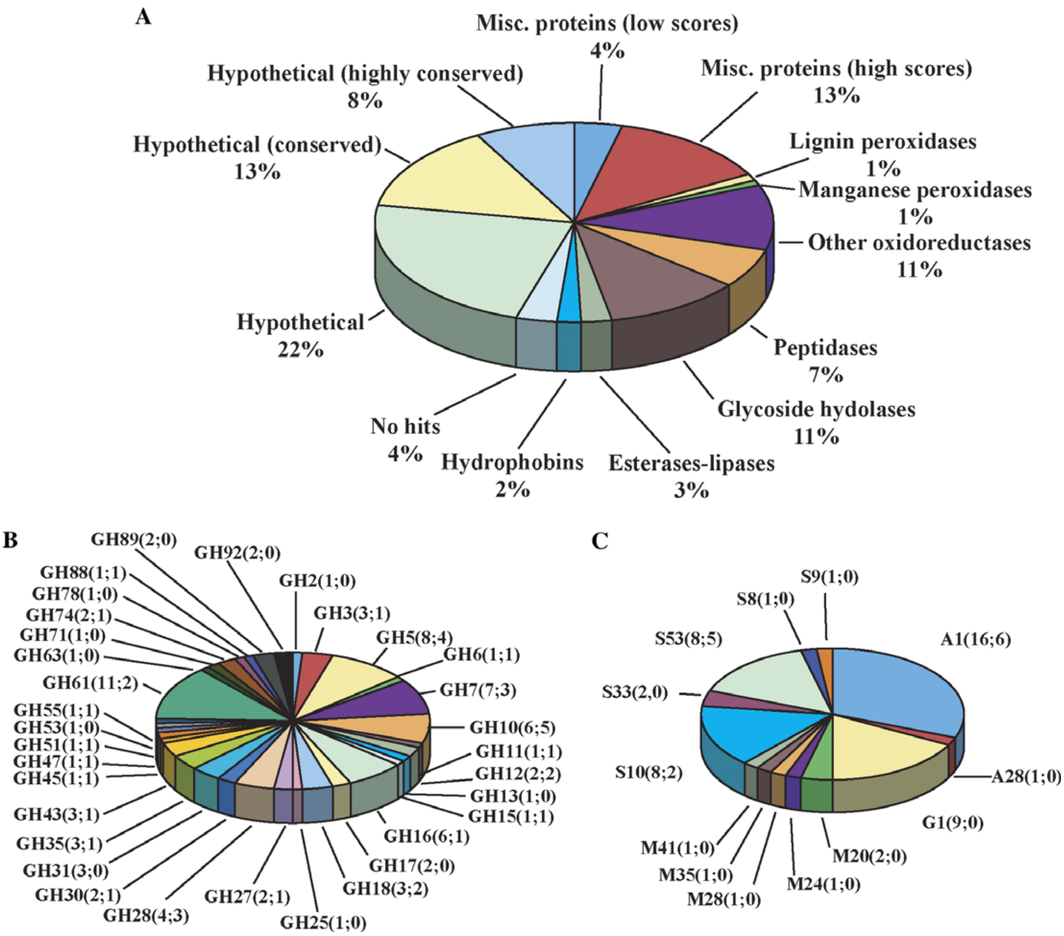

Fig. 1. Distribution of *P. chrysosporium* secretome models. (A) Proteins were analysed by BlastP using the BLOSUM62 matrix. The total 794 models include 25 LC–MS/MS-detected proteins not predicted by PHOBIUS due to incomplete N-terminal sequence. Designated "Hypothetical," 182 models were similar only to other hypothetical or putative proteins with relatively low Smith–Waterman (Smith and Waterman, 1981) scores (<50). Hypothetical proteins designated "conserved" and "highly conserved," showed increased similarity to such conceptual translations with scores of 50–100, and >100, respectively. In addition to those assigned to recognized structural and functional groupings (lignin peroxidases, manganese peroxidases, other oxidoreductases, peptidases, glycoside hydrolases, esterases–lipases, and hydrophobins), 144 models showed similarity to a wide range of proteins (Misc. proteins). Only 28 models (~4%) gave 'no hits' to the NCBI database. See Supplemental materials for detailed list. (B) Eighty-seven models (~11% of total) were assigned to specific glycoside hydrolase families (http://afmb.cnrs-mrs.fr/CAZY/) (Henrissat, 1991). (C) Fifty-two peptidases were classified by MEROPS server (interrefhttp://merops.sanger.ac.uk/urlhttp://merops.sanger.ac.uk/) (Rawlings et al., 2004). Family designations are followed parenthetically by the number of family members with predicted secretion signal and the number of expressed genes experimentally confirmed by mass spectroscopy.

cellulolytic cultures. A relatively low-scoring (12.6) peptide sequence was tentatively assigned to model 133799. Designated *pcs1*, the cDNA for this serine protease was previously isolated from a medium unlikely to support significant lignin- or cellulose-degrading activity (Faraco et al., 2005). Serine protease *scp1* and aspartyl protease *asp2*, both expressed in cellulolytic cultures (Vanden Wymelenberg et al., 2005), were not detected here.

Protease genes *asp1* and *prt53A* (Table 2) were expressed under all conditions and their corresponding cDNAs were cloned and sequenced. The *prt53A* cDNA sequence (GenBank DQ242648) confirmed the accuracy of model 133020 except for a 5 amino acid insertion within the first intron. Database searches showed the *prt53A*-encoded protein corresponds to the N-terminal sequence of a pepstatin insensitive protease derived from solid substrate cultures (Datta, 1992). Based on this experimentally determined N-terminal position, the mature protein is 364 amino acids with a molecular weight of 37 kDa. A 17 residue secretion signal and 184 propeptide precede the mature peptide. The full-length sequence is similar to numerous hypothetical proteins and to aorsin, a family S53 protease from *Aspergillus oryzae* (gi21321299; (Lee et al., 2003a)). The cDNA corresponding to *asp1* (GenBank DQ242649) precisely matched model 135608. The *asp1* sequence is most closely related (Smith–Waterman score = 452) to the *Irpex lacteus* aspartyl protease (Fujimoto et al., 2004), and by comparison with several related sequences, a preprosequence of approximately 67 residues is suspected.

Clustering of family A1 aspartyl proteases was observed (Fig. 2). Genes designated *asp3*, *asp4*, and *asp5* were tandemly oriented within 7.3 kb on scaffold 17, and all were expressed in carbon- and nitrogen-starved cultures (Table 2). No peptides corresponding to the adjacent gene *asp6*

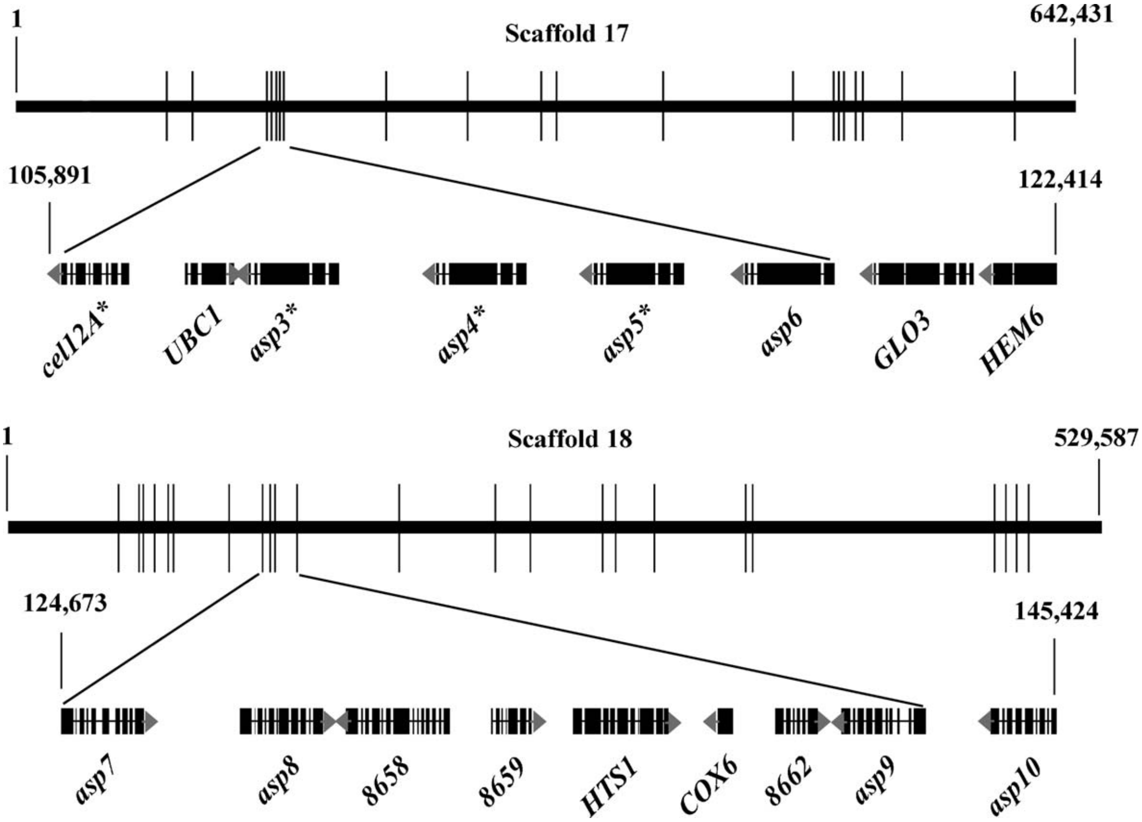

Fig. 2. Distribution of genes predicted to encode secreted proteins on scaffolds 17 and 18. Vertical crossbars show relative positions on scaffolds and protease-containing regions are expanded. The 16.5 kb region of scaffold 17 contains aspartyl proteases designated *asp3*, *asp4*, *asp5*, and *asp6*. *Saccharomyces cerevisiae* homologues flanking the cluster include those encoding putative ubiquitin-conjugating protease (*UBC1*), GTPase-activating protein (*GLO3*), and coproporphyrinogen oxidase (*HEM6*). A family 12 glycosyl hydrolase is encoded by *P. chrysosporium cel12A* (gi 51872339). The 20.8 kb region of scaffold 18 contains family A1 sequences designated *asp7*, *asp8*, *asp9*, and *asp10*. *S. cerevisiae* homologues within the cluster include histidyl tRNA synthase (*HTS1*) and subunit VI of cytochrome *C* oxidase (*COX6*). BlastP analysis of NCBI revealed no clear *S. cerevisiae* homologues for models 8658, 8659, and 8662. Model 8658 encodes a glycine-rich protein with clear secretion signal. The *asp10* model is N-terminally incomplete and lacks a clear secretion signal. Arrows show transcriptional orientation. Alternating fill and open spaces indicate exons and introns, respectively. LC–MS/MS-detected proteins are marked by asterisks.

were detected, but another nearby gene, *cel12A*, is expressed in cellulolytic cultures (Vanden Wymelenberg et al., 2005). ClustalW analysis of *asp3* through *asp10* revealed two well-defined clades apparently related to genome position. Pairwise comparisons among scaffold 17 sequences (*asp3-6*) ranged from 50 to 83%, whereas comparison of the same sequences to *asp7-10* showed <25% identity. Expressed genes *asp1*, *asp2*, and *asp11* (Table 2) lie on scaffolds 13, 25, and 7, respectively. The *asp1* gene was 47–58% identical to scaffold 18 family A1 genes, and substantially less similar (<28%) to those located on scaffold 17. The *asp2* and *asp11* sequences were more distantly related to other A1-encoding genes (18–36% identity) and to each other (28%). The *asp1* gene also lies adjacent to a closely related (65% identity) A1-like sequence (model 7287) on scaffold 13. In addition, three A1-like sequences (models 8008, 8010, and 8011) are positioned within a 10 kb region on scaffold 15.

In contrast to the A1 protease family, limited linkage was observed among the expressed serine proteases. *prt53C* and *prt53D* are adjacent on scaffold 2, and *prt53A* lies 6.6 kb from another S53-like protease (model 133398) on scaffold 1. Considering peptidase gene families whose expression is not yet established, extensive clustering was previously observed among glutamic acid proteases (family G1) (Sims et al., 2004). We also identified a cluster of 7 S33 family proteases within a 45 Kb region of scaffold 10.

3.2.2. Peroxidases and related enzymes

Peptides corresponding to lignin peroxidase genes *lipA*, *lipD*, and *lipE* were detected in carbon-starved cultures, and *lipC* peptides were identified in nitrogen-limited media (Table 3). These protein profiles are consistent with transcript patterns (Holzbaur and Tien, 1988; James et al., 1992; Reiser et al., 1993; Stewart and Cullen, 1999). Several peptides from carbon- and nitrogen-starved cultures were unambiguously assigned to manganese peroxidase gene *mnp2*. Owing to the close structural similarity among *P. chrysosporium* peroxidases, several peptides could not be attributed to a single gene, e.g., *lipA*/*lipH*, *mnp1*/*mnp4* (Table 3). Consistent with a close physiological connection to peroxidases (Kersten and Kirk, 1987; Kersten, 1990) and with transcript patterns (Janse et al., 1998; Kersten and Cullen, 1993; Stewart et al., 1992), glyoxal oxidase peptides

Table 2
Phanerochaete chrysosporium protease peptides detected in defined media

Model[a]	Family[b]	Peptide sequence[c] (medium, high score)	Probable cleavage[d]	Comments[e]

8470	A1 (*asp5*)	KIFQTGQSSTAVDQHKT (CL, 16.8; NL, 16.7);
		RAQDAIVDTGTTLLIVDPTSATAIHRQ (NL, 21.2);
		RNLYTEFDFGGERV (NL, 19.3)
			19/20: AVA-SP	

126189	A1 (*asp11*)	KNDGEITFGGLDESKF (CL, 13.7)	Incomplete	
8469	A1 (*asp4*)	KTFQTGSSSTAVDQRK (CL, 17.0; NL, 17.7); RVGFAPVVLK (CL, 10.1; NL, 12.4)	18/19: LAA-AS	
8468	A1 (*asp3*)	RTFNTGASSTAVDQKQ (CL, 17.3; NL, 18.8); RGSLAFTPVSIRN (NL, 14.9)	21/22: ASP-AP	
135608	A1 (*asp1*)	KATGATLDNNTGLLRL (CL, 17.2; NL, 13.8)	19/20: VAA-TP	7 peptides in avicel (Vanden Wymelenberg et al., 2005)
40125	A1 (*asp2*)		Uncertain	4 peptides in avicel (Vanden Wymelenberg et al., 2005)
1914	S10 (*scp1*)		22/23: AHA-RM	1 peptide in avicel (Vanden Wymelenberg et al., 2005)
133799	S10 (*pcs1*)	RLAFGTPLLRA (CL, 12.6)	20/21: ALA-AK	AJ748587 (Faraco et al., 2005)
130748	S53 (*prt53B*)	KGVSVLFSSGDGGVGGSQSTRC (CL, 18.5; NL, 19.8); RLGLATTPFTTATTN (CL, 12.4)	Incomplete	
1483	S53 (*prt53C*)	KQLNAVGYTPSAKS (CL, 12.7; NL, 15.0)	18/19: VAA-AP	
129261	S53 (*prt53D*)	RAYPDVSAQADNFRI (CL, 17.8)	18/19: AVA-VP	
26825	S53 (*prt53E*)	RFQPNFPASCPFVTTVGATTRV (CL, 17.4; NL, 19.0); RGSSIMFSSGDDGVGAGNCLTNDGKN (NL, 18)	Incomplete	

133020	S53 (*prt53A*)	KATQSSNTLGVSGFIDQFANQADLTTFLNRF (CL, 16.5);
		RGTSILFASGDGGVSGGQSQSCTKF (CL, 19.5; NL, 16.2);
		KGWDPVTGLGTPNFAALKA (CL, 19.9; NL, 15.3);
		RSLANNLCNAYAQLGARG (CL, 14.5; NL, 13.1)
			17/18: AFA-KP	3 peptides in avicel (Vanden Wymelenberg et al., 2005). Corresponds to wood-derived protease (Datta, 1992).

[a] To access protein information, end URL with model number, e.g., http://genome.jgi-psf.org/cgi-bin/dispGeneModel?db = Phchr1&id = 8470.

[b] Peptidase families identified by MEROPS (http://merops.sanger.ac.uk/) as described (Rawlings et al., 2004).

[c] Peptide sequences, media, and Spectrum Mill scores for *P. chrysosporium* strain RP78 cultivated in media designed for high production of lignin peroxidases (carbon-limited, CL; and nitrogen-limited, NL) as described in text.

[d] Most probable secretion signal cleavage site as determined by PHOBIUS and SignalP. Although predicted as secreted, the precise cleavage site could not be determined for the *asp2*-encoded protein. Models with incomplete N-terminals are noted.

[e] Peptides corresponding to four proteases were previously detected in medium containing avicel as sole carbon source (Vanden Wymelenberg et al., 2005).

were detected in carbon- and nitrogen-starved media. Using MALDI-MS fingerprinting, glyoxal oxidase was recently identified in similar media supplemented with vanillin (Shimizu et al., 2005). Clustering of lignin peroxidase genes has been reported (Stewart and Cullen, 1999), and *lipA*, *lipC*, and *lipE* are located within a 36 kb region on scaffold 19. The *lipD* gene is unlinked to all other peroxidases (Gaskell et al., 1994; Stewart et al., 1992).

Several peptides in carbon-starved cultures were matched to a model (3383) with similarity to the endonuclease/exonuclease/phosphatase family of proteins (pfam03372). A full-length cDNA clone was obtained (GenBank DQ242647), revealing several relatively minor errors related to intron–exon boundaries in the model. Subsequent analysis showed a precise match with the experimentally derived N-terminal sequence of an extracellular mannose-6-phosphatase (Rothschild et al., 1999). Designated *mpa1*, the gene encodes a 22 amino acid secretion signal, followed by a 5 residue propeptide, and a mature peptide of 327 residues. The predicted molecular weight of 35 kDa and p*I* of 5.2 are in good agreement with experimentally determined properties of the monomer. The enzyme has been shown to dephosphorylate lignin peroxidase isozyme H2, the product of *lipD* (Rothschild et al., 1999).

3.2.3. *Glycosyl hydrolases*

Eighteen unique peptide sequences were assigned to eight specific glycosyl hydrolase (GH) genes (Table 3), most of which have been implicated in degradation of hemicellulose or pectin. Expression of putative xylanase *xyn10D* and exoglucanase *exg55A* were previously observed in submerged cultures with ground wood (Abbas et al., 2004) or with avicel (Vanden Wymelenberg et al., 2005) as sole carbon sources. *xyn10A*- and *xyn10C*-encoded peptides were also detected in avicel medium (Vanden Wymelenberg et al., 2005), and a genomic clone of the former was successfully expressed in *Aspergillus niger* (Decelle et al., 2004).

Peptides corresponding to three previously unidentified GHs were detected in carbon-limited cultures. A GH family 28 sequence showed substantial sequence similarity (Smith–Waterman scores ~200) to several known exopolygalacturonases and was designated *epg28B*. The *P. chrysosporium* genome contains a minimum of five GH28-like sequences (Martinez et al., 2004), and the deduced *epg28B* sequence is 20 and 25% identical to *epg28A* and *rhg28*, respectively. The latter gene encodes a putative rhamogalacturonase (Vanden Wymelenberg et al., 2005). In contrast to the GH28 family, no representative GH35 or GH47 proteins had been previously observed in *P. chrysosporium* cultures.

Table 3
Peptides identified in *P. chrysosporium* culture filtrates

Protein model[a]	Putative identity	Peptides detected (medium, high score)[b]	Probable cleavage[c]	Comments[d]
Esterases–lipases				
38233	Carboxylesterase (pfam00135)	RTGCSGSADTLQCLRQ (CL, 18.0)	Incomplete	*P. sapidus* gi55466915 [583]
7398	Carboxylesterase (pfam00135)	RAAIFDSSTGPFKT (CL, 19.1); KAVGCTSGPGSFECLQRV (CL, 16.7; NL, 11.7); KTAPPASTYDEADKPFALLTKA (CL, 16.2)	23/24: AKA-GS	35.8% identical to protein 38233
2540	Lipase (pfam01764)	RINNKEDPIPIVPGRF (CL, 12.5)	19/20: ALA-AP	*Ustilagu maydis* ct gi71008942 [151]
8996	Lipase (pfam01764)	RINNESDPIPIVPGRF (CL, 15.0; NL, 13.2); RVGNPDFAALFDGEVSDFERI (CL, 15.3)	19/20: AHA-AA	49% identical to protein model 2540
10607	GDSL-like lipase (pfam00657)	RVLADGLGPNALGRI (CL, 14.8)	Incomplete	*Aspergillus fumigatus* ct gi66846850 [343]
126075	*axe1* (acetyl xylan esterase)	FAISNWGVDPNRV (CL, 13.0)	24/25: SQC-LP	4 peptides in avicel (Vanden Wymelenberg et al., 2005)
Glycosyl hydrolases				
10763	GH10 xylanase (*xyn10*C)	KLYWGTAADQNRF (CL, 14.9)	Incomplete	5 peptides in avicel (Vanden Wymelenberg et al., 2005)
30981	GH10 xylanase (*xyn10*D)	RGVFTFANADTIANLARN (CL, 23); KLYINDFNIEGTFAKS (CL, 19.4); RMTLPSTPALLQAQKA (CL, 16.2; NL, 16.1); KSTAMQNLVRS (CL, 14.1)	Incomplete	1 peptide in avicel (Vanden Wymelenberg et al., 2005)
138345	GH10 xylanase (*xyn10*A)	RMTLPSTPALLAQQKT (CL, 16.1)	19/20: VQA-QS	2 peptides in avicel (Vanden Wymelenberg et al., 2005)
4449	GH28 exopolygalacturonase (*epg28B*)	KVFGGNPSPTSTAGGGTGFVKN (CL, 17.1)	Incomplete	*Aspergillus tubingensis* gi1483221 [199]
9466	GH35 β-galactosidase (*lac35A*)	RFPVPVGILNPNGKK (CL, 18.4); RPNDTGAQFIIVRQ (CL, 13.7); RTLPGVATFAGVKL (CL, 13.9); KVILTDYTFGNPANANKL (CL, 19.6)	22/23: ANS-AV	*Penicillium emersonii* gi44844271 [499]
4550	GH47 α-mannosidase (*msd47A*)	KEFAFGHDDLEPVSKS (CL, 14.1)	21/22: VAA-GQ	*Aspergillus saitoi* gi1171477 [273]
8072	Exo-1,3-β-glucanase (*exg55A*)	KGDGNTDDTAAIQAAINAGGRC (CL, 20.9; NL, 14.1); KSHPQYTGYAPSDFVSVRS (CL, 21.4); RSNNPNGFADTITAWTRN (CL, 20.9); KVSSPLVVLYQTQLIGDAKN (CL, 20.2); RWSGASSGHLQGSLVLNNIQLTNVPVAVGVKG (CL, 19.5)	26/27: ASG-LG	12 peptides in avicel (Vanden Wymelenberg et al., 2005)
41123	GH61 endoglucanase (*cel61C*)	RVPPNNNPVTDVTSKD (CL, 14.8)	Incomplete	1 peptide in avicel
Peroxidases				
10957	Lignin peroxidase (*lipA*)	RAPATQPAPDGLVPEPFHTVDQIINRV (CL, 19.3)	21/22: ANA-AA	gi126285
10957 or 121806	Lignin peroxidase (*lipA* or *lipH*)	RGTAFPGSGGNQGEVESPLPGEIRI (CL, 21.5; NL, 19.4)	21/22: ANA-AA or VQG-AA	gi126285 or gi5669882

131738	Lignin peroxidase (*lipC*)	RAPATQPAPDGLVPEPFHSVDQIIDRV (NL, 19.5)	21/22: AQG-AA	gi3137
6811	Lignin peroxidase (*lipD*)	RLQTDHLFARD (CL, 15.4); KTGIQGTVMSPLKG (CL, 16.5)	21/22: TQA-AP	2 peptides in avicel (Vanden Wymelenberg et al., 2005)
11110	Lignin peroxidase (*lipE*)	RGTLFPGSGGNQGEVESGMAGEIRI (CL, 19.4); RKPATQPAPDGLVPEPFHTVDQIIARV (CL, 21.4)	21/22: ANA-AV	gi169271
140708 or 8191	Manganese peroxidase (*mnp1* or *mnp4*)	RFEDAGGFTPFEVVSLLASHSVARA (CL, 18.24)	18/19: VRA-AP	MnP1 and MnP4 differ by single aa. *mnp1* = gi13124450
3589	Manganese peroxidase (*mnp2*)	RFEDAGNFSPFEVVSLLASHTVARA (CL, 22.5; NL, 23); RSSLIDCSDVVPVPKPAVNKPATFPATKG (CL, 22.4; NL, 19.3); KDLDTLTCKA (CL, 14.7)	18/19: TRA-AP	*mnp2* = gi169292
140708 or 3589 or 8191 or 4636	Manganese peroxidase (*mnp1* or *mnp2* or *mnp4* or *mnp5*)	KHNTISAADLVQFAGAVALSNCPGAPRL (CL, 17.3)	18/19: VRA-AP or TRA-AP or VRA-AP or TLA-AP	*mnp5* is 90% identical to *mnp1*
Other				
5655	Acid phosphatase (*pho1*)	RFGIQTLSPKF (CL, 14.6; NL, 15.2); RLNWVNSFPVDAVRF(CL, 15.4)	18/19: AHS-QN	1 peptide in avicel (Vanden Wymelenberg et al., 2005)
964	Related to alginate lyase	KVPGLYGGNSDDEAVSCSGGRR (CL, 19.6)	20/21: VAA-RP	*Haliotis discus* gi34787299 (127)
3346	Amidase (pfam01425) (*amd1*)	RAVIETNPSALAQARV (NL, 18.4)	20/21: LSA-CL	*Stenotrophomonas maltophilia* gi19744118 (234). COOH′ TMH
140079	Glutaminase (*gta1*)	RAQFVNSGTLPNTQDTRF (CL, 14.7)	20/21: ASA-AV	21 peptides in avicel(Vanden Wymelenberg et al., 2005)
11068	Glyoxal oxidase (*glx1*)	RIETLDPPFMFRS (CL, 15.7); RISGLLSCFD (CL, 10.4); KNTETILPDIPNGVRV(CL, 11.9; NL, 11.3); RSRPALLTMPEKL (CL, 12.5); KVTVPITIPSDLKA (CL, 12.9; NL, 12.4)	22/23: ASD-AP*	gi399595
6854	Hypothetical protein	REFVVATVDPDAPTPQNPTVAQIRH (CL, 14.6)	19/20: VSA-QD	*Magnaporthe grisea* ct gi39977919 [65]
3328	Hypothetical protein	KVAIFGGKPGEQLQYKG (CL, 14.5); RISGTDFSQRF (NL, 14);	25/26: APA-AS	*Neurospora crassa* ct gi32416302 [64]
7809	Glucolactonase (COG3386) (*gnl1*)	RVVADGFDKPNGIIAFSEDGKT (CL, 16.1)	25/26: ASA-QT	*Gibberella fujikuri* gi52430041 [248]
3383	Phosphatase (pfam03372) (*mpa1*)	RANVGGNFATFTGFNSPGDTASFTRI (CL, 19.0); RDDGKQAGEFSAIFYNKN (CL, 15.4); RIDFVFGGSNGKW (CL, 16.8); KTGEQPWSTRR (CL, 15.9)	22/23: ASS-VV	Probable mannose-6-phosphatase (Rothschild et al., 1999)
8221	Similar to secreted antigens	RVNAQAEQVAASECGL (CL, 21.5)	19/20: ALA-AP	*Coccidioides posadasii* gi25528649 [116]

[a] To access protein information, end URL with model number, e.g., http://genome.jgi-psf.org/cgi-bin/dispGeneModel?db = Phchr1&id = 38233.

[b] Media designed for high production lignin and manganese peroxidases (carbon-limited, CL; and nitrogen-limited, NL). Spectrum mill scores >13 are considered significant.

[c] Most probable secretion signal cleavage site as determined by PHOBIUS and SignalP. Models with incomplete N-terminals are noted. *Experimentally determined (Kersten and Cullen, 1993).

[d] Similarity to other *P. chrysosporium* sequences or NCBI accessions [Smith–Waterman scores]. Abbreviations: ct, conceptual translation; COOH′ TMH, transmembrane helices located at carboxy terminus. Additional peptides corresponding to certain genes were previously identified in a medium containing avicel as sole carbon source (Vanden Wymelenberg et al., 2005), and the number of high-scoring peptides are noted. Peptides found exclusively in avicel media are listed in reference (Vanden Wymelenberg et al., 2005) and in Supplementary material online.

Table 4
Peptide sequences from avicel-containing medium assigned to v2.1 models, but not to earlier v1.0 models[a]

Model		Putative identity[b]	Peptides in avicel medium (score)[c]	Probable cleavage[d]	Comments[e]
V1.0	V2.1				
pc.15.107.1	140836	Aldose epimerase (*ale2*)	RLLTDPAHPVFNPIVGRY (16.6: #1)	Incomplete N-terminal	>75% identical to *ale1* (Vanden Wymelenberg et al., 2005)
pc.47.69.1	41641	Xylanase (*xyn10E*)	WDATENTRGVFTRSQAD (17.0: #4)	Incomplete N-terminal	91% identical to *xyn10C*
pc.4.68.1	139777	Hypothetical protein	RLFENTFPNTLDTTVKY (14.9: #1); PDLARLFENTFPNTLD (13.3: #2)	20/21: AGA-QC	Conserved DUF1237 domain
pc.10.47.1	138739	Hypothetical protein	KVVQNVAGSPSTNSEDFHVGILRI (13.7: #1)	19/20: TKA-GT	~*A. fumigatus* gi66853515 [238]
None	131440	Hypothetical protein	PDAAGNKLLFVNLGPYD (15.7: #4)	Incomplete N-terminal	CBM1 domain
None	129310	CBH1 *(cel7G)*	KYGTGYCDSQCPKD (16.5: #1); NDAAAFTPHPCTTTGQTRCSGD (18.6: #4)	18/19: AVG-QQ	*cel7F* duplication. CBM1 domain.

[a] LC–MS/MS data from earlier investigation (Vanden Wymelenberg et al., 2005) analyzed using v2.1 database (www.jgi.doe.gov/whiterot). These peptide sequences could not be assigned to earlier v1.0 models (Martinez et al., 2004). None of these peptides were detected in ligninolytic media (Tables 2 and 3).

[b] Putative identity determined by BlastP NCBI searches.

[c] Peptides and parenthetical Spectrum Mill scores for *P. chrysosporium* strain RP78 cultivated in cellulolytic medium containing avicel as sole carbon source.

[d] Most probable secretion signal cleavage site as determined by PHOBIUS and SignalP.

[e] The DUF1237 domain (pfam06824) is widely distributed and of unknown function. Family 1 carbohydrate binding modules (CBM1) (http://afmb.cnrs-mrs.fr/CAZY/) bind to crystalline cellulose. Peptides previously assigned to *cel7F* (v1.0, pc. 95.47.1, (Vanden Wymelenberg et al., 2005)) also match model 129310 (*cel7G*).

Designated *lac35A* and *msd47A*, those detected here are closely related to microbial β-galactosidases and α-mannosidases, respectively. Based on blast searches of the genome, the GH35 family contains three sequences, and *lac35A* is <37% identical to models 9590 and 134404. The GH47 family includes at least six genes, and model 2107 is most closely related to *msd47A* (28% identity). With few exceptions (Covert et al., 1992b), extended clusters have not been observed among glycosyl hydrolase genes.

3.2.4. Esterases and lipases

Several esterases and lipases were detected in carbon- and nitrogen-starved cultures and, with the exception of an acetyl xylan esterase (*axe1*), none were previously known (Table 3). Models 38233 and 7398 belong to a family of 10 carboxylesterases (pfam00135) and they are 63 and 32% identical, respectively, to a *Pleurotus sapidus* esterase (Zorn et al., 2005). The gene encoding protein 7398 and 2 other members of this carboxylesterase gene family are clustered within a 20 kb region on scaffold 13. Class 3 lipases (pfam01764) also occur as a family of related gene models, two of which were secreted under ligninolytic conditions. Pairwise comparisons within this family of six sequences range from 31 to 76% identity. Three class 3 lipases, including expressed protein model 2540, were clustered within a 20 kb region on scaffold 3. Finally, 1 peptide was assigned to protein model 10607, a putative lipolytic enzyme (i.e., Interpro family IPR001087, NCBI conserved domain cd01830, pfam00657). Blast analysis of *P. chrysosporium* v2.1 models suggests that model 10607 is unique.

3.2.5. Other proteins

Peptides corresponding to glutaminase *gta1* and acid phosphatase *pho1* were detected in avicel medium (Vanden Wymelenberg et al., 2005; Table 3). Not previously known were a putative amidase (model 3346) and a gluconolactonase (model 7809) detected in nitrogen- and carbon-limited cultures, respectively. A closely related ascomycete lactonohydrolase with lactone ring cleaving ability has been characterized (Honda et al., 2005; Kobayashi et al., 1998). The *gnl1* gene is unique within the *P. chrysosporium* genome, whereas the *amd1* is structurally related (62% identity) but unlinked to model 3719. A single peptide from carbon-limited cultures was assigned to model 8221, a small protein (<13 kDa) distantly related to secreted proteins of various plant and animal pathogens. Finally, 3 peptides corresponded to hypothetical models 6854 and 3328. These proteins were only distantly related to GenBank accessions most of which were conceptual translations.

3.3. Protein identifications in cellulolytic media revisited

Owing to improvements in the v2 assembly and v2.1 gene models, the expression of six new genes in cellulolytic medium was demonstrated by re-analyzing archived spectra (Vanden Wymelenberg et al., 2005; Table 4). In the case of a putative aldose epimerase gene *ale2*, a fifth GH10 endoxylanase gene *xyn10E*, and highly conserved hypothetical protein models 139777 and 138739, intron–exon junctions were corrected in v2.1 versus v1.0 models. Another hypothetical protein containing a highly conserved cellulose binding domain (model 131440) was not predicted in v1.0. This gene lies adjacent to another CBM1-containing sequence, model 3717. Overall, the two predicted proteins are 78% identical. Excluding their binding domains, the proteins show no significant similarity to any known carbohydrate-active enzymes.

Two peptide sequences were matched to gene model 129310, which is an exact duplication of *cel7G* gene encoding CBH1 (= model 129072). Designated *cel7F*, the sequence was truncated in the v1.0 assembly, lying at the terminus of scaffold 95 (Fig. 3). The two genes are located within 7 kb.

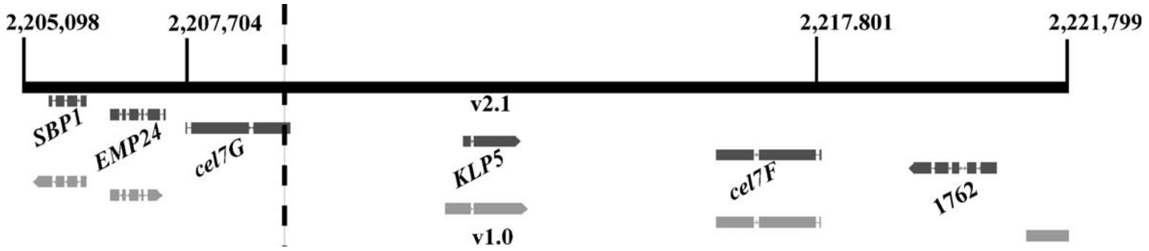

Fig. 3. Region of scaffold 2 containing *cel7* duplication. Earlier assembly (v1.0) scaffold break shown as dashed line. No *cel7G* model was observed in assembly v1.0. Gene designations are tentative and based on similarity to corresponding *S. cerevisiae* sequences. Abbreviations: *SBP1*, Ran-specific GTPase-activating protein 1; *EMP24*, p24 protein component of COPII-coated vesicles; *KLP5*, Kinesin-like protein.

4. Discussion

White-rot basidiomycetes, such as *P. chrysosporium*, are the only microbes convincingly shown to efficiently degrade all the major components of wood including cellulose, hemicellulose, and lignin. Owing to the large molecular weight of these polymers, the first step in the decay process is necessarily extracellular. Initial fragmentation of cellulosic and hemicellulosic materials typically involves hydrolytic attack by an array of glycoside hydrolases possibly combined with a limited number of esterases (acetyl xylan esterase) and oxidative enzymes (cellobiose dehydrogenase). In contrast, lignin depolymerization is generally believed to involve oxidative systems, the major components being lignin peroxidases, manganese peroxidases, and a peroxide-generating enzyme, glyoxal oxidase. The availability of a high quality draft genome and automated annotations provide an opportunity to define the *P. chrysosporium* secretome by computational and experimental approaches.

Employing PHOBIUS and TargetP, a 769-member 'computational secretome' was subdivided into broad categories based on BlastP analysis of the non-redundant NCBI database. This listing surely underestimates the actual number of secreted proteins, in large part due to incomplete N-termini in some models. Illustrating this shortcoming, 166 glycosyl hydrolases were predicted by conserved catalytic domains (Martinez et al., 2004) while 87 were predicted by PHOBIUS. Similarly, expression of 83 genes encoding extracellular proteins has been established by LC–MS/MS, but only 63 (76%) had been predicted. Manual inspection of the expressed, but unpredicted, gene models suggest inaccurate N-termini generally caused by introns punctuating short exons.

While the computational secretome is clearly incomplete, the predicted proteins include many interesting sequences that provide a framework for future investigations. For example, protein model 197 of *P. chrysosporium* is similar (bit score 111) to a riboflavin-oxidizing enzyme from *Schizophyllum commune* which has been proposed to play a role in removing nutrients essential for competing organisms (Chen and McCormick, 1997). The *S. commune* enzyme is specific for riboflavin and shows no activity with simple mono- or polyhydric alcohols, sugars or nucleosides (Tachibana and Oka, 1981). The *Phanerochaete* model 197 is similar to model 196 (bit score 92) and model 5495 (bit score 70), all of which have predicted secretion signals and regions of low complexity rich in Pro/Ser in the C-terminal region. Significantly, model 197 is adjacent to 196 on the genome, suggesting a functional relationship. Other interesting gene models within the computational secretome include recognizable genes such as those encoding putative oxalate decarboxylase isozymes, and a large number of structurally related hypothetical proteins.

Protein patterns in ligninolytic cultures differed sharply from cellulolytic medium (Vanden Wymelenberg et al., 2005). Hydrolytic enzymes such as cellobiohydrolase I isozymes, cellobiohydrolase II, endoglucanase isozymes, and β-glucosidase were detected in avicel medium, as was cellobiose dehydrogenase. None of these enzymes were observed in carbon- and nitrogen-starved cultures. Several carbohydrate-active enzymes broadly characterized as hemicellulases, e.g., xylanases, exopolygalacturonase, β-galactosidase, α-mannosidase, acetyl xylan esterase were present along with the expected oxidative enzymes lignin peroxidases, manganese peroxidase, and glyoxal oxidase. Simultaneous expression of these genes, as well as the putative carboxylesterases and lipases, may be related to the covalent linkages between hemicellulose and lignin in plant cell walls (Williamson et al., 1998).

While several of the enzymes in carbon- or nitrogen-starved cultures are likely involved in lignin and hemicellulose depolymerization, others may be involved in nutrient scavenging and recycling during idiophase. Proteinases, amidase, and glutaminase could play a role in nitrogen recycling in nutrient-starved defined media as well as in natural woody substrates. Derepression of protease expression under nutrient limitation is well known in Ascomycetes (Cohen, 1973). Consistent with the phylogenetic distribution of trypsin proteases, no family S1 sequences were detected in the *P. chrysosporium* genome (Hu and Leger, 2004). On the other hand, the genome features a minimum of 10 S10- and 9 G1-family sequences, yet none were detected in the media examined to date. Possibly, these proteins are targeted to vacuoles or endosomes. Future investigations may identify conditions under which these serine and glutamic acid proteases are secreted.

In addition to nitrogen acquisition, the expressed proteases may play an important role in the modification of extracellular protein. For example, proteases partially purified from *P. chrysosporium* cultures will cleave cellobiose dehydrogenase into separate functional domains (Eggert et al., 1996; Habu et al., 1993). In this connection, putative proteases *asp1* and *prt53A* have been identified in avicel cultures along with cellobiose dehydrogenase (Table 2) and we have detected their transcripts in colonized wood (data not shown). The simultaneous occurrence of proteases and lignin peroxidases is well established, but their physiological relationship remains unclear (Dass et al., 1995; Dosoretz et al., 1990a,b).

In contrast to the peptidases, the post-translational modifications catalyzed by mannose-6-phosphatase are well characterized (Kuan and Tien, 1989; Rothschild et al., 1997; Rothschild et al., 1999). Our results show this phosphatase is encoded by a single gene (*mpa1*) with no apparent paralogs in the *P. chrysosporium* genome. To date, dephosphorylation of lignin peroxidase H2 has been demonstrated only in nutrient-starved cultures, and the physiological role in lignin degradation, if any, is unknown. Possibly relevant to this question, we have identified *mpa1* transcripts in colonized wood (data not shown).

Expressed hypothetical proteins merit further investigation. By definition, ‘hypothetical’ protein sequences reveal little about function, but the conserved CBM1 in model 131440 strongly suggests interaction with crystalline cellulose (Table 4). Expressed hypothetical protein 13977 contains a highly conserved domain of unknown function (DUF1237) and EST analysis (contig 267, www.jgi.doe.gov/whiterot) demonstrates expression in colonized wood. Protein 138739 is highly homologous to several GenBank accessions, all of which were derived from filamentous Ascomycetes. In contrast, the sequences of proteins 6854, 3328, and 2035 (Table 3; (Vanden Wymelenberg et al., 2005)) offer few distinguishing features. A previously reported (Vanden Wymelenberg et al., 2005) expressed hypothetical protein (v.1 model pc.140.25.1) was substantially corrected and extended in v.2 (model 5607) and can now be assigned to GH family 30. Such model refinement, either automated or experimental, is expected to continually reduce the number of hypothetical proteins.

We observed extensive clustering of structurally related genes encoding genes with predicted secretion signals. Among these were 15 separate families of hypothetical proteins, several protease clusters, and 3-gene clusters for carboxylesterase and for lipases. Previous investigations had elucidated lignin peroxidases gene clusters (Stewart and Cullen, 1999) and more recent examination of the genome has revealed additional clustering of sequences encoding putative cytochrome P450s (Doddapaneni et al., 2005) and glutamic acid proteases (Sims et al., 2004). Additional clustering may be obscured by inaccurate models and by occasional positioning at scaffold termini.

In several instances, expression patterns seem related to genomic organization. For example, of five family A1 peptidases detected in nutrient-starved media, three are clustered on scaffold 17. Similarly, of four lignin peroxidases detected, three are clustered within 36 kb on scaffold 19. More surprisingly, structurally unrelated genes encoding extracellular proteins were closely linked in several instances. For example, a family 12 GH gene (*cel12A*), expressed in avicel medium, is located near the peptidase family A1 cluster on scaffold 17 (Fig. 2). Another example occurs within the A1 cluster on scaffold 18 where a glycine-rich hypothetical protein lies immediately adjacent to *asp8* (Fig. 2). Also, the CBM1 containing expressed hypothetical protein 131440 (Table 4) borders the expressed amidase, *amd1* (Table 3). The genome organization of the secretome and its regulation requires additional investigation.

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by the University of Wisconsin under Grant No. DE-FG02-87ER13712, the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fgb.2006.01.003.

References

Abbas, A., Koc, H., Liu, F., Tien, M., 2004. Fungal degradation of wood: initial proteomic analysis of extracellular proteins of *Phanerochaete chrysosporium* grown on oak substrate. Curr. Genet. 47, 49–56.

Birney, E., Durbin, R., 2000. Using GeneWise in the *Drosophila* annotation experiment. Genome Res. 10, 547–548.

Bonnarme, P., Asther, M., Asther, M., 1993. Influence of primary and secondary proteases produced by free or immobilized cells of the white rot fungus *Phanerochaete chrysosporium* on lignin peroxidase activity. J. Biotechnol. 30, 271–282.

Brown, A., Sims, P.F.G., Raeder, U., Broda, P., 1988. Multiple ligninase-related genes from *Phanerochaete chrysosporium*. Gene 73, 77–85.

Chen, H., McCormick, D.B., 1997. Riboflavin 5′-hydroxymethyl oxidation. Molecular cloning, expression, and glycoprotein nature of the 5′-aldehyde-forming enzyme from *Schizophyllum commune*. J. Biol. Chem. 272, 20077–20081.

Cohen, B.L., 1973. The neutral and alkaline proteases of *Aspergillus nidulans*. J. Gen. Microbiol. 77, 521–528.

Covert, S., Vanden Wymelenberg, A., Cullen, D., 1992b. Structure, organization and transcription of a cellobiohydrolase gene cluster from *Phanerochaete chrysosporium*. Appl. Environ. Microbiol. 58, 2168–2175.

Cullen, D., Kersten, P.J., 2004. Enzymology and molecular biology of lignin degradation. In: Brambl, R., Marzulf, G.A. (Eds.), The Mycota III Biochemistry and Molecular Biology. Springer-Verlag, Berlin, pp. 249–273.

Dass, S.B., Dosoretz, C.G., Reddy, C.A., Grethlein, H.E., 1995. Extracellular proteases produced by the wood-degrading fungus *Phanerochaete chrysosporium* under ligninolytic and non-ligninolytic conditions. Arch. Microbiol. 163, 254–258.

Datta, A., 1992. Purification and characterization of a novel protease from solid substrate cultures of *Phanerochaete chrysosporium*. J. Biol. Chem. 267, 728–732.

Decelle, B., Tsang, A., Storm, R., 2004. Cloning, functional expression and characterization of three *Phanerochaete chrysosporium* endo-1,4-*b*-xylanases. Curr. Genet. 46, 166–175.

Doddapaneni, H., Chakraborty, R., Yadav, J.S., 2005. Genome-wide structural and evolutionary analysis of the P450 monooxygenase genes (P450ome) in the white rot fungus *Phanerochaete chrysosporium*: evidence for gene duplications and extensive gene clustering. BMC Genomics 6, 92.

Dosoretz, C., Dass, B., Reddy, C.A., Grethlein, H., 1990a. Protease-mediated degradation of lignin peroxidase in liquid cultures of *Phanerochaete chrysosporium*. Appl. Environ. Microbiol. 56, 3429–3434.

Dosoretz, C.D., Chen, H.-C., Grethlein, H.E., 1990b. Effect of environmental conditions on extracellular protease activity in lignolytic cultures of *Phanerochaete chrysosporium*. Appl. Environ. Microbiol. 56, 395–400.

Eggert, C., Habu, N., Temp, U., Eriksson, K.-E.L., 1996. Cleavage of *Phanerochaete chrysosporium* cellobiose dehydrogenase (CDH) by three endogenous proteases. In: Srebotnik, E., Messner, K. (Eds.), Biotechnology in the Pulp and Paper Industry. Fakultas-Universitatsverlag, Vienna, pp. 551–554.

Eriksson, K.-E., Pettersson, B., 1982. Purification and partial characterization of two acidic proteases from the white rot fungus *Sporotrichium pulverulentum*. Eur. J. Biochem. 124, 635–642.

Faraco, V., Palmieri, G., Festa, G., Monti, M., Sannia, G., Giardina, P., 2005. A new subfamily of fungal subtilases: structural and functional analysis of a *Pleurotus ostreatus* member. Microbiology 151, 457–466.

Fujimoto, Z., Fujii, Y., Kaneko, S., Kobayashi, H., Mizuno, H., 2004. Crystal structure of aspartic proteinase from *Irpex lacteus* in complex with inhibitor pepstatin. J. Mol. Biol. 341, 1227–1235.

Gaskell, J., Stewart, P., Kersten, P., Covert, S., Reiser, J., Cullen, D., 1994. Establishment of genetic linkage by allele-specific polymerase chain reaction: application to the lignin peroxidase gene family of *Phanerochaete chrysosporium*. Bio/technology 12, 1372–1375.

Habu, N., Samejima, M., Dean, J.F.D., Eriksson, K.-E., 1993. Release of the FAD domain from cellobiose oxidase by proteases from cellulolytic cultures of *Phanerochaete chrysosporium*. FEBS Lett. 327, 101–106.

Henrissat, B., 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. 280 (Pt. 2), 309–316.

Holzbaur, E., Tien, M., 1988. Structure and regulation of a lignin peroxidase gene from *Phanerochaete chrysosporium*. Biochem. Biophys. Res. Commun. 155, 626–633.

Honda, K., Tsuboi, H., Minetoki, T., Nose, H., Sakamoto, K., Kataoka, M., Shimizu, S., 2005. Expression of the *Fusarium oxysporum* lactonase gene in *Aspergillus oryzae*: molecular properties of the recombinant enzyme and its application. Appl. Microbiol. Biotechnol. 66, 520–526.

Hu, G., Leger, R.J., 2004. A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. J. Evol. Biol. 17, 1204–1214.

James, C.M., Felipe, M.S.S., Sims, P.F.G., Broda, P., 1992. Expression of a single lignin peroxidase-encoding gene in *Phanerochaete chrysosporium* strain ME446. Gene 114, 217–222.

Janse, B.J.H., Gaskell, J., Akhtar, M., Cullen, D., 1998. Expression of *Phanerochaete chrysosporium* genes encoding lignin peroxidases, manganese peroxidases, and glyoxal oxidase in wood. Appl. Environ. Microbiol. 64, 3536–3538.

Kall, L., Krogh, A., Sonnhammer, E.L., 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338, 1027–1036.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32, D277–D280.

Kersten, P.J., 1990. Glyoxal oxidase of *Phanerochaete chrysosporium*; its characterization and activation by lignin peroxidase. Proc. Natl. Acad. Sci. USA 87, 2936–2940.

Kersten, P., Cullen, D., 1993. Cloning and characterization of a cDNA encoding glyoxal oxidase, a peroxide-producing enzyme from the lignin-degrading basidiomycete *Phanerochaete chrysosporium*. Proc. Natl. Acad. Sci. USA 90, 7411–7413.

Kersten, P.J., Kirk, T.K., 1987. Involvement of a new enzyme, glyoxal oxidase, in extracellular H_2O_2 production by *Phanerochaete chrysosporium*. J. Bacteriol. 169, 2195–2201.

Kirk, T.K., Schultz, E., Conners, W.J., Lorentz, L.F., Zeikus, J.G., 1978. Influence of culture parameters on lignin metabolism by *Phanerochaete chrysosporium*. Arch. Microbiol. 117, 277–285.

Kobayashi, M., Shinohara, M., Sakoh, C., Kataoka, M., Shimizu, S., 1998. Lactone-ring-cleaving enzyme: genetic analysis, novel RNA editing, and evolutionary implications. Proc. Natl. Acad. Sci. USA 95, 12787–12792.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Rogozin, I.B., Smirnov, S., Sorokin, A.V., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5, R7.

Kuan, I.C., Tien, M., 1989. Phosphorylation of lignin peroxidase from *Phanerochaete chrysosporium*. J. Biol. Chem. 264, 20350–20355.

Lee, B.R., Furukawa, M., Yamashita, K., Kanasugi, Y., Kawabata, C., Hirano, K., Ando, K., Ichishima, E., 2003a. Aorsin, a novel serine proteinase with trypsin-like specificity at acidic pH. Biochem. J. 371, 541–548.

Lee, S.A., Wormsley, S., Kamoun, S., Lee, A.F., Joiner, K., Wong, B., 2003b. An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. Yeast 20, 595–610.

Martinez, D., Larrondo, L.F., Putnam, N., Sollewijn Gelpke, M.D., Huang, K., Chapman, J., Helfenbein, K.G., Ramaiya, P., Detter, J.C., Larimer, F., Coutinho, P.M., Henrissat, B., Berka, R., Cullen, D., Rokhsar, D., 2004. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat. Biotechnol. 22, 695–700.

Rawlings, N.D., Tolle, D.P., Barrett, A.J., 2004. MEROPS: the peptidase database. Nucleic Acids Res. 32, D160–D164.

Reiser, J., Walther, I., Fraefel, C., Fiechter, A., 1993. Methods to investigate the expression of lignin peroxidase genes by the white-rot fungus *Phanerochaete chrysosporium*. Appl. Environ. Microbiol. 59, 2897–2903.

Rothschild, N., Hadar, Y., Dosoretz, C., 1997. Lignin peroxidase isozymes from *Phanerochaete chrysosporium* can be enzymatically dephosphorylated. Appl. Environ. Microbiol. 63, 857–861.

Rothschild, N., Levkowitz, A., Hadar, Y., Dosoretz, C., 1999. Extracellular mannose-6-phosphatase of *Phanerochaete chrysosporium*: a lignin peroxidase-modifying enzyme. Arch. Biochem. Biophys. 372, 107–111.

Salamov, A.A., Solovyev, V.V., 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. 10, 516–522.

Shimizu, M., Yuda, N., Nakamura, T., Tanaka, H., Wariishi, H., 2005. Metabolic regulation at the tricarboxylic acid and glyoxylate cycles of the lignin-degrading basidiomycete *Phanerochaete chrysosporium* against exogenous addition of vanillin. Proteomics 5, 3919–3931.

Sims, A.H., Dunn-Coleman, N.S., Robson, G.D., Oliver, S.G., 2004. Glutamic protease distribution is limited to filamentous fungi. FEMS Microbiol. Lett. 239, 95–101.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Stewart, P., Cullen, D., 1999. Organization and differential regulation of a cluster of lignin peroxidase genes of *Phanerochaete chrysosporium*. J. Bacteriol. 181, 3427–3432.

Stewart, P., Gaskell, J., Cullen, D., 2000. A homokaryotic derivative of a *Phanerochaete chrysosporium* strain and its use in genomic analysis of repetitive elements. Appl. Environ. Microbiol. 66, 1629–1633.

Stewart, P., Kersten, P., Vanden Wymelenberg, A., Gaskell, J., Cullen, D., 1992. The lignin peroxidase gene family of *Phanerochaete chrysosporium*: complex regulation by carbon and nitrogen limitation, and the identification of a second dimorphic chromosome. J. Bacteriol. 174, 5036–5042.

Tachibana, S., Oka, M., 1981. Occurrence of vitamin B-2 aldehyde forming enzyme in *Schizophyllum commune*. J. Biol. Chem. 256, 6682–6685.

Tien, M., Kirk, T.K., 1984. Lignin-degrading enzyme from *Phanerochaete chrysosporium*: purification, characterization, and catalytic properties of a unique H_2O_2-requiring oxygenase. Proc. Natl. Acad. Sci. USA 81, 2280–2284.

Williamson, G., Kroon, P.A., Faulds, C.B., 1998. Hairy plant polysaccharides: a close shave with microbial esterases. Microbiology 144 (Pt. 8), 2011–2023.

Vanden Wymelenberg, A.V., Sabat, G., Martinez, D., Rajangam, A.S., Teeri, T.T., Gaskell, J., Kersten, P.J., Cullen, D., 2005. The *Phanerochaete chrysosporium* secretome: database predictions and initial mass spectrometry peptide identifications in cellulose-grown medium. J. Biotechnol. 118, 17–34.

Xu, Y., Uberbacher, E.C., 1997. J. Comput. Biol. 4, 325–338.

Zorn, H., Bouws, H., Takenberg, M., Nimtz, M., Getzlaff, R., Breithaupt, D.E., Berger, R.G., 2005. An extracellular carboxylesterase from the basidiomycete *Pleurotus sapidus* hydrolyses xanthophyll esters. Biol. Chem. 386, 435–440.