Genetic Bases of Fungal White Rot Wood Decay Predicted by Phylogenomic Analysis of Correlated Gene-Phenotype Evolution

László G. Nagy,*[1] Robert Riley,[2] Philip J. Bergmann,[3] Krisztina Krizsán,[1] Francis M. Martin,[4] Igor V. Grigoriev,[2] Dan Cullen,[5] and David S. Hibbett[3]

[1]Synthetic and Systems Biology Unit, Institute of Biochemistry, BRC-HAS, Szeged, Hungary

[2]US Department of Energy (DOE) Joint Genome Institute, Walnut Creek, CA

[3]Biology Department, Clark University, Worcester, MA

[4]Institut National de la Recherche Agronomique, Unité Mixte de Recherche 1136 Université Henri Poincaré, Interactions Arbres/Micro-organismes, 02854 Champenoux, France

[5]US Department of Agriculture (USDA) Forest Products Laboratory, Madison, WI

*Corresponding author: E-mail: lnagy@brc.hu

Associate editor: Iñaki Ruiz-Trillo

Abstract

Fungal decomposition of plant cell walls (PCW) is a complex process that has diverse industrial applications and huge impacts on the carbon cycle. White rot (WR) is a powerful mode of PCW decay in which lignin and carbohydrates are both degraded. Mechanistic studies of decay coupled with comparative genomic analyses have provided clues to the enzymatic components of WR systems and their evolutionary origins, but the complete suite of genes necessary for WR remains undetermined. Here, we use phylogenomic comparative methods, which we validate through simulations, to identify shifts in gene family diversification rates that are correlated with evolution of WR, using data from 62 fungal genomes. We detected 409 gene families that appear to be evolutionarily correlated with WR. The identified gene families encode well-characterized decay enzymes, e.g., fungal class II peroxidases and cellobiohydrolases, and enzymes involved in import and detoxification pathways, as well as 73 gene families that have no functional annotation. About 310 of the 409 identified gene families are present in the genome of the model WR fungus *Phanerochaete chrysosporium* and 192 of these (62%) have been shown to be upregulated under ligninolytic culture conditions, which corroborates the phylogeny-based functional inferences. These results illuminate the complexity of WR and suggest that its evolution has involved a general elaboration of the decay apparatus, including numerous gene families with as-yet unknown exact functions.

Key words: **comparative genomics, bioinformatics, protein of unknown function, wood-decay, fungal enzymes**

Article

Fast Track

Introduction

Plant cell walls (PCW) are complex mixtures of biopolymers, including cellulose, lignin, and hemicellulose. Enzymatic degradation of PCW is a hallmark trait of fungi and has enormous potential for industrial applications such as biofuel production (Martinez et al. 2005; Hofrichter et al. 2010; Horn et al. 2012; Grigoriev et al. 2014; Rytioja et al. 2014). Wood (secondary xylem) is enriched in lignin and may also contain highly variable secondary metabolite extractives. The major decomposers of wood are Agaricomycetes (mushroom-forming fungi), which manifest two main modes of decay, white rot (WR) and brown rot (BR), although species with intermediate or unclassified nutritional strategies exist (e.g., *Schizophyllum commune*) (Floudas et al. 2015). In WR, all components of PCW are degraded, including the recalcitrant lignin fraction, whereas in BR lignin is modified but remains largely intact in decayed residues (Martinez et al. 2005; Floudas et al. 2012). Agaricomycetes also include species with wood decay mechanisms that are intermediate between WR and BR (e.g., *Schizophyllum commune*) (Floudas et al. 2015), as well as soil and litter decomposers, ectomycorrhizal symbionts (ECM) and plant pathogens.

Phylogenomic analyses of known decay-related gene families have provided insight into the diversity and evolution of the decay apparatus in Agaricomycetes (Floudas et al. 2012; Riley et al. 2014; Nagy et al. 2016). The sister group of Agaricomycetes is a clade of BR fungi, the Dacrymycetes, and the ancestor of the Agaricomycetes appears to have been a saprotroph with a modest repertoire of PCW-degrading enzymes (Nagy et al. 2016). Gene families encoding enzymes that attack crystalline cellulose, including cellobiohydrolases (glycoside hydrolase families GH6 and GH7) and lytic polysaccharide monooxygenases (LPMO) and other carbohydrate-active enzymes (CAZys), diversified early in the evolution of Agaricomycetes, prior to the divergence of lineages

leading to Sebacinales and Cantharellales (Nagy et al. 2016). The best-known ligninolytic enzymes, fungal class II peroxidases (PODs), began to diversify around the time of divergence of Auriculariales, leading to the evolution of true WR fungi, capable of degrading both crystalline cellulose and lignin. Independent, and perhaps irreversible, origins of BR and ECM occurred in multiple lineages and were associated with parallel reductions in PCW-degrading enzymes (Floudas et al. 2012; Kohler et al. 2015). Molecular clock analyses suggest that the origin of WR occurred around 300 Ma, albeit with very broad confidence intervals on ages, which is consistent with the view that evolution of WR contributed to the Permo-Carboniferous decline in organic carbon sequestration. However, the possibility that the evolution of WR—specifically lignin degradation by PODs—affected coal deposition patterns has been questioned on the basis of paleobotanical and geological evidence (Hibbett et al. 2016; Nelsen et al. 2016).

The studies cited above have focused on a limited suite of enzymes that have been functionally characterized, with the ligninolytic PODs having received much of the attention. Cellobiohydrolases and LPMOs have also been focal enzymes, in part because of their industrial applications. However, PCW decay is a complex process and despite intense research in this area (Martinez et al. 2004, 2009; Eastwood et al. 2011; Vanden Wymelenberg et al. 2010) its enzymatic mechanisms are still incompletely known (Riley et al. 2014).

To identify the complete set of gene families that could have been important in the evolution of WR, we applied the COMPARE method (Nagy et al. 2014), which harnesses the power of phylogenetic correlations to make systematic, unbiased predictions of the genes that underlie trait evolution. Here, we extend this approach to identify gene families in which shifts in diversification rates are correlated with the gains and losses of traits. We evaluated the performance of the extended method using simulated gene family evolution, and then applied it to an empirical dataset of 62 genomes, focusing on Agaricomycetes with WR, BR, ECM, and other nutritional modes. Results confirmed the importance of enzymes with known roles in decay (PODs, GH6, GH7, LPMOs, etc), and other genes with known products, but also identified 73 gene families with no PFAM annotation that may also be important in WR.

Results and Discussion

Prediction of enzyme families involved in fungal wood-decomposition

To identify gene families potentially involved in WR, we assembled a dataset of 62 fungal genomes [supplementary table S1, Supplementary Material online, based on the species list from (Nagy et al. 2016)], including 22 species that primitively lack WR (15 Basidiomycota, five Ascomycota, one chytrid, and one zygomycete), 24 WR species of Agaricomycetes (representing a single origin of WR), and 16 BR and ECM species of Agricomycetes (which comprise eight derived lineages that have lost the ability to produce WR) (fig. 3; Floudas et al. 2012, 2015; Kohler et al. 2015). We used a species tree from our previous study (Nagy et al. 2016) and reconstructed gene family duplication/loss histories in each gene family using ortholog coding, which delimits "orthogroups" of genes with no more than one copy per species. We mapped orthogroups onto the species tree using Dollo parsimony optimization and compared gene duplications and losses to the reconstructed evolutionary history of WR.

We predicted that gene families encoding proteins involved in WR should exhibit background rates of duplication and loss in regions of the phylogenetic tree preceding the origin of WR, elevated duplication rates and reduced loss rates in WR lineages, and reduced duplication rates and increased loss rates in lineages that have secondarily lost WR. This pattern of diversification has been demonstrated based on genomic comparisons and gene tree-species tree reconciliations for key gene families which are known to function in WR, including PODs, and a number of GH families (Floudas et al. 2012), which provided benchmarks for our analysis. We analyzed the inferred gene duplication/loss patterns of each gene family against this model using a permutation ANOVA that we implemented to test for gene-phenotype co-gain and co-loss patterns and thus take phylogenetic history into account.

We found 3410 and 1606 families showing a significant correlation with WR at $P \leq 0.05$ and $P \leq 0.001$, respectively. We further screened the 3410 families for gene losses in secondarily nonWR clades as would be expected under a model of convergent loss of WR. We required the gene families to show gene losses in three or more of the secondarily nonWR lineages, resulting in a set of 409 families.

We also evaluated the 409 candidate gene families identified with COMPARE based on results of three published expression studies on the model WR species, *Phanerochaete chrysosporium* (Vanden Wymelenberg et al. 2010; Gaskell et al. 2014; Korripally et al. 2015), in which the fungus was grown in liquid cultures with wood as the sole carbon source or in glucose medium. The *P. chrysosporium* genome contains 310 of the 409 gene families that we predicted are functionally related to WR. Of these, 192 gene families (62%) have at least one gene copy that was significantly upregulated ($P \leq 0.05$) when wood was the sole carbon source in at least one of the published datasets (Vanden Wymelenberg et al. 2010; Gaskell et al. 2014; Korripally et al. 2015) (supplementary dataset S1, Supplementary Material online). On the other hand, we found 55 gene families containing at least one significantly downregulated *P. chrysosporium* gene ($P \leq 0.05$). Thus, functional predictions made with COMPARE based on the gene and organismal phylogenies and reconstructed patterns of trait evolution are supported by independent transcriptomic evidence.

Inferred Components of the WR Toolkit

Enzymes with Known or Suspected Lignocellulolytic Functions

We analyzed enrichment of gene ontology terms in the 409 detected gene families, relative to gene families that did not show a correlation with WR evolution. As expected, significant enrichment was found for terms related to polysaccharide

binding and metabolism and extracellular enzymatic pathways, as well as several enzyme activities including peroxidase, oxidoreductases, and hydrolase activities (supplementary table S3, Supplementary Material online). These corresponded to various peroxidases with known or suspected roles in lignin degradation, including PODs, heme-thiolate peroxidases, dye decolorizing peroxidases, and laccases, as well as diverse glycoside hydrolases (families GH5, GH6, GH7, GH10, GH12, GH28, GH43, GH76, and GH92), including cellulases, xylanases, endoglucanases, mannanases, oxidases and esterases, and LPMOs (supplementary dataset S1, Supplementary Material online). This suite of enzymes is consistent with the range of activities known to be required for the enzymatic attack of PCW components (Martinez et al. 2005; Ruiz-Duenas and Martinez 2009; Eastwood et al. 2011; Floudas et al. 2012) (Courty et al. 2009; Dashtban, et al. 2010). Remarkably, GH6, a cellobiohydrolase important in attack of crystalline cellulose, shows small copy-number changes (on average 1 to 0) across WR and nonWR species, yet its association with wood-decay could be detected ($P = 0.007$, fig. 3*c*).

Our analyses also recovered ($P < 0.001$) three gene families, PODs, DyPs, and Laccases, suggested as key players of lignin decomposition (Courty et al. 2009; Dashtban, et al. 2010). We also found 21 gene families containing cellulose-binding modules (CBM1, PF00734), which increase the affinity of enzymes for cellulosic substrates (Varnai et al. 2013; Riley et al. 2014). This represents a significant enrichment of CBM1 domains relative to gene families for which we did not detect a correlation with WR ($P = 5.59 \times 10^{-16}$, hypergeometric test, supplementary dataset S1, Supplementary Material online). Furthermore, we detected the gene family containing GLP1, a secreted glycoprotein that has been implicated in the production of reactive hydroxyl radicals during wood decay (Tanaka et al. 2007).

Cellular Detoxification and Import Pathways

We predict a role in wood decomposition for nine protein clusters of the Major Facilitator Superfamily MFS-1 (domain enrichment $P = 2.02 \times 10^{-5}$, hypergeometric test), six of which were also detected in expression studies (Eastwood et al. 2011; Olson et al. 2012; Korripally et al. 2015) and may be involved in transporting decomposition intermediates (e.g., sugars, lignin metabolites) into the cell. Components of intracellular antioxidant and detoxification pathways have also been detected and are enriched in the 409 gene families relative to other families, including eight clusters of the cytochrome P450 superfamily ($P = 2.62 \times 10^{-4}$, hypergeometric test) and three clusters of glutathione-S-transferases ($P = 5.33 \times 10^{-2}$, hypergeometric test), suggesting a role in the transformation of toxic compounds released during lignin degradation (Mathieu et al. 2013; Morel et al. 2009, 2013). In contrast, nonphylogenetic methods failed to find significant overrepresentation of glutathione-S-transferases in saprotrophic fungi as compared with ECM and parasitic ones (Morel et al. 2013), highlighting the power of phylogeny-based methods for predicting gene function.

Proteins of Unknown Function

Our analyses also predicted a role in wood decay for 73 gene families containing no known PFAM domains and 49 containing domains of unknown function (DUF). About 26 of the gene families that lack PFAM domains and 24 of the gene families with DUFs contain genes that were found to be significantly upregulated in the expression studies used for comparisons (Vanden Wymelenberg, et al. 2010; Gaskell et al. 2014; Korripally et al. 2015) (supplementary dataset S1, Supplementary Material online). Of these, the duplication-loss history of a conserved fungal gene family with hitherto unknown function is shown on figure 3*d* (DUF3455, PF11937). This family shows a positive net diversification and in WR species and losses that are concurrent with reductions in wood-decay capabilities of BR, ECM and parasitic species, although in contrast to PODs some copies are retained in most species (fig. 3). This and similar families containing DUFs represent worthy targets for experimental studies.

Extension and Validation of the COMPARE Pipeline Using Simulated Gene Family Evolution

The COMPARE method produces a mapping of gene duplications and losses onto an organismal phylogeny (Nagy et al. 2014). Here, we used this mapping to detect gene families that evolve in a significantly correlated fashion with the phenotype (fig. 2). To this end, we converted inferred numbers of duplications and losses to duplication and loss rates for each branch and analyzed the correlation between the evolution of WR and duplication/loss patterns for each gene family using a permutation ANOVA (Mitchell and Bergmann 2015).

We assessed the performance of this method using simulations with five different simulated organismal phylogenies each with 35 terminals. We modeled 15 trait histories on each simulated organismal phylogeny, and then simulated gene family evolution with rates of gene duplications (λ) and losses (μ) correlated with presence or absence of the trait (see "Materials and Methods" section for details). To assess type I and type II error rates, we analyzed gene trees simulated under equal-rate and phenotype-dependent variable-rate models, respectively. Variable rate models emulated the evolution of a complex trait imposing selective pressure on the accumulation and functional diversification of paralogs in a subset of species (fig. 2). Gene trees evolved under equal-rate models were used to estimate type I error rates.

We obtained 294,634 and 92,900 gene trees under variable-rates and equal-rates models, respectively. Across all simulated gene trees, COMPARE detected 73% of phenotype-induced rate differences [at $P \leq 0.05$, permutation ANOVA; see Mitchell and Bergmann (2015)], whereas on a narrower, biologically more realistic set of gene trees (Floudas et al. 2012) (see "Materials and Methods" section), COMPARE successfully detected rate differences in >96% of the variable rate gene trees ($P \leq 0.05$, ANOVA, fig. 2). To obtain type I error rates, we imposed the 60 trait histories on each of the equal-rate gene trees, yielding a type I error rate of 10.6% ($P \leq 0.05$, ANOVA).

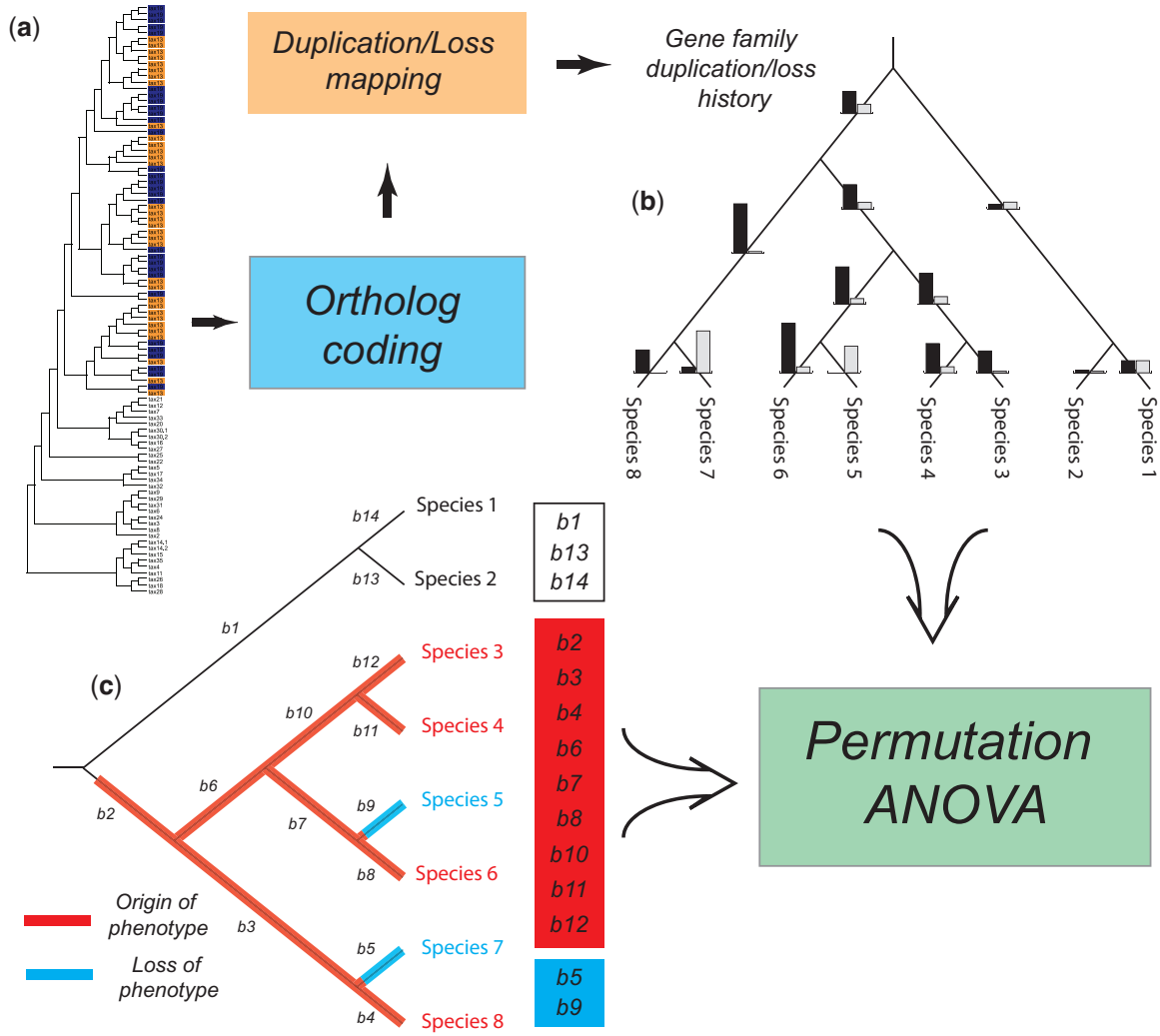

FIG. 1. Overview of the COMPARE strategy. The pipeline starts with a (reconciled) gene tree (*a*) on which functionally equivalent groups of genes (orthogroups) are identified by means of the ortholog-coding algorithm (Nagy et al. 2014). The origin and losses of the resulting orthogroups are then mapped onto the species tree, providing information on the rate of duplication and loss (black and grey bars, respectively) for each gene family for each branch of the species tree (*b*). These are then compared with the evolution of the phenotypic trait of interest (*c*) using a custom permutation ANOVA analysis.

Type II error rates negatively correlate with gene tree size, with the vast majority of the false negatives observed in small trees (<50 terminals) and in trees where the proportion of genes of species with increased diversification rate is low (<0.4) relative to the entire gene tree (referred to as θ, see "Materials and Methods" section and fig. 2), probably due to the smaller amounts of information contained in smaller trees. Thus, the method performs better under lower gene turnover rates and on larger trees. Nevertheless, it can detect duplication rate differences on gene trees as small as ten terminals, which shows that it performs well on a range of biologically realistic settings (fig. 2, supplementary fig. S1, Supplementary Material online). It should be noted that gene trees were assumed to be known without error in our simulations, in contrast to real datasets, where gene trees are always associated with some degree of error. Gene tree error is a general source of uncertainty in genome-wide studies of phylogenetic relationships and evolution (Wapinski et al. 2007; Galtier and Daubin 2008; Boussau et al. 2013; Wu et al. 2013) and is thus likely to affect COMPARE analyses as well. Gene tree–species tree reconciliations can be used to mitigate this effect (see below; Chen et al. 2000; Bansal et al. 2010; Wu et al. 2013; Szollosi et al. 2015).

Conclusions

Predicting the Genetic Bases of Eukaryotic Phenotypes

Leveraging rapidly accumulating whole genome data for understanding the genetic bases of complex phenotypes is a grand challenge of bioinformatics. Advances in sequencing technologies have led to a dramatic increase in the number of available genomes, but the development of appropriate bioinformatic tools has lagged behind, making bioinformatics the bottleneck in many genomics studies (Perkel 2013). Whereas effective methods are available for identifying the genetic bases of polygenic traits within populations (e.g., QTL mapping; Mackay et al. 2009), predicting protein function over larger evolutionary timescales poses fundamentally

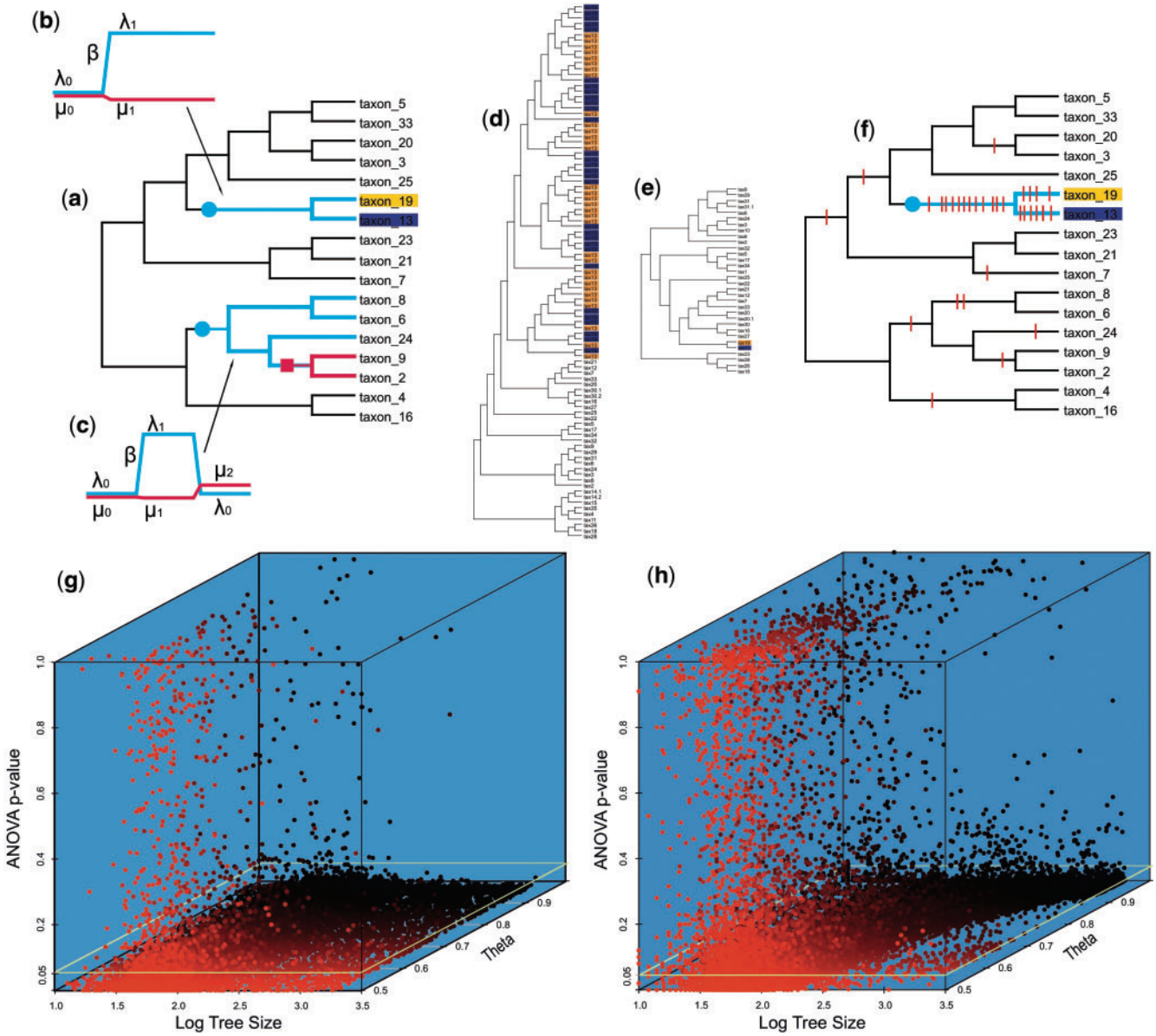

Fig. 2. Simulation studies used to assess the performance of COMPARE. A trait emerging in a group of species (blue) along a species tree (*a*) influences the diversification of genes involved in the trait. We model this process by increasing gene duplication (λ_1) and decreasing gene loss rates (μ_1) by a factor (β) upon the evolution of the trait (*b*) and/or decreased duplication (λ_2) and increased loss rate (μ_2) upon its loss (*c*) relative to background duplication and loss rates (λ_0, μ_0). This results in higher copy numbers in species, which evolved the trait in gene families functionally associated with it (*d*) as compared with families which are not (*e*). Such differences in gene duplication/loss histories can be detected by mapping duplications and losses (red bars) onto the species tree and statistical comparisons with the evolutionary history of the trait of interest (*f*). (*g*–*h*) shows the distribution of *P*-values obtained for simulated gene trees evolved under phenotype-dependent variable rate models with two input gene turnover rates (g: 0.2, h: 0.9). Yellow frame marks the 0.05 significance level. The *z*-axis corresponds to the proportion of terminals belonging to the subtree with altered duplication rate (θ, see "Materials and Methods" section). Markers are colored according to their value on the *z*-axis (from black to red) and show that most false negative detections occur at low θ values.

different challenges. Gains and losses of phenotypic traits may be correlated with expansion or contraction of functionally associated gene families (Ohno 1970; Zhang 2003; Conant and Wolfe 2008). Thus, analyses of gene family diversification patterns in relation to traits should provide insights into the genetic mechanisms of phenotypic evolution.

Methods for analyzing gene–gene rather than gene–phenotype associations have been proposed previously; phylogenetic profiling (Pellegrini et al. 1999) and different flavors thereof (Cokus et al. 2007; Antonov and Mewes 2008; Gonzalez et al. 2009; Simonsen et al. 2012; Lin et al. 2013; Psomopoulos et al. 2013) infer functional linkages between proteins based on their co-occurrence patterns (phylogenetic profiles) in extant prokaryotic genomes, an idea that has been extended to gene–phenotype associations as well (Antonov and Mewes 2008; Gonzalez et al. 2009). One common limitation of phylogenetic profiling and its derivatives is that these approaches are based on co-occurrence of proteins rather

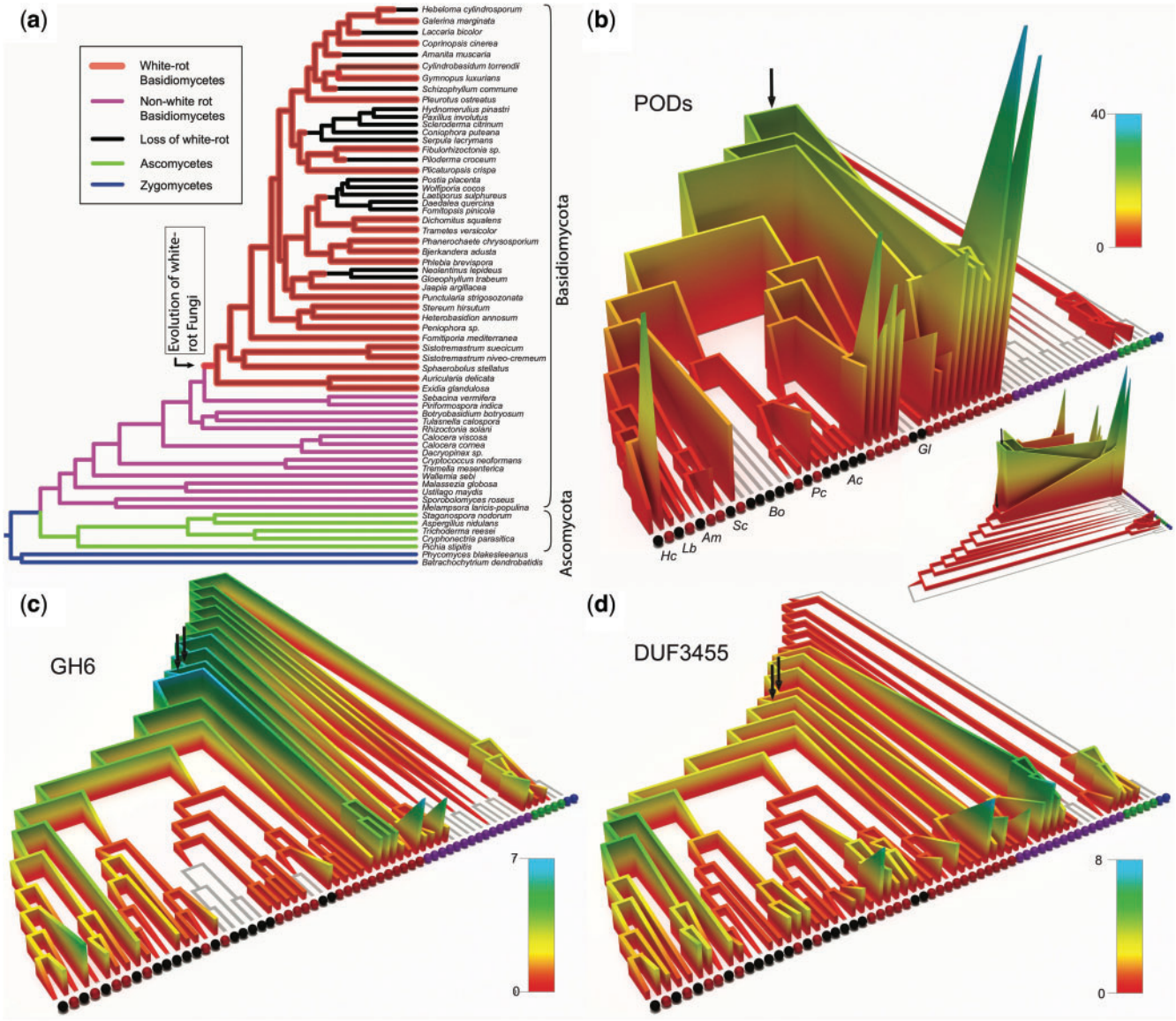

Fig. 3. The evolution of fungal lignocellulose decomposition ability and corresponding changes in gene family structure. (*a*) Phylogenetic tree of fungi showing the origin (red) and losses (black) of lignolytic capabilities. NonWR clades are condensed. The tree was obtained from our previous study (Nagy et al. 2016) and is based on concatenated ML analyses of 623 genes (91,981 amino acids). Reconstructed duplication/loss histories are shown for three gene families for which our analyses predicted a role in lignocellulose decomposition (*b*–*d*). Class-II-peroxidase copy number (*b*, PODs, with posterior view of the same tree shown as inset) shows an abrupt increase from 3 to 21 copies coinciding with the evolution of white rot and significant reductions upon its loss, whereas glycoside hydrolase family 6 (*c*, GH6) and a family of conserved fungal proteins (*d*, DUF3455) show more modest, yet consistent changes concurrent with the evolution of wood decomposition ability. Branch color corresponds to reconstructed ancestral and observed extant copy numbers in the species' genomes, vertical arrows denote the evolution of white rot. Highlighted clades in which white rot was lost or modified: *Hebeloma cylindrosporum* (Hc), *Laccaria bicolor* (Lb), *Amanita muscaria* (Am), *Schizophyllum commune* (Sc), Boletales (Bo), *Piloderma croceum* (Pc), Antrodia clade (Ac), and Gloeophyllales (Gl).

than phylogenetic patterns of gains and losses through evolution and thus cannot discriminate correlated gene family expansions and contractions from other processes generating overlapping profiles of genes, including shared inheritance (Barker et al. 2007; Barker and Pagel 2005). Further, phylogenetic profiling works best when intricate duplication histories in multigene families and deep paralogy are rare or absent, which might be the reason why it is less suited to study eukaryotes (Snitkin et al. 2006; Jothi et al. 2007; Singh and Wall 2008).

Here we extend the previously published COMPARE pipeline to formally screen for gene families, which evolve in a correlated fashion with the phenotypic trait of interest. COMPARE includes two major improvements over previous methods: (i) it uses a phylogenetically informed orthology detection algorithm (Nagy et al. 2014), and (ii) it predicts functional associations between protein families and a phenotype by analyzing phylogenetic correlations between gene family duplication/loss histories and phenotypic characters.

Identification of Components of the Cellular Pathways Underlying Wood-Decay

Comparative phylogenomic studies on the evolution of wood decay mechanisms have tended to focus on gene families encoding enzymes that are known to play a role in lignocellulolysis, particularly PODs and CAZys active on lignin and crystalline cellulose (Floudas et al. 2012; Riley et al. 2014; Kohler et al. 2015; Nagy et al. 2016). At the same time, empirical studies of wood decay in model systems routinely detect, in addition to the expected enzymes, large numbers of uncharacterized proteins that are upregulated under ligninolytic conditions (Vanden Wymelenberg et al. 2010; Gaskell et al. 2014; Korripally et al. 2015). In this study, we attempted to look beyond the enzymatic "known knowns" and assess the general importance of uncharacterized proteins in wood decay. Our results, which are validated by simulation analyses, comparisons to prior gene tree-species tree reconciliation studies, and gene expression analyses in model systems, reveal over 400 gene families that appear to be evolutionarily correlated with WR. Among the candidate genes are some encoding the expected enzymes that directly attack the major substrates of PCW, such as PODs, cellobiohydrolases, and LPMOs. Other classes of apparently important enzymes include those potentially involved in transport or detoxification of decay byproducts, such as lignin derivatives, or perhaps plant secondary metabolites, and many other proteins of unknown function.

WR is often equated with the presence of lignin degradation mediated by PODs. However, WR fungi also possess a robust complement of cellulolytic enzymes, which appear to have diversified prior to the origin of ligninolytic PODs (Floudas et al. 2012; Riley et al. 2014; Nagy et al. 2016). Enzymes attacking hemicellulose and pectin are also part of the WR apparatus. The findings reported here contribute to the view that the evolution of WR was marked by a general elaboration of the decay apparatus, not only the ligninolytic PODs (Nagy et al. 2016). These results highlight the need to take a holistic view of the evolution and functional biology of WR and its potential impact on the carbon cycle, considering the synergistic effects of diverse enzymes on all PCW components.

Materials and Methods

Simulation Studies

We evolved sets of gene trees within each of the five synthetic species trees (supplementary fig. S1, Supplementary Material online) under equal and variable rate models and a range of gene duplication and loss rates. Under the equal rate models, each branch of the species tree had the same set of duplication and loss rates, emulating the case when there is no evolutionary innovation along the tree that impacts gene family diversification rates. The impact of an evolved phenotype on the phylogenetic structure of gene families was simulated by defining points on the species tree where the duplication/loss rates change (fig. 2). A gain of the trait increased the background gene duplication rate λ_0 by some factor β and decreased the background gene loss rate μ_0 by β. Thus, the duplication and loss rate under the trait are $\lambda_1 = \lambda_0\beta$ and $\mu_1 = \mu_0/\beta$, respectively. Similarly, if the trait is lost, duplication rate (λ_1) drops back to the background duplication rate (λ_0), whereas the gene loss rate increases as $\mu_2 = \mu_0\beta$, corresponding to the lack of selection to maintain existing gene copies. Note that gene family contraction will only happen if $\lambda < \mu_2$.

We used five different species trees each with 35 species for evolving gene trees within them. Comparative genomics studies frequently sample one or a few species per order or family, resulting in very low overall sampling frequency. We accommodated this into our study by simulating species trees under the Yule model with sampling in Mesquite 3.0 (Maddison and Maddison 2009). First, species trees were evolved until they reached one thousand terminals, then taxa were randomly pruned from the tree until the number of terminals reached 35. The initial speciation rate was set to 0.01 and the tree depth was set to 1.0. We initially simulated 10000, 35-taxon species trees, of which we chose five (supplementary fig. S1, Supplementary Material online) based on values of their γ-statistic, which measures the temporal distribution of internal nodes and depends on the assumptions of the taxon sampling strategy (Pybus and Harvey 2000). For instance, a diversifying sampling, where researchers try to sample every major (e.g., ordinal level) clade—a common strategy in genomics—would result in most of the internal nodes being closer to the root (small γ value), whereas with random sampling the distribution of internal nodes would be even (γ value close to 0). We sorted species trees according to their γ-statistic and chose 5 (trees no. 735, 913, 1135, 3313, and 4247) out of 10,000 at even intervals on the range of obtained γ-values (-7.29 to -2.14). Thus, our five species trees represent a range of taxon sampling strategies that researchers may apply in comparative genomics studies.

We chose five different background duplication rates (0.2, 0.4, 0.6, 0.8, 1) and five different β values (2.5*x*, 5*x*, 10*x*, 15*x*, and 20*x*). Gene loss rates were defined as a fraction of the background duplication rate, either $0.9\lambda_0$ or $0.2\lambda_0$, reflecting different views on the fate of newly duplicated genes. Scenarios of phenotype evolution were defined as one gain of the phenotype and 0 or more losses across the species tree.

We defined 15 different trait histories along each species tree by hand, including seven histories with one gain of the trait and no losses, resulting in one subtree of the species tree with increased duplication rate and decreased loss rate, as well as four histories with one gain of the trait and one loss and four with one gain and two losses. One or more losses potentially add valuable signal to the analyses, because a contraction of the gene family is an additional source of information. Thus, real datasets with multiple rate change points might be more informative than those with a single origin of the trait of interest, although it should be noted that traits with multiple nonhomologous origins (i.e., that are not Dollo-like) might be problematic (i.e., if convergent origins of the trait have different genetic bases). The trait histories were set up so as to affect various proportions of the total length of the species tree, because the longer the affected path, the more pronounced the impact of rate changes on the gene tree can be. Proportions of the total tree lengths impacted by rate

changes range from ~1% to ~50%. The 60 histories (15 for each species tree) are summarized in supplementary table S2, Supplementary Material online.

We used the *rgenetree* function of the HyPhy R package (Hallinan 2015) to simulate 100 gene trees for each parameter combination. Thus, with five duplication rates, five β values and two loss ratios, we had 50 parameter combinations for each trait history. With 15 trait histories, this resulted in 750 × 100 gene trees per species tree. In addition, we simulated 400 trees under the equal rate model (no rate change) for each parameter combination. During the simulations, we excluded gene trees >2000 terminals (45,668 trees), since trees above this size are rarely encountered in real datasets. We further excluded 17,224 variable rate trees in which branches with increased duplication rates were missing due to one of its ancestors having gone extinct (these trees would have misled downstream analyses).

To characterize simulated gene trees, we defined a new parameter, θ, as the proportion of terminals belonging to the subtree with an altered duplication rate. We found that this parameter described the simulated gene trees better than input parameters due to the stochastic nature of the simulations. To estimate biologically reasonable values of θ, we surveyed 12 gene families shown to be related to white-rot in fungi (Floudas et al. 2012). We found that θ values across these gene trees ranged from 0.57 to 0.94, with a mean of 0.73, therefore, we restricted the analyses of simulated gene trees to the biologically realistic range of $\theta = 0.5–1$.

Ortholog-Coding and Dollo Reconstruction of Gene Family Duplication/Loss Histories

We identified sets of orthologous genes in each of the gene trees and recoded these sets as presence/absence characters using the ortholog coding algorithm (Nagy et al. 2014), then reconstructed the duplication/loss history by mapping the presence of orthologs on the respective species tree using Dollo parsimony. We reconstructed the duplication/loss history of one gene tree at a time and recorded the number of gains (duplications) and losses along each branch of the species tree. This resulted in a matrix of number of duplications and losses for each branch which, when normalized by branch lengths gave duplication and loss rates for each of the branches of the species tree for each gene tree.

Obtaining Type I and II Error Rates

P-Values measuring the extent of correlated evolution between gene families and phenotypes for simulated gene trees were obtained by using a permutation ANOVA implemented with custom R code (Mitchell and Bergmann 2015). For each gene tree, the ANOVA tested whether the set of branches of the species tree impacted by the trait history had significantly higher rates of gene duplication and loss than the rest of the species tree. Thus, the species tree was divided into three groups of branches, one including branches of the tree unaffected by trait gain or loss (background duplication and loss rates, λ_0, μ_0), one on which the trait evolved (increased duplication rates, λ_1, and decreased loss rates, μ_1) and a third in which the trait was lost (increased loss rates, μ_2, and decreased duplication rates, λ_2). For all ANOVA analyses, permutation was performed with 100 replicates and gene trees with $P \leq 0.05$ considered significant detections. Type I error rates were inferred by obtaining *P*-values for the trees evolved under equal rates models, by imposing each of the 15 trait histories to each equal-rate gene tree. A fully Maximum Likelihood-based method (Barker and Pagel 2005; Barker et al. 2007) has also been considered, but the nature of the data (duplication/loss histories) precluded its application on our datasets.

Availability

COMPARE is implemented in Perl and is composed of independent scripts that can be executed sequentially, making it easy to modify or incorporate into existing pipelines. Source code is available at https://github.com/laszlognagy/COMPARE (last accessed July 11, 2016).

Prediction of Gene Families Involved in WR

Representative white rot and brown rot Agaricomycetes and biotrophic fungi (ECM mutualists and pathogens) were assembled to cover all major clades for which genomic data were available. An ortholog database was constructed by performing all-vs.-all blast searches on predicted proteomes, followed by MCL clustering, multiple sequence alignment and gene tree inference. Predicted protein sequences were downloaded for 62 genomes from the Joint Genome Institute (JGI) MycoCosm pages (supplementary table S1, Supplementary Material online). All vs. all Blast searches and similarity-based clustering of protein sequences were performed using mpiBlast v.1.6.9 and MCL v.1.3.7, respectively. For clustering, an inflation parameter of 2.0 was chosen. A maximum likelihood phylogenomic species tree, based on 623 single-copy genes, was obtained from our previous study (Nagy et al. 2016) and used as the organismal phylogeny to map orthogroup gains and losses.

Next, each cluster of proteins was aligned by using PRANK v.140603 (Loytynoja and Goldman 2008) (default parameters). Maximum likelihood gene trees were estimated from the resulting alignments in RAxML v.8.1.2 (Stamatakis 2006) under the CAT model for clusters with >50 proteins, whereas for smaller clusters we used the computationally more demanding GTRGAMMA model. To mitigate topological error, gene trees were refined by optimizing likelihood and duplication/loss costs in Treefix v.1.1.10 (Wu et al. 2013). Gene orthogroups were identified using ortholog coding on the reconciled gene trees and the origin and losses for each orthogroup were mapped on the organismal phylogenetic tree using Dollo parsimony. Gene family duplication/loss histories were obtained by merging orthogroup gain/loss data across a gene family.

Gene family histories were tested for correlations with evolution of the ability to decompose lignocellulosic plant cell walls by fungi. The ancestral state reconstruction of decay ability was taken from (Nagy et al. 2016) and comprised a single origin of white rot and seven loss events in the

Gloeophyllales clade, the *Antrodia* clade, Atheliales, and Boletales, as well as *Hebeloma*, *Laccaria* and *Amanita* in the Agaricales. Our dataset includes three taxa (*Jaapia*, *Schizophyllum*, and *Cylindrobasidum*) with wood-decomposition mechanisms that do not conform to typical WR or BR (Riley et al. 2014). We performed sensitivity analyses, in which we assigned these taxa to different categories in the analyses of variance. These alternative assignments showed that our results are robust to alternative assignments of these taxa (data not shown). We tested for correlations between the evolution of decay mode and gene family duplication/loss histories using a permutation ANOVA as described above. For analyses of variance, extant species and internal branches were grouped into three categories, white-rot (WR), nonWR preceding the evolution of WR (n-WR) and those, which have lost WR (l-WR). Numbers of duplications and losses were converted to duplication and loss rates (λ and μ) by taking into account branch lengths of the organismal phylogenetic tree. Gene families showing a correlation with the evolution of white rot at $P \leq 0.05$ were further analyzed. We obtained a *P*-value for 140,039 out of 140,137 protein clusters. For the remaining 98 clusters, reconstruction of duplication/loss histories failed mostly due to alignment problems or prohibitive computational burden of tree inference and reconciliation.

Enrichment of functional annotation terms was done using Gene Ontology terms mapped using the 2015/02/14 version Pfam2GO database on predicted PFAM domains obtained using HMMER3 v.3.1. We used the GO::TermFinder perl package to analyze enrichment using the hypergeometric test with Bonferroni correction for multiple hypothesis testing. The 2015/02/12 release of the GO hierarchy was used (go-basic v1.2).

Supplementary Material

Supplementary tables S1–S3 and figures S1–S2 are available at *Molecular Biology and Evolution* online.

Author Contributions

L.G.N. and D.S.H. conceived and perfromed the study, L.G.N. and P.J.B. implemented the extension of the COMPARE pipeline and performed simulation studies, L.G.N., R.R. and K.K. performed analyses of biological data. D.C. analyzed and interpreted expression data, F.M.M. and I.V.G. contributed unpublished resources and whole-genome sequences. All authors have read and revised the paper.

Acknowledgments

We thank Dimitrios Floudas for helpful discussions on wood-decay related gene families and Sándor Kocsubé for his help with designing the figures. This work was supported by the Lendület Programme of the Hungarian Academy of Sciences (grant no. LP2014/12, to LGN), NSF awards DEB-0933081 and IOS-1456958 (to DSH) and by the Laboratory of Excellence ARBRE (ANR-11-LABX-0002-01) (to FMM).

References

Antonov AV, Mewes HW. 2008. Complex phylogenetic profiling reveals fundamental genotype-phenotype associations. *Comput Biol Chem.* 32:412–416.

Bansal MS, Burleigh JG, Eulenstein O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC Bioinformatics* 11:S42.

Barker D, Meade A, Pagel M. 2007. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* 23:14–20.

Barker D, Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol.* 1:e3.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–330.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol.* 7:429–447.

Cokus S, Mizutani S, Pellegrini M. 2007. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics* 8: S7.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, 9:938–950.

Courty PE, Hoegger PJ, Kilaru S, Kohler A, Buee M, Garbaye J, Martin F, Kues U. 2009. Phylogenetic analysis, genomic organization, and expression analysis of multi-copper oxidases in the ectomycorrhizal basidiomycete Laccaria bicolor. *New Phytol.* 182:736–750.

Dashtban M, Schraft H, Syed TA, Qin W. 2010. Fungal biodegradation and enzymatic modification of lignin. *Int J Biochem Mol Biol.* 1:36–50.

Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, et al. 2011. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science* 333:762–765. [pii]10.1126/science.1205411

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336:1715–1719.

Floudas D, Held BW, Riley R, Nagy LG, Koehler G, Ransdell AS, Younus H, Chow J, Chiniquy J, Lipzen A, et al. 2015. Evolution of novel wood decay mechanisms in Agaricales revealed by the genome sequences of Fistulina hepatica and *Cylindrobasidium torrendii*. *Fungal Genet Biol.* 76:78–92.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Phil. Trans. R. Soc. B Biol. Sci.* 363:4023–4029.

Gaskell J, Marty A, Mozuch M, Kersten PJ, BonDurant SS, Sabat G, Azarpira A, Ralph J, Skyba O, Mansfield SD, et al. 2014. Influence of populus genotype on gene expression by the wood decay fungus *Phanerochaete chrysosporium*. *Appl Environ Microbiol.* 80:5828–5835.

Gonzalez NA, Vazquez A, Ortiz Zuazaga HG, Sen A, Olvera HL, Pena de Ortiz S, Govind NS. 2009. Genome-wide expression profiling of the osmoadaptation response of *Debaryomyces hansenii*. *Yeast* 26:111–124.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42:D699–D704.

Hallinan NM. 2015. Package 'HyPhy'. v1.0, downloaded from CRAN.

Hibbett D, Blanchette R, Kenrick P, Mills B. 2016. Climate, decay, and the death of the coal forests. *Curr Biol.* 26:R563–R567

Hofrichter M, Ullrich R, Pecyna MJ, Liers C, Lundell T. 2010. New and classic families of secreted fungal heme peroxidases. *Appl Microbiol Biotechnol.* 87:871–897.

Horn SJ, Vaaje-Kolstad G, Westereng B, Eijsink VGH. 2012. Novel enzymes for the degradation of cellulose. *Biotechnol Biofuels* 5:45.

Jothi R, Przytycka TM, Aravind L. 2007. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics* 8:173.

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canbäck B, Choi C, Cichocki N, Clum A, Colpaert J, et al. 2015. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat Genet*. 47:410–415.

Korripally P, Hunt C, Houtman CJ, Jones DC, Kitin PJ, Cullen D, Hammel KE. 2015. Regulation of gene expression during the onset of ligninolytic oxidation by *Phanerochaete chrysosporium* on spruce wood. *Appl Environ Microbiol*. 81:7802–7812.

Lin TW, Wu JW, Chang DT. 2013. Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins. *PLoS One* 8:e75940.

Loytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320:1632–1635.

Mackay TF, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*. 10:565–577.

Maddison WP, Maddison DR. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.6. http://mesquiteproject.org.

Martinez AT, Speranza M, Ruiz-Duenas FJ, Ferreira P, Camarero S, Guillen F, Martinez MJ, Gutierrez A, del Rio JC. 2005. Biodegradation of lignocellulosics: microbial, chemical, and enzymatic aspects of the fungal attack of lignin. *Int Microbiol*. 8:195–204.

Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, et al. 2009. Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. *Proc Natl Acad Sci U S A*. 106:1954–1959.

Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, et al. 2004. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat Biotechnol*. 22:695–700.

Mathieu Y, Prosper P, Favier F, Harvengt L, Didierjean C, Jacquot JP, Morel-Rouhier M, Gelhaye E. 2013. Diversification of fungal specific class A glutathione transferases in saprotrophic fungi. *PLoS One* 8:e80298.

Mitchell A, Bergmann PJ. 2015. Thermal and moisture habitat preferences do not maximize jumping performance in frogs. *Funct Ecol*. 30:733–742.

Morel M, Meux E, Mathieu Y, Thuillier A, Chibani K, Harvengt L, Jacquot JP, Gelhaye E. 2013. Xenomic networks variability and adaptation traits in wood decaying fungi. *Microb Biotechnol*. 6:248–263.

Morel M, Ngadin AA, Droux M, Jacquot JP, Gelhaye E. 2009. The fungal glutathione S-transferase system. Evidence of new classes in the wood-degrading basidiomycete *Phanerochaete chrysosporium*. *Cell Mol Life Sci*. 66:3711–3725.

Nagy GL, Riley R, Tritt A, Adam C, Daum C, Floudas D, Sun H, Yadav J, Pangilinan J, Larsson K-H, et al. 2016. Comparative genomics of early-diverging mushroom-forming fungi provides insights into the origins of lignocellulose decay capabilities. *Mol Biol Evol*. 33:959–970.

Nagy LG, Ohm RA, Kovacs GM, Floudas D, Riley R, Gacser A, Sipiczki M, Davis JM, Doty SL, de Hoog GS, et al. 2014. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun*. 5:4471.

Nelsen MP, DiMichele WA, Peters SE, Boyce CK. 2016. Delayed fungal evolution did not cause the Paleozoic peak in coal production. *Proc Natl Acad Sci U S A*. 113:2442–2447.

Ohno S. 1970. Evolution by gene duplication. Springer, New York.

Olson A, Aerts A, Asiegbu F, Belbahri L, Bouzid O, Broberg A, Canback B, Coutinho PM, Cullen D, Dalman K, et al. 2012. Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen. *New Phytol*. 194:1001–1013.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*. 96:4285–4288.

Perkel J. 2013. Finding the true $1000 genome. *Biotechniques* 54:71.

Psomopoulos FE, Mitkas PA, Ouzounis CA. 2013. Detection of genomic idiosyncrasies using fuzzy phylogenetic profiles. *PLoS ONE* 8:e52854.

Pybus OG, Harvey PH. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc Biol Sci*. 267:2267–2272.

Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, Levasseur A, Lombard V, Morin E, Otillar R, et al. 2014. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proc Natl Acad Sci U S A*. 111, 9923–9928.

Ruiz-Duenas FJ, Martinez AT. 2009. Microbial degradation of lignin: how a bulky recalcitrant polymer is efficiently recycled in nature and how we can take advantage of this. *Microb Biotechnol*. 2:164–177.

Rytioja J, Hilden K, Yuzon J, Hatakka A, de Vries RP, Makela MR. 2014. Plant-polysaccharide-degrading enzymes from basidiomycetes. *Microbiol Mol Biol Rev*. 78:614–649.

Simonsen M, Maetschke SR, Ragan MA. 2012. Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. *Bioinformatics* 28:851–857.

Singh S, Wall DP. 2008. Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function. *Evol Bioinformatics* 4:217–223.

Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C. 2006. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics* 7:420.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Szollosi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst Biol* 64:E42–E62.

Tanaka H, Yoshida G, Baba Y, Matsumura K, Wasada H, Murata J, Agawa M, Itakura S, Enoki A. 2007. Characterization of a hydroxyl-radical-producing glycoprotein and its presumptive genes from the white-rot basidiomycete *Phanerochaete chrysosporium*. *J Biotechnol* 128:500–511.

Vanden Wymelenberg A, Gaskell J, Mozuch M, Sabat G, Ralph J, Skyba O, Mansfield SD, Blanchette RA, Martinez D, Grigoriev I, et al. 2010. Comparative transcriptome and secretome analysis of wood decay fungi *Postia placenta* and *Phanerochaete chrysosporium*. *Appl Environ Microbiol*. 76:3599–3610.

Varnai A, Siika-aho M, Viikari L. 2013. Carbohydrate-binding modules (CBMs) revisited: reduced amount of water counterbalances the need for CBMs. *Biotechnol Biofuels* 6:30.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.

Wu YC, Rasmussen MD, Bansal MS, Kellis M. 2013. TreeFix: statistically informed gene tree error correction using species trees. *Syst Biol*. 62:110–120.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol*. 18:292–298.