Journal of Wood Science

ORIGINAL ARTICLE Open Access

Acetylxylan esterase is the key to the host specialization of wood-decay fungi predicted by random forest machine-learning algorithm

Natsuki Hasegawa[1], Masashi Sugiyama[2,3] and Kiyohiko Igarashi[1,3*]

Abstract

Wood-decay fungi produce extracellular enzymes that metabolize wood components such as cellulose, hemicellulose and lignin. Each fungus has a preference of wood species as the host, but identification of these preferences requires a huge amount of cultivation data. Here, we developed a method of predicting the wood species preference, Angiosperm specialist or Gymnosperm specialist or generalist, of wood-decay fungi using the random forest machine-learning algorithm, trained on the numbers of families associated with host specialization in the Carbohydrate-Active enZymes database. The accuracy of the prediction was about 80%, which is lower than that of the classification of white- and brown-rot fungi (more than 98%) by the same method, but the reason for this may be the ambiguity of the definition of "preference" and "generalists". Carbohydrate esterase (CE) family 1 acetylxylan esterase was the most significant contributor to the prediction of host specialization, followed by family 1 carbohydrate-binding module and CE family 15, mainly containing glucuronoyl esterases. These results suggest that the ability to degrade glucuronoacetylxylan, a major hemicellulose of Angiosperm, is the key factor determining the host specialization of wood-decay fungi.

Keywords Acetylxylan esterase, Wood-decay fungi, Carbohydrate-Active enZymes, Machine learning, Random forest algorithm

Introduction

Wood-decay fungi are a unique group of organisms on Earth that exclusively metabolize wood [1]. Their ecological impact extends beyond the local decay process, influencing global carbon cycling. Moreover, they serve as a critical source of biomass-converting enzymes for building a decarbonized society [2]. However, wood-decay fungi can have both positive and negative effects. While they contribute to nutrient recycling and carbon sequestration, their activity can compromise the structural integrity of wood, especially in construction-grade softwood materials [3]. This degradation not only reduces the lifespan and value of wooden structures, but also poses risks during natural disasters such as earthquakes [4–6]. Understanding the fundamental principles underlying wood decay is therefore essential for achieving a sustainable economy.

Research on wood-decay fungi dates back to the early nineteenth century when scientists classified them based on the post-decay wood color, distinguishing between white-rot and brown-rot (formerly known as red-rot) fungi [7]. Advances in chemistry led to compositional analyses of decayed wood, linking enzymatic activity to wood degradation [8]. Subsequent studies focused on enzyme purification and characterization, connecting

*Correspondence:
Kiyohiko Igarashi
aquarius@mail.ecc.u-tokyo.ac.jp
[1] Department of Biomaterial Sciences, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-Ku, Tokyo 113-8657, Japan
[2] Center for Advanced Intelligence Project, RIKEN, Chuo-Ku, Tokyo, Japan
[3] UT7 Next Life Research Group, The University of Tokyo, Bunkyo-Ku, Tokyo, Japan

specific enzymes to chemical changes in wood components [2]. In recent decades, molecular biology and bioinformatics have enabled comprehensive investigations into the genomics, transcriptomics, proteomics, and metabolomics of various wood-decay fungi [9–12]. However, these studies have mostly been conducted in controlled laboratory environments, which may not fully represent natural decay processes in real-world settings.

Our recent research leveraged machine learning, specifically the random forest (RF) algorithm, to predict whether a given fungus is a white- or brown-rot species based on its Carbohydrate-Active enZymes (CAZymes) family composition [13]. By analyzing genomic data, we achieved over 98% accuracy in distinguishing between these two decay types [14]. Notably, the lytic polysaccharide monooxygenase (LPMO) from the auxiliary activity (AA) family 9 emerged as the most influential enzyme for this classification. LPMO, which was discovered in 2010 [15], enhances cellulase activity by oxidatively damaging cellulose surfaces [16, 17]. Other enzymes such as cellulobiohydrolases from glycoside hydrolase (GH) family 7 and Class II peroxidases (from the AA2 family) also contributed to distinguishing white-rot from brown-rot fungi, in accordance with previous biochemical and other analyses [11, 18].

Furthermore, white- and brown-rot fungi exhibit different host specializations [19]. Brown-rot fungi include more generalists that can infect both Gymnosperms, generally called softwood, and Angiosperms, hardwood. In contrast, white-rot fungi are specialists, with over half of them specifically occurring on hardwoods [20]. In the present study, we trained the RF algorithm [21] on the number of CAZymes families associated with host specialization and identified acetylxylan esterase (AcXE) from carbohydrate esterase (CE) family 1 as playing a key role in host specialization.

Materials and methods

Creation of a host specialization data set for wood-decay basidiomycetes

The data set related to host specialization of wood-decay basidiomycetes was created based on prior research by Krah and colleagues [20]. First, we utilized the R package "rusda" to retrieve data from the Fungus-Host Distribution Database and Specimens Database (https://fungi.ars.usda.gov) maintained by the United States Department of Agriculture (USDA) [22]. We used the species names reported for each fungus in our previously created decay type data set as queries. For some fungi, automatic collection using "rusda" was not feasible, and in these cases, we manually collected the data.

We then scraped data from the National Center for Biotechnology Information (NCBI) taxonomy using the genus names as queries to determine whether the collected hosts were classified under the Acrogymnospermae phylum Gymnosperms or softwood, or the Magnoliopsida phylum, Angiosperms or hardwood. For hosts with discrepancies between the USDA Fungus Databases and NCBI registration names, we supplemented the classification by manually searching for synonyms based on data from the Global Biodiversity Information Facility (GBIF) [23], International Plant Names Index (IPNI) [24], and Tropicos [25] databases. Host species not belonging to either Gymnosperms or Angiosperms were removed from the data set.

Subsequently, we examined whether each host species was woody or herbaceous using the woodiness data set [26]. Herbaceous hosts were excluded, and for hosts without information in the woodiness data set, we researched using genus names. If all species within the same genus were either woody or herbaceous, we extrapolated the classification. In cases where both woody and herbaceous hosts coexisted, we considered them indeterminate and removed those host species from the data. These steps resulted in a data set comprising wood-decay fungi associated only with Gymnosperm and Angiosperm hosts.

From this data set, the "Gymnosperm association" was defined by dividing the number of gymnosperm host tree species (N_G) by the sum of the number of angiosperm (N_A) and gymnosperm host tree species: gymnosperm associations [%] $= \frac{N_G}{N_G+N_A}$. Based on these values, we categorized fungi into three groups: angiosperm specialists (0–10%), generalists (10–90%), and Gymnosperm specialists (90–100%), following the approach by Krah and colleagues. [20].

Construction and evaluation of RF models

We next conducted two analyses: classification to predict which of the three host specialization groups, Angiosperm specialist, Gymnosperm specialist, or generalist, a given genome sample belongs to, and regression to directly predict the Gymnosperm association value. In line with our previous report, we split the data set into training and test data (70% and 30%, respectively), corrected for data set imbalance by means of oversampling using the synthetic minority over-sampling technique (SMOTE) [27] for classification and the synthetic minority over-sampling technique for regression with Gaussian noise (SMOGN) [28] for regression, and evaluated model performance on the test data. We used the RandomForestClassifier and RandomForestRegressor from the Python library scikit-learn for model construction. We also performed the same tasks using LightGBM, an ensemble learning algorithm based on decision trees [29]. LightGBM adjusts data weights based on previous

tree predictions, creating trees sequentially in a gradient boosting fashion. Compared to RF, LightGBM generally achieves higher accuracy [29]. We built models using the Python package "lightgbm" and automatically tuned hyperparameters using Optuna [30].

For both tasks and algorithms, we incorporated numbers of all enzyme families/subfamilies as explanatory variables. Model construction was randomized and repeated 1000 times with oversampling and data splitting. We calculated performance metrics and averaged the Gini importance of each explanatory variable.

Results and discussion

The host specialization of wood-decay basidiomycetes differs between white-rot and brown-rot fungi, i.e., white-rot fungi predominantly specialize in Angiosperms (hardwood), while brown-rot fungi exhibit a higher proportion of generalists that infect both Gymnosperms (softwood) and hardwood [20]. Despite this knowledge of decay modes, the genetic basis of host specialization remains largely unexplored. In ascomycetes, changes in host range have been associated with gene duplications or losses [31, 32], but host range reduction in mycorrhizal fungi does not always involve gene loss [33]. The mechanisms underlying host specialization likely relate closely to the organism's nutritional strategy. White- and brown-rot fungi, being part of the Basidiomycota lineage, decompose dead plant cells differently from ascomycetes that can interact with living plants. As saprotrophs, they derive nutrients from decaying organic matter. Therefore, they may employ unique mechanisms diverging from those observed in ascomycetes that form mycorrhizal associations with living plants.

In this study, we applied comparative genomics methodology using the RF algorithm, which was validated for its utility in the context of decay types in our previous report [14], to study host specialization. The aim of this work was to gain insights into the genetic mechanisms underlying host specialization in wood-decay basidiomycetes. This approach also serves as an illustrative example of how machine learning can predict candidate genes by systematically exploring genomes in uncharted research areas.

Data set for the random forest machine-learning algorithm

In the host specialization data set used for the experiment, there were 88 samples of angiosperm specialists, 64 samples of generalists, and 30 samples of gymnosperm specialists (Fig. 1). The composition of this data set aligned with previous studies by Krah and colleagues [20]: among white-rot fungi, angiosperm specialists constituted more than half, while among brown-rot fungi, generalists were the most prevalent. Except for a single case (*Fistulina hepatica*), Agaricales, where only white-rot fungi were found, lacked Gymnosperm specialists, and Gloeophyllales, composed solely of brown-rot fungi, had no Angiosperm specialists.

Model performance

Classification predictions using RF improved with oversampling of the data set, resolving the discrepancy between recall and precision, but the accuracy remained only around 80% (Fig. 2a). In regression tasks, the coefficient of determination (R^2) was approximately 0.6, and both the mean absolute error (MAE) and root-mean-square error (RMSE) exceeded 0.2 (Fig. 2c). While these values ensured some predictive ability for host specialization traits, they were lower than the decay type predictions obtained in the previous report [14]. Even when using LightGBM, which is generally considered more accurate than RF, there was minimal change in precision metrics for both classification and regression tasks (Fig. 2b, d). The Gymnosperm association value, based on reporting frequency of host relationships, could be affected by sampling bias and probabilistic errors. Additionally, the conversion of its continuous value into categorical labels introduced artificial boundaries such as "preference" and "generalists", potentially contributing to the limitations in predictive accuracy due to data set imperfections. To address this, alternative indicators for accurately and precisely evaluating fungal host specialization would be needed.

Nevertheless, the model still achieved reasonable accuracy in predicting wood-decay fungi's host specialization. Leveraging machine learning in this noisy data set may be advantageous for identifying trends that would be challenging to discern manually. Furthermore, the host specialization trait in wood-decay fungi remains understudied and poorly understood, making it an area ripe for exploration. Our methodology enabled us to pioneer this uncharted field by predicting noteworthy genes through comprehensive exploration and illustrates an effective application of machine learning to comparative genomics. Averaging the class probabilities predicted by the RFs in the classification task, 40 samples were misclassified, but in most of them the class probability of the correct host specificity group was the second highest value (Fig. 3a). In addition, both the angiosperm specialists and gymnosperm specialists were misclassified as generalists more often than they were misclassified as specialist in the other category (Fig. 3b). Therefore, the prediction by RF was not entirely misplaced, and the model was considered to reflect, at least to some extent, the relationship between the host specificity of wood rot fungi and the number of CAZymes genes.

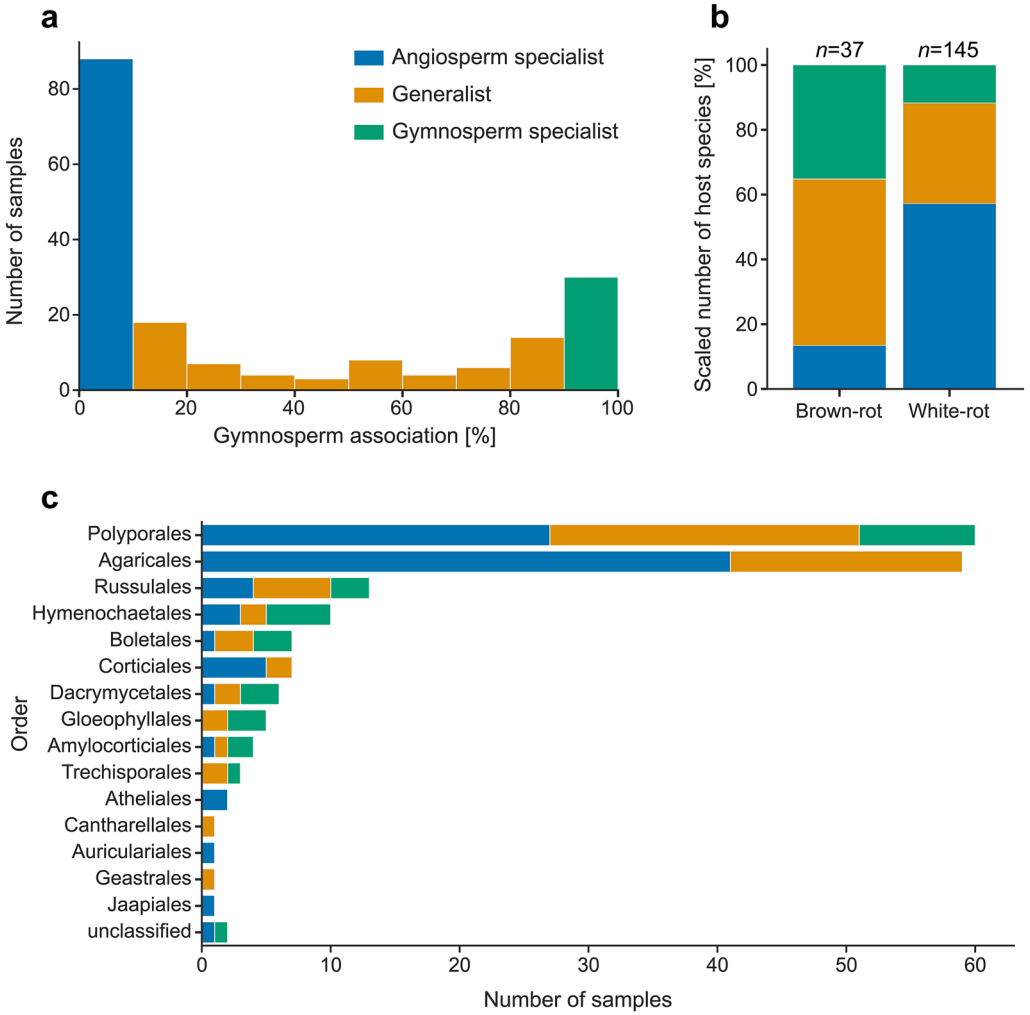


Fig. 1 Composition of the host specificity data set. **a** Histogram of gymnosperm association values for each sample in the host specificity data set. **b** Proportion of host specificity for different decay styles. **c** Sample count per order

The high number of prediction errors between specialists and generalists, as described above, suggested that the prediction accuracy of the model could be improved by adjusting the value of the boundary separating specialists and generalists. However, the main focus of comparative genomics is to understand the genetic basis of traits, and prediction accuracy is only a guarantee of the reliability of the model. Therefore, such minor tuning for accuracy was not done in this case because it would only have increased the fit to the data set of this experiment and might have reduced its generality for wood-rotting fungi as a whole. In addition, it is generally possible to improve the prediction performance in RF and LightGBM by carefully selecting only those explanatory variables with large contributions, but for the same reason, we did not follow-up using this approach.

CAZymes contributing to host specialization prediction

CE1 stood out across all four patterns, showing more than twice the importance of other families (Fig. 4, Table S1). CE1 includes AcXE, which degrades acetyl side chains protecting xylan. Notably, CE1 gene numbers significantly differed between gymnosperm specialists and the other two host specialization groups in white-rot fungi, suggesting its critical role as a bottleneck in hardwood xylan degradation (Fig. 5). Although CE1 also includes feruloyl esterases, which disconnect feruloyl side chain from herbaceous arabinoxylan, this activity might not be the target of the classification in this study considering that the content of feruloyl moiety is limited in arboreous plant. The GH10 and GH11 families involving xylanase activity also exhibited substantial difference in importance, correlating with their differences in acetyl xylan degradation activity [34]. The major difference of

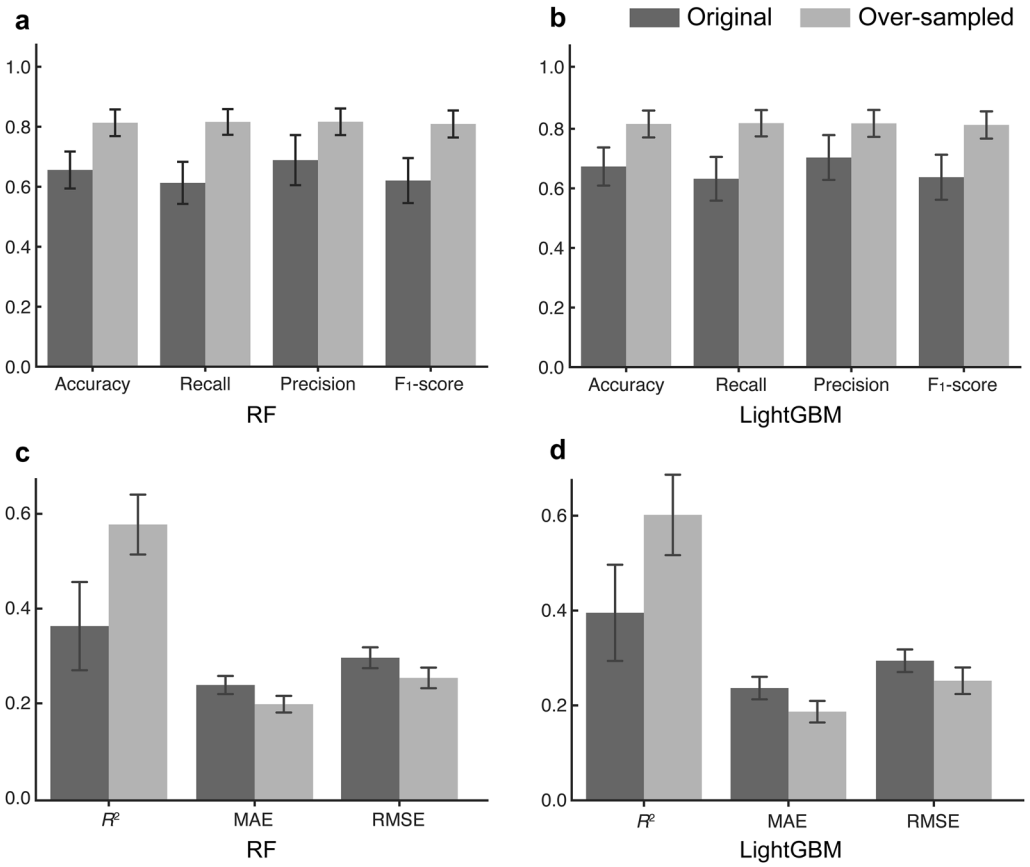
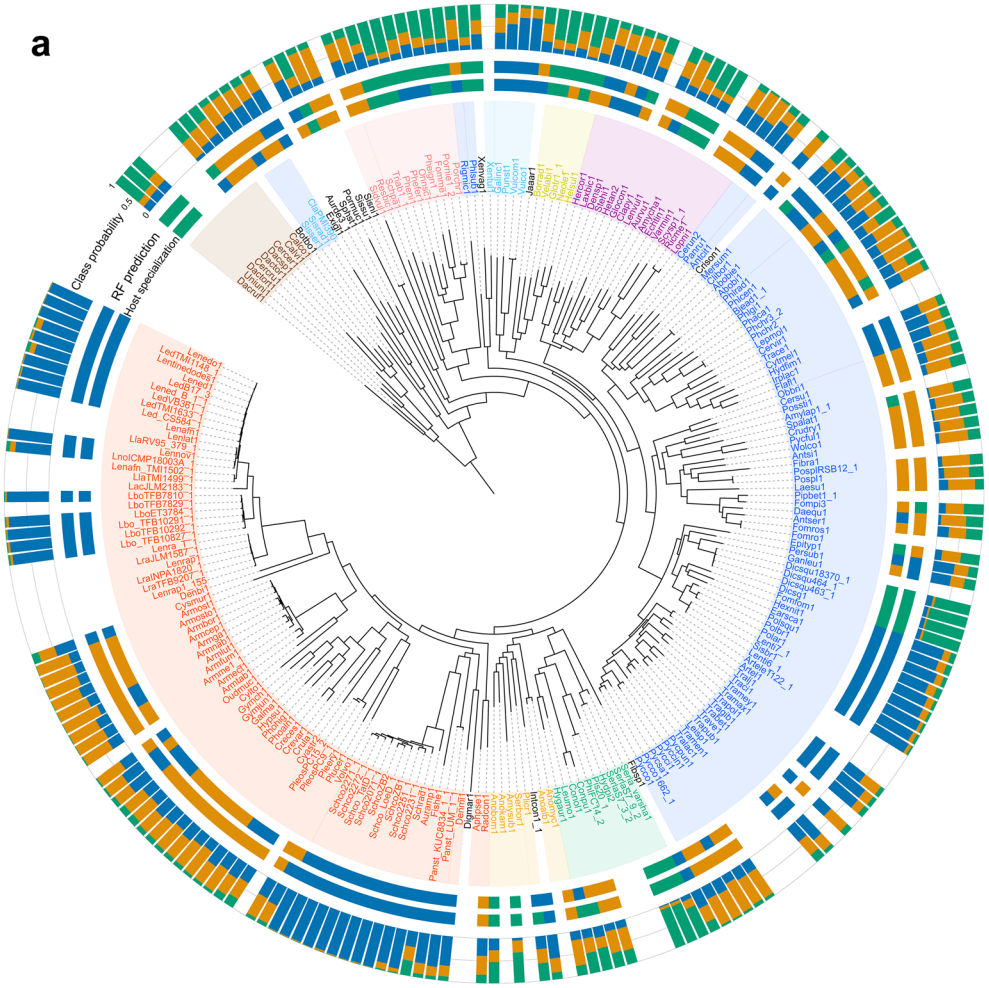


Fig. 2 Model prediction accuracy. **a**, **c** Prediction accuracy for host specificity using RF. **b**, **d** Prediction accuracy for host specificity using LightGBM. Error bars represent standard deviation. For classification tasks (**a**, **b**) the following metrics were adopted: accuracy: percentage of all test samples that were correctly predicted; recall: percentage of samples predicted to be gymnosperm specialists that actually are gymnosperm specialists; precision: percentage of gymnosperm specialist samples correctly predicted to be gymnosperm specialist; F1-score: harmonic mean of the reproducibility and goodness-of-fit rates. For regression tasks (**c**, **d**), the following metrics were adopted: coefficient of determination (R^2): proportion of variance explained by the model relative to the total variance of the dependent variable; mean absolute error (MAE): average of absolute errors; root-mean-square error (RMSE): square root of the average squared error

carbohydrates between softwood and hardwood is hemicelluloses, *O*-acetyl-glucuronoxylan is the major hemicellulose in hardwood, while *O*-acetyl-galactoglucomannan for softwood [35]. Both hemicelluloses contain acetyl side chains, but the acetyl content of softwood is limited (~1.5%) [36] and deacetylase activity of galactoglucomannan has not been discovered in CE1, suggesting that AcXE is the key for the classification.

The second most important family was commonly CBM1 in all four patterns. CBM1 is generally used for the adsorption on crystalline cellulose and typical module for cellulases. However, the domain is also connected to esterase domains such as CE1 and CE15, glucuronoyl esterases, in the genome of the white-rot fungus *Phanerochaete chrysosporium* [37], suggesting higher significance next to CE1 is reasonable. The order in third place and later differed among methods and especially among tasks. For example, CE15 was third in importance in the classification task, whether using RF or LightGBM, but was as low as 20th or below in the regression (Table S1). On the other hand, PL1 was about 10–15th in importance in classification, but was in the top 5 in regression. This may be because in regression, the contribution to prediction is evaluated uniformly in any range, while in classification, a wide range of gymnosperm associations (10–90%) is collectively considered generalist, so that only the contribution to prediction of the extreme values is evaluated as important.

While CE1 and other families related to hemicellulose and pectin degradation dominated the host specialization prediction model compared to the decay type predictions in the previous report [14], families associated with crystalline cellulose and lignin showed relatively low importance, except for CBM1 and AA9 LPMO. Lignin is known to differ in content between softwoods and hardwoods, but both AA2 peroxidases and AA1_1 laccases



b

Predicted host specialization

Actual host specialization	A	Gen	G	Overall
A		18 (20.5%)	6 (6.8%)	24 (27.3%)
Gen	7 (10.9%)		6 (9.4%)	13 (20.3%)
G	1 (3.3%)	2 (6.7%)		3 (10.0%)

Fig. 3 RF model predictions for classification tasks. **a** Comparison between true host specialization and RF predictions. Predictions are based on the average class probabilities. **b** Misclassified sample count and percentage for each host specificity group (total samples: A = 88, Gen = 64, G = 30)

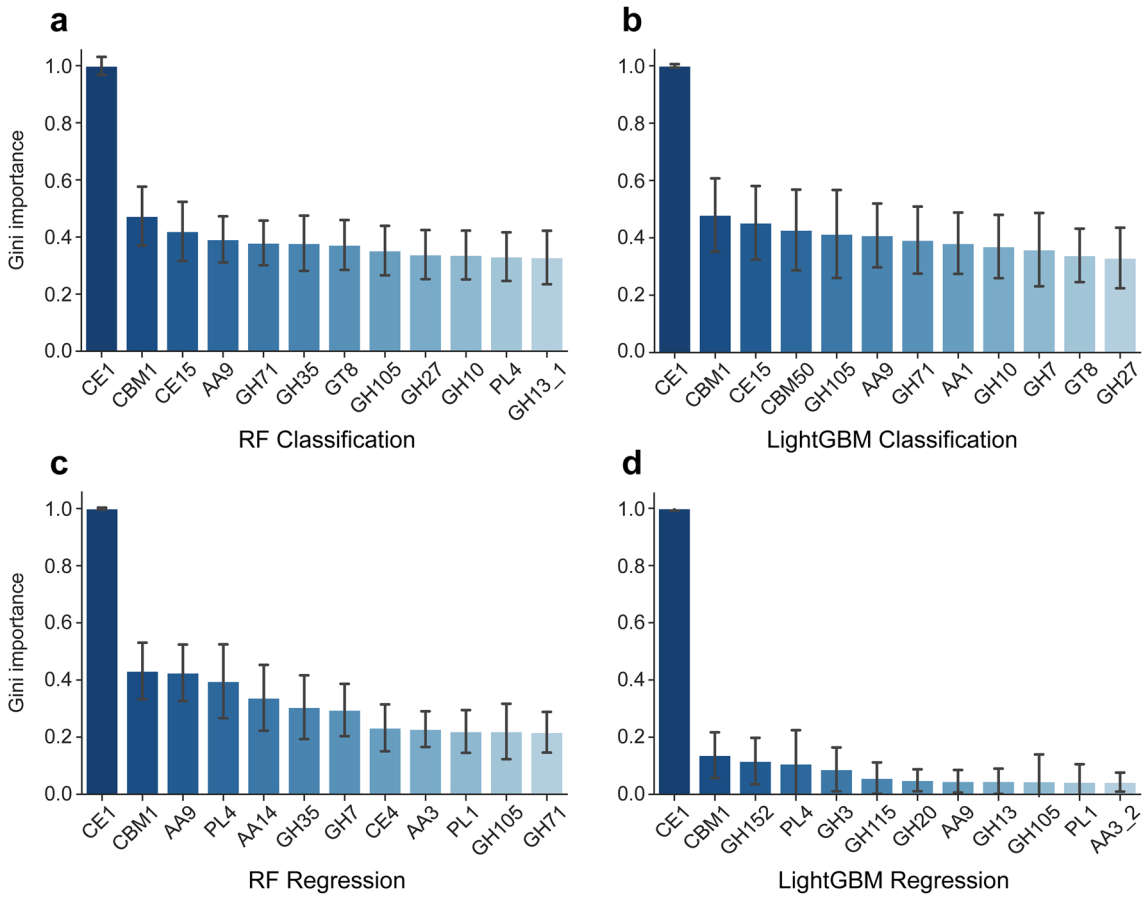


Fig. 4 Importance of each CAZy family in the model. Relative importance of each CAZy family in the model for both classification and regression tasks, using RF and LightGBM. Only the top 12 families are shown. Error bars represent standard deviation

were below the 25th rank. For wood-decay fungi, the primary target could be sugar components that are relatively easy to metabolize, and lignin should be broken down obligatorily to access these components. Interestingly, host specialization group comparisons revealed significant gene number distribution differences in white-rot fungi, but not in brown-rot fungi. This suggests that mechanisms beyond CAZymes may contribute to brown-rot fungal host specialization. The contrast between white-rot fungi, which flexibly use various CAZymes according to their targets, and brown-rot fungi, which employ alternative mechanisms, highlights the complexity of wood decay strategies.

In white-rot fungi, there were families without significant gene number distribution differences between Angiosperm specialists and generalists. This finding, coupled with the higher misclassification rates between these two groups compared to other combinations (Fig. 3b), suggests that the differences in CAZymes between Angiosperm specialists and generalists are small compared to those between these two groups and Gymnosperm specialists. Specifically, the presence of CE1 genes related to acetyl xylan degradation appears to be the critical factor limiting hardwood availability in white-rot fungal host specialization.

Conclusion

In this study, we used machine learning to establish that white-rot fungi exhibit significant differences in CAZymes composition between specialists for softwoods and those for hardwoods, with acetylxylan degradation capacity being a major distinguishing factor. In contrast, brown-rot fungi showed no significant gene number differences among host specialization groups, and the CAZymes families with high importance for decay style predictions were consistently more abundant in white-rot fungi. These findings suggest that brown-rot fungi rely on different mechanisms beyond CAZymes. To address the diverse decay styles of wood-decay fungi and to further understand brown-rot fungal decay systems, high-resolution experimental methods

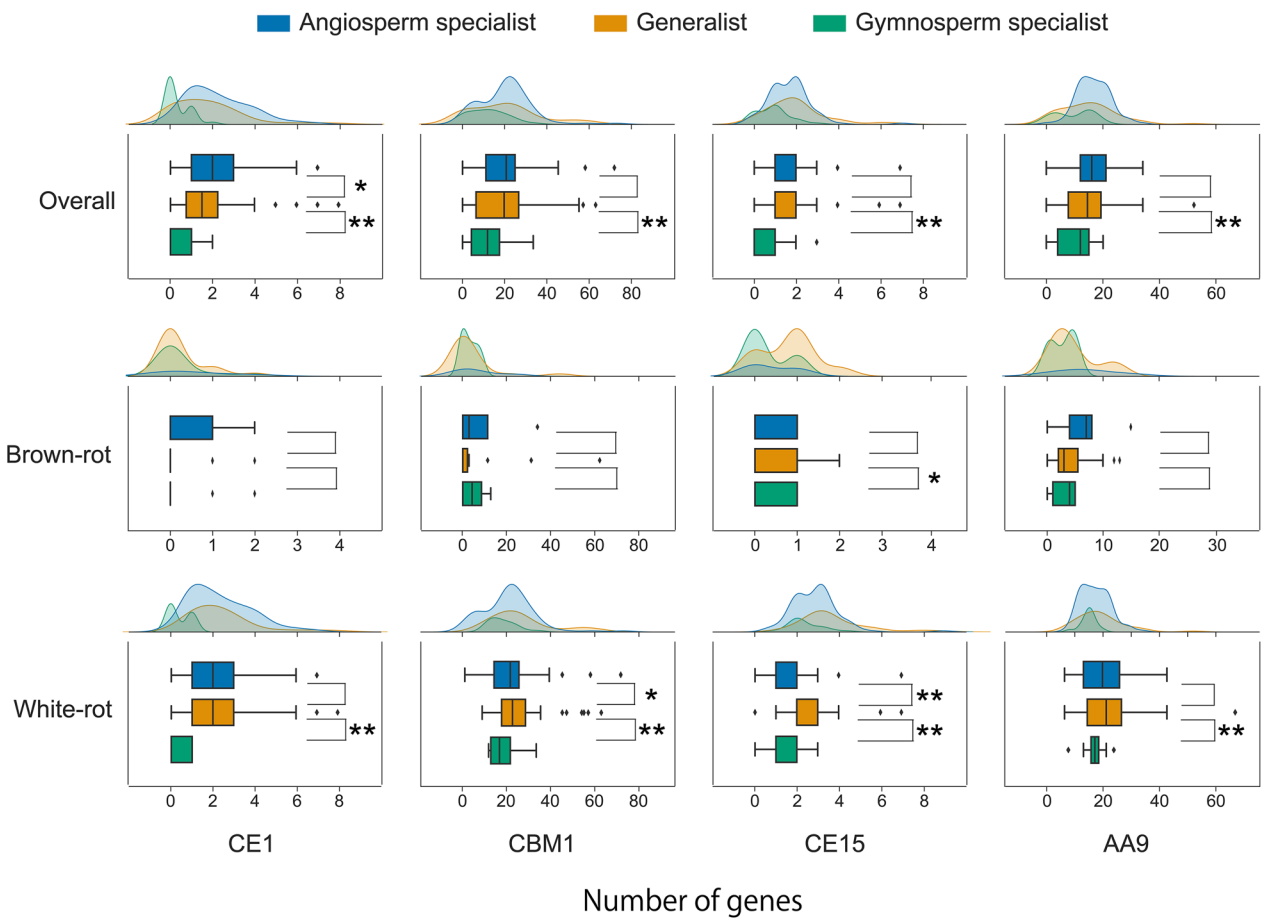


Fig. 5 Gene count distribution across host specificity groups. Box plots and kernel density estimation graphs showing gene count distribution for the top 4 families in the RF model for each host specificity group. Welch's *t*-test was performed to compare gene numbers between angiosperm specialists and generalists, and between generalists and gymnosperm specialists (denoted as *: $p < 0.05$, **: $p < 0.01$)

are needed. While our study focused on gene numbers, it will be necessary to bridge the gap between genomic data and actual decay processes by dissecting decay mechanisms temporally and spatially through various omics analyses in the future.

Abbreviations

AA	Auxiliary activity
AcXE	Acetylxylan esterase
CAZy	Carbohydrate-active enzyme
CE	Carbohydrate esterase
GH	Glycoside hydrolase
LPMO	Lytic polysaccharide monooxygenase
MAE	Mean absolute error
NCBI	National Center for Biotechnology Information
RF	Random forest
RMSE	Root-mean-square error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s10086-024-02159-9.

Supplementary Material 1.

Acknowledgements
We thank Dr. Akio Nakabayashi at Yokogawa Electric Corporation for technical discussions about the analysis.

Author contributions
NH contributed to the data analysis and draft writing, MS supported selection of machine learning algorithms, and KI designed the experiments and wrote the manuscript.

Funding
This study received financial support in the form of Grants-in-Aid for Scientific Research (A) from the Japan Society for the Promotion of Science (JSPS, No. 23H00341) to KI.

Availability of data and materials
For detailed information on the data sets used and scripts, as well as all experimental results, please refer to https://github.com/UTForestChemistryLab/rf-comparative-genomics.

Declarations

Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Received: 19 June 2024 Accepted: 25 September 2024
Published online: 03 October 2024

References

1. Eriksson K-E, Blanchette RA, Ander P (1990) Microbial and enzymatic degradation of wood and wood components. Springer series in wood science. Springer, Berlin
2. Singh AP, Singh T (2014) Biotechnological applications of wood-rotting fungi: a review. Biomass Bioenerg 62:198–206. https://doi.org/10.1016/j.biombioe.2013.12.013
3. Schultz TP, Nicholas DD, Preston AF (2007) Perspective—a brief review of the past, present and future of wood preservation. Pest Manag Sci 63(8):784–788. https://doi.org/10.1002/ps.1386
4. Momohara I, Ota Y, Nishimura T (2010) Assessment of decay risk of airborne wood-decay fungi. J Wood Sci 56(3):250–255. https://doi.org/10.1007/s10086-009-1093-6
5. Momohara I, Ota Y, Sotome K, Nishimura T (2012) Assessment of decay risk of airborne wood-decay fungi II: relation between isolated fungi and decay risk. J Wood Sci 58(2):174–179. https://doi.org/10.1007/s10086-011-1224-8
6. Momohara I, Ota Y, Yamaguchi T, Ishihara M, Takahata Y, Kosaka H (2013) Assessment of the decay risk of airborne wood-decay fungi III: decay risks at different sampling sites. J Wood Sci 59(5):442–447. https://doi.org/10.1007/s10086-013-1355-1
7. Hartig R (1878) Die Zersetzungserscheinungen des Holzes der Nadelholzbäume und der Eiche in forstlicher, botanischer und chemischer Richtung. Verlag von Julius Springer, Berlin
8. Kirk TK, Highley TL (1973) Quantitative changes in structural components of conifer woods during decay by white-and brown-rot fungi. Phytopathology 63:1338–1342. https://doi.org/10.1094/phyto-63-1338
9. Fernandez-Fueyo E, Ruiz-Dueñas FJ, Ferreira P, Floudas D, Hibbett DS, Canessa P, Larrondo LF, James TY, Seelenfreund D, Lobos S, Polanco R, Tello M, Honda Y, Watanabe T, Watanabe T, San RJ, Kubicek CP, Schmoll M, Gaskell J, Hammel KE, St John FJ, Vanden Wymelenberg A, Sabat G, BonDurant SS, Syed K, Yadav JS, Doddapaneni H, Subramanian V, Lavín JL, Oguiza JA, Perez G, Pisabarro AG, Ramirez L, Santoyo F, Master E, Coutinho PM, Henrissat B, Lombard V, Magnuson JK, Kües U, Hori C, Igarashi K, Samejima M, Held BW, Barry KW, LaButti KM, Lapidus A, Lindquist EA, Lucas SM, Riley R, Salamov AA, Hoffmeister D, Schwenk D, Hadar Y, Yarden O, de Vries RP, Wiebenga A, Stenlid J, Eastwood D, Grigoriev IV, Berka RM, Blanchette RA, Kersten P, Martinez AT, Vicuna R, Cullen D (2012) Comparative genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis. P Natl Acad Sci USA 109(14):5458–5463. https://doi.org/10.1073/pnas.1119912109
10. Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, John FS, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev IV, Hibbett DS (2012) The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336(6089):1715–1719. https://doi.org/10.1126/science.1221748
11. Hori C, Gaskell J, Igarashi K, Samejima M, Hibbett D, Henrissat B, Cullen D (2013) Genomewide analysis of polysaccharides degrading enzymes in 11 white- and brown-rot Polyporales provides insight into mechanisms of wood decay. Mycologia 105(6):1412–1427. https://doi.org/10.3852/13-072
12. Hori C, Yoshida M, Igarashi K, Samejima M (2019) Origin and diversity of wood decay fungi revealed by genome-based analyses. Mokuzai Gakkaishi 65(4):173–188
13. Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N (2022) The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50(D1):D571–D577. https://doi.org/10.1093/nar/gkab1045
14. Hasegawa N, Sugiyama M, Igarashi K (2024) Random forest machine-learning algorithm classifies white- and brown-rot fungi according to the number of the genes encoding Carbohydrate-Active enZyme families. Appl Environ Microbiol. https://doi.org/10.1128/aem.00482-24
15. Vaaje-Kolstad G, Westereng B, Horn SJ, Liu ZL, Zhai H, Sorlie M, Eijsink VGH (2010) An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. Science 330(6001):219–222. https://doi.org/10.1126/science.1192231
16. Uchiyama T, Uchihashi T, Ishida T, Nakamura A, Vermaas JV, Crowley MF, Samejima M, Beckham GT, Igarashi K (2022) Lytic polysaccharide monooxygenase increases cellobiohydrolases activity by promoting decrystallization of cellulose surface. Sci Adv 8(51):eade5155. https://doi.org/10.1126/sciadv.ade5155
17. Westereng B, Ishida T, Vaaje-Kolstad G, Wu M, Eijsink VGH, Igarashi K, Samejima M, Stahlberg J, Horn SJ, Sandgren M (2011) The putative endoglucanase *Pc*GH61D from *Phanerochaete chrysosporium* is a metal-dependent oxidative enzyme that cleaves cellulose. PLoS ONE. https://doi.org/10.1371/journal.pone.0027807
18. Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, Hammel KE, Wymelenberg AV, Gaskell J, Lindquist E, Sabat G, BonDurant SS, Larrondo LF, Canessa P, Vicuna R, Yadav J, Doddapaneni H, Subramanian V, Pisabarro AG, Lavín JL, Oguiza JA, Master E, Henrissat B, Coutinho PM, Harris P, Magnuson JK, Baker SE, Bruno K, Kenealy W, Hoegger PJ, Kües U, Ramaiya P, Lucash S, Salamov A, Shapiro H, Tu H, Chee CL, Misra M, Xie G, Teter S, Yaver D, James T, Mokrejs M, Pospisek M, Grigoriev IV, Brettin T, Rokhsar D, Berka R, Cullen D (2009) Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. P Natl Acad Sci USA 106(6):1954–1959. https://doi.org/10.1073/pnas.0809575106
19. Purhonen J, Ovaskainen O, Halme P, Komonen A, Huhtinen S, Kotiranta H, Læssoe T, Abrego N (2020) Morphological traits predict host-tree specialization in wood-inhabiting fungal communities. Fungal Ecol. https://doi.org/10.1016/j.funeco.2019.08.007
20. Krah FS, Bässler C, Heibl C, Soghigian J, Schaefer H, Hibbett DS (2018) Evolutionary dynamics of host specialization in wood-decay fungi. Bmc Evol Biol. https://doi.org/10.1186/s12862-018-1229-7
21. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
22. Cline ET, Farr DF (2006) Access to web-based information about fungal geographic distribution, host range, and scientific names using the USDA-ARS Systematic Botany and Mycology (SBML) databases: What can SBML do for you? Phytopathology 96(6):S190
23. GBIF.org (2022) GBIF Home Page. https://www.gbif.org
24. IPNI (2022) International Plant Names Index. The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Botanic Gardens. http://www.ipni.org
25. Tropicos.org (2022) Tropicos. https://tropicos.org
26. FitzJohn RG, Pennell MW, Zanne AE, Stevens PF, Tank DC, Cornwell WK (2014) How much of the world is woody? J Ecol 102(5):1266–1272. https://doi.org/10.1111/1365-2745.12260
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Int Res 16(1):321–357
28. Branco P, Torgo L, Ribeiro R (2017) SMOGN: a pre-processing approach for imbalanced regression. Proc Mach Learn Res 74:36–50
29. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and, Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, December 2017, 3149–3157.
30. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: Kdd'19: proceedings of the 25th Acm Sigkdd international conference on knowledge

discovery and data mining. p. 2623–2631. https://doi.org/10.1145/3292500.3330701
31. Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, Harrison RJ, Holub E, Sukno SA, Sreenivasaprasad S, Thon MR (2016) Gene family expansions and contractions are associated with host range in plant pathogens of the genus. BMC Genomics. https://doi.org/10.1186/s12864-016-2917-6
32. Yoshida K, Saunders DGO, Mitsuoka C, Natsume S, Kosugi S, Saitoh H, Inoue Y, Chuma I, Tosa Y, Cano LM, Kamoun S, Terauchi R (2016) Host specialization of the blast fungus *Magnaporthe oryzae* is associated with dynamic gain and loss of genes linked to transposable elements. BMC Genomics. https://doi.org/10.1186/s12864-016-2690-6
33. Lofgren LA, Nguyen NH, Vilgalys R, Ruytinx J, Liao HL, Branco S, Kuo A, LaButti K, Lipzen A, Andreopoulos W, Pangilinan J, Riley R, Hundley H, Na HS, Barry K, Grigoriev IV, Stajich JE, Kennedy PG (2021) Comparative genomics reveals dynamic genome evolution in host specialist ectomycorrhizal fungi. New Phytol 230(2):774–792. https://doi.org/10.1111/nph.17160
34. Kojima K, Sunagawa N, Yoshimi Y, Tryfona T, Samejima M, Dupree P, Igarashi K (2022) Acetylated xylan degradation by glycoside hydrolase family 10 and 11 xylanases from the white-rot fungus *Phanerochaete chrysosporium*. J Appl Glyosci 69(2):35–43. https://doi.org/10.5458/jag.jag.JAG-2021_0017
35. Perez J, Munoz-Dorado J, de la Rubia T, Martinez J (2002) Biodegradation and biological treatments of cellulose, hemicellulose and lignin: an overview. Int Microbiol 5(2):53–63. https://doi.org/10.1007/s10123-002-0062-3
36. Pawar PM, Koutaniemi S, Tenkanen M, Mellerowicz EJ (2013) Acetylation of woody lignocellulose: significance and regulation. Front Plant Sci 4:118. https://doi.org/10.3389/fpls.2013.00118
37. Vanden Wymelenberg A, Minges P, Sabat G, Martinez D, Aerts A, Salamov A, Grigoriev I, Shapiro H, Putnam N, Belinky P, Dosoretz C, Gaskell J, Kersten P, Cullen D (2006) Computational analysis of the Phanerochaete chrysosporium v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins. Fungal Genet Biol 43(5):343–356. https://doi.org/10.1016/j.fgb.2006.01.003

Publisher's Note