nature biotechnology

Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium strain RP78*

Diego Martinez[1], Luis F Larrondo[2], Nik Putnam[1,3], Maarten D Sollewijn Gelpke[1], Katherine Huang[1], Jarrod Chapman[1,3], Kevin G Helfenbein[8], Preethi Ramaiya[4], J Chris Detter[1], Frank Larimer[5], Pedro M Coutinho[6], Bernard Henrissat[6], Randy Berka[4], Dan Cullen[7] & Daniel Rokhsar[1]

White rot fungi efficiently degrade lignin, a complex aromatic polymer in wood that is among the most abundant natural materials on earth. These fungi use extracellular oxidative enzymes that are also able to transform related aromatic compounds found in explosive contaminants, pesticides and toxic waste. We have sequenced the 30-million base-pair genome of *Phanerochaete chrysosporium* strain RP78 using a whole genome shotgun approach. The *P. chrysosporium* genome reveals an impressive array of genes encoding secreted oxidases, peroxidases and hydrolytic enzymes that cooperate in wood decay. Analysis of the genome data will enhance our understanding of lignocellulose degradation, a pivotal process in the global carbon cycle, and provide a framework for further development of bioprocesses for biomass utilization, organopollutant degradation and fiber bleaching. This genome provides a high quality draft sequence of a basidiomycete, a major fungal phylum that includes important plant and animal pathogens.

Lignin, a major 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 (the most abundant natural polymers), is not susceptible to hydrolytic attack. Only a small group of fungi are able to completely degrade lignin to carbon dioxide and thereby gain access to the carbohydrate polymers of plant cell walls, which they use as carbon and energy sources. Selective degradation of lignin by these fungi leaves behind crystalline cellulose with a bleached appearance that is often referred to as "white rot"[1,2]. These filamentous wood decay fungi are common inhabitants of forest litter and fallen trees, and have potential in a wide range of biotechnological applications including hazardous waste remediation and the industrial processing of paper and textiles. All white rot fungi are basidiomycetes, a diverse fungal phylum that accounts for over one-third of fungal species, including edible mushrooms, plant pathogens such as smuts and rust, mycorrhizae and opportunistic human pathogens.

The most intensively studied white rot basidiomycete, *P. chrysosporium*[3], is phylogenetically distant from other sequenced fungi, all of which are members of the Ascomycotina, *e.g.*, *Saccharomyces cerevisiae*[4], *Schizosaccharomyces pombe*[5] and *Neurospora crassa*[6–8]. Like most basidiomycetes, the vegetative mycelium of *P. chrysosporium* contains two distinct haploid nuclei, a condition known as dikaryosis. Restriction-fragment length polymorphism (RFLP) mapping and low-resolution pulsed field gels suggest it contains seven to nine chromosomes with a haploid genome size of approximately 30 million base pairs (Mbp)[9,10]. To reveal the genetic repertoire of white rot fungi we have sequenced the genome of *P. chrysosporium*, a representative fungus, using a whole genome shotgun strategy. Here we present an initial analysis of the draft genome sequence. An interactive web portal to the white rot genome can be found at http://www.jgi. doe.gov/whiterot.

RESULTS

Assembly and general characteristics

Using a pure whole genome shotgun approach, we sequenced the *P. chrysosporium* genome to >10.5 coverage (Methods). The net length of assembled contigs totaled 29.9 Mbp, which is in excellent agreement with pulsed-field gel estimates of a 30-Mbp genome size[9,10]. Genome statistics are presented in **Table 1**.

Identifying genes in the *P. chrysosporium* genome is particularly challenging, because it is the first genome representing the phylum Basidiomycota. Fungi are believed to have appeared approximately one billion years ago, and the divergence of the Basidiomycota and Ascomycota occurred over 500 million years ago[11]. To reveal the gene repertoire of *P. chrysosporium*, we combined comparative methods with an *ab initio* gene-finding approach, bootstrapping the gene prediction process by first obtaining high confidence homologs and subsequently using these models to identify less conserved genes (see

[1]US DoE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [2] Departmento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontifica Universidad Católica de Chile and Millennium Institute for Fundamental and Applied Biology, Santiago, Chile. [3]Department of Physics, University of California, Berkeley 94720, USA. [4]Novozymes Biotech, 1445 Drew Avenue, Davis, California 95616, USA. [5]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [6]Architecture et Fonction des Macromolécules Biologiques, UMR 6098, CNRS and Universités d'Aix-Marseille I & II, 31 Chemin Joseph Aiguier, 13402 Marseille, France, [7]USDA, Forest Service, Forest Products Laboratory, One Gifford Pinchot Dr., Madison, Wisconsin 53726, USA. [8]Invertebrate Zoology, American Museum of Natural History, New York, New York 10027, USA. Correspondence should be addressed to D.C. (dcullen@facstaff.wisc.edu) or D.R. (DSRokhsar@lbl.gov).

Published online 2 May 2004; doi:10.1038/nbt967

Methods). This modeling approach yielded a prediction of 11,777 genes in the *P. chrysosporium* genome.

Of the 11,777 predicted models, 72.1% (8,486) have significant sequence similarity to GenBank proteins (Smith-Waterman alignment is at least 60% of either model or hit length), and 6.4% (757) have <60% alignment but contain conserved protein domains according to InterPro scanning[12]. Of the remaining 2,534 models, 2,499 are GRAIL[13] predictions (99.7%), representing highly divergent genes, previously unrecognized genes that are potentially unique to filamentous fungi, basidiomycetes, or white rot fungi, or spurious gene predictions. Within the subset of 8,486 gene models with significant sequence similarity to known genes or domains, 28.7% (2,439) are in essence full-length homologs of known proteins (*i.e.*, the alignment is at least 80% of both the predicted model and its homolog). Lastly, 16.7% (1,418) of the predicted gene products show sequence similarity only to proteins with 'unknown' or 'hypothetical' annotations. Their existence corroborates the presence of these functionally ambiguous proteins in both white rot fungi and other genomes, but provides no additional information. The taxonomic distribution of gene model homologies is summarized in **Figure 1**.

Typical of shotgun sequencing of eukaryotes, extended repeats, telomeres and rRNA clusters were excluded from the assembly. Nevertheless, substantial numbers of noncoding repetitive sequences and putative mobile elements were assembled. Short repeats (<3 kbp) not clearly associated with transposons varied in copy number from >40 (GenBank no. Z31724) to 4 (GenBank nos. AF134289–AF134291). Several putative transposase-encoding sequences resemble class II transposons of ascomycetous fungi such as *Aspergillus niger Ant*, *Cochiobolus carbonum Fot1*, *Nectria* "Restless," *Fusarium oxysporum Tfo1* and *Cryphonectria parasitica Crypt1* (for review see refs. 14, 15). Additional transposase-encoding sequences included *EN/Spm*- and *TNP*-like elements (gx.25.15.1; pc.90.8.1) that are common in higher plants (pfam02992) but hitherto unknown in fungi[14,15]. Fungal class II elements often exceed 50–100 copies per genome, but the corresponding *P. chrysosporium* transposases are each represented by only 1–4 copies.

Numerous multicopy retrotransposons were identified, and several seem likely to affect expression of genes related to lignin degradation (**Supplementary Table 1** online). Most of these elements appear truncated and/or rearranged, and the long terminal repeats (LTRs), typical of retroelements, often lie apart as "solo LTRs"[16,17]. Several non-LTR retrotransposons, similar to other fungal long interdispersed nuclear elements (LINEs)-like retroelements, were also identified. *Copia*-like retroelements are particularly abundant, and one such element interrupts a putative member of the cytochrome P450 gene family within the seventh exon (**Supplementary Table 2** online). A similar situation was observed for a multicopper oxidase gene *mco3*, where a Skippy-like *gypsy* retroelement has been inserted within the twelfth intron. Intact coding regions flanked these inserts suggesting recent transpositions and/or splicing of the elements. Another *gyspy*-like element is inserted 100 bp upstream of a 'hybrid' peroxidase gene pc.91.32.1 (**Supplementary Table 1** online). The occurrence of intact transposons and other highly conserved repetitive elements is in marked contrast to the recently sequenced *N. crassa* genome, where repeat-induced point mutations (RIP) have greatly reduced the frequency of repeats greater than 400 bp[6–8].

Table 1 General features of the *P. chrysosporium* genome

Assembly size	29.9 Mb
GC content overall	57%
GC content (coding)	59%
Protein coding genes	11,777
tRNAs	200
% coding	45%
Average intergenic distance	1,339 bp
Intron size (average)	117 bp
Intron size (mode)	54 bp
Exon size (average)	232 bp
Exon size (mode)	89 bp

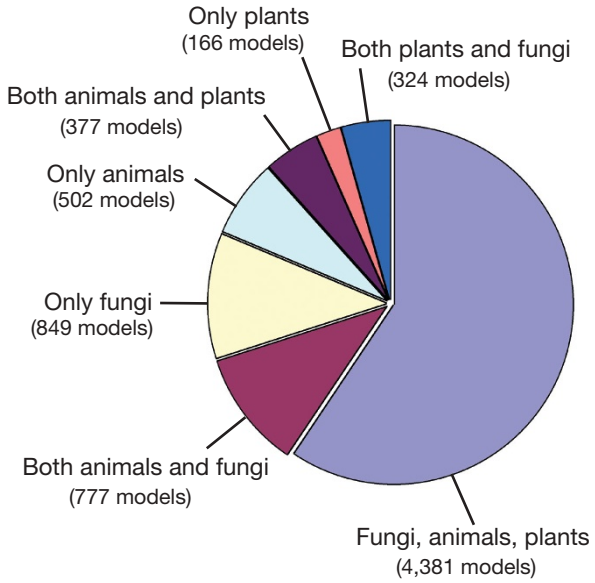

Figure 1 Taxonomic distribution of gene models that correspond to Smith-Waterman double-affine alignments with BLOSUM62 scores >100. There are 7,336 total gene models within this minimum score.

Protein families and domains in *P. chrysosporium*

InterPro[12] identification of conserved domains among predicted genes of *P. chrysosporium* provides an overview of the coding capability of this filamentous fungus (**Fig. 2**; **Supplementary Table 3** online). A large expansion in the InterPro category corresponding to the cytochrome P450 superfamily may reflect the complexity of metabolizing lignin derivatives and related aromatic compounds (see below). Also consistent with efficient depolymerization and degradation of lignin is the relatively high number of putative glucose methanol choline reductases (IPR000172). This family includes extracellular alcohol oxidases and cellobiose dehydrogenase[18], enzymes directly involved in lignocellulose degradation (below). Short chain dehydrogenase/reductases (IPR002198), aspartyl proteases (IPR001461) and Ras small GTPase (IPR003579) domains are more abundant in filamentous fungi (*P. chrysosporium*, *N. crassa*) relative to the sequenced yeasts (*S. cerevisiae*, *S. pombe*). The filamentous fungal genomes are considerably larger and encode more genes than yeast. Increased size and complexity of filamentous fungal genomes are likely due, in part, to their hyphal morphology, elongation and penetration of complex substrata.

Degradation of lignin and related aromatic compounds

White rot fungi catalyze the initial depolymerization of lignin by secreting an array of oxidases and peroxidases that generate highly reactive and nonspecific free radicals, which in turn undergo a

complex series of spontaneous cleavage reactions[19]. Major components of the *P. chrysosporium* lignin depolymerization system include multiple isozymes of lignin peroxidase (LiP) and manganese-dependent peroxidase (MnP). Consistent with previous studies, ten *lip* genes were identified. The sequence of scaffold 85 confirms and extends earlier genetic data[9,20] providing a detailed view of the principal lignin peroxidase gene cluster. Genes encoding *lipC* and *lipI* lie 45 kb apart within this cluster, which is in good agreement with the observed 1% recombination between these genes[21] and the estimate of one crossover per 60 kb inferred from RFLP mapping[9].

Additional analyses exposed five *mnp* genes, including two previously unknown members of this gene family (**Supplementary Table 1** online). One of these was designated *mnp5*; its predicted peptide corresponds to the N-terminal sequence of a peroxidase partially purified from colonized wood pulp (GenBank no. A61147). Unexpectedly, the other new MnP gene (*mnp4*) was located only 5.7 kb from the well-characterized gene, *mnp1*. The predicted proteins of *mnp4* and *mnp1* are nearly identical, with a single amino acid substitution. Clustering of *mnp* genes has been observed in other white rot fungi but not in *P. chrysosporium*. An interesting peroxidase gene, model pc.91.32.1, is unlinked to all peroxidases, but shares residues common to both *mnps* and *lips*. The pc.91.32.1 sequence is most closely related to the 'hybrid' peroxidase of *Pleurotus eryngii* (GenBank no. AF007224), but not all catalytic and manganese-binding residues are conserved[22].

LiP and MnP require extracellular H_2O_2 for their *in vivo* catalytic activity, and one likely source is the copper radical oxidase, glyoxal oxidase (GLOX)[23–25]. In addition to *glx*, the genome sequence reveals at least six other sequences predicted to encode copper radical oxidases (*cro1* through *cro6*). Moreover, three highly homologous genes, designated *cro3, cro4* and *cro5,* were uncovered within the lignin peroxidase gene cluster on scaffold 85. The position of these new genes strongly suggests a relationship between genomic organization and the proposed dependency[25] between lignin peroxidases and copper radical oxidases. Additional GLOX-like sequences were detected on scaffolds 46, 72 and 120, apparently unlinked to any peroxidases (**Supplementary Table 1** online).

Beyond copper radical oxidases, extracellular FAD-dependent oxidases are likely candidates for generating H_2O_2, but such genes had not been previously characterized in *P. chrysosporium*. Genes encoding FAD oxidases in related white rot fungi include aryl alcohol oxidases (AAO) of *P. eryngii* (GenBank nos. AF064069, AF143814) and a pyranose oxidase from *Coriolus versicolor* (GenBank no. D73369). Until now, the only extracellular FAD oxidase sequence known for *P. chrysosporium* was within the oxidoreductase domain of cellobiose dehydrogenase, an enzyme implicated in both cellulose and lignin degradation[18]. Nevertheless, at least four distinct AAO-like sequences, a pyranose oxidase-like sequence and a glucose oxidase-like sequence have been identified in the genome data. The precise roles and interactions of these genes in lignin degradation remains to be determined[26], but when viewed together with the copper radical oxidase genes, it is clear that *P. chrysosporium* possesses an impressive array of genes encoding extracellular oxidative enzymes.

In addition to extracellular peroxidase systems, laccases had been implicated in lignin degradation. These blue copper oxidases catalyze one-electron oxidation of phenolics, aromatic amines and other electron-rich substrates with the concomitant reduction of O_2 to H_2O[27]. Genome searches revealed no conventional laccases. Instead, four multicopper oxidase (MCO) sequences are found clustered within a 25-kb segment on scaffold 56. Signal P identified a putative secretion signal in gene *mco1*, and subsequent analysis has shown that it encodes a ferroxidase-like protein[28]. Thus it appears that *P. chrysosporium* does not have the capacity to produce laccases although distantly related multicopper oxidases may have a role in extracellular oxidations.

Degradation of cellulose and hemicellulose

In addition to lignin, *P. chrysosporium* completely degrades all major components of plant cell walls including cellulose and hemicellulose. The genome harbors the genetic information to encode more than 240 putative carbohydrate-active enzymes (ref. 29; http://afmb.cnrs-mrs.fr/CAZY/) including 166 glycoside hydrolases, 14 carbohydrate esterases and 57 glycosyltransferases, comprising at least 69 distinct families (**Supplementary Table 1** online). A global correlation between the number of carbohydrate active enzymes and the total number of open reading frames in bacterial and eukaryotic genomes has been observed[30]. The number of glycoside hydrolases and glycosyltransferases predicted in the *P. chrysosporium* genome matches this correlation accurately. In other sequenced eukaryotic genomes (*S. cerevisiae*, *S. pombe, Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Homo sapiens*), the overall number of glycosyltransferases exceeds, sometimes by a large factor, that of the glycoside hydrolases (B.H., unpublished observation), consistent with the greater relative importance of constructing rather than breaking down polysaccharides. In *P. chrysosporium*, the situation is reversed, pointing both to the development of a large repertoire of glycosidases (in accord with its lifestyle) and to a comparatively smaller catalog of glycosyltransferases (57 in *P. chrysosporium*, 61 in *S. cerevisiae*, 64 in *S. pombe*, over 140 in *D. melanogaster*, over 230 in *C. elegans* and *H. sapiens*, and over 410 in *A. thaliana*).

As in the case of the extracellular oxidases (above), many of the glycoside hydrolases appear within large families of closely related genes that may encode redundant (or partially overlapping) functions. Among the cellulases, endoglucanases are thought to hydrolyze cellulose at random positions in less crystalline regions whereas exocellobiohydrolases act processively on the chains to liberate the disaccharide cellobiose. Lastly, β-glucosidases (cellobiases) cleave cellobiose to yield glucose. The complement of cellulases in *P. chrysosporium* includes at

Table 2 Sequence summary by library

Library	Insert size[a]	Reads[b]	Trimmed	Weight (%)	Avg. trim length	Passing rate	Passing clones	Total seq. (Mb)	Fold seq. cover (x)	Clone cover (Mb)	Fold clone cover (X)
LQI	3,500.0 ± 300.0	313,201	222,926	35.48%	465.69	71.20%	139,403	103.81	3.46	487.91	16.26
YAI	3,500.0 ± 300.0	370,664	219,996	32.52%	432.52	59.40%	171,112	95.15	3.17	598.89	19.96
YAH	3,500.0 ± 350.0	164,707	123,512	22.38%	530.23	75.00%	76,575	65.49	2.18	268.01	8.93
YAE	3,500.0 ± 300.0	72,137	53,369	9.63%	527.99	74.00%	33,885	28.18	0.94	118.6	3.95
Total		**920,709**	**619,803**	**100.00%**	**472.14**	**67.35%**	**420,975**	**292.63**	**9.75**	**1,473.41**	**49.1**

[a]Insert size was estimated based on agarose gel size selection targets. [b]Reads were screened for vector and quality trimmed. Only reads with more than 100 bp remaining after filtering were considered passing. Clones that had both forward and reverse sequences that passed the above screening are reported in 'Passing clones.' Fold and clone coverage was calculated using a genome size of 30 Mb. The percentage of reads from each library that sum to the total 'Trimmed' reads were used to calculate the final 'Weighted Average Trim Length' of 472.14 nucleotides. This information does not include the cosmid end-sequences that were used for assembly quality verification.

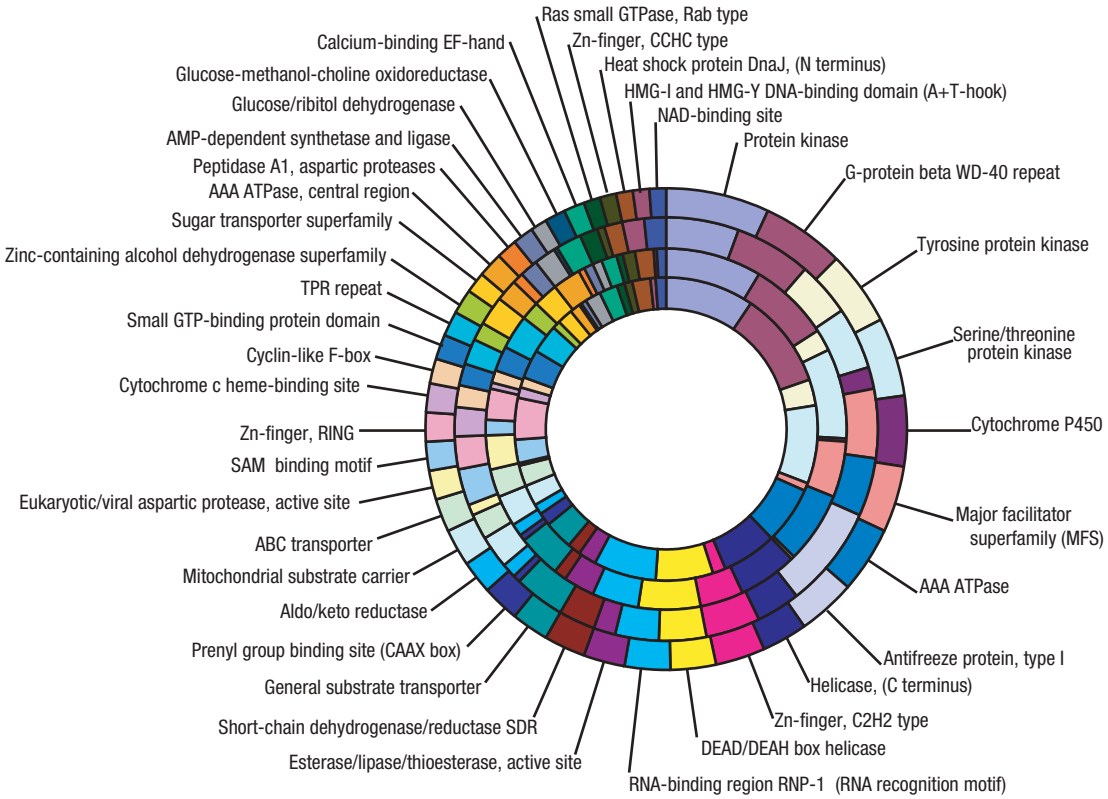

Figure 2 Schematic showing the InterPro version 5.3 (ref. 12) domains in *P. chrysosporium* (outer ring), *N. crassa*, *S. pombe* and *S. cerevisiae* (inner ring). InterPro contains the combined protein signature databases of Swiss-prot, TrEMBL, PROSITE, PRINTS, Pfam, ProDom, SMART and TIGRFAM. Note the large difference between the three organisms in the size of bands corresponding to the cytochrome P450 domain (IPR001128) indicating the expansion of this gene family in *P. chrysosporium*. Type I Antifreeze domains (IPR000104) are artifacts attributed to a common tripeptide repeat, Ala-Ala-Thr. The number of sequences per domain is listed separately (**Supplementary Table 3** online).

least 40 putative endoglucanases, seven exocellobiohydrolases, and at least nine β-glucosidases (**Supplementary Table 1** online). The list of putative endoglucanases comprises five glycoside hydrolase families (GH5, GH9, GH12, GH61 and GH74). Scanning the genome for exocellobiohydrolase genes revealed only those that had been previously described, encompassing six members of family GH7 (CBH1 isozymes) and a single member of type GH6 (CBH2). Multiple β-glucosidase genes that code for two enzymes in family GH1 and seven members of GH3 were also found. Unlike the genes for ligninolytic enzymes, the cellulase genes of *P. chrysosporium* do not appear to be tightly clustered in the genome, with the noted exception of the previously known case of neighboring *cel7A* and *cel7B* genes[31].

In addition to the enzymes responsible for hydrolysis of cellulose, numerous other polysaccharide-degrading enzymes are predicted in the *P. chrysosporium* genome, including catalytic activities for degradation of hemicellulose, starch and glycogen, mutan, chitin and β-glucans (**Supplementary Note** and **Supplementary Table 1** online).

Secondary metabolism

Examination of the genome suggests potential for production of an array of biologically active compounds (**Supplementary Note** online). Among these are numerous putative polyketide synthases and nonribosomal peptide synthases (**Supplementary Table 4** online). In addition, a minimum of 148 cytochrome P450 sequences representing 12 cytochrome P450 families[32] were observed (**Supplementary Table 2** and **Supplementary Fig. 1** online).

Mating type loci

Two multigenic mating type loci, A and B, were identified in the *P. chrysosporium* genome. The Aα locus is similar to orthologs in other homobasidiomycetes[33–35] in that it features two divergently transcribed homeodomain-encoding genes immediately adjacent to the gene encoding a mitochondrial intermediate peptidase (*mip*) on scaffold 7. In addition, five sequences were identified with similarity to the pheromone receptor genes of mating type B loci. Three receptors are clustered within a 12-kb region on scaffold 110, and these are most similar to *Coprinus cinereus rcb2* and *rcb3*[36,37]. Members of the same receptor family were also identified on two separate scaffolds (255 and 88), both of which were similar to *C. cinereus rcb3*. Mating type loci are typically associated with fleshy fruiting bodies and believed to govern the fusion of compatible homokaryons, the migration of nuclei, and the formation of a morphological structure known as the clamp connection. *P. chrysosporium* does not form clamp connections and the sexual basidiospores are formed in a simple, resupinate layer on the substrate. Thus, the conservation of mating type genes in *P. chrysosporium* was unexpected.

DISCUSSION

It has been proposed that colonization of land by eukaryotes was facilitated by a symbiotic partnership between a photosynthetic organism and a fungus, and that this relationship had far-reaching effects on climate and atmosphere, possibly contributing to the rapid evolution of animals in the Precambrian[38]. The oldest fossil evidence for land

plants and fungi suggests their appearance 480–460 million years ago; however, molecular clock estimates indicate an earlier colonization of land at about 600 million years ago. Divergence of the major fungal taxa Basidiomycotina and Ascomycotina occurred during the Paleozoic period (~550 million years ago)[11]. The ascomycetes include the sequenced yeasts *S. cerevisiae* and *S. pombe*, and the filamentous fungus *N. crassa*. Draft genomes of other filamentous fungi such as *Aspergillus nidulans, C. cinereus, Ustilago maydis, Fusarium graminearum* and *Magnaporthe grisea* are in various stages of completion.

Our objective was to generate high quality draft coverage of the genomic DNA sequence for the white rot fungus *P. chrysosporium*, a filamentous basidiomycete. In addition to the lignin-degrading white rot fungi, basidiomycetes include commercially important species (mushrooms), mycorrhizae, as well as pathogens of plants (smuts, rusts) and animals. Providing further insight into this large and diverse phylum, the genomes of the human pathogen, *Filobasidiella neoformans* (= *Cryptococcus neoformans*) (http://www-sequence.stanford. edu/group/C.neoformans/index.html), the maize pathogen *Ustilago maydis* (http://www-genome.wi.mit.edu/seq/fgi/candidates.html), soybean pathogens *Phakospora pachyrhizi* and *Phakospora meibomiae* (J. Boore, Joint Genome Institute, personal communication) and an inky cap mushroom *C. cinereus* (http://www-genome.wi.mit.edu/seq/ fgi/candidates.html), will soon be available. Comparative analyses of these genomes will undoubtedly provide valuable information about the genetic features that distinguish pathogenic and saprophytic lifestyles among the basidiomycetes. In addition, comprehensive comparisons with the forthcoming genome data from other fungal taxa may yield important clues about their origins and influences on other organisms in the tree of life.

Fungi are the only eukaryotic organisms that digest their food components extracellularly through a process that typically involves the secretion of degradative and/or oxidative enzymes before absorption of the nutrients. As the primary degraders of lignin in wood, white rot fungi play an important role in the global carbon cycle. *P. chrysosporium* has been the most intensively studied species and is capable of efficient degradation of all wood components through the production of an array of oxidative and hydrolytic enzymes. Extensive genetic diversity was observed within complex gene families encoding peroxidases, oxidases, glycosyl hydrolases and cytochrome P450s. Reasons for multiple genes in the white rot fungus are largely unknown, but their presence may reflect that multiple specificities are essential for the effective hydrolysis of these complex wood polymers. The redundant activities might suggest a need for biochemical diversity for optimum growth under varying environmental conditions (*e.g.*, temperature, pH, ionic strength), or simply that similar but not identical enzyme functions are necessary to effectively break down these complex carbohydrate polymers whose structure, physical state and accessibility vary widely upon the botanical source or upon the extent of decay. The occurrence of *P. chrysosporium* gene families with closely related members is in clear contrast to the genome of *N. crassa* in which the repeat-induced point mutation system appears to have restricted the sizes of these gene families.

With the *P. chrysosporium* genome in hand, we are poised to achieve a deeper understanding of the processes by which white rot fungi colonize wood, interact with other organisms in their ecosystem and perform a vital function in the carbon cycle. Large-scale processes for the hydrolysis of plant cell-wall polysaccharides may one day expand the utility of plant biomass for fuels and biochemicals through industrial fermentation. Modification or degradation of specific carbohydrate components in wood is especially attractive to the textile, fuel and paper industries. Determining the choreography of expression and secretion of the oxidative and hydrolytic enzymes and their individual and collective contributions in the breakdown of lignocellulose will be greatly facilitated by the availability of the white rot fungus genome data.

METHODS

Genome sequencing. Genomic DNA was purified from a homokaryotic strain RP-78 (ref. 39; available from USDA Forest Mycology Center) and was sheared to give an approximate fragment size of 3 kbp. The DNA fragments were blunt-end repaired and then size selected on an agarose gel. Four principal small insert (3–4 kbp) libraries were generated by blunt-end ligation into pUC18. These libraries were end-sequenced with dye terminator chemistry using standard m13-40 and m13-28 primers[40]. Details of the libraries generated are presented in **Table 2**. Approximately 15% of the sequence was mitochondrial contaminant, which (after initial detection as a large 115 kb circular contig) was assembled separately to produce a finished mitochondrial sequence, which will be analyzed in depth elsewhere (J.C., N.P., D.R. unpublished data). After vector, mitochondrial and quality screening, 619,803 end sequences, representing approximately 9.75 coverage, were produced.

Genome assembly. These paired sequence fragments were assembled using the JAZZ suite of assembly tools (ref. 41; J.C., N.P., D.R. unpublished data), yielding a high quality draft assembly. Ninety percent of the assembled sequence was reconstructed in 165 scaffolds (414 contigs); half of the assembled sequence was arranged in 45 scaffolds longer than 203 kbp and 109 contigs longer than 79 kbp. The fidelity of the assembly is supported by the high degree (96%) of plasmid-end pairs preserved in contigs and scaffolds, as well as the ends of a sampling of cosmid ends. The net length of assembled contigs totaled 29.9 Mbp and included substantial numbers of repetitive elements. Completeness of the assembly with regard to coding region was supported by analysis of 1,390 unique ESTs derived from colonized wood. Approximately 98% of these ESTs as well as all 39 previously known *P. chrysosporium* genes in GenBank were recovered. Part of this group was a cluster of linked genes that notably includes tandemly repeated copies of lignin peroxidases, which we would have expected to be challenging to assemble. Based on an analysis of high confidence gene models (see below) the rate of short inserts and deletions in coding sequence is expected to be less than one per sixteen kilobases, providing an excellent substrate for annotation.

To further corroborate the long-range structure of the assembled genome we end-sequenced 888 clones from a pWE15 cosmid library with average inserts of 40Kb[10,31] (1.5 clone coverage). This analysis confirmed that 785 (88%) of the end-sequences from clones successfully sequenced at both ends were found in opposing orientations within 3 s.d. of each other or split between scaffolds with each read pointing outward from the end of a scaffold, consistent with new linking information not found in the short insert data. In the absence of a large-scale physical or genetic mapping effort (as is available for many model systems), we used genetic methods[20] to test the assembly on the longest length scales. Ten pairs of polymorphic markers were identified from opposite ends of long scaffolds generated by an earlier assembly (**Supplementary Table 5** online). Scoring of recombinant progeny confirmed that these markers were indeed linked as predicted.

Gene identification. The genome assembly was compared to all known proteins in GenBank (release 131) at low stringency using ungapped BLASTX[42] (Blosum62, score > 30), with significant hits indicating potential exons. Alignments were parsed to derive one or more 'optimal' colinear set of hits for each protein (S. Rash, Joint Genome Institute, personal communication) and the best-scoring putative homolog was submitted with the surrounding genomic sequence to Genewise[43], which predicts gene structure based on homology, recognizing splicing signals and the potential for short insertions and deletions that occur in draft sequence.

The resulting homology-based gene models exhibit a low rate of single-base insertions and deletions (395 models out of 9,638 (4%), or one per 16,000 bases of coding sequence) corroborating the accuracy of the assembled draft sequence. These models were often incomplete, though they appeared to possess accurate intron-exon boundaries compared with known *P. chrysosporium* genes. To obtain a more complete set of gene models from this phylogenetically

distant fungus, the homology-derived transcripts were submitted as "faux" ESTs (along with 1,134 real ESTs) to GrailEXP[13], an *ab initio* gene finder that incorporates expressed sequence information and has tunable modules for the detection of splice boundaries and other sequence features that could be bootstrapped from the homology models. The outputs of GeneWise and GrailEXP were post-processed to select the best gene model at each locus. To identify tRNAs we used tRNAscan-SE version 1.23 (ref. 44). Accuracy of modeling was assessed as described previously[45,46] for 19 full-length cDNA sequences. The correlation coefficient, sensitivity and specificity, all measures of predictive accuracy, were 0.73, 0.75 and 0.96, respectively (**Supplementary Table 6** online). Manually curated models appearing on the browser are designated with the prefix "ug" (user gene).

Genome data availability. The annotated genome is available on an interactive web portal at http://www.jgi.doe.gov/whiterot/. The genome sequence has been deposited and assigned GenBank accession number AADS00000000.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was performed under the auspices of the US Department of Energy by (i) the University of California, Lawrence Livermore National Laboratory under contracts No. W-7405-Eng-48 and AC03-76SF00098, (ii) the University of California, Los Alamos National Laboratory under contract No. W-7405-ENG-36, (iii) UT-Battelle, Oak Ridge National Laboratory under contract DE-AC05-00OR22725 and (iv) the University of Wisconsin under grant no. DE-FG02-87ER13712. L.F.L. is supported by Fundacion Andes Predoctoral Fellowship, FONDECYT-Chile grant no. 20000076 and Millennium Institute for Fundamental and Applied Biology.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 24 December 2003; accepted 16 March 2004
Published online at http://www.nature.com/naturebiotechnology/

1. Blanchette, R. Delignification by wood-decay fungi. *Ann. Rev. Phytopath.* **29**, 381–398 (1991).
2. Eriksson, K.-E.L., Blanchette, R.A. & Ander, P. Microbial and enzymatic degradation of wood and wood components. (Springer-Verlag, Berlin, 1990).
3. Burdsall, H.H. & Eslyn, W.E. A new *Phanerochaete* with a *chrysosporium* imperfect state. *Mycotaxon* **1**, 123–133 (1974).
4. Goffeau, A. *et al.* Life with 6,000 genes. *Science* **274**, 546, 563–547 (1996).
5. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
6. Galagan, J.E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868 (2003).
7. Mannhaupt, G. *et al.* What's in the genome of a filamentous fungus? Analysis of the *Neurospora genome* sequence. *Nucleic Acids Res.* **31**, 1944–1954 (2003).
8. Schulte, U., Becker, I., Mewes, H.W. & Mannhaupt, G. Large scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol.* **94**, 3–13 (2002).
9. Raeder, U., Thompson, W. & Broda, P. RFLP-based genetic map of *Phanerochaete chrysosporium* ME446: lignin peroxidase genes occur in clusters. *Mol. Microbiol.* **3**, 911–918 (1989).
10. Gaskell, J., Dieperink, E. & Cullen, D. Genomic organization of lignin peroxidase genes of *Phanerochaete chrysosporium*. *Nucleic Acids Res.* **19**, 599–603 (1991).
11. Berbee, M. & Taylor, J. in *Systematics and Evolution*, vol. VIIB. (eds. McLaughlin, D., McLaughlin, E. & Lemke, P.) 229–246 (Springer, Berlin, 2001).
12. Mulder, N.J. *et al.* The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318 (2003).
13. Xu, Y. & Uberbacher, E.C. *J. Comput. Biol.* **4**, 325–338 (1997).
14. Kempken, F. & Kuck, U. Transposons in filamentous fungi—facts and perspectives. *BioEssays* **20**, 652–659 (1998).
15. Wostemeyer, J. & Kreibich, A. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. *Curr. Genet.* **41**, 189–198 (2002).
16. Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A. & Voytas, D.F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8**, 464–478 (1998).
17. Goodwin, T.J. & Poulter, R.T. Multiple LTR-retrotransposon families in the asexual yeast *Candida albicans*. *Genome Res.* **10**, 174–191 (2000).
18. Henriksson, G., Johansson, G. & Pettersson, G. A critical review of cellobiose dehydrogenases. *J. Biotechnol.* **78**, 93–113 (2000).
19. Kirk, T.K. & Farrell, R.L. Enzymatic "combustion": the microbial degradation of lignin. *Ann. Rev. Microbiol.* **41**, 465–505 (1987).
20. Gaskell, J. *et al.* 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 (1994).
21. Stewart, P. & Cullen, D. Organization and differential regulation of a cluster of lignin peroxidase genes of *Phanerochaete chrysosporium*. *J. Bact.* **181**, 3427–3432 (1999).
22. Martinez, A.T. Molecular biology and structure-function of lignin-degrading heme peroxidases. *Enzyme Microb. Technol.* **30**, 425–444 (2002).
23. Kersten, P.J. & Kirk, T.K. Involvement of a new enzyme, glyoxal oxidase, in extracellular H_2O_2 production by *Phanerochaete chrysosporium*. *J. Bacteriol.* **169**, 2195–2201 (1987).
24. Whittaker, M.M., Kersten, P.J., Cullen, D. & Whittaker, J.W. Identification of catalytic residues in glyoxal oxidase by targeted mutagenesis. *J. Biol. Chem.* **274**, 36226–36232 (1999).
25. Kersten, P.J. Glyoxal oxidase of *Phanerochaete chrysosporium*; its characterization and activation by lignin peroxidase. *Proc. Natl. Acad. Sci. USA* **87**, 2936–2940 (1990).
26. Ander, P. & Marzullo, L. Sugar oxidoreductases and veratryl alcohol oxidase as related to lignin degradation. *J. Biotechnol.* **53**, 115–131 (1997).
27. Thurston, C.F. The structure and function of fungal laccases. *Microbiol.* **140**, 19–26 (1994).
28. Lorrondo, L., Salas, L., Melo, F., Vicuna, R. & Cullen, D. A novel extracellular multicopper oxidase from *Phanerochaete chrysosporium* with ferroxidase activity. *Appl. Environ. Microbiol.* **69**, 6257–6263 (2003).
29. Henrissat, B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **280** (Pt 2), 309–316 (1991).
30. Coutinho, P.M., Stam, M., Blanc, E. & Henrissat, B. Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends Plant Sci.* **8**, 563–565 (2003).
31. Covert, S., Vanden Wymelenberg, A. & Cullen, D. Structure, organization and transcription of a cellobiohydrolase gene cluster from *Phanerochaete chrysosporium*. *Appl. Environ. Microbiol.* **58**, 2168–2175 (1992b).
32. Nelson, D.R. *et al.* P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6**, 1–42 (1996).
33. Kues, U., James, T.Y., Vilgalys, R. & Challen, M.P. The chromosomal region containing *pab-1, mip*, and the A mating type locus of the secondarily homothallic homobasidiomycete *Coprinus bilanatus*. *Curr. Genet.* **39**, 16–24 (2001).
34. Isaya, G. *et al.* Mammalian mitochondrial intermediate peptidase: structure/function analysis of a new homologue from *Schizophyllum commune* and relationship to thimet oligopeptidases. *Genomics* **28**, 450–461 (1995).
35. Casselton, L.A. *et al.* Mating type control of sexual development in *Coprinus cinereus*. *Can. J. Bot.* **73**, S266–S272 (1995).
36. Halsall, J.R., Milner, M.J. & Casselton, L.A. Three subfamilies of pheromone and receptor genes generate multiple B mating specificities in the mushroom *Coprinus cinereus*. *Genetics* **154**, 1115–1123 (2000).
37. O'Shea, S.F. *et al.* A large pheromone and receptor gene complex determines multiple B mating type specificities in *Coprinus cinereus*. *Genetics* **148**, 1081–1090 (1998).
38. Heckman, D.S. *et al.* Molecular evidence for the early colonization of land by fungi and plants. *Science* **293**, 1129–1133 (2001).
39. Stewart, P., Gaskell, J. & Cullen, D. A homokaryotic derivative of a *Phanerochaete chrysosporium* strain and its use in genomic analysis of repetitive elements. *Appl. Environ. Microbiol.* **66**, 1629–1633 (2000).
40. Detter, J.C. *et al.* Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80**, 691–698 (2002).
41. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301–1310 (2002).
42. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
43. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
44. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
45. Burset, M. & Guigo, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
46. Guigo, R., Agarwal, P., Abril, J.F., Burset, M. & Fickett, J.W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).

Erratum: Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78

Diego Martinez, Luis F Larrondo, Nik Putnam, Maarten D Sollewijn Gelpke, Katherine Huang, Jarrod Chapman, Kevin G Helfenbein, Preethi Ramaiya, J Chris Detter, Frank Larimer, Pedro M Coutinho, Bernard Henrissat, Randy Berka, Dan Cullen & Daniel Rokhsar
***Nat. Biotechnol.* 22, 695–700 (2004)**

The credit on the contents page for the cover image of the June issue read "Photo courtesy of T. Kuster (USDA Forest Products Laboratory, Madison, WI)." The credit should have included L. Hornick (US DoE Joint Genome Institute, Walnut Creek, CA).

Corrigendum: Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78

Diego Martinez, Luis F Larrondo, Nik Putnam, Maarten D Sollewijn Gelpke, Katherine Huang, Jarrod Chapman, Kevin G Helfenbein, Preethi Ramaiya, J Chris Detter, Frank Larimer, Pedro M Coutinho, Bernard Henrissat, Randy Berka, Dan Cullen & Daniel Rokhsar
***Nat. Biotechnol.* 22, 695–700 (2004)**

On page 695, column 1, line 4, the phrase "hemicellulose and cellulose (the most abundant natural polymers)" should have read "cellulose (the most abundant natural polymer) and hemicellulose"; column 2, line 4, "*P. chrysosporium*, a representative fungus" should have read "*P. chrysosporium*, a representative white rot fungus."

On page 699, column 1, paragraph 3, last sentence, the phrase "system appears to have restricted" should have had a reference, so that it read, "system appears[6] to have restricted."

Corrigendum: Intellectual property—the dispute between research institutions and voluntary health agencies

Lichtman, MA, Hunter, MD & Liders, GJ
***Nat. Biotechnol.* 22, 385–386 (2004)**

In Table 1, the sixth entry under "Agency" should have read "National Multiple Sclerosis Society (New York, NY, USA)," not "National Multiple Sclerosis Foundation." The agency listed exists but it is not the one funding the research.