AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

| Applied and Industrial Microbiology | Research Article

Whole genome structural predictions reveal hidden diversity in putative oxidative enzymes of the lignocellulose-degrading ascomycete *Parascedosporium putredinis* NO1

Conor J. R. Scott,[1] Daniel R. Leadbeater,[1] Nicola C. Oates,[1] Sally R. James,[2] Katherine Newling,[2] Yi Li,[2] Nicholas G. S. McGregor,[3] Susannah Bird,[1] Neil C. Bruce[1]

AUTHOR AFFILIATIONS See affiliation list on p. 18.

ABSTRACT Economic valorization of lignocellulose is paramount to realizing a true circular bioeconomy; however, this requires the development of systems and processes to expand the repertoire of bioproducts beyond current renewable fuels, chemicals, and sustainable materials. *Parascedosporium putredinis* NO1 is an ascomycete that thrived at the later stages of a wheat-straw composting community culture, indicating a propensity to degrade recalcitrant lignin-enriched biomass, but exists within an underrepresented and underexplored fungal lineage. This strain has been proven to be an exciting candidate for the identification of new enzymes targeting recalcitrant components of lignocellulose following the recent discovery of a new lignin β-ether linkage cleaving enzyme. The first genome for the genus *Parascedosporium* for *P. putredinis* NO1 genome was sequenced, assembled, and annotated. The genome is 39 Mb in size, consisting of 21 contigs annotated to contain 9.998 protein-coding sequences. The carbohydrate-active enzyme (CAZyme) repertoire was compared to 2570 ascomycete genomes and in detail with *Trichoderma reesei*, *Fusarium oxysporum*, and sister taxa *Scedosporium boydii*. Significant expansion in the oxidative auxiliary activity class of CAZymes was observed in the *P. putredinis* NO1 genome, resulting from increased sequences encoding putative lytic polysaccharide monooxygenases (LPMOs), oxidative enzymes acting within LPMO redox systems, and lignin-degrading laccases. *P. putredinis* NO1 scored above the 95th percentile for AA gene density across the ascomycete phylum, suggesting a primarily oxidative strategy for lignocellulose breakdown. Novel structure-based searching approaches were employed, revealing 17 new sequences with structural similarity to LPMO, laccase, and peroxidase sequences and which are potentially new lignocellulose-degrading enzymes.

IMPORTANCE An annotated reference genome has revealed *P. putredinis* NO1 as a useful resource for the identification of new lignocellulose-degrading enzymes for biorefining of woody plant biomass. Utilizing a "structure-omics"-based searching strategy, we identified new potentially lignocellulose-active sequences that would have been missed by traditional sequence searching methods. These new identifications, alongside the discovery of novel enzymatic functions from this underexplored lineage with the recent discovery of a new phenol oxidase that cleaves the main structural β-O-4 linkage in lignin from *P. putredinis* NO1, highlight the underexplored and poorly represented family Microascaceae as a particularly interesting candidate worthy of further exploration toward the valorization of high value biorenewable products.

KEYWORDS *Parascedosporium*, ascomycete, CAZymes, auxiliary activity, oxidative, lignocellulose, lignin, AlphaFold, structural, structure-omics

Editor Chang-Jun Cha, Chung-Ang University, Anseong, Yeonggi-do, South Korea

Address correspondence to Conor J. R. Scott, cs1535@york.ac.uk.

The authors declare no conflict of interest.

See the funding table on p. 18.

Received 31 March 2023
Accepted 22 August 2023
Published 9 October 2023

Energy consumption continues to grow rapidly alongside improvements in living standards, and fossil fuels continue to play a major role in industrial and agricultural sectors. With their widely accepted environmentally damaging effects, the need to move away from the use of fossil fuels and toward a net zero carbon fuel source is ever more pressing. Lignocellulosic residues consisting of cellulose, hemicellulose, and lignin with minor amounts of pectins and nitrogen compounds offer the largest source of biomass for liquid fuel, chemicals, and energy (1). However, biorefining of lignocellulose has so far been limited by the recalcitrant nature of the intricate and insoluble lignin network (2, 3).

Fungi are exceptional wood degraders and are predominantly used to produce an array of bioproducts, including commercial enzyme cocktails used in biological processing of lignocellulosic biomass. Ascomycetes, known as soft-rot fungi, degrade lignocellulose by penetration of plant secondary cell walls with hyphae that secrete complex enzyme cocktails in abundance at the site of attack (4). *Parascedosporium putredinis* NO1 is a soft-rot ascomycete identified previously as dominant in the later stages of a mixed microbial compost community grown on wheat straw (5). This behavior suggests that the fungus can efficiently deconstruct and potentially metabolize the more recalcitrant carbon sources in the substrate. Indeed, the recent discovery of a new oxidase enzyme that cleaves the major β-ether units in lignin in the *P. putredinis* NO1 secretome, which releases the pharmaceutically valuable compound tricin from wheat straw while simultaneously enhancing digestibility of the biomass (5), promotes a requirement for further exploration of this taxa.

Here, an annotated reference genome for *P. putredinis* NO1 reveals a repertoire of carbohydrate-active enzymes (CAZymes) and oxidative enzymes focused on degrading the most recalcitrant components of lignocellulose. Comparisons across the ascomycete tree of life suggest an increased proportion of oxidative enzymes within the CAZyme repertoire of *P. putredinis* NO1. Further investigation through CAZyme repertoire comparison with two other industrially relevant wood-degrading ascomycetes, *Trichoderma reesei* and *Fusarium oxysporum*, as well as sister taxa *Scedosporium boydii*, reveals expansion in families of enzymes with roles in the oxidative dissolution of lignocellulose and demonstrates this fungus to be an exciting candidate for the identification of new lignocellulose-degrading activities.

Novel approaches were used to search the *P. putredinis* NO1 genome for potentially unannotated enzyme sequences with relation to three types of classic oxidative lignocellulose degraders: lytic polysaccharide monooxygenases (LPMOs), laccases, and peroxidases. Predicted structures were obtained for >96% of the protein-coding sequences in the genome. Structural searches were found to be effective at identifying multiple sequences for potentially novel proteins involved in lignocellulose breakdown, which had low levels of structural similarity to the classic oxidative lignin and crystalline cellulose-degrading enzymes. These sequences were also missed by sequence and domain-oriented searches. Further investigation and comparison of structures revealed varying levels of structural overlap despite the lack of sequence similarity. This strategy of combining search approaches can be adopted to identify divergent enzyme sequences which may have alternate lignocellulose-degrading activity, variation in substrate specificity, and different temperature and pH optima.

Further investigation and characterization of such lignocellulose-degrading enzymes add to the wealth of enzymes which can be incorporated into commercial enzyme cocktails to improve their effectiveness and boost the efficiency at which biomass is converted to renewable liquid fuel and value-added chemicals.

RESULTS AND DISCUSSION

The genome of *P. putredinis* NO1 suggests a strategy to degrade the most recalcitrant components of lignocellulose

The *P. putredinis* NO1 genome was sequenced using nanopore sequencing with the Oxford Nanopore Technologies' (ONT) MinION system to avoid errors in the assembly and annotation of coding regions resulting from long regions of repetitive DNA in

eukaryotic genomes (6). The genome is 39 Mb in size, and the assembly consists of 21 contigs, containing 9998 protein-coding sequences. To investigate the lignocellulose-degrading enzyme repertoire of the *P. putredinis* NO1 genome, all protein-coding sequences were annotated for CAZyme domains using the dbCAN server (7). In total, 795 CAZyme domains were predicted in the *P. putredinis* NO1 genome, and the distribution of these domains across the CAZyme classes can be seen in Fig. 1A. Glycoside hydrolases (GHs) are the most abundant CAZyme class with 290 identified. While auxiliary activities (AAs) also make a large contribution with 162 domains. Glycosyl transferases contribute 113 domains, and carbohydrate esterases contribute 51. Polysaccharide lyases contribute the fewest, with only 18 domains. In addition to these catalytic classes, 161 carbohydrate-binding modules (CBMs) were also identified.

An interesting observation was the proportionally high number of AA class CAZymes observed in the *P. putredinis* NO1 genome. To investigate this further and more broadly within the scope of the ascomycete tree of life, CAZyme profiles of all available ascomycete genomes were elucidated (Fig. 2). To filter poorly represented lineages, genera and families with less than three and eight species-level representatives, respectively,

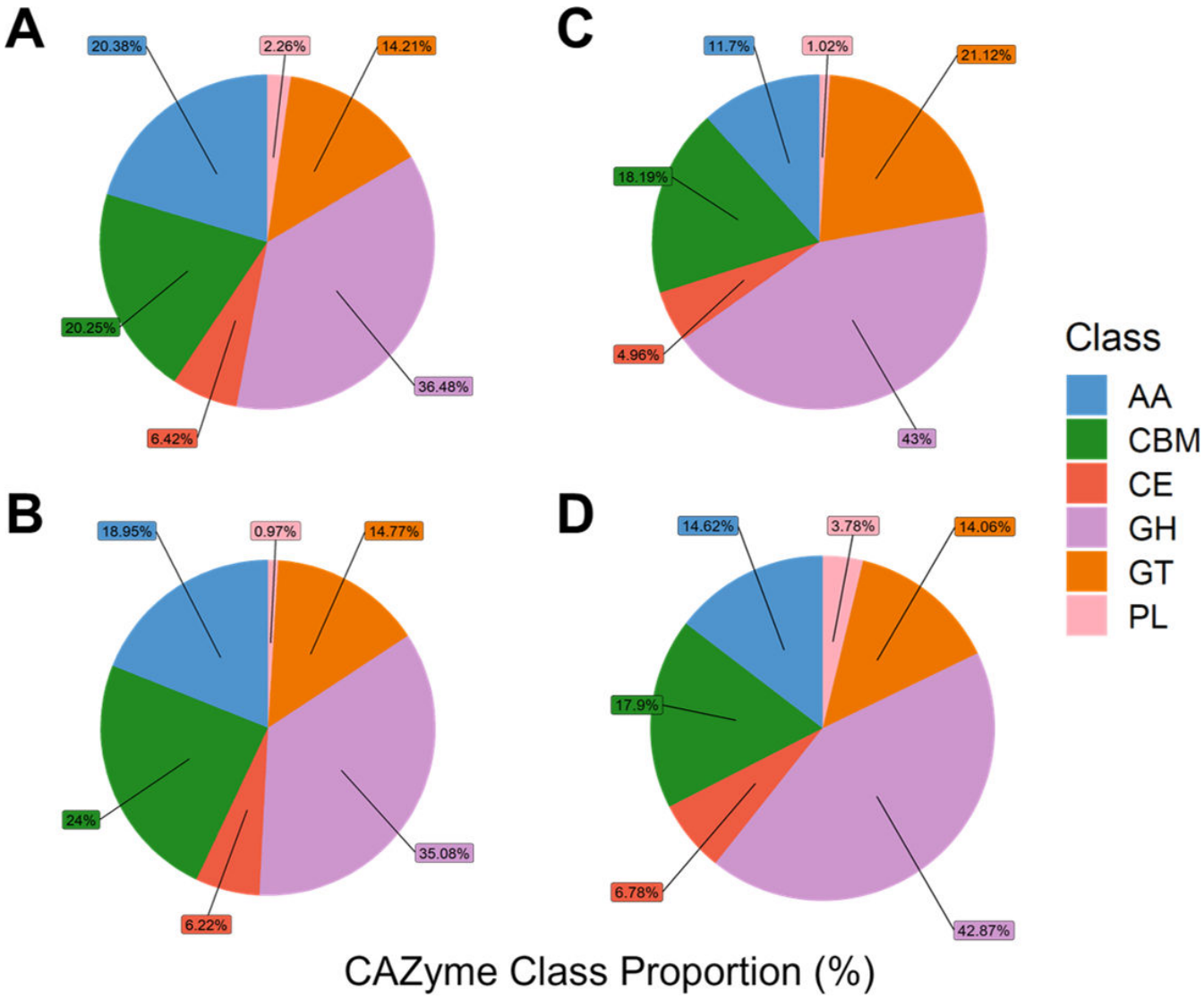

FIG 1 Comparison of CAZyme class repertoire. The proportions of each class of CAZyme contributing to CAZyme repertoire for *P. putredinis* NO1 (**A**), *S. boydii* (**B**), *T. reesei* (**C**), and *F. oxysporum* (**D**). AA, auxiliary activity; CBM, carbohydrate-binding module; CE, carbohydrate esterase; GH, glycoside hydrolase; GT, glycosyl transferase; PL, polysaccharide lyase.

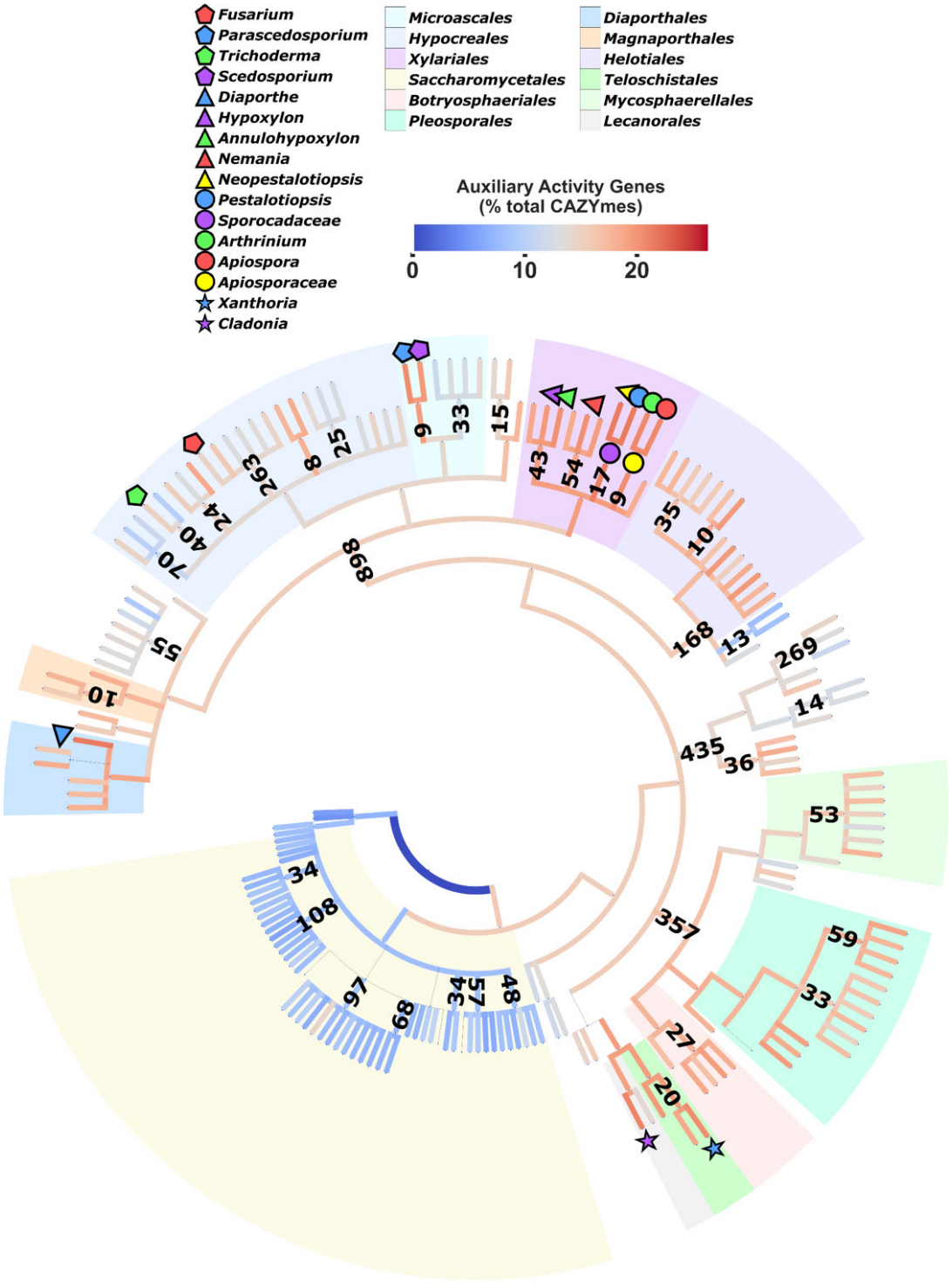

FIG 2 Auxiliary activity distribution and density across the ascomycete tree of life. Genes predicted for ascomycete genome assemblies were annotated for CAZymes to explore patterns in the distribution and density of auxiliary activities (n = 2,570) within the ascomycete phylogenetic tree. Branch colors indicate the mean proportion of auxiliary activities within only the CAZyme annotations accounted for by all descendant taxa. Numerical clade annotations represent the number of sequenced genomes available. Key taxa, including lineages above the 95th percentile for AA proportion, have been highlighted. Genera and families with less than three and eight species-level representatives, respectively, have been pruned for clarity (n = 462 taxa). Nodes of taxonomic ranks below genus have been pruned (n = 93).

were filtered. It was clear that *P. putredinis* NO1 has one of the highest proportions of AA class CAZymes within its repertoire relative to total CAZymes among ascomycete fungi. *P. putredinis* NO1 was a substantial outlier within the order Microascales (14.19 ± 3.15) and belonged above the 95th percentile for AA gene density among the highest AA populated genomes (20.38%), behind the genera *Diaporthe* (21.67 ± 1.3%) of the order Diaporthales, *Cladonia* (20.9 ± 1.74%) of the order Lecanorales, *Xanthoria* (20.55 ± 1.33%) of the order Teloschistales, and members belonging to the highly AA-enriched order Xylariales (19.36 ± 1.98%) containing contributions from densely populated genera *Hypoxylon* (19.78 ± 1.79%), *Annulohypoxylon* (20.09 ± 1.1%), *Nemania* (20.25 ± 1.82%), *Neopestalotiopsis* (21.56 ± 0.66%), *Pestalotiopsis* (21.49 ± 0.97%), *Arthrinium* (20.67 ± 2.13%) and *Apiospora* (20.87 ± 1.82%). The *Parascedosporium* sister taxa *Scedosporium* (19.73 ± 1.07%) exhibited slightly lower AA density and belonged above the 90th percentile. Interestingly, Saccharomycetales (8.95 ± 4.63%), often associated with lignocellulose deconstruction, displayed significantly reduced AA abundance in stark contrast to neighboring phylogenies such as Teloschistales (19.37 ± 1.6%) and Pleosporales (17.97 ± 1.98%). Members of the orders Helotiales (17.32 ± 2.36%), Botryosphaeriales (17.86 ± 1.3%), Magnaporthales (17.47 ± 0.98), and Diaporthales (18.78 ± 2.86) displayed a degree of enrichment of AAs, while members belonging to Hypocreales (13.93 ± 4.55%) and Mycosphaerellales (15.21 ± 2.89%) displayed lower abundances. Considering how the AA class of enzymes is predominantly associated with the degradation of lignin and crystalline cellulose, it highlights a potential strategy of the fungus to target these components. Indeed, in a mixed microbial community grown on wheat straw, the fungus was observed to become more dominant in the later stages of the culture, potentially due to its capacity to modify the more difficult to degrade components of lignocellulose for growth (5). The gene density distribution here highlights promising candidate lineages for further exploration for additional insights into lignin and cellulose turnover.

Within white- and brown- (basidiomycete), and soft-rot (ascomycete) fungi, it has been demonstrated that the CAZyme repertoire can vary greatly from species to species (8). To investigate the repertoire of *P. putredinis* NO1 in more detail, CAZyme domains were compared to that of three other wood-degrading ascomycetes. *Scedosporium boydii* is located within the sister taxon of *Parascedosporium* and has a genome of 43 Mb containing 1029 CAZyme domains. The genome and CAZyme complement of the soft-rot *P. putredinis* NO1 are larger than that of *Trichoderma reesei*, which contains 786 domains in 34 Mb of DNA. *T. reesei* is a mesophilic soft-rot fungus known for its ability to produce high titres of polysaccharide-degrading enzymes that are used in biomass-degrading enzyme cocktails (9). The genome of *P. putredinis* NO1 is slightly smaller than that of *Fusarium oxysporum* at 47 Mb, a phytopathogenic fungus containing an expanded CAZyme repertoire of 1430 domains (10). The lignocellulose-degrading activities of *F. oxysporum* have been well investigated in part due to its pathogenicity and ability to ferment sugars from lignocellulose breakdown directly into ethanol (11, 12).

Examining the distribution of predicted CAZyme domains revealed that despite the similar overall number of CAZyme domains for *P. putredinis* NO1 and *T. reesei*, the proportion of AA class CAZyme domains is much higher in the genome of *P. putredinis* NO1 (Fig. 1A). Proportionally, AA class CAZymes make the largest contribution to the CAZyme repertoire in the genome of *P. putredinis* NO1 compared to the other ascomycetes (Fig. 1). This again could suggest an oxidative strategy to target to lignin and crystalline cellulose. Although analysis of fungal secretomes would be required to confirm an improved ability of *P. putredinis* NO1 to deconstruct lignocellulosic components, the high potential capacity for degradation of lignin and crystalline cellulose within the genome suggests that this is an important fungus to explore for new lignocellulose-degrading enzymes. This is especially relevant considering that this is the first genome assembly of the genus *Parascedosporium*.

The increased contribution of AA class CAZymes is mirrored by a reduced proportion of GH class CAZymes in the *P. putredinis* NO1 genome compared to *T. reesei* and *F.*

oxysporum. This reduced GH contribution is also visible in the genome of *S. boydii*, a close relative of *P. putredinis* NO1. Despite the reduced number of the hydrolytic GH class CAZymes, the repertoires of *P. putredinis* NO1 and *S. boydii* contain the highest proportions of CBMs, domains typically associated with hydrolytic CAZymes such as GHs (13), but which have also been observed in oxidative LPMOs (14, 15). The increased proportion of CBMs in the genome of *P. putredinis* NO1 could aid the catalytic CAZymes in accessing and binding to these substrates. Indeed, examining the CBM domains at the family level shows a high number of crystalline cellulose-binding domains in the genome of both *P. putredinis* NO1 and *S. boydii*, much higher than the number of domains assigned to any of the other CBM families (Fig. S1).

Closer investigation of the AA CAZyme repertoire reveals more about the lignocellulose-degrading strategy of *P. putredinis* NO1

The high number of AA domains, a functional class that notably contains LPMOs, peroxidases, and laccases (2), in the genome of *P. putredinis* NO1 is likely to endow this fungus with the ability to degrade recalcitrant components of the plant cell wall through a primarily oxidative mechanism. LPMOs are copper-containing enzymes that enhance polysaccharide degradation by generating new sites for attack by hydrolytic CAZymes (16). LPMOs have been shown to act on all major polysaccharide components of lignocellulose. Their oxidative action relies on exogenous electron donors provided by other AA family CAZymes, small molecule reductants and even lignin (2, 16). It has recently been demonstrated that LPMOs readily utilize hydrogen peroxide (H_2O_2) as a co-substrate also (17, 18).

Investigating the distribution of AA domains across the AA families revealed AA9 family members to be the most abundant in the *P. putredinis* NO1 genome with 35 domains, the highest in the four ascomycetes investigated here (Fig. S2). This family contains the cellulose, xylan, and glucan-active LPMOs described above (19). AA3 and AA3_2 domains are the second and third most abundant families in the *P. putredinis* NO1 genome with 29 and 27 domains, respectively. These are flavoproteins of the glucose-methanol-choline oxidoreductase family, which includes activities such as cellobiose dehydrogenase, glucose-1-oxidase, aryl alcohol oxidase, alcohol oxidase, and pyranose oxidase (20). It is proposed that flavin-binding oxidative enzymes of this family play a central role in spatially and temporally supplying H_2O_2 to LPMOs and peroxidases or to produce radicals that degrade lignocellulose through Fenton chemistry (17). The *P. putredinis* NO1 genome also contains 12 AA7 family domains, the family of glucooligosaccharide oxidase enzymes. These have recently been demonstrated to transfer electrons to AA9 LPMOs, which boosts cellulose degradation (21). Altogether, the apparent expansion of these LPMO system families suggests a potentially increased capacity for *P. putredinis* NO1 to oxidatively target crystalline cellulose.

The genome of *P. putredinis* NO1 also contains 12 AA1 family CAZyme domains. This family includes laccase and multi-copper oxidase enzymes which catalyze the oxidation of various aromatic substrates while simultaneously reducing oxygen to water (22). It has also been demonstrated that laccases can boost LPMO activity through the release of low molecular weight lignin polymers from biomass, which can in turn donate electrons to LPMOs (23). Additionally, seven domains belonging to the AA8 family were identified, a family of iron reductase domains initially identified as the N-terminal domain in cellobiose dehydrogenase enzymes but also found independently and appended to CBMs (2, 24, 25). These domains are believed to be involved in the generation of reactive hydroxyl radicals that can indirectly depolymerize lignin. There are six AA4 domains in the genome of *P. putredinis* NO1, the highest number of the four ascomycetes investigated here. These are vanillyl-alcohol oxidase enzymes with the ability to catalyze the conversion of a wide range of phenolic oligomeric compounds (26). These may act downstream of the lignin depolymerization catalyzed by other members of the AA class. There is a clear capacity in the *P. putredinis* NO1 genome for lignin depolymerization and metabolism through the multiple domains identified belonging to these families. The *P.*

putredinis NO1 genome also contains two AA16 domains, a recently identified family of LPMO proteins with an atypical product profile compared to the traditional AA9 family of LPMOs and a potentially different mode of activation (27).

Gene expression of CAZymes in the *P. putredinis* NO1 genome has been explored previously during growth on glucose, compared to growth on wheat straw with samples taken at days 2, 4, and 10 (5). This transcriptomic data give a view of the potential strategy by which *P. putredinis* NO1 utilizes its expanded repertoire of AA class CAZymes. Upregulation of AA class CAZymes during growth on wheat straw compared to growth on glucose was observed predominantly at day 4 and then gave way to upregulation instead of mainly GH class hydrolytic CAZymes at day 10. This could represent a strategy where the recalcitrant lignin and crystalline cellulose are targeted first by LPMOs and lignin degraders such as laccases, making the polysaccharide substrates of hydrolytic GH enzymes more accessible.

Searching the *P. putredinis* NO1 genome for new oxidative lignocellulose-degrading enzymes with sequence-, domain-, and structural-based strategies

Due to the evidence of a strategy for *P. putredinis* NO1 to target the most recalcitrant components of lignocellulose and the recent discovery of a new oxidase with the ability to cleave the major linkage in lignin from this strain (5), it was hypothesized that the genome of this fungus contains additional new enzymes for the breakdown of plant biomass. Particularly, this fungus could contain new enzymes with roles in degrading the lignin and crystalline cellulose components and which have not been annotated as CAZymes in this analysis.

Traditionally, homolog searching has been performed using a sequence-based approach (28). This approach uses either the primary amino acid sequence of an example protein to search an unknown database for similar sequences, or uses hidden Markov models (HMMs) to search for domains of interest (29). However, both techniques rely on primary amino acid sequence homology and neglect that proteins with distantly related sequences may have similar three-dimensional structures and therefore activity. The recent emergence of AlphaFold provides a resource for the fast and accurate prediction of unknown protein structures (30). Using this tool, structures were predicted for >96% of the protein-coding regions of the *P. putredinis* NO1 genome. These structures were used to create a database of protein structures into which structures of interesting enzymes such as those for LPMOs, laccases, and peroxidases could be searched. These structural searches for new enzymes were performed alongside sequence- and domain-based searches for comparison of the ability to identify interesting new candidates.

LPMO-related sequences were searched for in the *P. putredinis* NO1 genome using the sequence of an AA9 family LPMO from *Aspergillus niger* with the default *E*-value cut off of 1×10^{-5}, with the AA9 HMM from Pfam and considering domain hits that fell within the default significance inclusion threshold of 0.01 (31), and the structure of the same *A. niger* LPMO with a tailored 'lowest percentage match' parameter. In total, 49 sequences were identified across the three searching strategies, and 33 of these sequences were also annotated by dbCAN as AA9 family of LPMOs (Table S1). With the objective of identifying new enzymes, the remaining 16 sequences were investigated further, and the distribution of the identification of these sequences across the three search strategies can be seen in Table 1. Two of the sequences, PutMoI and PutMoM, were identified by all three search approaches. These sequences both had conserved signal peptides with a conserved N-terminal histidine after the cleavage site, a characteristic feature of LPMOs (32).

When creating the structure database, it was tempting to filter predicted structures by pLDDT score, the AlphaFold metric for prediction confidence, to create a database solely of "high confidence" structures (30). However, pLDDT scores reflect local confidence and should instead be used for assessment of individual domains (33). The majority of the structures generated here had pLDDT scores of over 60%; however,

TABLE 1 Identifying LPMO-related proteins encoded in the *P. putredinis* NO1 genome[a]

Coding region	GenBank accession	Protein ID	Identified by searching approach			InterPro annotation
			Sequence	Domain	Structure	
FUN_000653-T1	CAI7987917.1	PutMoA			✔	AA16 LPMO
FUN_000713-T1	CAI7987978.1	PutMoB			✔	Rho factor associated
FUN_002573-T1	CAI7991617.1	PutMoC		✔		–
FUN_002890-T1	CAI7992277.1	PutMoD			✔	AA16 LPMO
FUN_002962-T1	CAI7992399.1	PutMoE			✔	–
FUN_003190-T1	CAI7992922.1	PutMoF			✔	Ferritin-like
FUN_003535-T1	CAI7993628.1	PutMoG		✔		AA13 LPMO
FUN_003783-T1	CAI7994168.1	PutMoH			✔	–
FUN_006366-T1	CAI7999797.1	PutMoI	✔	✔	✔	AA9 LPMO
FUN_006413-T1	CAI7999893.1	PutMoJ		✔		AA9 LPMO
FUN_006553-T1	CAI8000144.1	PutMoK			✔	–
FUN_007242-T1	CAI8001774.1	PutMoL		✔		AA9 LPMO
FUN_007666-T1	CAI8002525.1	PutMoM	✔	✔	✔	AA9 LPMO
FUN_008106-T1	CAI8003467.1	PutMoN	✔	✔		AA9 LPMO
FUN_009239-T1	CAI7992001.1	PutMoO			✔	–
FUN_010012-T1	CAI8003342.1	PutMoP			✔	–

[a]Coding regions of proteins related to LPMOs identified through genome searching approaches with the sequence of an *A. niger* AA9 LPMO (*E*-value cutoff = 1×10^{-5}), the Pfam AA9 HMM (significance threshold = 0.01), and the structure of the *A. niger* AA9 LPMO (lowest percentage match = 50%) and which were not annotated as AA9 CAZymes by dbCAN. InterPro annotations were retrieved where possible

pLDDT scores lower than 70% are considered low confidence (Fig. S3). Extracellular enzymes are of particular interest here, but these often have disordered N-terminal signal peptides which can reduce the overall pLDDT scores. Therefore, for secreted enzyme identification from AlphaFold structures, it is inappropriate to filter by pLDDT score. Indeed, the PutMoI structure mentioned above had a pLDDT score of 62%, considered to be low confidence (30), but which had characteristic features of LPMOs and which demonstrated structural similarity to the *A. niger* AA9 LPMO used for structural searches (Fig. 3A and B). The central beta-sheet structures align well to the *A. niger* AA9 LPMO for both PutMoI and PutMoM, but both also have additional loops of disordered protein which likely explains the relatively low PDBefold alignment confidence scores (*Q*-scores) of 0.23 and 0.34 for PutMotI and PutMoM, respectively. This again highlights the unreliability of structural confidence scores alone and demonstrates how manual inspection of structural alignments may prove more useful. Despite not being annotated as AA9 LPMOs by the dbCAN server for CAZyme annotation (7), both sequences were identified using the Pfam AA9 HMM and appear to be conserved AA9 LPMOs and, therefore, are not of interest in the discovery of new enzymes.

By utilizing multiple searching approaches, potentially new sequences with LPMO-related activities can be identified. When searching for LPMO-related sequences, domain-based approaches identified all coding regions also identified by sequence-based searching as well as additional coding regions (Table S1). This pattern of domain-based searching identifying more coding regions than sequence-based searching was also observed for the other activities investigated (Tables S2 and S3). For structure-based searching, parameters of the searches could be tailored to identify additional coding regions with lower overall structural similarity but which may still be interesting. For example, searching the against the *P. putredinis* NO1 genome structure database with the structure of the *A. niger* AA9 LPMO and with the "lowest acceptable match" parameter, which is the cutoff at which secondary structures must overlap between a query and a target set at 50%, yielded 30 coding regions (Table S1). Of these sequences, nine were not identified by the sequence or domain-based searching approaches and were investigated in more detail (Table 1). To investigate these further, sequences were searched against the National Center for Biotechnology Information (NCBI) non-redundant protein database to identify related sequences (34); conserved domains were

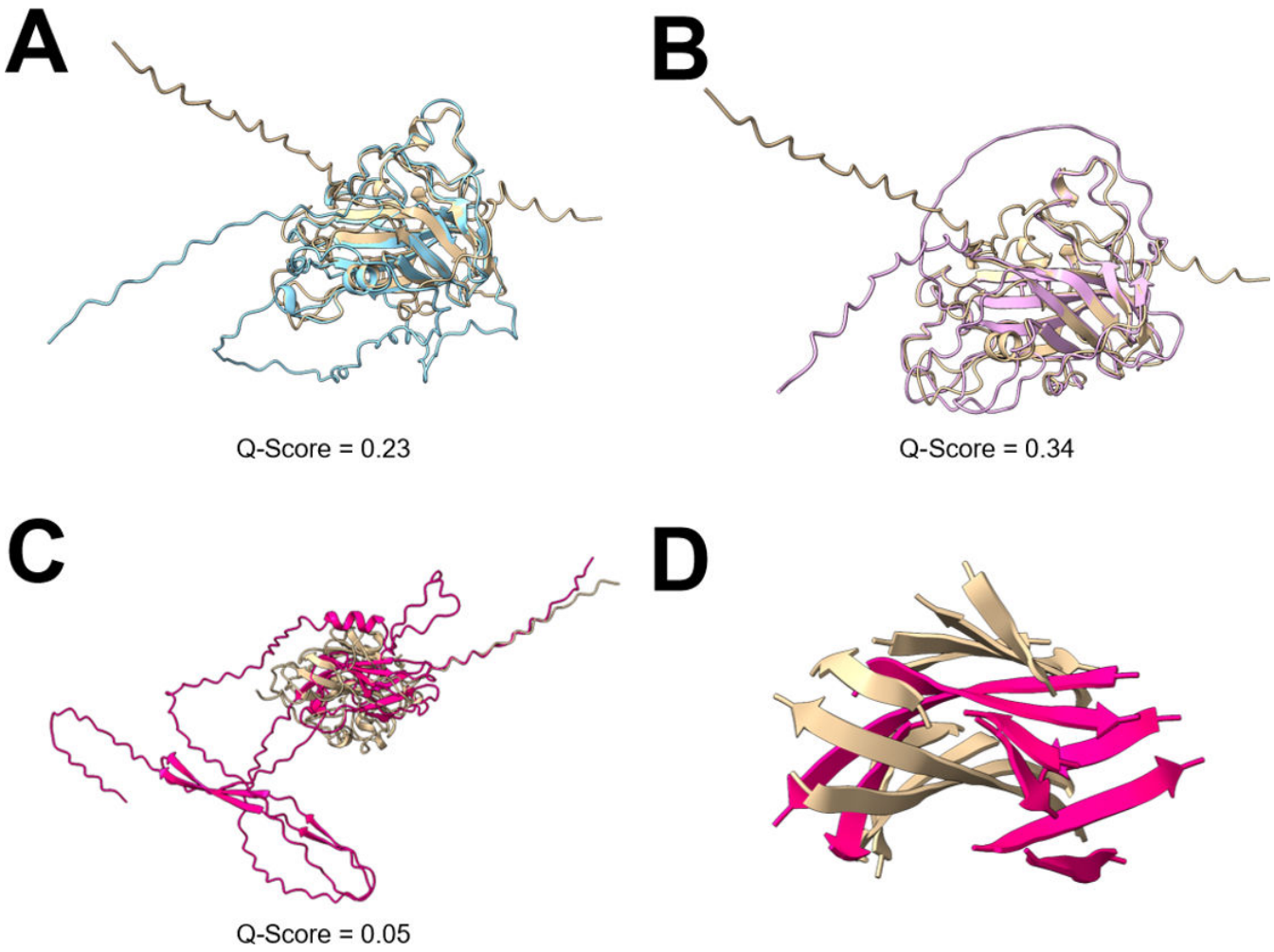

FIG 3 Structural comparison of LPMO-related proteins. The AlphaFold predicted structures of three sequences, PutMoI (A), PutMoM (B), and PutMoP (C and D) from the *P. putredinis* NO1 genome structurally aligned to the *A. niger* AA9 LPMO used in sequence and structure-based searching (UniProt ID: A2QZE1). *A. niger* AA9 LPMO, beige; PutMoI, blue; PutMoM, pink; PutMoP, hot pink. *Q*-score is a quality function of Cα alignment from PDBefold.

predicted with InterPro; any CAZyme domains were annotated with dbCAN (7); the predicted structures were compared with structures in the PDB database (35); and secretion signal peptides were predicted with SignalP (36) in an attempt to elucidate the potential functions. Two of the sequences, PutMoA and PutMoD, are the two predicted AA16 LPMOs identified in the *P. putredinis* NO1 CAZyme repertoire earlier (Fig. S2). Another two sequences, PutMoH and PutMoK, were not annotated as CAZymes but had conserved BIM1-like domains. BIM1-like proteins are LPMO_auxilliary-like proteins, function in fungal copper homeostasis, and share a similar copper coordination method to the LPMOs which they are related to (37). Although not likely to be involved in lignocellulose breakdown, this highlights how structurally related proteins in terms of active site or co-factor coordination structures can be identified with structural approaches where sequence- and domain-based approaches fail. Three of the nine sequences were also identified as being upregulated when *P. putredinis* NO1 was previously grown on wheat straw and compared to growth on glucose (Supplementary File 1) (5). Although this does not confirm the role of these proteins in lignocellulose breakdown, it provided another layer of information for the selection of interesting candidate sequences to investigate further. PutMoP was the most interesting sequence identified solely by the structural searching and showed upregulation during growth on wheat straw compared to glucose. It was not annotated as a CAZyme; no conserved

domains were identified; and sequence homology was only observed in hypothetical proteins in the NCBI non-redundant protein database (34). Comparing the AlphaFold predicted structure of PutMoP to the *A. niger* AA9 LPMO revealed similarity at the central beta-sheet structure despite a very low *Q*-score of 0.05 (Fig. 3C and D). A secretion signal peptide was also predicted for this protein, suggesting an extracellular role. This immunoglobulin-like distorted β-sandwich fold is a characteristic structural feature of LPMOs and is shared across the LPMO CAZyme families (38). The similarity of this central structure is likely the reason for identification of this sequence by structural comparison. This structural similarity at the protein center, the lack of amino acid sequence similarity, and the conserved secretion signal make this protein an interesting candidate for further investigation. Searching the PutMoP structure against the whole PDB structure database returned many diverse proteins not linked to lignocellulose breakdown; however, the *Q*-score was very low for all the structures and did not help to discern the potential activity of this protein. The sequence lacks the N-terminal histidine after the signal peptide cleavage site, which is conserved in LPMOs, so this protein is unlikely to be an LPMO. However, a secreted unknown protein with some central structural similarity to an important class of oxidative proteins that degrade crystalline cellulose is of definite interest.

In addition to searching for LPMO-related sequences, classes of enzymes involved in the breakdown of lignin are important targets for the biorefining of plant biomass. The recalcitrance of lignin is a limiting factor hindering the industrial use of lignocellulose as a feedstock to produce biofuels. Lignin itself is also a historically underutilized feedstock for valuable chemicals (39). Laccases are multicopper oxidase family of enzymes that catalyze oxidation of phenolic compounds through an electron transfer reaction that simultaneously reduces molecular oxygen to water (23). They modify lignin by depolymerization and repolymerization, Cα oxidation, and demethylation and are particularly efficient due to their use of readily available molecular oxygen as the final electron acceptor (40, 41).

Laccase-related sequences were searched for in the *P. putredinis* NO1 genome using the sequence of an AA1 family laccase from *A. niger*, a bespoke HMM constructed from ascomycete laccase and basidiomycete multi-copper oxidase sequences downloaded from the laccase engineering database (42) and with the structure of the *A. niger* AA1 laccase. In total, 32 sequences were identified across the three searching strategies, and only 9 of these were annotated by dbCAN as AA1 family of CAZymes (Table S2). The bespoke HMM allowed for more divergent sequences for these enzymes to be incorporated into the model's construction. The result was the identification of sequences that when explored further looked like laccase enzymes but were missed by traditional CAZyme annotation, highlighting how searching for CAZymes alone is a limited method for identifying lignocellulose-degrading enzymes. However, for the identification of new lignocellulose-degrading enzymes, more divergent sequences are of interest. A single coding sequence, PutLacJ, was identified by the structural searching approach with a 30% lowest acceptable match parameter that was not identified by sequence or domain-based searching (Table 2).

PutLacJ was not annotated as a CAZyme by dbCAN but does have a predicted cupredoxin domain, a feature of laccase enzymes (43). Structural comparisons against the PDB structure database revealed alignments with moderate confidence scores to copper-containing nitrite reductases from *Neisseria gonorrhoeae* which are suggested to play a role in pathogenesis (44). In fungi, it is more likely that these are playing a role in denitrification (45). The lack of a signal peptide makes it unlikely that this protein is involved in lignin depolymerization, despite the structural similarity to the beta-sheet regions of the *A. niger* laccase (Fig. 4).

Peroxidases (PODs) also play a major role in lignin deconstruction by white-rot fungi. PODs are lacking in brown-rot species, presumably due to their non-ligninolytic specialization of substrate degradation (46). The identification of new putative peroxidases in *P. putredinis* NO1 is of interest. Fungal class II peroxidases are divided into

TABLE 2 Identifying laccase-related proteins encoded in the *P. putredinis* NO1 genome[a]

			Identified by searching approach			
Coding region	GenBank accession	Protein ID	Sequence	Domain	Structure	InterPro annotation
FUN_000263-T1	CAI7987524.1	PutLacA		✔		–
FUN_000580-T1	CAI7987844.1	PutLacB		✔		Phospodiesterase
FUN_000646-T1	CAI7987911.1	PutLacC		✔		–
FUN_000759-T1	CAI7988026.1	PutLacD	✔	✔		Multi-copper oxidase
FUN_000832-T1	CAI7988099.1	PutLacE		✔		–
FUN_001183-T1	CAI7988671.1	PutLacF		✔		–
FUN_001583-T1	CAI7989479.1	PutLacG		✔		Salt tolerance regulator
FUN_002249-T1	CAI7990863.1	PutLacH	✔	✔	✔	AA1 multi-copper oxidase
FUN_002874-T1	CAI7992258.1	PutLacI		✔		–
FUN_003732-T1	CAI7994085.1	PutLacJ			✔	–
FUN_003828-T1	CAI7994234.1	PutLacK		✔		–
FUN_004259-T1	CAI7995254.1	PutLacL		✔		Nucleoside hydrolase
FUN_004616-T1	CAI7995870.1	PutLacM		✔		–
FUN_004739-T1	CAI7996089.1	PutLacN		✔		Fumarylacetoacetate hydrolase family
FUN_005132-T1	CAI7997298.1	PutLacO		✔	✔	AA1 multi-copper oxidase
FUN_005520-T1	CAI7998008.1	PutLacP		✔		Diacylglycerol acyltransferase
FUN_006244-T1	CAI7999594.1	PutLacQ		✔		–
FUN_006620-T1	CAI8000270.1	PutLacR		✔		Fumarylacetoacetate hydrolase family
FUN_006720-T1	CAI8000684.1	PutLacS		✔		Glycosyltransferase 90
FUN_007228-T1	CAI8001746.1	PutLacT		✔		–
FUN_007508-T1	CAI8002246.1	PutLaU		✔		Pex2
FUN_008329-T1	CAI8004041.1	PutLacV		✔		ATPase-related
FUN_009491-T1	CAI7995256.1	PutLacW		✔		Helicase

[a]Coding regions of proteins related to laccases identified through genome searching approaches with the sequence of an *A. niger* AA1 laccase (*E*-value cutoff = 1×10^{-5}), the bespoke laccase and multicopper oxidase HMM constructed from sequences from the laccase engineering database (significance threshold = 0.01), and the structure of the *A. niger* AA1 laccase (lowest percentage match = 30%) and which were not annotated as AA1 CAZymes by dbCAN. InterPro annotations were retrieved where possible

three lignolytic forms: lignin peroxidase (LiP), manganese peroxidase (MnP), and versatile peroxidase (VP) (47).

Sequence searches into the *P. putredinis* NO1 genome using sequences of MnP from *Aureobasidium subglaciale*, LiP from *F. oxysporum*, and VP from *Pyronema confluens* yielded only two sequences (Table S3). Both peroxidase-related sequences were also identified by domain searching using a bespoke HMM constructed from sequences of MnPs, LiPs, and VPs downloaded from the fPoxDB database of peroxidase sequences (48). This domain-based approach identified only three sequences in total, all of which were annotated as AA2 family of CAZymes also (Table S3). However, structural-based searching using the structures of the same three peroxidases and with the lowest acceptable match parameter of 30% used in sequence-based searches identified nine coding regions in total (Table S3), seven of which were not identified by sequence- or domain-based searching approaches and were not annotated as AA2 CAZymes (Table 3) but were all found to be upregulated previously when *P. putredinis* NO1 was grown on wheat straw compared to growth on glucose (Supplementary File 1) (5).

Investigating these sequences further revealed two sequences to be the most interesting, PutPoxA and PutPoxG, both with low *Q*-scores of 0.01 and 0.04, respectively. PutPoxA was not annotated as a CAZyme but does have a predicted domain of unknown function family 3632 (DUF3632). Genes encoding DUF3632 domains were previously found to be upregulated in the filamentous ascomycete *Neurospora crassa* when the CLR-2 transcription factor, important for growth on cellulose, was constitutively expressed (49). The protein does, however, lack a signal peptide, and structural comparison to the *A. subglaciale* MnP shows similar helical structures, but these secondary structures do not appear to overlap very well (Fig. 5A). PutPoxG was not annotated as a CAZyme, and no conserved domains were identified, although the helical

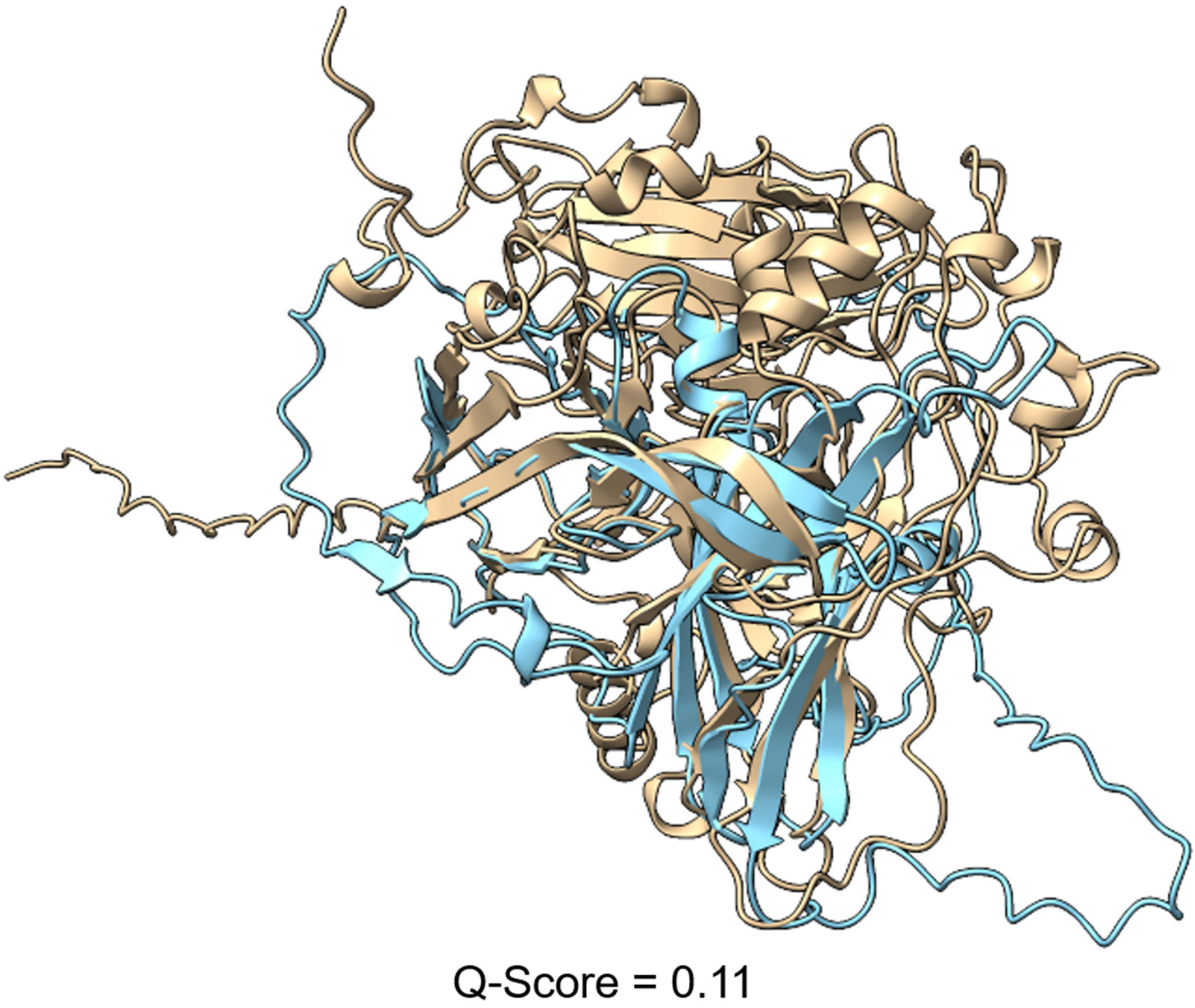

FIG 4 Structural comparison of PutLacJ laccase-related protein. The AlphaFold predicted structures of the sequence PutLacJ from the *P. putredinis* NO1 genome structurally aligned to the *A. niger* laccase used in sequence and structure-based searching (UniProt ID: A2QB28). *A. niger* laccase, beige; PutLacJ, blue. *Q*-score is a quality function of Cα alignment from PDBefold.

structures do seem to align better with the *A. subglaciale* MnP than PutPoxA (Fig. 5B). Furthermore, searching of both structures against the PDB database was performed, but all alignments had very low *Q*-scores of less than 0.1.

As with the candidates identified by LPMO and laccase searching approaches, it is hard to be confident on sequence and structural investigation alone that these proteins are involved in lignocellulose breakdown. However, by utilizing multiple searching approaches, more divergent and varied sequences with potential relation to industrially important enzymes have been identified here. This strategy of searching for new enzymes involved in the breakdown of the most recalcitrant components of lignocellulose would work well when combined with additional layers of biological data, e.g., transcriptomic or proteomic data. Many of the coding regions investigated here show structural similarity to the interesting classes of enzymes with which they were identified but lack the sequence similarity and therefore the functional annotation. Transcriptomic data showing upregulation of these genes or proteomic data showing increased abundances of these proteins when the organism in question is grown on lignocellulosic

TABLE 3 Identifying peroxidase-related proteins encoded in the *P. putredinis* NO1 genome[a]

Coding region	GenBank accession	Protein ID	Identified by searching approach			InterPro annotation
			Sequence	Domain	Structure	
FUN_002995-T1	CAI7992466.1	PutPoxA			✔	DUF3632
FUN_003542-T1	CAI7993642.1	PutPoxB			✔	Arabinofuranosidase
FUN_003618-T1	CAI7993895.1	PutPoxC			✔	-
FUN_004484-T1	CAI7995643.1	PutPoxD			✔	Cell division control
FUN_008413-T1	CAI8004205.1	PutPoxE			✔	SIT4 phosphatase-associated
FUN_008923-T1	CAI7988420.1	PutPoxF			✔	-
FUN_009329-T1	CAI7993214.1	PutPoxG			✔	-

[a]Coding regions of proteins related to peroxidases identified through genome searching approaches with the sequences of an MnP from *A. subglaciale*, LiP from *F. oxysporum*, and VP from *P. confluens* (*E*-value cut-off = 1×10^{-5}), the bespoke peroxidase HMM contricted from MnP, LiP, and VP sequences in the fPoxDB database (significance threshold = 0.01), and the structure of the same three peroxidases used for sequence searches (lowest percentage match = 30%) and which were not annotated as AA2 CAZymes by dbCAN. InterPro annotations were retrieved where possible

substrates would inspire more confidence in the role of these proteins in the degradation of plant biomass. Therefore, we used sequence similarity to identify the corresponding transcripts for these coding regions in the transcriptomic time course data set of *P. putredinis* NO1 grown for 10 days in cultures containing wheat straw published previously (5). The transcriptomic data were explored for all sequences which were identified solely by structural searches and therefore considered interesting (Supplementary File 1). For the four sequences explored in more detail, we found that three of the four, PutMoP, PutPoxA, and PutPoxG, were found to be significantly upregulated on at least one timepoint when grown on wheat straw compared to growth on glucose (Fig. 6). Expression of PutLacJ was instead found to be significantly higher during growth on glucose compared to growth on wheat straw. However, structural investigation revealed that PutLacJ had similarity to copper-containing nitrite reductase proteins, and it was concluded that it is unlikely to be involved in lignocellulose breakdown. Characterization would be required to confirm the role of these candidates in lignocellulose breakdown and to understand whether these activities are new. However, the implication in lignocellulose-degrading processes through the analysis of transcriptomic data provides another source of information by which candidates identified through the described strategy can be investigated. It is hoped that adoption of a similar strategy for analysis of the wealth of sequence data now publicly available will allow identification of novel enzyme sequences for many important processes to be made simpler.

Conclusions

P. putredinis NO1 was revealed here to contain a diverse repertoire of lignocellulose-degrading enzymes in its genome. The newly annotated reference genome is a potentially useful resource, considering the potential of *P. putredinis* NO1 for the identification of industrially valuable enzymes (5). Among ascomycetes, *P. putredinis* NO1 exists within the 95th percentile for abundant auxiliary activity gene density, implying potential specialism regarding mechanisms of lignocellulose degradation, and belongs to a substantially underrepresented and underexplored lineage. Investigating CAZyme families in more detail revealed an increased capacity to target the most recalcitrant components of lignocellulose when compared to three other biomass-degrading ascomycetes. For crystalline cellulose degradation, expansions were observed in families of LPMOs and in families associated with LPMO systems. Multiple domains encoding lignin-degrading laccase proteins were also identified. Considering the context in which *P. putredinis* NO1 was identified, thriving at the late stages of a mixed microbial community grown on wheat straw, we found it is feasible that the genome of this fungus contains new ligninolytic activities. By utilizing a strategy of searching genomic data for new enzymes with simultaneous sequence-, domain-, and structural-based approaches, multiple interesting sequences were identified.

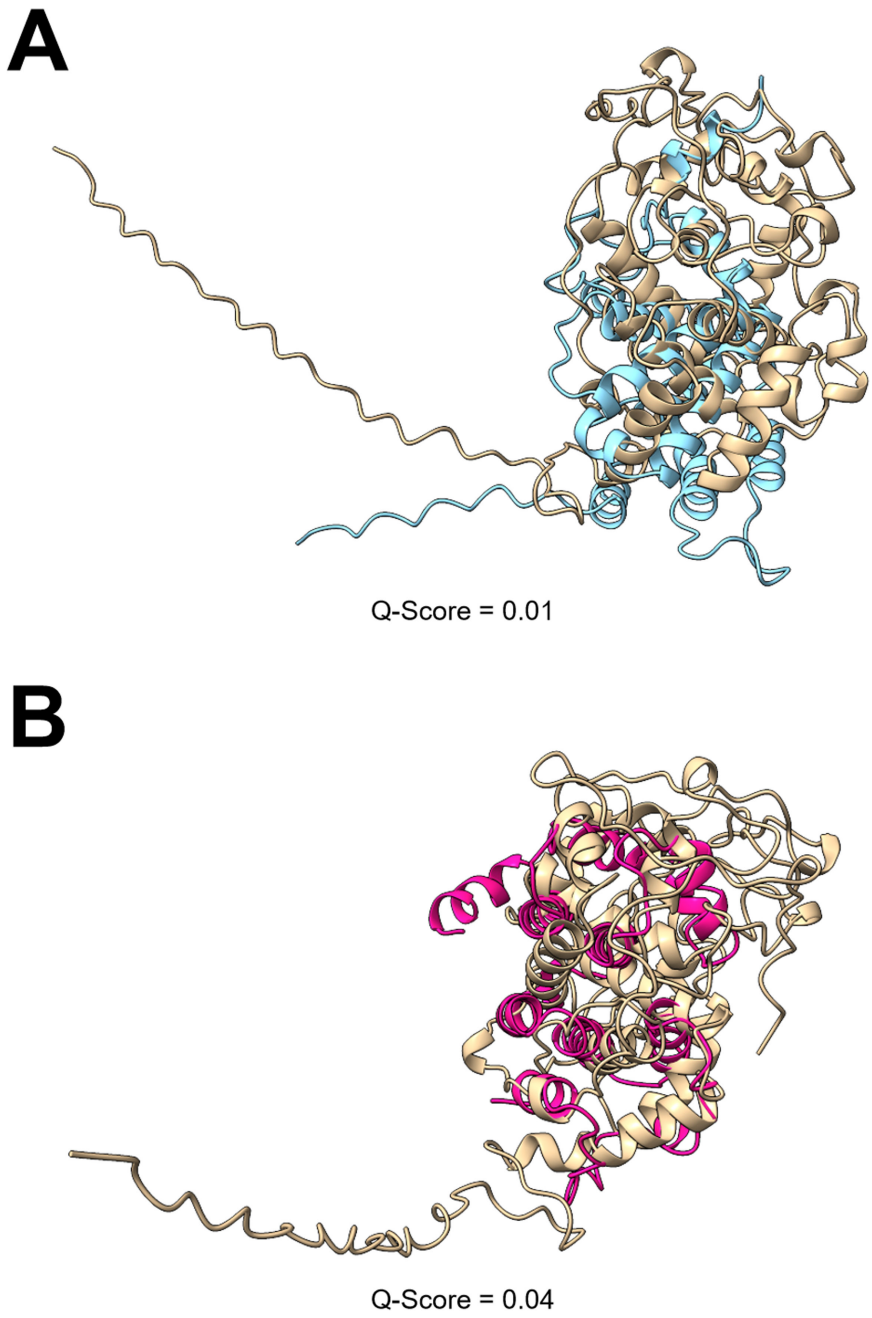

FIG 5 Structural comparison of peroxidase-related proteins. The AlphaFold predicted structures of two sequences, PutPoxA (A) and PutPoxG (B), from the *P. putredinis* NO1 genome structurally aligned to the *A. subglaciale* MnP used in sequence and structure-based searching. A predicted structure was unavailable, and so a predicted structure was generated with AlphaFold. *A. subglaciale* MnP, beige; PutPoxA, blue; PutPoxG, hot pink. *Q*-score is a quality function of Cα alignment from PDBefold.

MATERIALS AND METHODS

Strain isolation

P. putredinis NO1 was isolated from a wheat straw enrichment culture and maintained as reported previously (5).

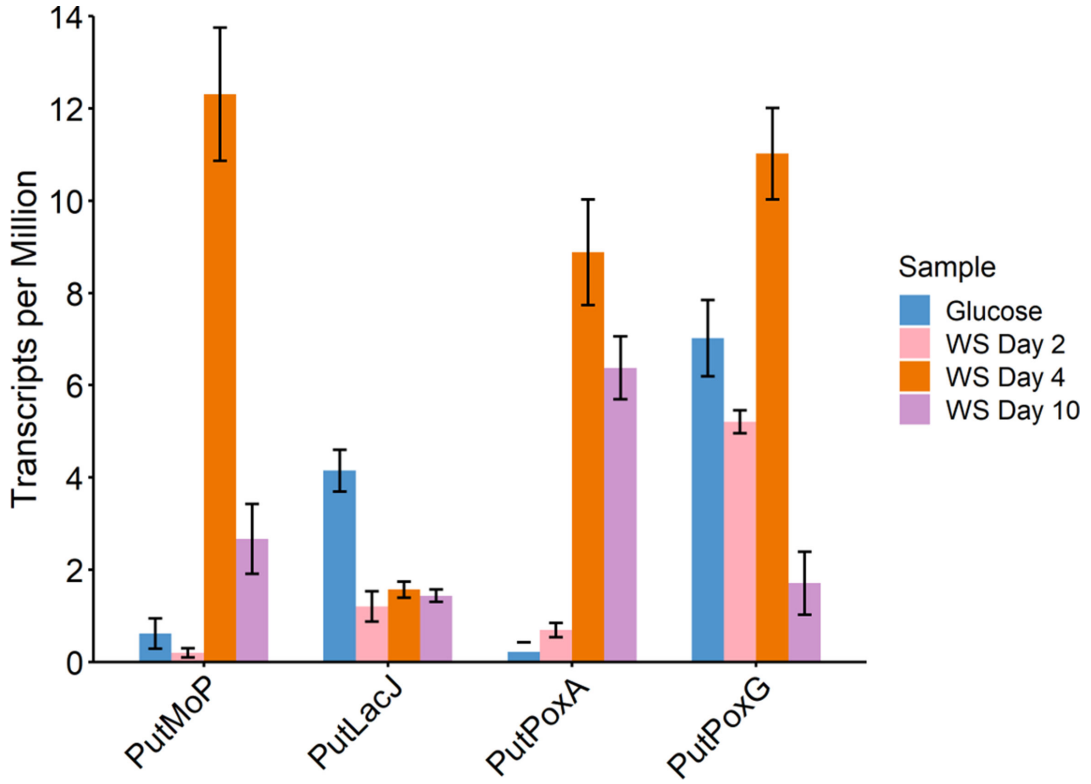

FIG 6 Gene expression of interesting candidates. Transcripts per million values for each of the four candidates explored, during growth on glucose, or on days 2, 4, and 10 of growth in liquid cultures containing WS. WS, wheat straw.

Genomic DNA extraction and sequencing

For DNA extraction, *P. putredinis* NO1 was grown in optimized media containing 10% (wt/vol) sucrose at 30°C with shaking at 140 rpm for 14 days. Wet fungal biomass was washed in deionized water before pelleting in 50 mL falcon tubes at 4,500 rpm for 15 minutes, and 10 technical replicates of 100 mg of biomass were then prepared in 1.5-mL tubes. Fungal biomass was then digested by adding 100 µL of 1-mg/mL chitinase from *Streptomyces griseus* (Merck) and 200 µL of 50-mM EDTA and incubating at 37°C for 3 hours. DNA extraction was then performed with the Wizard Genomic DNA Purification Kit (Promega). Digested samples were centrifuged at 18,000 × *g* at 4°C for 2 minutes, and the supernatant was discarded. Pellets were resuspended with 300 µL of nuclei lysis solution and 100 µL of protein precipitation solution and were rotated for 5 minutes before a 5-minute incubation on ice. Samples were then centrifuged at 18,000 × *g* at 4°C for 3 minutes, and the supernatant was transferred to fresh tubes containing 300 µL of cold isopropanol, gently mixed by inversion, and centrifuged again. The supernatant was discarded, and the pellet was washed in 70% ice-cold ethanol before centrifugation followed by air-drying the DNA pellet. The pellet was then resuspended in 50 µL of DNA rehydration solution with the addition of 1.5 µL of RNase solution. Samples were than incubated at 37°C for 15 minutes followed by rehydration at 4°C overnight. Replicate DNA samples were run on 0.75% agarose Tris-acetate-EDTA gel alongside GeneRuler 1-kb Plus DNA Ladder (Thermo Scientific) at 120 V for 40 minutes. The gel was then visualized in the Uvitec Gel-Documentation System to confirm the presence of long-strand DNA.

Genomic DNA was subject to an additional cleanup step using a 0.6:1.0 ratio of AMPure XP beads:sample prior to long-read sequencing using the ONT's MinION system. The sequencing library was prepared using ONT's ligation sequencing kit SQK-LSK109, as per the manufacturer's guidelines with modifications as follows: incubation times for end repair steps were increased from 5 to 30 minutes; ligation reactions were performed at

room temperature for 1 hour; and elution steps were performed at 37°C for 15 minutes. The resulting DNA libraries were sequenced on MinION R9.4.1 flow cells with a 48-hour run time. Basecalling was performed using Guppy v.3.5.2 software.

Genome assembly and annotation

Oxford Nanopore Technologies reads were filtered to those of length over 5 kb with SeqKit v.0.11.0 (50) before being assembled with Canu v.2.0 (51). The resulting genome assembly was filtered with Tapestry v.1.0.0 (52) to 39 Mb, 21 contigs, before being polished with Medaka v.0.11.3. Previously obtained Illumina reads were used to polish the assembly. Short-read Illumina sequencing libraries were prepared using the NEBNext Ultra DNA library prep kit for Illumina (New England Biolabs) and sequenced on an Illumina HiSeq 2500, with paired end 100-bp reads, by the University of Leeds Next Generation Sequencing Facility. The Illumina reads were quality-checked with FastQC v.0.11.7 (53) and adapter trimmed with Cutadapt v.2.10 (54) and used for three rounds of Pilon v.1.23 (55) polishing of the genome assembly. A previously obtained transcriptome assembly from NO1 grown on six lignocellulosic substrates (wheat straw, empty fruit bunches from palm oil, wheat bran, sugar cane bagasse, rice straw, and kraft lignin) was used for genome annotation with FunAnnotate v.1.8.1 and InterproScan v.5.46 (56, 57).

Ascomycete genome annotation and CAZyme prediction

All available genome assemblies (n = 2,635) of ascomycota origin were retrieved from the NCBI genome assembly database. Genome assemblies with N50 values of >1,000 were retained and gene prediction was performed with FunAnnotate v.1.8.1 (58), BUSCO (59), and AUGUSTUS (60), generating a final data set of 2,570 genomes. Predicted genes for each genome were annotated with the CAZyme database (v.09242921), and mean gene densities were then calculated for each taxonomic level for comparative analysis. Unique taxonomy identifiers for each genome were retrieved from the NCBI taxonomy database using the Entrez NCBI API (61). No filtering was undertaken and a phylogenetic tree was reconstructed using ETE3 to retrieve the tree topology (get_topology) without intermediate nodes at a rank limit of genus (62) (Fig. 2). Gene densities from annotations were mapped to the corresponding genomes on the tree. Genome metadata and annotations are available in Supplementary File 2.

The number and proportion of CAZyme domains in the genomes of *P. putredinis* NO1 (GCA_938049765.1), *S. boydii* (GCA_002221725.1), *T. reesei* (GCA_016806875.1), and *F. oxysporum* (GCA_023628715.1) were plotted using the 'ggplot2' package of R studio v.3.6.3 (58, 63).

Sequence-based searches for LPMOs, laccases, and peroxidases

The sequences for an ascomycete AA9 family of LPMO and for an AA1 family of laccase were obtained from the CAZy database (59). An AA9 LPMO from *Aspergillus niger* (GenBank: CAK97151.1) and an AA1 Laccase from *A. niger* (GenBank: CAK37372.1) were used. For peroxidase sequences, individual sequences for three types of reported lignin-degrading peroxidases were obtained from the fPoxDB database (48). A manganese peroxidase from *Aureobasidium subglaciale* (GenBank: EJD50148.1), a lignin peroxidase from *F. oxysporum* f. sp. *lycopersici* (NCBI RefSeq: XP_018248194.1), and a versatile peroxidase from *Pyronema confluens* (Locus: PCON_11,254 m.01) available only from the fPoxDB database were used.

These sequences were searched against the *P. putredinis* NO1 genome protein sequences through command line BLAST with an *E*-value cut off of 1×10^{-5} (60). Results were compiled for the three classes of peroxidase.

Domain-based searches for LPMOs, laccases, and peroxidases

Due to the lack of online databases for LPMO sequences, the genome was searched for LPMO-related sequences using the Pfam AA9 HMM (31).

Sequences for basidiomycete laccases and ascomycete multi-copper oxidases were downloaded from the Laccase Engineering Database v.7.1.11 (42). These were aligned using Kalign v.3.0, and this alignment was subsequently used to generate a bespoke HMM using the HMMER v.3.2.1 program (62, 64).

Sequences for manganese peroxidases, lignin peroxidases, and versatile peroxidases were downloaded from the fPoxDB database (48). These were aligned and used to construct a bespoke HMM model as before.

These models were used to search the *P. putredinis* NO1 genome using HMMER v.3.2.1 (64) and domain hits falling within the default significance inclusion threshold of 0.01.

Structure-based searches for LPMOs, laccases, and peroxidases

Predicted structure for >96% (n = 9611) of coding regions in the *P. putredinis* NO1 genome was modeled using AlphaFold v.2.0.0 on the VIKING computer cluster (30).

The 9611 models of coding sequences were compiled into "tarball" databases and compressed into ".tar.gz" files on the VIKING cluster. These files were uploaded to the PDBefold online server to search against (65). Structures for the same sequences used in sequence-based searching were obtained from UniProt database, if available (66), or modeled using AlphaFold v.2.00 on the VIKING computing cluster. These structures were searched against the *P. putredinis* NO1 structure database using PDBefold to identify similar structures in the *P. putredinis* NO1 genome. The lowest acceptable match parameter was adjusted, depending on the activity being searched with until coding regions not identified using sequence- or domain-based searching strategies were identified.

In silico investigation of candidate sequences

Sequences which were identified by structural searching solely were considered potentially interesting and warranted further investigation to attempt to elucidate function. Sequences were searched against the NCBI non-redundant protein database with default search parameters and an *E*-value cutoff of 1×10^{-5} to investigate proteins with similar sequence (34). Domains were predicted using the primary amino acid sequence with the InterPro tool for domain prediction with default parameters (67). CAZyme domains were predicted with the online dbCAN prediction tool with default search parameters (7). Interesting candidate structures were further investigated with PDBefold by searching the structures against the whole PDB database to identify structurally similar proteins using the lowest acceptable match parameter of 70% (35, 65). Secretion signals were predicted using SignalP v.6.0 with default parameters (36). Altogether, this annotation information was used to investigate the potential functions of interesting sequences.

Transcriptomic data for interesting sequences

A previously published transcriptomic data set for *P. putredinis* NO1 was used to validate expression of sequences of interest identified here during growth on lignocellulosic substrates (5). Gene expression data, in transcripts per million, was investigated for all sequences identified solely by structural approaches and not by sequence- or domain-based searching. Gene expression data are available in Supplementary File 1.

ACKNOWLEDGMENTS

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC), UK (grant BB/1018492/1, BB/P027717/1, and BB/W000695/1). C.S. was

supported by a CASE studentship from the BBSRC Doctoral Training Programme (BB/M011151/1) with Prozomix Ltd.

This project was undertaken on the Viking Cluster, which is a high-performance compute facility provided by the University of York. We are grateful for computational support from the University of York High Performance Computing service, Viking and the Research Computing team.

Special thanks to Sally James for performing the nanopore sequencing in her kitchen in the first weeks of the COVID-19 pandemic, and to Katherine Newling for her immense help with all bioinformatic work and my endless questions.

C.J.R.S. conceptualized the investigation carried out in this paper; extracted the genomic DNA from *P. putredinis* NO1; performed CAZyme repertoire comparison analysis; structurally modeled the *P. putredinis* NO1 genome; performed sequence-, domain-, and structure-based searches of the genome; analyzed the search strategy results and was the major contributor in writing the manuscript. D.R.L. carried out the annotation of ascomycete genomes and CAZyme repertoire comparison analysis and was a major contributor to the writing of the manuscript. N.C.O. was involved in maintaining *P. putredinis* NO1 and extracting genomic DNA. S.R.J. library prepped and sequenced the *P. putredinis* NO1 genomic DNA. K.N. assembled the *P. putredinis* NO1 genome, performed initial annotation, and aided deposition of sequence data. Y.L. assembled the transcriptome that was used for annotation of the *P. putredinis* NO1 genome. N.G.S.M. was a contributor to the writing of the paper. S.B. carried out the Illumina sequencing, which was used to polish the *P. putredinis* NO1 genome assembly. N.C.B. was a major contributor to the conceptualization and supervision of the study in addition to making a major contribution to the writing of the manuscript.

AUTHOR AFFILIATIONS

[1]Department of Biology, Centre for Novel Agricultural Products, University of York, York, United Kingdom
[2]Department of Biology, Bioscience Technology Facility, University of York, York, United Kingdom
[3]Department of Chemistry, York Structural Biology Laboratory, The University of York, York, United Kingdom

AUTHOR ORCIDs

Conor J. R. Scott http://orcid.org/0000-0001-7404-7619

FUNDING

Funder	Grant(s)	Author(s)
UKRI | Biotechnology and Biological Sciences Research Council (BBSRC)	BB/1018492/1	Conor J. R. Scott
UKRI | Biotechnology and Biological Sciences Research Council (BBSRC)	BB/P027717/1	Conor J. R. Scott
UKRI | Biotechnology and Biological Sciences Research Council (BBSRC)	BB/W000695/1	Daniel R. Leadbeater
UKRI | Biotechnology and Biological Sciences Research Council (BBSRC)	BB/M011151/1	Neil C. Bruce

AUTHOR CONTRIBUTIONS

Conor J. R. Scott, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Daniel R. Leadbeater, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing | Nicola C. Oates, Conceptualization, Formal analysis | Sally R. James, Data curation, Resources | Katherine

Newling, Data curation, Formal analysis, Resources, Software | Yi Li, Data curation, Formal analysis | Nicholas G. S. McGregor, Writing – original draft | Susannah Bird, Data curation, Formal analysis | Neil C. Bruce, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing

DATA AVAILABILITY

The sequence data generated and analyzed during the current study are available in the European Nucleotide Archive, project code PRJEB60285, secondary accession ERP145344. The whole genome sequence set for the genome assembly is available in the European Nucleotide Archive, Accession CASHTG010000000.1. The assembly is also available through the NCBI database, accession GCA_949357655.1.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental file 1 (Spectrum01035-23-S0001.csv). Transcriptomic data for all interesting sequences identified in LPMO, laccase, and peroxidase searches.
Supplemental file 2 (Spectrum01035-23-S0002.csv). All ascomycete genome annotation data.
Legends of supplemental files and tables (Spectrum01035-23-S0003.docx). Legends of supplemental files 1 and 2 and Tables S1 to S3.
Supplemental Table S1 (Spectrum01035-23-S0003.csv). All sequences identified through LPMO genome searching approaches.
Supplemental Table S2 (Spectrum01035-23-S0004.csv). All sequences identified through laccase genome searching approaches.
Supplemental Table S3 (Spectrum01035-23-S0005.csv). All sequences identified through peroxidase genome searching approaches.
Supplemental Figure S1 (Spectrum01035-23-S0007.tif). Supplemental Figure S1, visualizing the CBM family allocations across four asocmycetes.
Supplemental Figure S2 (Spectrum01035-23-S0008.tif). Supplemental Figure S2, visualizing the AA family allocations across four ascomycetes.
Supplemental Figure S3 (Spectrum01035-23-S0009.tif). Supplemental Figure S3, vizualizing pLDDT score distribution for genome predicted structures.

REFERENCES

1. Andlar M, Rezić T, Marđetko N, Kracher D, Ludwig R, Šantek B. 2018. Lignocellulose degradation: an overview of fungi and fungal enzymes involved in lignocellulose degradation. Eng Life Sci 18:768–778. https://doi.org/10.1002/elsc.201800039
2. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. 2013. Expansion of the enzymatic repertoire of the Cazy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6:41. https://doi.org/10.1186/1754-6834-6-41
3. Yamamoto M, Tomiyama H, Koyama A, Okuizumi H, Liu S, Vanholme R, Goeminne G, Hirai Y, Shi H, Nuoendagula, Takata N, Ikeda T, Uesugi M, Kim H, Sakamoto S, Mitsuda N, Boerjan W, Ralph J, Kajita S. 2020. A century-old mystery unveiled: sekizaisou is a natural lignin mutant. Plant Physiol 182:1821–1828. https://doi.org/10.1104/pp.19.01467
4. Kameshwar AKS, Qin W. 2018. Molecular networks of *Postia placenta* involved in degradation of lignocellulosic biomass revealed from metadata analysis of open access gene expression data. Int J Biol Sci 14:237–252. https://doi.org/10.7150/ijbs.22868
5. Oates NC, Abood A, Schirmacher AM, Alessi AM, Bird SM, Bennett JP, Leadbeater DR, Li Y, Dowle AA, Liu S, Tymokhin VI, Ralph J, McQueen-Mason SJ, Bruce NC. 2021. A multi-omics approach to lignocellulolytic enzyme discovery reveals a new ligninase activity from *Parascedosporium putredinis* No1. Proc Natl Acad Sci U S A 118:18. https://doi.org/10.1073/pnas.2008888118
6. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res 47:10994–11006. https://doi.org/10.1093/nar/gkz841
7. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95–W101. https://doi.org/10.1093/nar/gky418
8. Sista Kameshwar AK, Qin W. 2018. Comparative study of genome-wide plant biomass-degrading cazymes in white rot, Brown rot and soft rot fungi. Mycology 9:93–105. https://doi.org/10.1080/21501203.2017.1419296
9. Qian Y, Zhong L, Sun Y, Sun N, Zhang L, Liu W, Qu Y, Zhong Y. 2019. Enhancement of cellulase production in *Trichoderma reesei* via disruption of multiple protease genes identified by comparative secretomics. Front Microbiol 10:2784. https://doi.org/10.3389/fmicb.2019.02784

10. Demers JE, Gugino BK, Jiménez-Gasco MDM. 2015. Highly diverse endophytic and soil *Fusarium oxysporum* populations associated with field-grown tomato plants. Appl Environ Microbiol 81:81–90. https://doi.org/10.1128/AEM.02590-14
11. Anasontzis GE, Kourtoglou E, Villas-Boâs SG, Hatzinikolaou DG, Christakopoulos P. 2016. Metabolic engineering of *Fusarium oxysporum* to improve its ethanol-producing capability. Front Microbiol 7:632. https://doi.org/10.3389/fmicb.2016.00632
12. Nirmaladevi D, Venkataramana M, Srivastava RK, Uppalapati SR, Gupta VK, Yli-Mattila T, Clement Tsui KM, Srinivas C, Niranjana SR, Chandra NS. 2016. Molecular phylogeny, pathogenicity and toxigenicity of *Fusarium oxysporum* f. sp. lycopersici. Sci Rep 6:21367. https://doi.org/10.1038/srep21367
13. Zhao ZT, Liu HQ, Wang CF, Xu JR. 2013. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14:274. https://doi.org/10.1186/1471-2164-14-274
14. Hansson H, Karkehabadi S, Mikkelsen N, Douglas NR, Kim S, Lam A, Kaper T, Kelemen B, Meier KK, Jones SM, Solomon EI, Sandgren M. 2017. High-resolution structure of a lytic polysaccharide monooxygenase from *Hypocrea jecorina* reveals a predicted linker as an integral part of the catalytic domain. J Biol Chem 292:19099–19109. https://doi.org/10.1074/jbc.M117.799767
15. Bennati-Granier C, Garajova S, Champion C, Grisel S, Haon M, Zhou S, Fanuel M, Ropartz D, Rogniaux H, Gimbert I, Record E, Berrin J-G. 2015. Substrate specificity and regioselectivity of fungal AA9 lytic polysaccharide monooxygenases secreted by *Podospora anserina*. Biotechnol Biofuels 8:90. https://doi.org/10.1186/s13068-015-0274-3
16. Kracher D, Scheiblbrandner S, Felice AKG, Breslmayr E, Preims M, Ludwicka K, Haltrich D, Eijsink VGH, Ludwig R. 2016. Extracellular electron transfer systems fuel cellulose oxidative degradation. Science 352:1098–1101. https://doi.org/10.1126/science.aaf3165
17. Bissaro B, Várnai A, Røhr ÅK, Eijsink VGH. 2018. Oxidoreductases and reactive oxygen species in conversion of lignocellulosic biomass. Microbiol Mol Biol Rev 82. https://doi.org/10.1128/MMBR.00029-18
18. Wang BJ, Walton PH, Rovira C. 2019. Molecular mechanisms of oxygen activation and hydrogen peroxide formation in lytic polysaccharide monooxygenases. ACS Catal 9:4958–4969. https://doi.org/10.1021/acscatal.9b00778
19. Monclaro AV, Petrović DM, Alves GSC, Costa MMC, Midorikawa GEO, Miller RNG, Filho EXF, Eijsink VGH, Várnai A. 2020. Characterization of two family AA9 LPMos from *Aspergillus tamariiwith* distinct activities on *Xyloglucan* reveals structural differences linked to cleavage specificity. PLoS One 15:e0235642. https://doi.org/10.1371/journal.pone.0235642
20. Sützl L, Laurent CVFP, Abrera AT, Schütz G, Ludwig R, Haltrich D. 2018. Multiplicity of enzymatic functions in the Cazy AA3 family. Appl Microbiol Biotechnol 102:2477–2492. https://doi.org/10.1007/s00253-018-8784-0
21. Haddad Momeni M, Fredslund F, Bissaro B, Raji O, Vuong TV, Meier S, Nielsen TS, Lombard V, Guigliarelli B, Biaso F, Haon M, Grisel S, Henrissat B, Welner DH, Master ER, Berrin J-G, Abou Hachem M. 2021. Discovery of fungal *Oligosaccharide-Oxidising* flavo-enzymes with previously unknown substrates, redox-activity profiles and interplay with lpmos. Nat Commun 12. https://doi.org/10.1038/s41467-021-22372-0
22. Ferraroni M, Westphal AH, Borsari M, Tamayo-Ramos JA, Briganti F, Graaff LH de, Berkel WJH van. 2017. Structure and function of *Aspergillus niger* laccase mcog. Biocatalysis 3:1–21. https://doi.org/10.1515/boca-2017-0001
23. Brenelli L, Squina FM, Felby C, Cannella D. 2018. *Laccase*-derived lignin compounds boost cellulose oxidative enzymes AA9. Biotechnol Biofuels 11:10. https://doi.org/10.1186/s13068-017-0985-8
24. Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, Blumentritt M, Coutinho PM, Cullen D, de Vries RP, Gathman A, Goodell B, Henrissat B, Ihrmark K, Kauserud H, Kohler A, LaButti K, Lapidus A, Lavin JL, Lee YH, Lindquist E, Lilly W, Lucas S, Morin E, Murat C, Oguiza JA, Park J, Pisabarro AG, Riley R, Rosling A, Salamov A, Schmidt O, Schmutz J, Skrede I, Stenlid J, Wiebenga A, Xie X, Kües U, Hibbett DS, Hoffmeister D, Högberg N, Martin F, Grigoriev IV, Watkinson SC. 2011. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science 333:762–765. https://doi.org/10.1126/science.1205411
25. Henriksson G, Johansson G, Pettersson G. 1993. Is Cellobiose oxidase from *Phanerochaete chrysosporium* a one-electron reductase? Biochim Biophys Acta 1144:184–190. https://doi.org/10.1016/0005--2728(93)90171-b
26. Xu C, Su X, Wang J, Zhang F, Shen G, Yuan Y, Yan L, Tang H, Song F, Wang W. 2021. Characteristics and functional bacteria in a microbial consortium for rice straw lignin-degrading. Bioresour Technol 331:125066. https://doi.org/10.1016/j.biortech.2021.125066
27. Filiatrault-Chastel C, Navarro D, Haon M, Grisel S, Herpoël-Gimbert I, Chevret D, Fanuel M, Henrissat B, Heiss-Blanquet S, Margeot A, Berrin J-G. 2019. AA16, a new lytic polysaccharide monooxygenase family identified in fungal secretomes. Biotechnol Biofuels 12:55. https://doi.org/10.1186/s13068-019-1394-y
28. Pearson WR. 2013. "An introduction to sequence similarity ("homology") searching". Curr Protoc Bioinformatics Chapter 3:3. https://doi.org/10.1002/0471250953.bi0301s42
29. Johnson LS, Eddy SR, Portugaly E. 2010. Hidden markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11. https://doi.org/10.1186/1471-2105-11-431
30. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with alphafold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
31. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995
32. Petrović DM, Bissaro B, Chylenski P, Skaugen M, Sørlie M, Jensen MS, Aachmann FL, Courtade G, Várnai A, Eijsink VGH. 2018. Methylation of the N-terminal histidine protects a lytic polysaccharide monooxygenase from auto-oxidative inactivation. Protein Sci 27:1636–1650. https://doi.org/10.1002/pro.3451
33. Varadi M, Velankar S. 2022. The impact of alphafold protein structure database on the fields of life sciences. Proteomics, November:e2200128. https://doi.org/10.1002/pmic.202200128
34. Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504. https://doi.org/10.1093/nar/gki025
35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235
36. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. Signalp 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40:1023–1025. https://doi.org/10.1038/s41587-021-01156-3
37. Garcia-Santamarina S, Probst C, Festa RA, Ding C, Smith AD, Conklin SE, Brander S, Kinch LN, Grishin NV, Franz KJ, Riggs-Gelasco P, Lo Leggio L, Johansen KS, Thiele DJ. 2020. A lytic polysaccharide monooxygenase-like protein functions in fungal copper import and meningitis. Nat Chem Biol 16:337–344. https://doi.org/10.1038/s41589-019-0437-9
38. Arora R, Bharval P, Sarswati S, Sen TZ, Yennamalli RM. 2019. Structural dynamics of lytic polysaccharide monoxygenases reveals a highly flexible substrate binding region. J Mol Graph Model 88:1–10. https://doi.org/10.1016/j.jmgm.2018.12.012
39. Ragauskas AJ, Beckham GT, Biddy MJ, Chandra R, Chen F, Davis MF, Davison BH, Dixon RA, Gilna P, Keller M, Langan P, Naskar AK, Saddler JN, Tschaplinski TJ, Tuskan GA, Wyman CE. 2014. Lignin valorization: improving lignin processing in the biorefinery. Science 344:1246843. https://doi.org/10.1126/science.1246843
40. Lassouane F, Aït-Amar H, Amrani S, Rodriguez-Couto S. 2019. A promising *Laccase* immobilization approach for bisphenol A removal from aqueous solutions. Bioresource Technology 271:360–367. https://doi.org/10.1016/j.biortech.2018.09.129

41. Hilgers R, Vincken JP, Gruppen H, Kabel MA. 2018. *Laccase*/mediator systems: their reactivity toward phenolic Lignin structures. ACS Sustain Chem Eng 6:2037–2046. https://doi.org/10.1021/acssuschemeng.7b03451
42. Sirim D, Wagner F, Wang L, Schmid RD, Pleiss J. 2011. The *Laccase* engineering database: a classification and analysis system for *Laccases* and related multicopper oxidases. Database (Oxford) 2011:bar006. https://doi.org/10.1093/database/bar006
43. Pardo I, Rodríguez-Escribano D, Aza P, de Salas F, Martínez AT, Camarero S. 2018. A highly stable *Laccase* obtained by swapping the second cupredoxin domain. Sci Rep 8. https://doi.org/10.1038/s41598-018-34008-3
44. Boulanger MJ, Murphy MEP. 2002. Crystal structure of the soluble domain of the major anaerobically induced outer membrane protein (Ania) from *Pathogenic neisseria*: a new class of copper-containing nitrite reductases. J Mol Biol 315:1111–1127. https://doi.org/10.1006/jmbi.2001.5251
45. Matsuoka M, Kumar A, Muddassar M, Matsuyama A, Yoshida M, Zhang KYJ. 2017. Discovery of fungal denitrification inhibitors by targeting copper nitrite reductase from *Fusarium oxysporum*. J Chem Inf Model 57:203–213. https://doi.org/10.1021/acs.jcim.6b00649
46. Zhu Y, Plaza N, Kojima Y, Yoshida M, Zhang J, Jellison J, Pingali SV, O'Neill H, Goodell B. 2020. Nanostructural analysis of enzymatic and non-enzymatic brown rot fungal deconstruction of the lignocellulose cell wall(dagger). Front Microbiol 11:1389. https://doi.org/10.3389/fmicb.2020.01389
47. 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, St John F, 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 336:1715–1719. https://doi.org/10.1126/science.1221748
48. Choi J, Détry N, Kim K-T, Asiegbu FO, Valkonen JPT, Lee Y-H. 2014. fPoxDB: fungal peroxidase database for comparative genomics. BMC Microbiol 14:117. https://doi.org/10.1186/1471-2180-14-117
49. Craig JP, Coradetti ST, Starr TL, Glass NL. 2015. Direct target network of the *Neurospora crassa* plant cell wall deconstruction regulators CLR-1, CLR-2, and XLR-1. mBio 6:e01452-15. https://doi.org/10.1128/mBio.01452-15
50. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962
51. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive K-MER weighting and repeat separation . Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116
52. Davey JW, Catta-Preta CMC, James S, Forrester S, Motta MCM, Ashton PD, Mottram JC. 2021. Chromosomal assembly of the nuclear genome of the endosymbiont-bearing *Trypanosomatid angomonas deanei*. G3 (Bethesda) 11:jkaa018. https://doi.org/10.1093/g3journal/jkaa018
53. Andrews S. FastQC: a quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute. 2011.
54. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j 17:10. https://doi.org/10.14806/ej.17.1.200
55. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963
56. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. Interproscan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031
57. Palmer JasonJMS2020. Funannotate V.1.8.1: eukaryotic genome annotation
58. Villanueva RAM, Chen ZJ. 2019. ggplot2: Elegant graphics for data analysis. Meas Interdiscip Res 17:160–167. https://doi.org/10.1080/15366367.2019.1565254
59. 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:D571–D577. https://doi.org/10.1093/nar/gkab1045
60. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST plus: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
61. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res 50:D20–D26. https://doi.org/10.1093/nar/gkab1112
62. Lassmann T. 2019. Kalign 3: multiple sequence alignment of large data SETS. Bioinformatics 36:1928–1929. https://doi.org/10.1093/bioinformatics/btz795
63. . R development core team. R: a language and environment for statistical computing; R Foundation for Statistical Computing, Vienna, Austria, 2022.
64. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448
65. Krissinel E, Henrick K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60:2256–2268. https://doi.org/10.1107/S0907444904026460
66. UniProt C. 2019. Uniprot: a worldwide Hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049
67. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunić I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. 2023. Interpro in 2022. Nucleic Acids Res 51:D418–D427. https://doi.org/10.1093/nar/gkac993