This is a repository copy of *A Bioinformatic Workflow for in silico Secretome Prediction with the Lignocellulose Degrading Ascomycete Fungus Parascedosporium putredinis NO1*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/205569/

Version: Published Version

Article:

Scott, Conor, Leadbeater, Daniel Raymond and Bruce, Neil Charles orcid.org/0000-0003-0398-2997 (2023) A Bioinformatic Workflow for in silico Secretome Prediction with the Lignocellulose Degrading Ascomycete Fungus Parascedosporium putredinis NO1. Molecular Microbiology. pp. 754-762. ISSN 0950-382X

https://doi.org/10.1111/mmi.15144

RESOURCES

WILEY

A bioinformatic workflow for in silico secretome prediction with the lignocellulose degrading ascomycete fungus *Parascedosporium putredinis* NO1

Conor J. R. Scott ⬥ | Daniel R. Leadbeater ⬥ | Neil C. Bruce ⬥

Centre for Novel Agricultural Products, Department of Biology, University of York, York, UK

Correspondence
Conor J. R. Scott, Centre for Novel Agricultural Products, Department of Biology, University of York, York, UK.
Email: cs1535@york.ac.uk

Abstract

The increasing availability of microbial genome sequences provides a reservoir of information for the identification of new microbial enzymes. Genes encoding proteins engaged in extracellular processes are of particular interest as these mediate the interactions microbes have with their environments. However, proteomic analysis of secretomes is challenging and often captures intracellular proteins released through cell death and lysis. Secretome prediction workflows from sequence data are commonly used to filter proteins identified through proteomics but are often simplified to a single step and are not evaluated bioinformatically for their effectiveness. Here, a workflow to predict a fungal secretome was designed and applied to the coding regions of the *Parascedosporium putredinis* NO1 genome. This ascomycete fungus is an exceptional lignocellulose degrader from which a new lignin-degrading enzyme has previously been identified. The 'secretome isolation' workflow is based on two strategies of localisation prediction and secretion prediction each utilising multiple available tools. The workflow produced three final secretomes with increasing levels of stringency. All three secretomes showed increases in functional annotations for extracellular processes and reductions in annotations for intracellular processes. Multiple sequences isolated as part of the secretome lacked any functional annotation and made exciting candidates for novel enzyme discovery.

KEYWORDS
ascomycete, bioinformatics, CAZymes, genome, *Parascedosporium*, secretome

1 | INTRODUCTION

Proteomics is the study of proteins within a sample and involves the use of techniques that provide high molecular specificity for a broad range of peptides in a single measurement (Alfaro et al., 2016). Proteins are predominantly responsible for biological functions, and therefore acquiring qualitative and quantitative data on proteins can

help us understand microbiological processes, such as lignocellulose breakdown (Nielsen, 2017).

Fungi are exceptional wood degraders and produce an array of bioproducts, including secreted enzymes used in commercial enzyme cocktails for the valorisation of lignocellulosic biomass. *Parascedosporium putredinis* NO1 is an ascomycete fungus from which new extracellular lignin-degrading enzymes have previously been

identified (Oates et al., 2021). With an available reference genome containing 9998 protein-coding sequences for *P. putredinis* NO1, more effective proteomics experiments can now be carried out to help to identify the full repertoire of enzymes secreted by the fungus to facilitate growth on lignocellulose (Scott, 2023).

A 'secretome' is defined as the set of proteins secreted by a cell or an organism at a given time (Alfaro et al., 2016). However, extracellular protein studies are not straightforward and often unavoidably identify contaminant intracellular proteins in abundance because of cell death and lysis (Rodrigues et al., 2011). For secretomic investigations, this adds an additional layer of complexity and redundancy, especially for novel enzyme identification. Bioinformatic techniques could instead be used to filter proteomic data to predict in silico secretomes, allowing extracellular processes to be more clearly and accurately understood and simplifying new enzyme identification.

Previous attempts to predict in silico fungal secretomes have used single prediction tools, such as the prediction of secretion signal peptides (Artzi et al., 2017; de Paula et al., 2019; Moremen & Haltiwanger, 2019). However, proteins with signal peptides may be targeted to secretory pathways, but not necessarily secreted (Nielsen, 2017). Additionally, fungal protein secretion is more complex as proteins can be secreted via conventional or unconventional pathways (Alfaro et al., 2016). For example, it has been demonstrated that various metabolic enzymes are secreted by fungi despite the absence of secretion signals (Miura & Ueda, 2018). Perhaps, the most well-investigated method of fungal unconventional protein secretion is through vesicles, which are utilised by fungi as efficient vehicles for the release of proteins into the extracellular environment, along with polysaccharides and pigments (Rodrigues et al., 2011). Other secretomic investigations attempt to create basic workflows for in silico secretome prediction (Alfaro et al., 2016; Gogleva et al., 2018). However, these often lack diversity in the tools used in each step and fail to confirm their effectiveness bioinformatically. As such, many available data sets are considered (meta-) exo-proteomes as they include contaminant intracellular proteins due to the lack of secretome identification pipelines.

Here, a bioinformatic workflow was designed to isolate sequences of the *P. putredinis* NO1 genome predicted to be secreted. The workflow is built around two initial strategies of prediction: localisation prediction and secretion prediction. In both strategies, more than one tool is used to capture secretome sequences which may be missed by a single tool alone. The effectiveness of each tool to predict a subset of sequences enriched in sequences encoding enzymes known to be extracellular and secreted was evaluated through annotation of sequences for COG category, CAZyme class and KEGG metabolic pathways. Three resulting secretomes were produced with increasing levels of stringency on the sequences included: relaxed, strict and super strict. All secretomes contained greatly reduced numbers of sequences compared to the total number of sequences in the *P. putredinis* NO1 genome and showed increases in annotations for extracellular functions. Subsets containing 1933, 812 or 509 sequences were produced for the relaxed, strict or super strict secretomes, respectively.

This will allow comparative investigations with proteomic data to be more accurate and the identification of enzymes and other new proteins to be made much simpler.

2 | RESULTS AND DISCUSSION

2.1 | Designing a workflow to isolate the *P. putredinis* NO1 secretome

The first genome of the genus *Parascedosporium* provides a unique resource to explore for potentially new enzymes. The genome assembly is 39 Mb, consists of 21 contigs, and contains 9998 protein-coding sequences. *P. putredinis* NO1 belongs to the Microascaceae family of ascomycete fungi and is the sister taxon of *Scedosporium* species, of which genomes for four species are available. The genomic repertoires of predicted lignocellulose degrading enzymes were compared previously between *P. putredinis* NO1 and *Scedosporium boydii* and were found to be very similar (Scott, 2023). To identify and isolate sequences in the *P. putredinis* NO1 genome predicted to encode proteins that are actively secreted into the extracellular space, a secretome isolation workflow was designed (Figure 1). Localisation prediction was performed using the tools DeepLoc and BUSCA to identify sequences predicted to encode extracellular proteins. SignalP, TargetP and SecretomeP were used to identify sequences predicted to contain proteins with secretion signal peptides. SecretomeP was simultaneously used to attempt to predict sequences encoding non-classically secreted proteins. All sequences were submitted to TMHMM for the prediction of sequences encoding proteins containing transmembrane helices and therefore transmembrane proteins. Multiple annotation strategies were used to build information on the potential functions of all sequences of the *P. putredinis* NO1 genome and were used for evaluation of the workflow.

2.2 | Investigating discrepancies in prediction tools of the secretome isolation workflow

Differences in the sequences captured by each tool were observed for both localisation and secretion branches of the workflow, highlighting the importance of utilising multiple tools during secretome prediction. This also reflects the differences in conventional and unconventional release of extracellular proteins in fungi. Proteins favouring vesicle-mediated release and without the presence of secretion signal peptides would be missed if only tools for conventionally secreted protein prediction (Miura & Ueda, 2018; Rodrigues et al., 2011). This is especially relevant here in the context of lignocellulose breakdown as vesicle-mediated secretion of lignocellulose-degrading enzymes has been demonstrated for *Trichoderma reesei*, another ascomycete degrader of plant biomass (de Paula et al., 2019). In total, 769 coding regions of the *P. putredinis* NO1 genome were predicted to be extracellular by DeepLoc. Considerably more sequences at 1588 were predicted to be extracellular by BUSCA and 622 of these sequences were identified
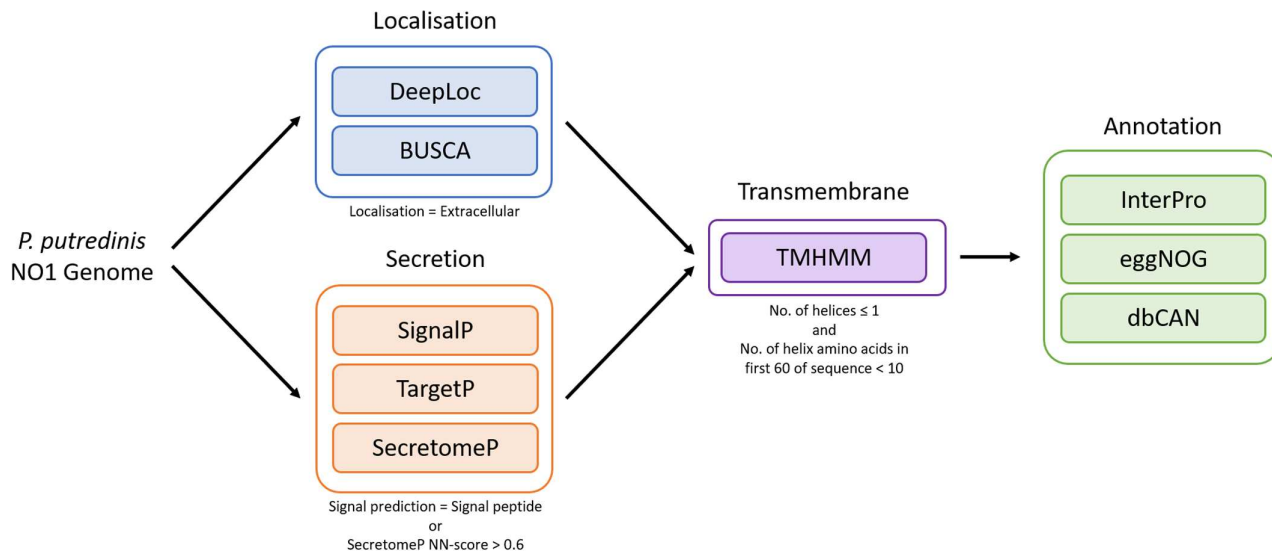
FIGURE 1 Secretome isolation workflow. The bioinformatic workflow was developed to identify and isolate coding regions of the *P. putredinis* NO1 genome predicted to encode actively secreted extracellular proteins.

by both tools (Figure 2a). Comparing prediction of classical secretion signal peptides across the three secretion tools revealed 798 protein sequences predicted to contain signal peptides by all three tools, 975 by at least two tools and 411 exclusively by SecretomeP (Figure 2b). SecretomeP also predicted an unusually large number of sequences, 3226, to encode non-classically secreted proteins. These predictions showed little overlap with sequences predicted to encode classically secreted proteins by SignalP and TargetP (Figure S1). This suggested overprediction for non-classically secreted proteins with this tool.

2.3 | Evaluating individual tools of the secretome isolation workflow

COG category assignments and KEGG metabolic pathway terms from the eggNOG mapper and CAZyme class annotations from dbCAN were used to evaluate the effectiveness of the workflow. Proportions of each annotation in the whole genome were compared to the subsets of proteins from each of the prediction tools. Both localisation prediction tools and all classical secretion prediction tools demonstrated expected increased proportions of functional annotations associated with proteins of fungal secretomes. These tools demonstrated increased proportions of sequences assigned to COG category G for carbohydrate metabolism, which is expected as carbohydrate breakdown begins outside of the cell. The tools showed reductions in the proportion of intracellular Glycosyl Transferase (GT) class CAZymes (Moremen & Haltiwanger, 2019). Finally, all tools except for classical secretion prediction by SecretomeP showed increased proportions of assignments to the KEGG metabolic pathway for carbohydrate metabolism. Other increases and reductions were observed for all validation methods respective to the whole-genome annotation, and these patterns varied by tool (Figures S2–4A–F). This is likely a reflection of the different sequences captured by each tool due to the different prediction methods.
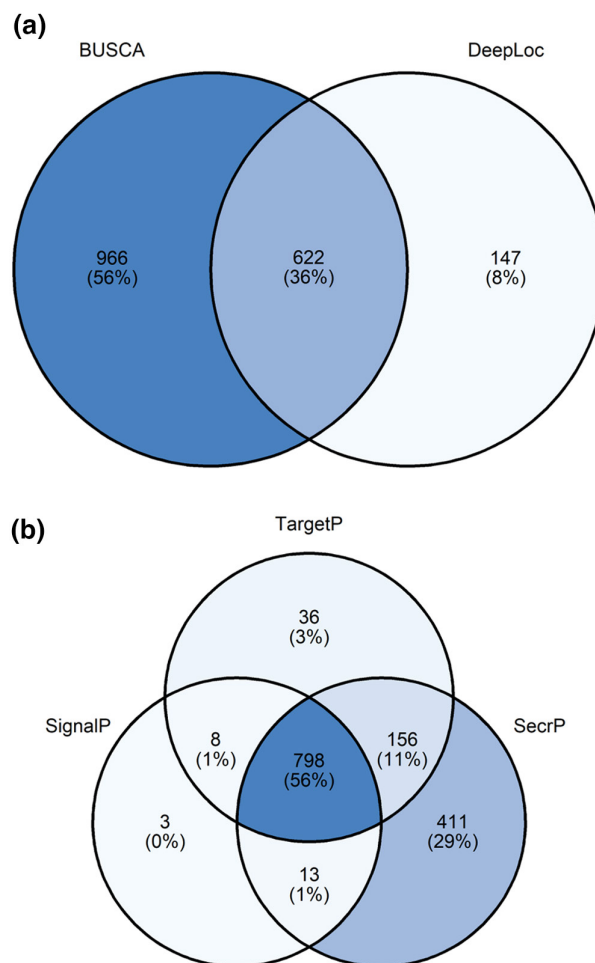

FIGURE 2 Visualising discrepancies between prediction tools. The differences in sequences predicted to encode extracellular proteins by each localisation prediction tool (a), and sequences predicted to encode classically secreted proteins by each secretion prediction tool (b).

For new enzyme identification, this is important for capturing as much of the secretome as possible. Sequences predicted to encode non-classically secreted proteins by SecretomeP did not show the expected increases in proportions. SecretomeP is designed for bacterial or mammalian sequences but has been used for fungal secretome prediction before and so was investigated here (Alfaro et al., 2016). However, its inability to accurately isolate sequences encoding secretome proteins, and due to the unusually large number of sequences captured by SecretomeP, non-classical secretion was omitted from secretome isolation. The tools used in this workflow could be readily applied to any sequence data from other fungal species or other microorganisms and the tools contain options for either eukaryotic or prokaryotic prediction. The removal or incorporation of transmembrane proteins could be altered based on the purpose of the in silico secretome isolation or based on the microorganism of interest. For example, in the context of lignocellulose breakdown some anaerobic fungi and bacteria have been demonstrated to assemble extracellular cell surface tethered constructs known as cellulosomes to enhance lignocellulose dissolution and product uptake (Artzi et al., 2017). Therefore, the membrane-associated proteins involved in these structures may be of interest for investigations of these microorganisms.

2.4 | Filtering the *P. putredinis* NO1 genome to isolate the secretome

Discrepancies between prediction tools inspired the creation of multiple 'secretome' subsets with different levels of stringency. Sequences predicted to encode extracellular proteins and sequences predicted to encode secretion signals were merged into three final subsets: relaxed, strict and super strict. For each subset, proteins predicted to encode transmembrane proteins by TMHMM were removed to give the final 'secretomes'. Conventionally, fungal secretomes consider membrane-bound extracellular proteins (Oates et al., 2021), however, for the purpose of investigating lignocellulose breakdown by aerobic fungi, these were omitted as the free extracellular proteins are more likely to be involved in depolymerisation reactions. If membrane-bound proteins are of interest, then this step could be removed. The 'relaxed' secretome subset contained coding regions predicted to encode extracellular proteins by at least one localisation prediction tool or predicted to encode secreted proteins by at least one secretion prediction tool, totalling 1933 sequences. The 'strict' secretome subset contained coding regions predicted to encode extracellular proteins by both localisation tools or sequences predicted to encode secreted proteins by all three secretion signal prediction tools, totalling 812 sequences. Finally, the 'super strict' secretome contains sequences predicted to encode extracellular proteins by both localisation prediction tools and which were also predicted to encode secretion signals by all secretion prediction tools, totalling 509 sequences.

To evaluate the secretomes, the COG category, CAZyme class and KEGG pathway proportions were again compared for each secretome against the whole genome. All secretome subsets demonstrated increased proportions of protein sequences assigned to COG categories G (carbohydrate metabolism), O (post-translational modification/turn-over/chaperone functions) and X (Unassigned) (Figure 3). For category G, the degree of increase correlated with the strictness of the secretome subset (genome: 5.8%, relaxed: 12.7%, strict: 20.0%, super strict: 23.0%), the same pattern was observed for category O (genome: 5.1%, relaxed: 6.4%, strict: 10.3%, super strict: 12.5%), whereas for category X, it was the strict secretome that had the largest increase (genome: 24.5%, relaxed: 34.0%, strict: 37.8%, super strict: 37.1%). Protein sequences with no clear functional annotation were abundant in the *P. putredinis* NO1 genome and were assigned to category X for comparison with other functional categories. The large proportion of these proteins in the *P. putredinis* NO1 genome reflects the novelty of this organism with this being the first genome assembly of the *Parascedosporium* genus. These proteins also represent a reservoir of potentially interesting new sequences and even new activities. For many of the other COG categories, reductions were observed compared to the genome for all secretome subsets and again the degree of reduction increased with how strictly the secretomes were filtered. Regarding some COG categories, differences between the secretomes were observed. For COG category E (amino acid metabolism and transport), a slight increase in proportions of assignments was observed for the relaxed and strict secretomes but a reduction was seen in the super strict secretome (genome: 4.5%, relaxed: 5.9%, strict: 4.7%, super strict: 3.9%). Only the relaxed secretome showed an increase in the proportion of assignments to category Q (secondary structure) (genome: 3.7%, relaxed: 4.0%, strict: 1.8%, super strict: 1.4%). The reasons for the increases in these COG categories that are not expected to include extracellular enzymes were not clear but may be due to misassignment of secretome proteins to these categories. Importantly, all secretomes showed increased proportions of proteins that lacked any functional annotations (i.e. assigned to category X). All secretomes also contained large proportions of protein sequences assigned to COG Category S (Function Unknown) (genome: 19.9%, relaxed: 20.2%, strict: 18.3%, super strict: 17.2%). Protein sequences assigned to this category were found to have predicted putative domains but lacked an overall functional annotation. Altogether, the sequences belonging to S and X categories represent an important subset of sequences for new enzyme identification and the persistence of these proteins in all secretomes demonstrates how such a bioinformatic workflow can isolate a subset of protein sequences that potentially contain new enzymes and activities. Considering the lignocellulose-degrading lifestyle of *P. putredinis* NO1 and the previous identification of a new secreted phenol oxidase enzyme involved in lignocellulose breakdown, it can be hypothesised that this fungus may contain other new lignocellulose-degrading enzyme activities (Oates et al., 2021). Indeed, the protein sequence encoding this new enzyme was isolated in all secretomes and assigned to COG category S, inspiring confidence that other new enzymes belonging to S and X categories have also been isolated.

When comparing CAZyme class annotation proportions across the secretomes and the whole genome, all secretomes had largely reduced proportions of intracellular Glycosyl Transferase (GT) class CAZymes (Figure 4) (Moremen & Haltiwanger, 2019). In particular, the super strict secretome contained no GT class CAZymes

(genome: 14.2%, relaxed: 3.1%, strict: 1.3%, super strict: 0.0%). Mirroring this GT reduction, all secretomes had increased proportions of Carbohydrate Esterase (CE) (genome: 6.4%, relaxed: 8.5%, strict: 10.0%, super strict: 11.8%), carbohydrate-binding module (CBM) (genome: 20.3%, relaxed: 25.6%, strict: 29.8%, super strict: 32.3%) and polysaccharide lyase (PL) class CAZyme sequences (genome: 2.3%, relaxed: 3.3%, strict: 3.6%, super strict: 3.9%). These increases

correlated with the strictness of the filtering used to obtain each secretome. All secretomes had increased proportions of auxiliary activity (AA) class CAZymes (genome: 20.4%, relaxed: 23.6%, strict: 25.4%, super strict: 21.1%), however, the super strict secretome had the smallest increase. As these CAZyme classes all predominantly act on extracellular substrates, this was expected. A reduction in Glycoside Hydrolase (GH) class proportions was seen for all
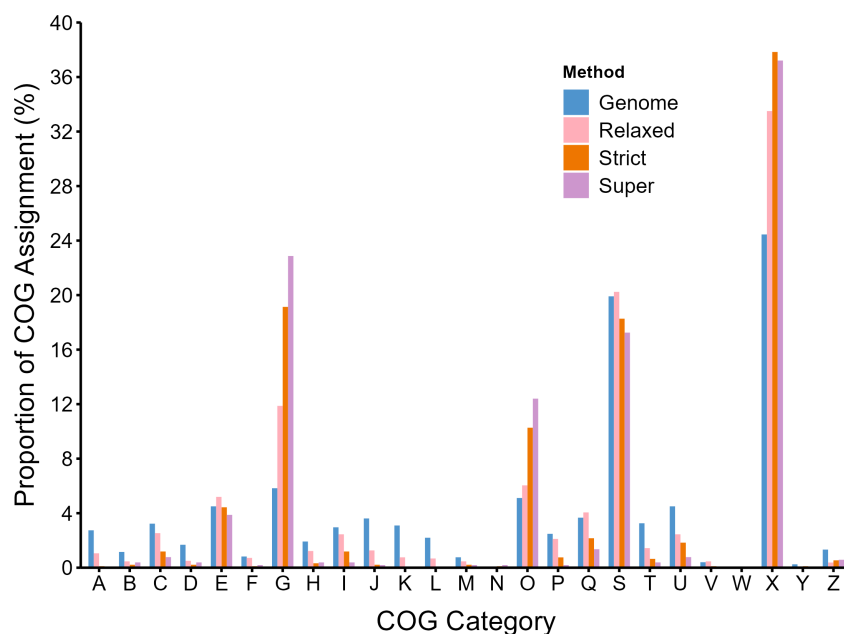

FIGURE 3 Investigating COG annotation proportions for secretome protein sequences. The proportion of each COG category in the whole-genome COG annotation compared to the relaxed, strict and super strict secretome subsets. A, RNA processing and modification; B, chromatin structure; C, energy production and conversion; D, cell cycle control and mitosis; E, Amino acid metabolism and transport; F, nucleotide metabolism and transport; G, carbohydrate metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; J, translation; K, transcription; L, replication and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, post-translational modification/turnover/chaperone functions; P, Inorganic ion transport and metabolism; Q, Secondary structure; T, signal transduction; U, intracellular trafficking and secretion; Y, nuclear structure; Z, cytoskeleton; S, function unknown; X, unassigned.
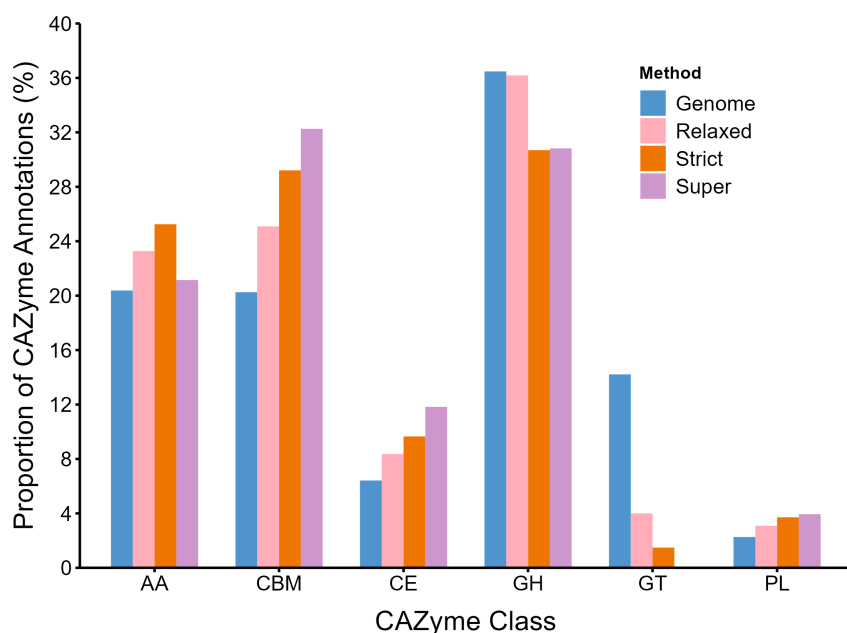

FIGURE 4 Investigating CAZyme annotation proportions for secretome protein sequences. The proportion of each CAZyme class in the whole-genome CAZyme annotation compared to the relaxed, strict and super strict secretome subsets. AA, auxiliary activity; CBM, carbohydrate-binding module; CE, carbohydrate esterase; GH, glycoside hydrolase; GT, glycosyl transferase; PL, polysaccharide lyase.

secretomes with the strict subset showing the largest reduction (genome: 36.5%, relaxed: 35.8%, strict: 29.6%, super strict: 30.8%). This is the largest class of CAZymes and many members of this class have been suggested to have intracellular activities previously (Park et al., 2018). Therefore, a reduction in the proportion of GH CAZymes may be expected.

KEGG metabolic pathway assignments were found to be less clear than COG categories and CAZyme class annotations and tended to differ more between tools. However, expected increases in the proportions of assignments to metabolic classes were still observed. All secretomes show increased proportions of assignments to the carbohydrate metabolism pathway (genome: 21.2%, relaxed: 21.5%, strict: 28.8%, super strict: 38.9%), however, this is only a slight increase for the relaxed secretome (Figure 5). A similar pattern was observed for the metabolism of other amino acid pathways (genome: 4.2%, relaxed: 4.6%, strict: 6.3%, super strict: 9.3%). For other pathways, differences were observed between the secretomes. For example, no assignments to the xenobiotics biodegradation and metabolism pathway were present in the super strict secretome, although the strict secretome only showed a slight reduction in proportion of assignments to this pathway compared to the genome, and the relaxed secretome even showed an increase (genome: 5.5%, relaxed: 9.7%, strict: 4.5%, super strict: 0.0%). Only the strict secretome showed an increased proportion of assignments to the metabolism of terpenoids and polyketide pathways (genome: 2.1%, relaxed: 1.7%, strict: 3.6%, super strict: 1.9%). The strict and super strict secretomes were observed to have increased proportions of assignments to the glycan biosynthesis and metabolism pathway (genome: 6.3%, relaxed: 4.3%, strict: 9.0%, super strict: 9.3%). These assignments may be the result of mis-assignment due to the action of many extracellular fungal enzymes on the β-glycosidic linkages that are present in glycans

(Krautter & Iqbal, 2021). The super strict secretome had a large increase in the proportion of assignments to the biosynthesis of other secondary metabolites, whereas the relaxed and strict secretomes only showed slight increases (genome: 3.4%, relaxed: 4.6%, strict: 4.5%, super strict: 9.3%). The relaxed and strict secretome showed increases in assignments to amino acid metabolism pathways compared to a reduction in assignments to these pathways for the super strict secretome (genome: 19.6%, relaxed: 21.8%, strict: 20.7%, super strict: 14.8%).

To determine and quantify the effect of the workflow on the *P. putredinis* NO1 protein data set, principal component analysis of K-Means clustered outputs for each of the tools was performed (Figure 6). The first, second and third components explained 40.25%, 20.69% and 11.29% of the total variance respectively. Cluster 1 (pink; $N = 3583$; SecretomeP outlier subset) contained proteins exclusively with positive identifications from SecretomeP which were proteins omitted from the original workflow. Cluster 2 (blue; $N = 1224$; secretome subset) clustered distinctly separately from the other clusters and consisted of proteins with positive secretion results from all tools and as such represents the identified secretome. Cluster 3 (orange; $N = 1169$; transmembrane subset) contained proteins weighted towards both positive DeepLoc outputs and a higher number of TMHMM helices suggesting this cluster contained predominantly transmembrane proteins. Cluster 4 (purple; $N = 4023$; intracellular) contained proteins with little to no positive identifications from any tools and as such represented the intracellular fraction. While further resolution in subcellular localisations could be achieved with this data, these results support the identification of the putative secretome with this workflow.

Overall, the expected patterns were observed for all three secretomes which suggests the ability of this workflow to capture secretome proteins.
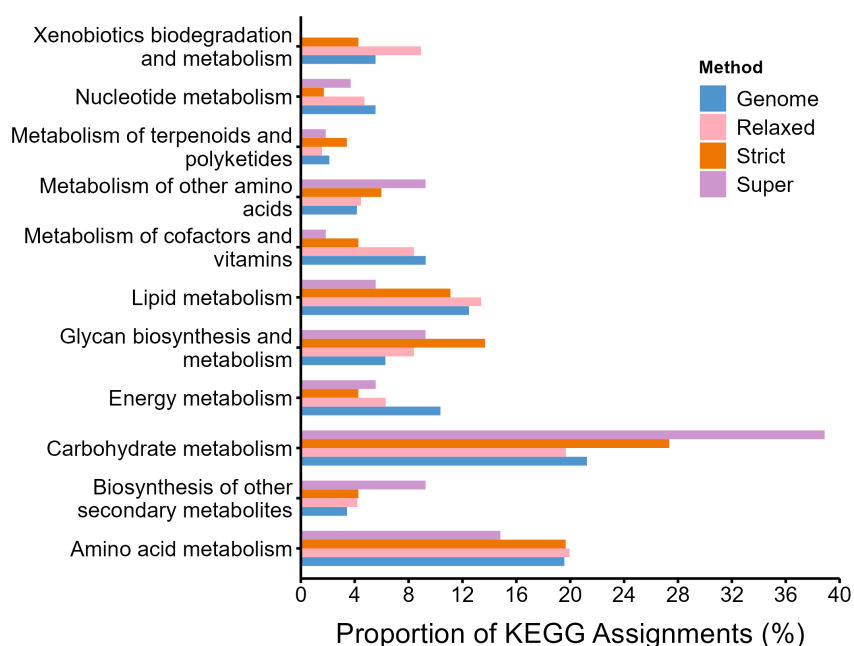

FIGURE 5 Investigating KEGG pathway annotation proportions for secretome protein sequences. The proportion of KEGG assignments to metabolic pathways in the whole genome compared to the relaxed, strict and super strict secretome subsets.
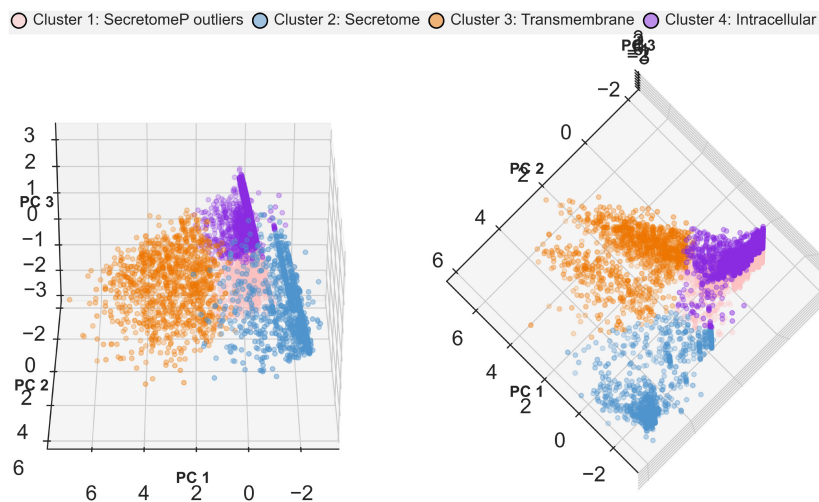
○ Cluster 1: SecretomeP outliers ○ Cluster 2: Secretome ○ Cluster 3: Transmembrane ○ Cluster 4: Intracellular

FIGURE 6 Principal component analysis of K-means clustered predictions for protein localisation. K-Means clustering was performed using protein localisation predictions generated for each protein ($n = 9998$) from workflow tools; BUSCA, DeepLoc, SignalP, SecretomeP, TargetP and TMHMM. Clusters ($n = 4$) are coloured by K-Means cluster. Subplots are ordinated for clarity. PC, principal component.

3 | CONCLUSIONS

Here, a workflow for predicting a fungal in silico secretome from genomic sequences was presented and evaluated. The workflow incorporated two strategies of localisation and secretion prediction both using multiple tools. Each tool demonstrated patterns in the proportions of annotations associated with secretome proteins compared to the whole genome annotation. This included increased proportions of assignments to pathways for carbohydrate metabolism and reduced proportions of intracellular GT class CAZymes. There were also differences in patterns of proportional increases and reductions across the tools for each of the evaluation methods. This demonstrated the importance of incorporating multiple tools into secretome prediction workflows. The expected patterns were also observed in the final filtered secretomes. Again, differences were observed between the secretomes, however, the main patterns were observed for all secretomes, and intensity correlated with the stringency of filtering.

Depending on the purpose of the in silico secretome, it is envisaged that different secretomes could be used. For example, for new enzyme identification, the relaxed secretome would be more appropriate to maximise captured sequences where novel sequences are unlikely to be captured by all secretome prediction tools. For the identification of new extracellular enzymes where functional annotation is impossible due to high sequence or structural divergence, bioinformatic workflows like that presented here can quickly and simply allow these enzyme sequences to be isolated. Indeed, sequences assigned to COG categories for unknown functions were present in all secretomes alongside the new phenol oxidase enzyme identified previously from this fungus (Oates et al., 2021). This suggested that this workflow can effectively isolate subsets of sequences encoding potentially new enzymes and activities. Combination with proteomic data would reduce the number of sequences further. In contrast, for comparative secretomic studies, the stricter secretomes would be favourable as confidence is increased that most of the sequences in these secretomes are truly secreted and extracellular. The workflow is readily adaptable across eukaryotic and prokaryotic organisms, as all tools used here have options for predictions from bacterial sequences. Modifications can be made through decisions on the incorporation or removal of protein sequences at each stage of the workflow, for example, whether to include extracellular transmembrane protein sequences as part of the predicted secretome. As easily as extracellular proteins can be identified they can also be removed if interest is instead focused on intracellular proteins. As new tools are developed and become popularised, they can easily be incorporated into the workflow.

4 | EXPERIMENTAL PROCEDURES

4.1 | Localisation prediction

Localisation prediction was carried out for all predicted coding regions of the *P. putredinis* NO1 genome using the online tool BUSCA (Savojardo et al., 2018) and with DeepLoc v1.0 on the high-performance computing cluster at the University of York (Almagro Armenteros et al., 2017).

4.2 | Secretion prediction

Secretion signal prediction was carried out for all predicted coding regions of the *P. putredinis* NO1 genome using the online tools SignalP v6.0 (Teufel et al., 2022) and TargetP v2.0 (Almagro Armenteros et al., 2019). Secretion signal prediction and simultaneous non-classical secretion prediction were performed using the online tool SecretomeP v2.0, where sequences not predicted to encode signal peptides but with an NN score >0.6 were predicted to be non-classically secreted (Bendtsen et al., 2004).

4.3 | Transmembrane helices prediction

Transmembrane helix prediction for the identification of transmembrane proteins was performed for all predicted coding regions of the *P. putredinis* NO1 genome using the online tool TMHMM v2.0 (Krogh et al., 2001). Sequences predicted to encode more than one

transmembrane helix were assumed to be transmembrane proteins. Sequences predicted to encode a single transmembrane helix, but with less than 10 amino acids of this helix occurring in the first 60 amino acids of the protein sequence (indicating a signal peptide) were also assumed to be transmembrane proteins and were also removed.

4.4 | Sequence annotation

Sequences of all predicted coding regions in the *P. putredinis* NO1 genome were annotated for COG categories and KEGG pathway annotations using the online tool eggNOG mapper v2 (Cantalapiedra et al., 2021). CAZyme domain annotation with dbCAN (Zhang et al., 2018) of all predicted coding regions was performed using the CAZyme database v09242921 as described previously (Scott, 2023).

4.5 | Prediction clustering

K-Means cluster principal component analysis (PCA) was performed with scikit-learn (SKlearn) (Pedregosa et al., 2011) on output data from each of the tools. Categorical variables were factorised into secretion positive results (Oates et al., 2021) and secretion negative results (0). Numerical outputs (TMHMM First60 and helices and SecretomeP NN values) were unchanged. Within-cluster sum of squares indicated a four-cluster solution was optimal.

4.6 | Secretome isolation

The database of localisation prediction, secretion signal prediction, transmembrane helix prediction and annotation for all predicted coding regions of the *P. putredinis* NO1 genome was assembled in R v4.2.0 (R Development Core Team, 2022). Evaluation of each tool used was performed using annotation information and plotted in R using the ggplot2 package ggplot2 (Villanueva & Chen, 2019).

AUTHOR CONTRIBUTIONS
Conor J. R. Scott: Conceptualization; investigation; writing – original draft; methodology; visualization; formal analysis. **Daniel R. Leadbeater:** Visualization; formal analysis. **Neil C. Bruce:** Supervision; project administration.

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT
The sequence data generated and analysed during the current study are available in the European Nucleotide Archive, project code PRJEB60285, secondary accession ERP145344 (https://www.ebi.ac.uk/ena/browser/view/PRJEB60285). The WGS Sequence Set for the genome assembly is available in the European Nucleotide Archive, Accession CASHTG010000000.1 (https://www.ebi.ac.uk/ena/browser/view/CASHTG010000000). The assembly is also available through the NCBI database, Accession GCA_949357655.1.

ETHICS STATEMENT
Ethics approval was not required for this study.

ORCID
Conor J. R. Scott https://orcid.org/0000-0001-7404-7619
Daniel R. Leadbeater https://orcid.org/0000-0002-2228-5604
Neil C. Bruce https://orcid.org/0000-0003-0398-2997

REFERENCES
Alfaro, M., Castanera, R., Lavin, J.L., Grigoriev, I.V., Oguiza, J.A., Ramirez, L. et al. (2016) Comparative and transcriptional analysis of the predicted secretome in the lignocellulose-degrading basidiomycete fungus *Pleurotus ostreatus*. *Environmental Microbiology*, 18(12), 4710–4726.

Almagro Armenteros, J.J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A. et al. (2019) Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2(5), e201900429.

Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H. & Winther, O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21), 3387–3395.

Artzi, L., Bayer, E.A. & Morais, S. (2017) Cellulosomes: b nanomachines for dismantling plant polysaccharides. *Nature Reviews Microbiology*, 15(2), 83–95.

Bendtsen, J.D., Jensen, L.J., Blom, N., Von Heijne, G. & Brunak, S. (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering, Design & Selection*, 17(4), 349–356.

Cantalapiedra, C.P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. (2021) eggNOG-mapper v2: functional annotation, Orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, 38(12), 5825–5829.

de Paula, R.G., Antonieto, A.C.C., Nogueira, K.M.V., Ribeiro, L.F.C., Rocha, M.C., Malavazi, I. et al. (2019) Extracellular vesicles carry cellulases in the industrial fungus *Trichoderma reesei*. *Biotechnology for Biofuels*, 12, 146.

Gogleva, A., Drost, H.G. & Schornack, S. (2018) SecretSanta: flexible pipelines for functional secretome prediction. *Bioinformatics*, 34(13), 2295–2296.

Krautter, F. & Iqbal, A.J. (2021) Glycans and glycan-binding proteins as regulators and potential targets in leukocyte recruitment. *Frontiers in Cell and Developmental Biology*, 9, 624082.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3), 567–580.

Miura, N. & Ueda, M. (2018) Evaluation of unconventional protein secretion by *Saccharomyces cerevisiae* and other fungi. *Cells*, 7(9), 128.

Moremen, K.W. & Haltiwanger, R.S. (2019) Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nature Chemical Biology*, 15(9), 853–864.

Nielsen, H. (2017) Predicting secretory proteins with SignalP. In: Kihara, D. (Ed.) *Protein function prediction: methods and protocols*. New York, NY: Springer New York, pp. 59–73.

Oates, N.C., Abood, A., Schirmacher, A.M., Alessi, A.M., Bird, S.M., Bennett, J.P. et al. (2021) A multi-omics approach to lignocellulolytic enzyme discovery reveals a new ligninase activity from *Parascedosporium putredinis* NO1. *Proceedings of the National Academy of Sciences of the United States of America*, 118(18), e2008888118.

Park, Y.J., Jeong, Y.U. & Kong, W.S. (2018) Genome sequencing and carbohydrate-active enzyme (CAZyme) repertoire of the white rot fungus *Flammulina elastica*. *International Journal of Molecular Sciences*, 19(8), 2379.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12(85), 2825–2830.

R Development Core Team. (2022) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rodrigues, M.L., Nosanchuk, J.D., Schrank, A., Vainstein, M.H., Casadevall, A. & Nimrichter, L. (2011) Vesicular transport systems in fungi. *Future Microbiology*, 6(11), 1371–1381.

Savojardo, C., Martelli, P.L., Fariselli, P., Profiti, G. & Casadio, R. (2018) BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Research*, 46(W1), W459–W466.

Scott, C.J.R., Leadbeater, D.R., Oates, N.C., James, S.R., Newlingm, K., Li, Y., et al. (2023) Whole genome structural predictions reveal hidden diversity in putative oxidative enzymes of the lignocellulose degrading ascomycete. *Parascedosporium putredinis*, NO1. bioRxiv. 2023. doi: https://doi.org/10.1101/2023.08.08.552407

Teufel, F., Armenteros, J.J.A., Johansen, A.R., Gislason, M.H., Pihl, S.I., Tsirigos, K.D. et al. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40(7), 1023–1025.

Villanueva, R.A.M. & Chen, Z.J. (2019) ggplot2: elegant graphics for data analysis, 2nd edition. *Measurement: Interdisciplinary Research and Perspectives*, 17(3), 160–167.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z. et al. (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 46(W1), W95–W101.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.