Beyond observation: genomic traits and machine learning algorithms for predicting fungal lifestyles

Yanpeng Chen
School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China

Pengwei Su
School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China

Marc Stadler
Rong Xiang
Kevin D. Hyde
Wenhui Tian
Sajeewa S. N. Maharachchikumbura (sajeewa83@yahoo.com)

Research Article

Keywords: FCWDEs, Genomics, machine learning, PCWDEs, secretome, TEs

Posted Date: June 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3118609/v1

Additional Declarations:
Supplementary Figures and Tables are not available with this version.

Beyond observation: genomic traits and machine learning algorithms for predicting fungal lifestyles

Yanpeng Chen[1], Pengwei Su[1], Marc Stadler[2,3], Rong Xiang[4], Kevin D. Hyde [5,6], Wenhui Tian[1], Sajeewa S. N. Maharachchikumbura[1*]

[1]School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]Helmholtz Centre for Infection Research GmbH, Department Microbial Drugs and German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, 38124 Braunschweig, Germany
[3]Institute of Microbiology, Technische Universität Braunschweig, Inhofenstraße 7, 38124 Braunschweig, Germany
[4]Precision Medicine Center, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 404100, China
[5]Center of Excellence in Fungal Research, Mae Fah Luang University, Chaing Rai 57100, Thailand
[6]Innovative Institute for Plant Health, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

***Corresponding author:**
Sajeewa S. N. Maharachchikumbura

Correspondence:
Sajeewa S. N. Maharachchikumbura: sajeewa83@yahoo.com or sajeewa@uestc.edu.cn

Abstract

Economically and agriculturally important fungal species have various lifestyles, and they may shift from mutualistic or saprobic to pathogenic depending on the habitat, host tolerance, and resource availability. Traditionally, the determination of fungal lifestyles has been based on observation at a particular host or habitat. Therefore, potential fungal pathogens have been neglected until they cause devastating impacts on human health, food security, and ecosystem stability. This study focused on the class Sordariomycetes to explore the genomic traits that could be used to determine the lifestyles of fungi and the possibility of predicting fungal lifestyles using machine learning algorithms. A total of 638 representative genomes covering five subclasses, 17 orders and 50 families were selected and annotated. Through an extensive literature survey, the lifestyles of 555 genomes were determined, including plant pathogens, saprotrophs, entomopathogens, mycoparasites, endophytes, human pathogens and nematophagous fungi. We evaluated the influence of sequencing technologies and concluded that second sequencing technologies have no influence on genome completeness but tend to generate a reduced size of transposable elements. We constructed three numerical matrices: a basic genomic feature matrix including 25 features; a functional protein matrix including 24 features; and a combined matrix. The most comprehensively comparative analysis to date across multiple lifestyles was conducted based on these matrices. Results indicate that basic genomic features reflect more on phylogeny rather than lifestyle, but the abundance of functional proteins displays relatively high discrimination not only in differentiating taxonomic groups at the higher levels but also in differentiating lifestyles. Genome size, GC content and gene number showed powerful discrimination for differentiating higher ranks, especially at the subclass level. Plant pathogens have the largest secretome; whereas entomopathogens have the smallest secretome; and the abundance of secretomes is a useful indicator to clearly differentiate plant pathogens from entomopathogens, mycoparasites, saprotrophs and entomopathogens,

and as well as differentiate entophytes from entomopathogens. Effectors have long been considered as disease determinants, and we did observe that plant pathogens have more effectors than saprotrophs and entomopathogens. However, we also observed a similar abundance of effectors in endophytes, suggesting that effectors maybe not a reliable indicator for pathogenic fungi. Single functional protein could not differentiate all lifestyles, but combinations of multiple numerical features of functional proteins result in accurate differentiation for most lifestyles. Furthermore, models of six machine learning algorithms were trained, optimized and evaluated, and the best-performance model was used to predict the lifestyle of 83 unlabeled genomes. Although the accuracy of the best machine learning model was limited by the inadequate genome number of several lifestyles and the inaccurate lifestyle assignments for some genomes, the predictive model still obtained a high degree of accuracy in differentiating plant pathogens. The predictive model can be further optimized with more sequenced genomes in the future, and provide a more reliable prediction. This can be used as an early warning system to identify potentially devastating fungi and take appropriate measures to prevent their spread.

Keywords: FCWDEs, Genomics, machine learning, PCWDEs, secretome, TEs

Introduction

The class Sordariomycetes, established by Eriksson and Winka (Eriksson OE 1997), is the second-largest class of the phylum Ascomycota (Hyde et al. 2020). Based on the latest outline of Wijayawardene et al. (2022), it comprises 7 subclasses, 46 orders and 172 families. The perithecial ascomata and inoperculate unitunicate asci are the main diagnostic features for distinguishing Sordariomycetes from other classes (Maharachchikumbura et al. 2015). Most Sordariomycete species have been introduced based solely on either the anamorph or teleomorph, and only a small number of them were characterized based on both anamorph and teleomorph (Wingfield et al. 2012; Maharachchikumbura et al. 2015; Réblová et al. 2016). Sordariomycete species are distributed worldwide and have been found in almost every ecosystem (Wang et al. 2018; Luo et al. 2019; Kwon et al. 2021; Maharachchikumbura et al. 2021). Although most Sordariomyetes are saprobic on organic matter from various plants, the class also includes several notorious plant pathogens. For instance, *Pyricularia oryzae* (syn. *Magnaporthe oryzae*; Magnaporthales, Pyriculariaceae), *Fusarium graminearum, F. oxysporum* (Hypocreales, Nectriaceae), and *Colletotrichum* species (Glomerellales, Glomerellaceae), are listed in the top 10 fungal plant pathogens (Dean et al. 2012). Moreover, several species, such as *Pyricularia grisea* and *Ophiostoma* spp, were recognized as invasive plant pathogens, which altered the local natural ecosystems (Anderson et al. 2004; Solla et al. 2005). Some species are related to human and animal diseases (Barros et al. 2011; Troy et al. 2013; Tortorano et al. 2014; Řehulka et al. 2016; Jenks et al. 2018), while other species are of great importance to medicine, agriculture, and industry (Crawford et al. 1952; Kaewchai et al. 2009; Xu et al. 2014).

Diverse lifestyles, including saprotrophic, necrotrophic, hemibiotrophic and biotrophic are present in Sordariomycetes, all of which represent distinct survival strategies evolved by fungi during their interactions with their hosts, companions and associated environments (Presti et al. 2015; Boddy 2016; Rai and Agarkar 2016). Saprotrophs live and feed on non-living organic matter from other organisms, contributing to the global carbon cycle by breaking down complex organic matter into simpler substances (Hobbie and Horton 2007; Mäkelä et al. 2014). Fungi of necrotrophic, hemibiotrophic or biotrophic lifestyles are important plant pathogens that pose a serious threat to economically important crops and are responsible for serious losses in quality and yield (Mapuranga et al. 2022). Necrotrophic fungi have a broad host range, and commonly produce diverse toxic molecules (e.g., lytic enzymes, metabolites) to kill host cells and subsequently derive nutrients from dead or dying tissues for growth (van Kan 2006; Mengiste 2012; Singh et al. 2014; Ismaiel and Papenbrock 2015; Newman and Derbyshire 2020). Biotrophic fungi are obligate parasites, which are completely dependent on the living host to complete their life cycles and therefore have to maintain host viability (Glazebrook 2005; Delaye et al. 2013). Hemibiotrophic fungi begin with an early biotrophic phase with their hosts, switching to a necrotrophic lifestyle after killing the host cells (Mendgen and Hahn 2002; Lee and Rose 2010). Endophytic fungi absorb nutrients from plant cells without causing visible symptoms of disease, sometimes in return benefiting plant growth via enhancing the plant's tolerance to abiotic (e.g., drought and salt) and biotic stresses (e.g., insects and other fungal pathogens) (Jia et al. 2016; Phurailatpam and Mishra 2020; Fontana et al. 2021; Wu et al. 2021). In accordance with differences in hosts and substrates, Sordariomycetes are also characterized as plant pathogens, animal pathogens, insect pathogens and mycoparasites. Some fungi are capable of switching between lifestyles. Transitions from the endophytic lifestyle to the pathogenic lifestyle and vice versa have been observed in some important fungal plant pathogens (O'Connell et al. 2012; Rai and Agarkar 2016; Liu et al. 2022).

Lifestyle-associated genomic traits are a particularly interesting area of research, as pathogenic transitions are highly relevant to gene gain and loss (Friesen et al. 2006; Spanu et al. 2010). *Pyrenophora tritici-repentis* (Pleosporaceae, Pleosporales, Dothideomycetes) becomes highly pathogenic on wheat (*Triticum aestivum*) by

obtaining the proteinaceous host-specific toxin *ToxA* from *Stagonospora nodorum* (Phaeosphaeriaceae, Pleosporales, Dothideomycetes), demonstrating that the transfer of the virulence gene is an essential source for the emergence of new pathogens (Friesen et al. 2006). An exclusively biotrophic lifestyle is related to gene losses of primary and secondary metabolic enzymes (Spanu et al. 2010). The convergent losses of decay-related genes and the expansion of symbiosis-related genes are the genetic bases for the evolution of mycorrhizal habits (Kohler et al. 2015). Transposable elements (TEs), also known as "jumping genes," are crucial genetic factors in both eukaryotic and prokaryotic genomes that shape the evolution of fungal genomes by altering genome plasticity and architecture, interrupting functional genes, generating novel genes or mediating horizontal gene transfer (Lorrain et al. 2021). TEs are critical contributors to fungal pathogenicity by facilitating the diversification of effector genes and even generating novel effector genes (Fouché et al. 2019). In addition, plant symbionts tend to have more TEs than animal parasites (Muszewska et al. 2017a).

To survive inside a host or a specific environment, fungi must be equipped with the necessary functional proteins to absorb nutrients or to overcome physical and chemical barriers posed by hosts (de Jonge et al. 2011; McCotter et al. 2016; Zeng et al. 2018). The term secretome refers to the complete secretory proteins of an organism, which are released outside the cells to decay substrates and interact with microbes, plants, animals, insects, and other fungi (Eastwood et al. 2011; Frey-Klett et al. 2011; Shang et al. 2015). The fungal secretome comprises various functional groups of protein, including carbohydrate-active enzymes (CAZymes), proteases, lipases, small-secreted proteins (SSPs) and other secretory proteins of unknown functions (Alfaro et al. 2014). Many comparative genomic studies have focused on fungal CAZymes, searching for possible connections between compositions of CAZymes and fungal lifestyles (Kubicek et al. 2014; Pellegrin et al. 2015; Kim et al. 2016; Knapp et al. 2018; Chang et al. 2022). CAZymes include many plant cell wall-degrading enzymes (PCWDEs), and their composition and abundance are often linked to a saprotrophic lifestyle, while this view has been challenged on the grounds that the highest number of CAZymes have been observed in plant pathogenic fungi (Zhao et al. 2013; Kubicek et al. 2014). Fungal effectors, also called virulence factors encoded by avirulence genes, are potent weapons used by fungal pathogens against plant and animal immunity (Stergiopoulos and Wit 2009; Kale and Tyler 2011). Most effectors are secreting cysteine-rich proteins and play an essential role in host-fungal interactions by suppressing host defense responses for promoting host colonization (Lu and Edwards 2016; Wang et al. 2020; Dasari et al. 2018). Some effectors are essential genetic factors in determining host species specificity, which help identify potential pathogenic fungi to certain plants (Li et al. 2020). Effector repositories have been considered to be potential markers for differentiating pathogenic and endophytic strains in the *Fusarium oxysporum* species complex (Czislowski et al. 2021).

Machine learning is a branch of artificial intelligence that is commonly subclassified into unsupervised and supervised methods (Deo 2015). The former is used to find naturally occurring connections or groupings within observations based on little knowledge or even no background information regarding the outcome of the results (Camacho et al. 2018). This is contrasted with the supervised method, which is the construction and optimization of model-based and well-constructed training data with observations and corresponding results (Bzdok et al. 2018). The model is then utilized to predict results of future instances. Both methods have been widely used for unearthing hidden information in big data or complex biological data (Ma et al. 2014; Xu and Jackson 2019). There are many applications of machine learning in species delimitation, such as in, successfully using unsupervised machine learning methods to assign arachnid taxa into species (Derkarabetian et al. 2019), developing a machine learning species identifier for the genus *Hebeloma* (Bartlett et al. 2022) and predicting fungal lifestyles of Dothideomycetes (Haridas et al. 2020). Moreover, machine learning has been used to

characterize and classify images of clinically and agriculturally important fungi, which avoids potentially subjective differences, reduces identification time, and lowers the costs (Zieliński et al. 2020; Tongcham et al. 2020).

To mine the association patterns of genomic traits and phylogeny and lifestyles, and further determine whether it is possible to predict lifestyles using machine learning approaches, we carried out a systematic bioinformatic analysis based on 638 Sordariomycete genomes. Firstly, we determined whether the sequencing technologies significantly influence genome assemblies and TE abundance, which exists theoretically and practically but has never been discussed in previous studies. Secondly, based on the study of Fijarczyk et al. (2022), we not only compared the basic genomic traits across multiple lifestyles but also the functional protein groups. Furthermore, we took the influence from phylogeny into account, and compared the difference of numerical genomic traits at different taxonomic levels for determining lifestyle and phylogeny, which is the most important determinant in shaping genomic traits. It is also an answer to resolve the long-standing controversy: whether differences in the secreted proteins reflect phylogeny or pathogenicity (Pellegrin et al. 2015). Finally, we explored whether it is possible to predict fungal lifestyles using machine learning algorithms.

Materials and Methods

Genome collection

The taxonomic scheme of the class Sordariomycetes has been updated continuously (Maharachchikumbura et al. 2015; Hyde et al. 2020; Wijayawardene et al. 2022), whereas the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=147550) does not keep up with the updates, and some genomes were assigned incorrect lineage information (Shen et al. 2020; Liu et al. 2022). To ensure the correctness of the taxonomic positions of selected genomes, a taxonomic framework table composed of all generic names in Sordariomycetes and the parent lineage information, was prepared according to the taxonomic outline (Wijayawardene et al. 2022), and some changes were added in keeping with the latest literature (Crous et al. 2021; Sun et al. 2021, Magyar et al. 2022, Sugita & Tanaka 2022). We used the “Ascomycota” as the search term in NCBI’s Genome Browser (https://www.ncbi.nlm.nih.gov/data-hub/genome/?taxon=4890, 12 August 2022) to obtain all records of Ascomycota genomes, and then a table, including assembly accession, organism name, strain identifiers, assemble level, and release date, was downloaded. Only records of the Sordariomycetes genome were retained according to the generic names, and the lineage information of the genus was also integrated into the table. These genomes were downloaded via NCBI command line tool datasets. Besides, we collected several genomes from JGI MycoCosm (Grigoriev et al. 2013) with written permission. More details, such as lifestyles, sources, and publication records, were determined by tracing the original literature, the sample details, and the description of the corresponding BioProject records. Strains isolated from soil were marked as saprotrophs. If the strains have two observed lifestyles, only one lifestyle was used as the training data, and the other lifestyle was used to check the predictions. For a small number of strains from certain habits or undetermined sources, we noted them according to the submitter’s description or as “Undetermined” in *Allantophomopsis lycopodina* ATCC 66958 (Leotiomycetes) was selected as the outgroup.

Assessment of genome completeness

Assessment of genome quality is the primary step in genomic studies, which is important to recognize potential

issues in subsequent analysis (Smits 2019). Benchmarking Universal Single-Copy Orthologs (BUSCO) is an ideal dataset for quantifying genome completeness (Simão et al. 2015) and conducting genome-scale phylogenetic inference (Shen et al. 2018; Shen et al. 2020; Manni et al. 2021). Here, we used BUSCO version 5.2.2 (Manni et al. 2021) with the ascomycota_odb10 database comprising 1,706 reference genes to assess the completeness of the genome assemblies. Only genomes with BUSCO gene content larger than 80% were retained for subsequent analyses.

Phylogenetic inference

The corresponding protein sequences of single-copy orthologs resulting from the BUSCO analysis were extracted and assembled into a single-locus dataset to conduct phylogenetic analysis. Each locus dataset was aligned using MAFFT version 7.310 (Katoh et al. 2002) with options "--auto --maxiterate 1000" that allow the program automatically to determine the approximate refinement strategy and conduct iterative refinement at most 1,000 times. Poorly aligned regions were removed using trimAl version 1.4 with the option "-gappyout", and the alignments with a length shorter than 100 were deleted. ModelFinder (Kalyaanamoorthy et al. 2017) implemented in IQ-TREE2 (Minh et al. 2020) was used to choose the best-fit evolution model of each alignment based on the Bayesian Information Criterion (BIC). All single-locus alignments were concatenated into a supermatrix using an in-house python script. A single evolution model was determined by the occurrence and used in concatenation-based phylogenetic analyses. Maximum-likelihood analysis was conducted using IQ-TREE2 with 1000 bootstrap replicates of the SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al. 2010) and 1000 bootstrap replicates of ultrafast bootstrap approximation (UFBoot) (Hoang et al. 2017) to estimate the reliability of each internal branch. The strain *Allantophomopsis lycopodina* ATCC 66958 served as an outgroup to root the phylogeny.

Identification and analysis of repetitive elements

A *de novo* library of repeat consensus sequences was generated for each genome using RepeatModeler version 2.0.2 with search engine NCBI-RMBLAST version 2.11.0+. Next, repetitive sequences in genomes were identified and soft-masked using RepeatMasker version 4.1.2 based on three repeat libraries including the *de novo* library, Dfam 2.0 (Hubley et al. 2015), and the Repbase-derived library (20181026) (Bao et al. 2015). The abundance of transposable element (TE) categories was summarized using an in-house python script, and further visualized using the package ggplot2 in R.

Recognition the influence of sequencing strategies

In this study, the selected genomes mainly were generated from second- and third-generation sequencing technologies. Given their differences in sequencing read length, we had to consider the impact of sequencing technology on the genome, especially in the genome completeness and TE sizes. Therefore, we first excluded the only one genome generated from the first-generation sequencing technology (Sanger sequencing), and divided the other genomes into two groups according their sequencing strategies. If the genome was generated using only the second-generation sequencing technologies, or with Sanger sequencing for improvement, we marked the sequencing strategy of the genome as second-generation sequencing strategy. If the genome was generated using only the third-generation sequencing technologies (Single-molecule real-time sequencing or Nanopore sequencing), or with second-generation sequencing for improvement, we marked the sequencing strategy of the genome as third-generation sequencing strategy. Comparative analyses of the completeness, continuity and TE

sizes of genomes generated from both different sequencing strategies, were conducted to figure out whether sequencing strategies impact the number of genes and the abundance of TEs. We also took the taxonomic position of the compared groups into consideration, to decrease the influence of phylogeny to the comparative results.

Gene prediction and functional annotation

Transfer RNA (tRNA) genes in each soft-masked genome were annotated using tRNAscan-SE version 2.0.9 with default parameters (Chan et al. 2021). Models of protein-coding genes were predicted using the BRAKER2 pipeline (Brůna et al. 2021), which combines robust features of GeneMark-EP+ (Brůna et al. 2020) and AUGUSTUS (Stanke et al. 2008). To improve gene prediction accuracy, fungal proteins with annotation scores above 3 in UniProtKB (Consortium 2020) were downloaded and further reduced by removing redundant protein sequences using CD-HIT version 4.8.1 (Fu et al. 2012). Sequence identity and alignment coverage were set to 0.8 to retain the representative sequences. Finally, a total of 95,251 protein sequences were used as external evidence for gene structure prediction. Protein hints of homologous regions in each genome were produced using ProtHint version 2.6.0 (Brůna et al. 2020) and further used in the BRAKER2 pipeline. Functional annotation, orthology assignments and domain prediction of all predicted proteins were conducted using eggNOG-mapper version 2.1.3 (Cantalapiedra et al. 2021).

Identification of secreted proteins and effectors

Secreted proteins were identified using a widely used pipeline described previously (Pellegrin et al. 2015; Miyauchi et al. 2020; Mesny et al. 2021). In brief, proteins with signal peptides were identified as candidate-secreted proteins using SignalP version 4.1 with default parameters (Petersen et al. 2011). Then, membrane proteins were removed using TMHMM version 2.0 (Melén et al. 2003) by detecting the presence of the transmembrane helix. Glycosylphosphatidylinositol (GPI)-anchored proteins were removed using NetGPI version 1.1 (Gíslason et al. 2021) online by detecting GPI-anchoring signals, and proteins residing in the endoplasmic reticulum lumen were removed using PS-SCAN (Nielsen et al. 1997) by detecting KDEL motif (Lys-Asp-Glu-Leu) in the C-terminal region. Two subcellular localization prediction tools, WoLF PSORT (Horton et al. 2007) and TargetP version 2.0 (Emanuelsson et al. 2007), were used to confirm that only proteins assigned extracellular tags were identified as secreted proteins.

Secreted CAZymes including auxiliary redox (AA) enzyme families were identified using run_dbCAN version 3.0.7 (Zhang et al. 2018). Proteases and lipases were identified by querying the MEROPS database (Rawlings et al. 2017) and LED database release 3.0 (http://www.led.uni-stuttgart.de), respectively, using BLASTp with a cut-off e-value of 1e-5. Other secreted proteins shorter than 300 amino acids were identified as SSPs and the remaining secreted proteins were marked as OTHER. Secreted effectors were identified using EffectorP version 3.0 (Sperschneider and Dodds 2022) with the option of fungal mode. There was no intersection between each group. Furthermore, we followed the grouping criteria in the study of Mesny et al. (2021), and further classified secreted CAZymes into the plant cell wall-degrading enzymes (PCWDEs), fungal cell wall-degrading enzymes (FCWDEs), Cellulose, Hemicellulose, Lignin, Pectin, Peptidoglycan, Mannan, Glucan and Sucrose.

Analyses of numerical traits

To explore which of the basic components of the genomes and the functional proteins determine the lifestyle, we classified the numerical traits of genome assemblies into two categories and constructed two numerical matrices: basic genomic features and functional protein features. The former includes 25 numerical features: genome size with TEs, genome size without TEs, TE size, GC content of genomes, GC content of genome without TE, GC content of TE, the number of genes, the number of tRNAs, the number of exons, the number of introns, the average length of genes, the average length of tRNAs, the average length of exons, the average length of introns, the average length of intergenic regions, the minimum length of genes, the minimum length of tRNAs, the minimum length of exons, the minimum length of introns, the minimum length of intergenic regions, the maximum length of genes, the maximum length of tRNAs, the maximum length of exons, the maximum of introns and the maximum length of intergenic regions. The latter includes 24 numerical features: total secreted proteins, the effectors, proteases, lipases, SSPs, CAZymes, GHs, GTs, PLs, CEs, AAs, CBMs, PCWDEs, FCWDEs, cellulose-, hemicellulose-, lignin-, pectin-, peptidoglycan-, mannan-, glucan-, chitin-, sucrose-degrading enzymes and other functional proteins. The numbers of these features were summarized using in-house python scripts.

Correlations were calculated for the two main categories, and details were characterized in the captions of the corresponding figures. To make the comparative analysis more reliable, we excluded the groups with fewer than 10 genomes. Overall comparisons were conducted to detect changes in these numerical traits across taxonomic ranks and lifestyles. Post hoc pairwise multiple comparisons were performed to discover how many pairwise comparisons were significantly different based on different grouping criteria and to explore which features were useful in differentiating taxonomic groups and lifestyles.

Predicting lifestyles using machine learning algorithms

Six commonly used machine learning algorithms for multi-class classification implemented in the python library scikit-learn (https://scikit-learn.org): Random Forests (abbreviated as RF), Decision Tree (DT), Naive Bayes (Bayes), Support Vector Machine (SVM), Logistic Regression (LR) and K-Nearest Neighbors (KNN). These algorithms were used to predict fungal lifestyles, and the predictive accuracies of these algorithms were compared to determine the best classifier. Three matrices including the basic genomic features (25 numerical traits), functional protein groups (24 numerical traits), and combined dataset of them (49 numerical traits) were used during the training and prediction stages for selecting the most suitable dataset. The genomes with undetermined lifestyles were excluded from the datasets. First, we standardized the values of features using the function StandardScaler. Next, features with low variances were detected and removed using the function VarianceThreshold with default parameters. Then, the dataset was split into the train (70%) and test subsets (30%) using the function train_test_split, and parameters of the best suitable estimator were determined using the function GridSearchCV. The performance of the estimator was evaluated using the function cross_val_score with 5 duplicates based on the test subset. Finally, we used the best estimator to predict the lifestyles of unlabeled genomes.

Results

Genome information

A total of 638 representative genomes from 5 subclasses, 17 orders, 50 families, 147 genera and 614 species, were selected in this study. More detailed information is described in Supporting Information Table S1. The most

numerous subclass is Hypocreomycetidae, which occupies 73.20% (n=467) of the total genomes (Table S2: sheet subclass-count). The 10 most-numerous orders are Hypocreales, Glomerellales, Miccroascales, Ophiostomatales, Diaporthales, Xylariales, Sordariales, Amphisphaeriales, Magnaporthales, and Coniochaetales in descending order, the number of which range from 3 to 363 (Fig. 1, Table S2: sheet order-count). The other six orders contain only one genome except for three genomes that have not yet been classified in any of the established orders with certainty. Through a comprehensive survey of scientific literature and related databases, we indirectly obtained lifestyle descriptions of most strains (86.99%, n=555) and further classified these strains into eight groups by their host and tropic mode (Table S2: sheet lifestyle-count). We marked the strains isolated from diseased plant tissues as plant pathogens, from decaying woods as saprobes, from insects as entomopathogens, from fungi as mycoparasite, from plant tissues without disease symptoms as endophytes and from diseased human tissues as human pathogens. Moreover, four carnivorous fungi that feed on nematodes were marked as nematophagous fungi, and other genomes that lacked descriptive information regarding lifestyle were marked as "Undetermined". The most common lifestyle is plant pathogen, which occupies 58.31% (n=372) of the total genomes, followed by saprotrophs at 12.38% (n=79), entomopathogens at 6.27% (n=40), mycoparasites at 3.61% (n=23), endophytes of 3.29% (n=21), human pathogens of 2.51% (n=16) and nematophagous fungi of 0.63% (n=4). The remaining 83 genomes were temporarily marked as "Undetermined". We also traced the sequencing technologies of these genomes (Fig. 1, Table S2: sheet wgs-count), and summarized that 74.92% (n=478) of them were sequenced using second-generation sequencing technologies, 24.92% (n=159) were sequenced using third-generation sequencing technologies and only one genome was sequenced using Sanger sequencing technology.

Lifestyle occurrences in Sordariomycetes groups

Based on the genome data in this study, seven kinds of lifestyles, *viz.* plant pathogens, saprotrophs, entomopathogens, mycoparasites, endophytes, human pathogens and nematophagous fungi determined across 555 Sordariomycete genomes but with different occurrences at the subclass, order and family levels (Fig. 1, Table S2: sheet subclass-lifestyle). More diverse lifestyle modes were observed in the more fully sampled groups. For instance, the most-sampled subclasses Hypocreomycetidae and the subordinate order Hypocreales comprise all seven lifestyles, whereas the subclass Sordariomycetidae and Xylariomycetidae only comprise four and three kinds of lifestyles, respectively. At the order level (Table S2: sheet order-lifestyle), the order Ophiostomatales comprise five kinds of lifestyles only inferior to the Hypocreales that includes seven lifestyles. We further compare the occurrence of lifestyles in these two orders at the family level. The family Ophiostomataceae (Ophiostomatales) features with pant pathogens; the family Nectriaceae (Hypocreales) features with plant pathogens; the family Hypocreaceae features with saprotrophs; the family Ophiocordycipitaceae and the family Clavicipitaceae feature with entomopathogens. We compared the distribution of lifestyles at different taxonomic levels (Table S2: sheets lifestyle-subclass, lifestyle-order and lifestyle-family). Endophytes, saprotrophs and plant pathogens are present in four subclasses, followed by human pathogens, present in three subclasses, and entomopathogens and mycoparasites, present in two subclasses. The insufficient sampling lifestyle of nematophagous fungi is only present in the subclass Hypocreomycetidae. At the order and family level, plant pathogen is the most common lifestyle in 11 orders and 29 families, followed by saprotrophs in 9 orders and 19 families, endophytes and in 5 orders and 11 families, and human pathogens in 5 orders and 5 families.

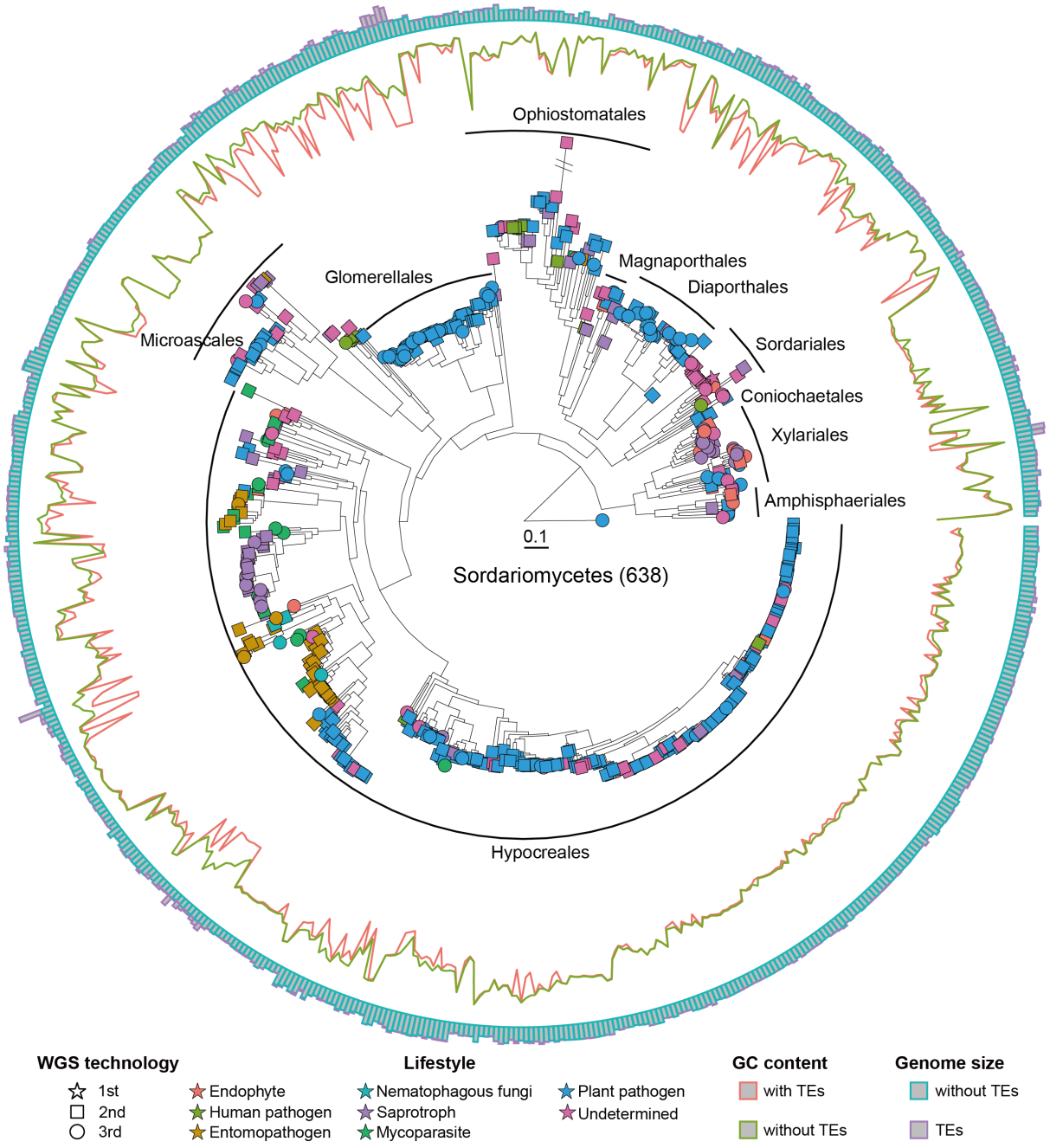

Fig. 1. Maximum likelihood (ML) phylogeny of 638 taxa in the class Sordariomycetes. The concatenation-based ML phylogeny (lnL = − 134,234,602.321) was reconstructed based on an amino acid dataset of 1,124 BUSCO genes (total of 884,972 sites) under the LG + G4 evolution model. The sequencing strategies are shown in different shapes (when multiple sequencing strategies were conducted for generating the genomes, we just marked the sequencing strategy by the most advanced technology). Lifestyles are indicated using different fill colors. Guanine-cytosine (GC) content of the genome and genome without transposable elements (TEs) are indicated by a line chart. Genome size and TE sizes are indicated using stacked bar charts. This figure was plotted using the packages ggtree version 3.4.4 (Yu et al. 2017) and ggtreeeExtra version 1.6.1 (Xu et al. 2021) in R (R Core Team 2022), with the dataset provided in Table S1.

Influence of sequencing technologies on TE size

The genomes were generated from first-generation, second-generation, and third-generation sequencing platforms, which account for 0.16% (n=1), 74.92% (n=478), and 24.92% (n=159) of the total number of genomes. To recognize the potential influences of sequencing technologies on subsequent numerical analysis, we compared the completeness, continuity, and TE sizes of genomes generated from second- and third-generation sequencing technologies (Table S3).There is no significant difference ($p = 0.08$) in BUSCO completeness (Fig. 2a), however, we observed significant differences in the number of contig/scaffold (Fig. 2b, $p < 2.2e\text{-}16$) and the N50 value (Fig. 2c, $p < 2.2e\text{-}16$), which suggests that the genomes generated from third-generation technologies are better in genomic continuity than that generated second-generation sequencing technologies. We also investigated whether the sequencing technologies influence the TE size, and found that the genomes generated from third-generation sequencing technologies have a larger size of TEs than second-generation sequencing technologies (Fig. 2d). We compared TE size between the two well-sampled families and the significant differences were also observed in Glomerellaceae genomes (Fig. 2e, $p = 0.0019$) and Nectriaceae genomes (Fig. 2f, $p = 6.1e\text{-}06$). Due to the non-negligible impact of sequencing technology on TE size, we did not explore further the relationships between lifestyles and the abundance of TEs. The abundance of TEs is provided in Table S1 and visualized in Fig. S1.

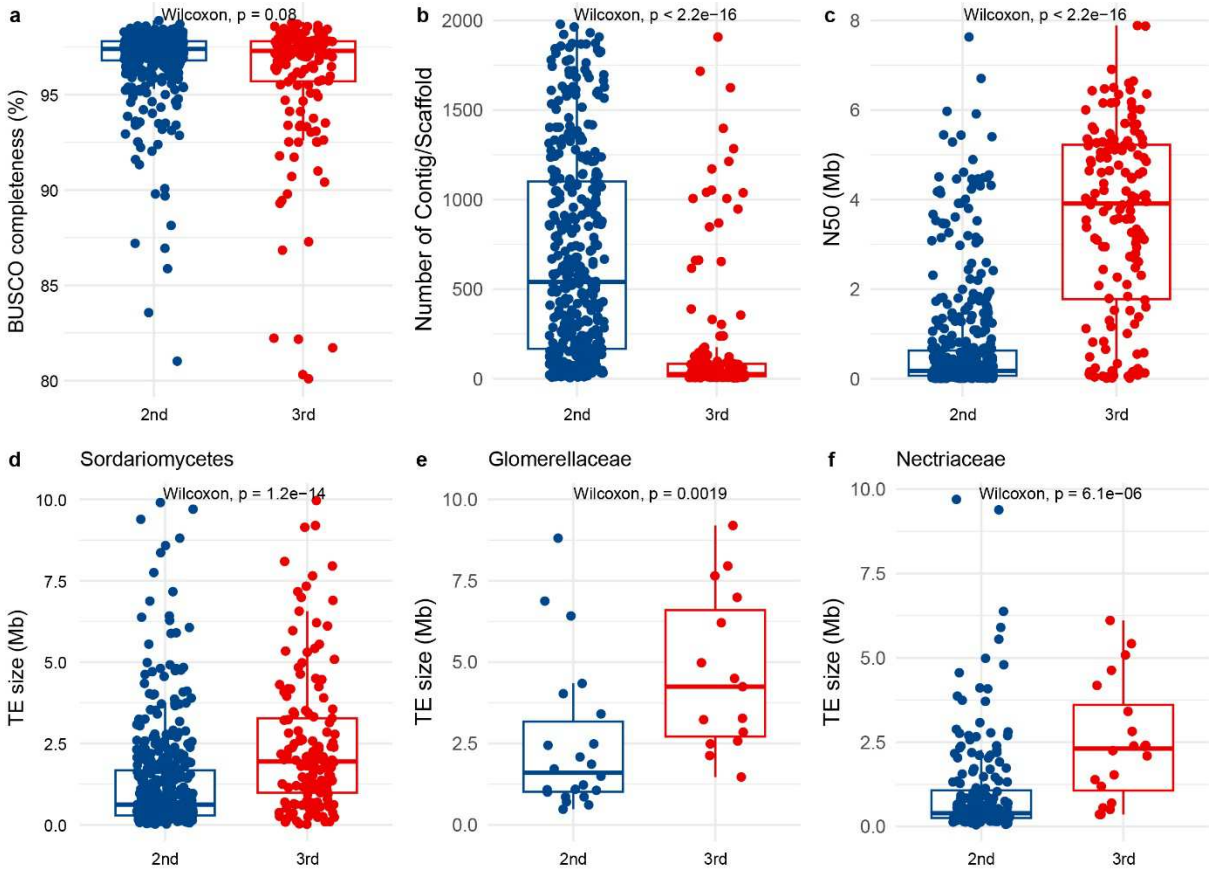

Fig. 2 Comparative analyses of genome completeness, continuities, and TE sizes of genomes generated by second- (2nd) and third-generation (3rd) sequencing strategies. a Bar plot of BUSCO completeness to represent the genome completeness; **b, c** Bar plots of the number of contigs/scaffolds and the value of N50 to represent the continuities. N50 is the shortest contig length that needs to be included for covering 50% of the genome, which is a measure to indicate the quality of assembled genomes that are fragmented in contigs of different lengths. The larger number of contigs/scaffolds means a more fragmented genome. The larger N50 value means a more contiguous genome. **d–f** Bar plots of TE size at the class and family levels to present the influence of sequencing technologies on TE size. Shapiro-Wilk test was conducted using the function shapiro.test (the

package stats) to check whether the compared datasets follow a normal distribution, and the results suggested that these datasets are not normally distributed. Thus, Wilcoxon Rank Sum and Signed Rank Tests were conducted using the function stat_compare_means (the package ggpubr) to test whether the compared datasets are significantly different ($p \leq 0.05$). All bar plots were plotted using the package ggpubr. For visualization, a small number of data points above 2,000 in subfigure b, data points above 8 Mb in subfigure c, and data points above 10 Mb in subfigure d and e, are not displayed. The input dataset is given in Table S1, and all resulting tables are given in Table S3. Statistical analyses and visualization were done in R (R Core Team 2022).

Variations of basic genomic features

We counted a total of 25 basic genomic features, which are summarized in Table S1. Results of correlation analyses among these features suggested that some features are highly correlated (Fig. 3, Table S4). Genome size is positively correlated with TE size with a Pearson's correlation coefficient of 0.63, which is smaller than its correlation coefficient with the genome size without TEs ($r = 0.86$), suggesting that the TEs can increase the genome size but not the dominant factor. GC content is positively correlated to the GC content without TEs ($r = 0.85$) but negatively related to the TE size ($r = -0.46$). In addition, GC content with TEs or without TEs is influenced by TE size, the larger TE size caused the larger difference between them, suggesting that TEs decrease the GC content of genomes. Genome size without TE is positively correlated to the number of genes ($r = 0.91$), the number of exons ($r = 0.90$), and the number of introns ($r = 0.88$). The latter two features, exons, and introns are important structural components of genes, the numbers of which reasonably displayed high correlations with the number of genes ($r = 0.97$; $r = 0.93$). The average length of genes is correlated to the average length of introns ($r = 0.78$) and the exons ($r = 0.48$), indicating that changes in intron length are the main cause of the variation of gene length compared to the exon. TE size is positively correlated to the average and maximum lengths of intergenic regions ($r = 0.60$; $r = 0.47$), but not displays significant correlations with gene structures including gene length, exon length, and intron length, suggesting that TEs are the main factor to change the distance between genes without significant influence on the gene structures. The minimum and maximum length of multiple features (genes, intergenic regions, introns, exons) exhibit relatively low correlation with other features, or correlations are not significant, except for the maximum length and the average length of intergenic regions ($r = 0.70$), the maximum length and the average length of introns ($r = 0.6$) and the minimum length of introns and genes ($r = 0.7$). Overall, most basic genomic features display a low correlation with each other, suggesting some of which are stable and independent in evolution.

We also compared the group means of these 25 genomic features over all groups of different taxonomic ranks and lifestyles (Table S5). We observed overall statistically significant differences in most genomic features (22/25) at the subclass level, excluding the minimum length of exons, TE sizes, and the minimum length of tRNAs (Table S5: sheet subclass). The minimum length of exons is the only feature that does not show a significant difference at the order level (Table S5: sheet order). And at the family, all features display significant differences (Table S5: sheet family). Considering the groups with different lifestyles, there are 6 genomic features without significant difference (Table S5: sheet lifestyle), which are the minimum length of exons, the average length of intergenic regions, the minimum length of intergenic regions, the size of TEs, the GC content of TEs and the maximum length of tRNAs. In paired comparison analysis (Fig. 4), we included 4 subclasses Diaporthomycetidae, Hypocreomycetidae, Sordariomycetidae and Xylariomycetidae, which formed 6 pairwise comparisons, 5 of which are significantly different in most of the features (Table S5: sheet pairwise-subclass). Specially, the number of genes and the number of exons display the most powerful resolution to differentiating the taxonomic groups at

the subclass level. At the order level (36 pairwise comparisons in total) and family level (91 pairwise comparisons in total), we observed a clear downward trend of significant differences, suggesting that all features lack resolutions at lower taxonomic levels (Table S5: sheets pairwise-order and pairwise-family). However, fairly low proportions of significantly different comparisons (15 pairwise comparisons in total) were observed across all features in the groups with different lifestyles (Table S5: sheet pairwise-lifestyle). Moreover, clustering analysis shows that several features (TE size, the minimum length of tRNAs, the minimum length of exon, and the minimum length of gene) display little usefulness in distinguishing different taxonomic groups, and most features are useless in differentiating different lifestyles.

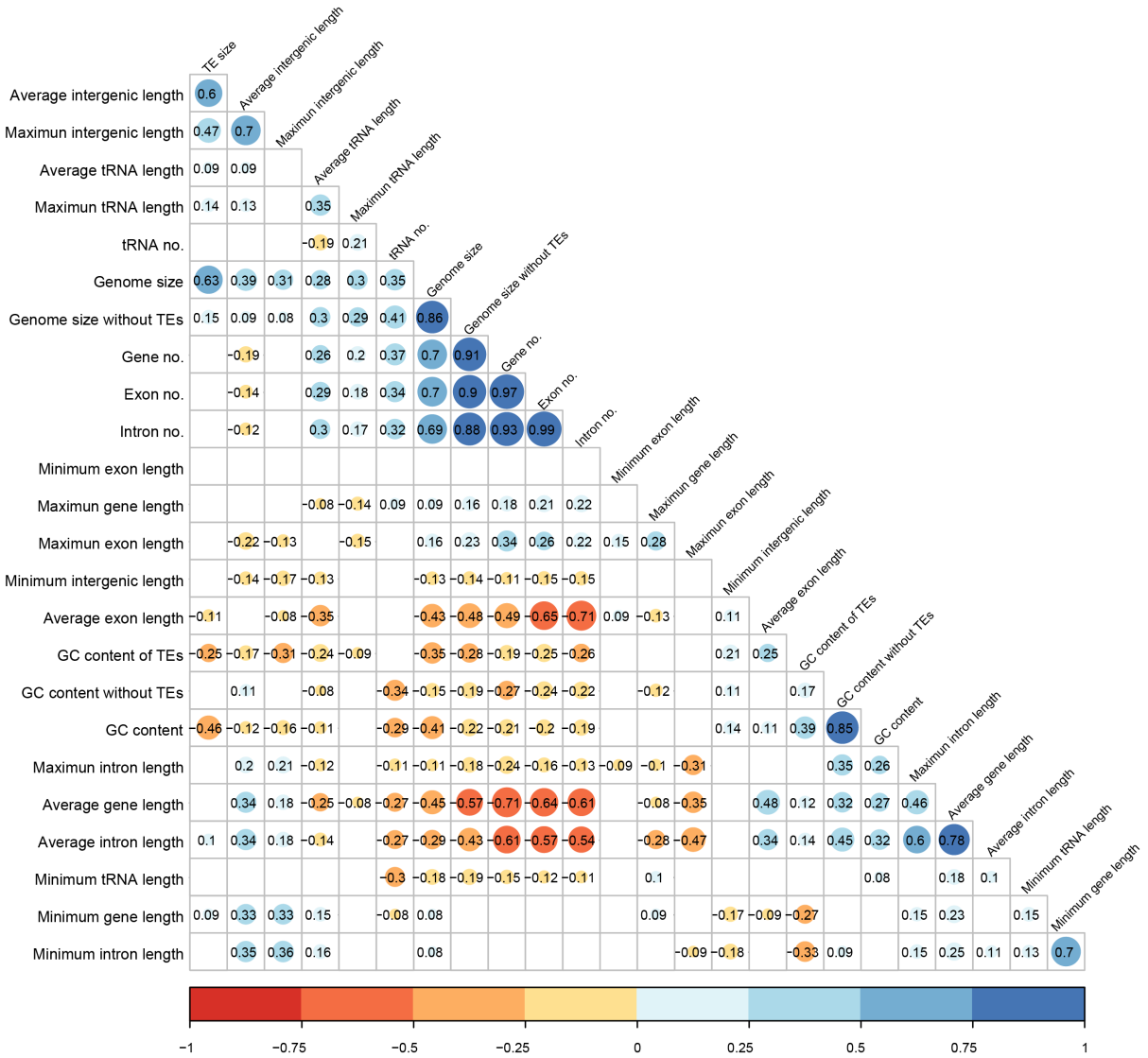

Fig. 3 Correlation analysis of 25 basic genomic features. Ladder heatmap of Pearson correlation coefficients of all pairwise genomic features. The colors and values in small squares indicate the degree of positive correlation (red) or negative correlation (blue). No significant correlated comparisons ($p > 0.05$) were displayed in white and blank squares. Pearson correlation coefficients were calculated using the function cor (the package stats), and the significance test was conducted using the function cor.mtest (the package corrplot). The figure was plotted using the package corrplot with the resulting datasets in Table S4. Values of these 25 features are provided in Table S1.

Although, not all features show strong discrimination in distinguishing one group from the other groups, a high proportion of significant differences of some genomic features was observed in specified comparisons. For instance, at the subclass level (Table S5: sheet class-class), there are 18, 17, 15, 15 and 15 significantly different features present in the pairwise comparisons of Hypocreomycetidae-Xylariomycetidae, Hypocreomycetidae-Sordariomycetidae, Diaporthomycetidae-Hypocreomycetidae, Diaporthomycetidae-Xylariomycetidae and Sordariomycetidae-Xylariomycetidae. Likewise, a high proportion of some features also were observed at the Order and Family levels (Table S5: sheets order-order and family-family). These results suggest that some features are useful in differentiating specified taxonomic groups, especially in phylogenetic distant comparisons. As for lifestyles, the largest difference in genomic features was only observed in the comparisons of entomopathogens-plant pathogens (14/25), followed by entomopathogens-endophytes (10/25), and the rest of comparisons display no difference or very small differences, especially in the comparisons of endophytes-saprotrophs (0/25, entomopathogens-mycoparasites (0/25), mycoparasites-saprotrophs (0/25), endophytes-mycoparasites (1/25), human pathogens-mycoparasites (1/25), endophytes-saprotrophs, human pathogens-plant pathogens (1/25), and human pathogens-saprotrophs (1/25) (Table S4: sheet lifestyle-lifestyle). It suggests that based on these basic genomic features it is difficult to differentiate compared lifestyles. In another word, we could not correctly assign a lifestyle label to a new taxon with very similar genomic features, to endophytes, saprotrophs, mycoparasites and entomopathogens.

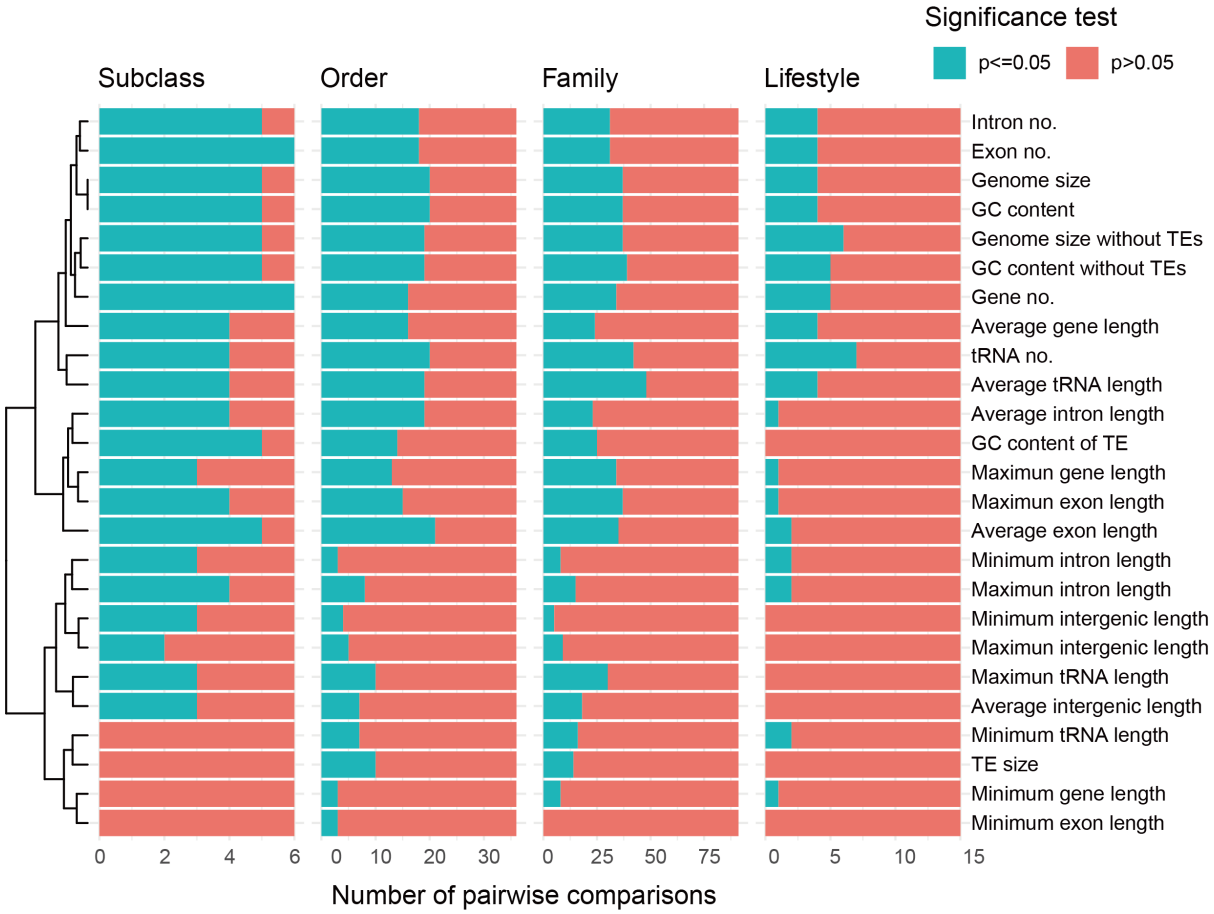

Fig. 4 Resolution powers of 25 basic genomic features in differentiating different taxonomic groups and lifestyles. Stacked bar plots of the number of significantly (orange; p <= 0.05) and non-significantly (green; p > 0.05) different comparisons across all features based on their taxonomic ranks and lifestyles. The cluster analysis was performed using the function dist (the package stats) with the dataset in Table S4 sheet: clustering-matrix to

obtain a Euclidean distance matrix, then using the function hclust (the package stats) to cluster these features with the "complete" agglomeration method. All datasets are given in corresponding sheets in Table S5.

Overview of functional protein groups

A total of 24 functional protein groups were summarized in Table S1 and visualized in Fig. S2. To explore the correlation between the number of the proteome and the number of each functional protein group we include the feature of proteomes, which is equivalent to the number of protein-coding genes in the last part during correlation analysis (Fig. 5; Table S6). The result shows that 66.67% (16/24) of protein groups are highly positively correlated ($r > 0.6$) with the total number of the proteome. The main subgroups of the secretome, the number of CAZymes, protease, lipase, SSPs, secreted effectors and other functional proteins are highly positively correlated with the total number of secretomes with the Pearson correlation coefficient of 0.95, 0.93, 0.86, 0.87, 0.96 and 0.97, respectively. The six subgroups of CAZymes display varying degrees of correlation with the total number of CAZymes. The numbers of AAs, GHs, CEs and PLs display high correlation with the Pearson correlation coefficient of 0.97, 0.97, 0.88 and 0.88, respectively. The number of CBMs displays a relatively high correlation ($r = 0.57$) with CAZymes, whereas the GTs display a low correlation ($r = 0.29$) with CAZymes. As for the more specified functional subgroups of CAZymes, the numbers of PCWDEs, pectin-degrading enzymes, hemicellulose-degrading enzymes, and cellulose-degrading enzymes, are highly correlated with the total number of CAZymes with the Pearson correlation coefficients of 0.97, 0.90, 0.89 and 0.87, respectively, followed by lignin-degrading enzymes and glucan-degrading enzymes with relatively high correlation coefficients of 0.54 and 0.51. The numbers of FCWDEs, chitin-degrading enzymes and mannan-degrading enzymes display relatively low correlation with CAZymes, the correlation coefficients of which are 0.41, 0.31 and 0.22 respectively, and no significant correlation was observed between peptidoglycan-degrading enzymes and CAZymes. We also noticed the high correlations between several specified functional subgroups of CAZymes, such as FCWDEs and chitin-degrading enzymes with correlation coefficients of 0.9, FCWDEs and glucan-degrading enzymes with correlation coefficients of 0.82, which are mainly due to the overlapping functional proteins (Table S6). Compared with the correlation matrix of genomic features (Fig. 3), most functional proteins are more stable in number, showing a trend of co-evolution except for mannan-degrading enzymes, GTs, and peptidoglycan-degrading enzymes.

The discrimination of these 24 functional protein groups was visualized by comparing the numbers of significantly different pairwise comparisons and not significantly different pairwise comparisons (Fig. 6, Table S7). Compared with the discrimination of 25 basic genomic features, clear increases in functional protein groups are observed at the taxonomic levels and lifestyles. At the subclass level, more than half (15/24) of these protein groups are powerful in differentiating subclasses ($n > 3$, TableS7: sheet cluster-matrix), especially the number of CBMs and mannan-degrading enzymes with 100% resolution (Table S7: sheet pairwise-subclass). However, CEs, hemicellulose-degrading enzymes and PCWDEs display very poor resolution, especially the latter two. At the order and family levels (Table S7: sheets pairwise-order and pairwise-family), the numbers of significantly different pairwise comparisons increase with the total number of pairwise comparisons, but the proportion of significantly different pairwise comparisons for each protein group decreases, most notably in CBMs and mannan-degrading enzymes. Although the numbers of PCWDEs and hemicellulose-degrading enzymes are useless in differentiating subclasses, we notice that PCWDEs can distinguish more than half of the pairwise comparisons at the order level (23/36) and the family level (48/91), and hemicellulose-degrading enzymes can distinguish more than half of the pairwise comparisons at the order level (19/36) and nearly half at the family level (39/91). On the subject of lifestyles (Table S7: sheet pairwise-lifestyle), we observed clear drops in the proportion of significantly

different pairwise comparisons for some protein groups, and also noticed some increased proportions, such as the glucan-, cellulose- and hemicellulose-degrading enzymes.

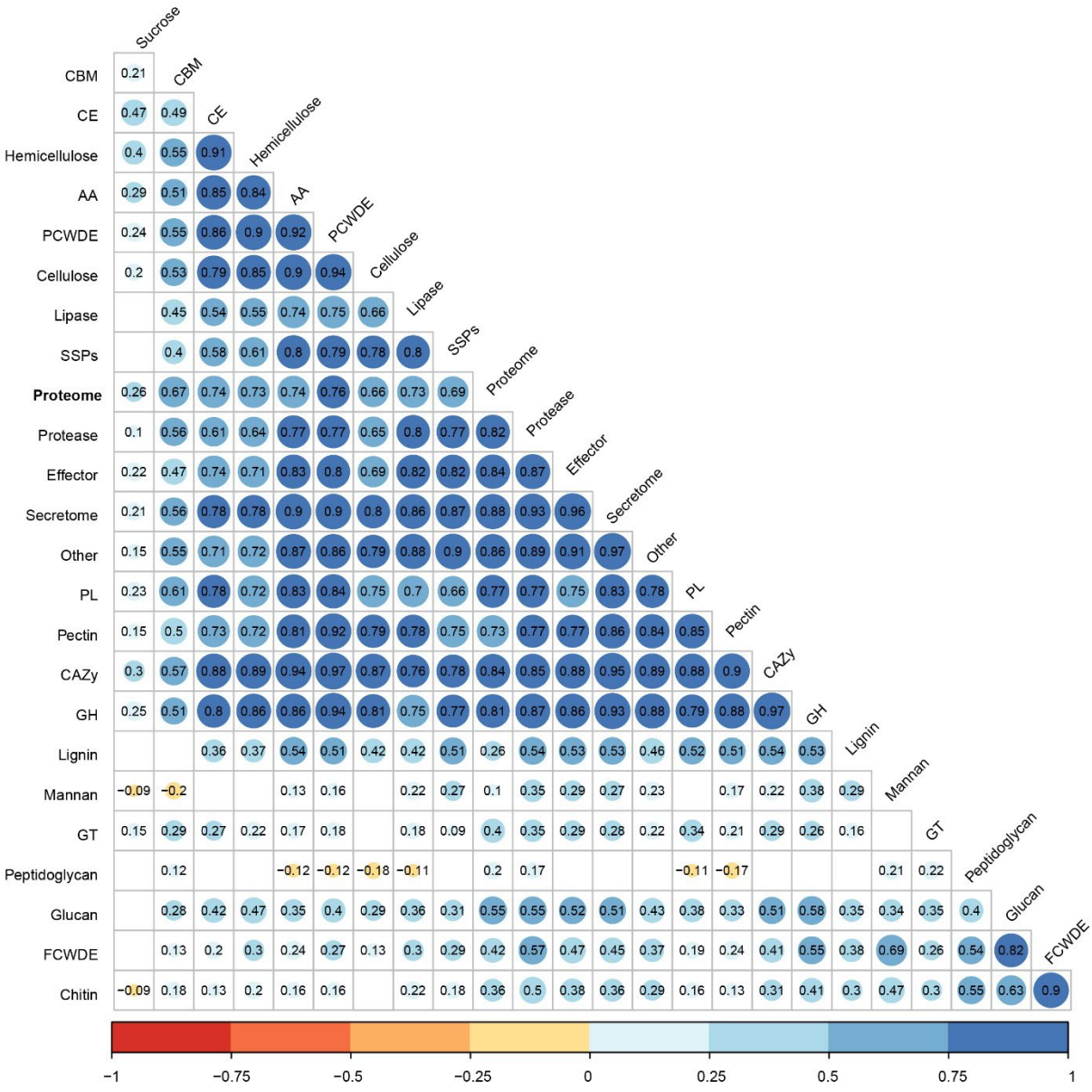

Fig. 5 Correlation analysis of 24 functional protein groups and proteomes. Ladder heatmap of Pearson correlation coefficients of all pairwise genomic features. The colors and values in small squares indicate the degree of positive correlation (red) or negative correlation (blue). No significant correlated comparisons ($p > 0.05$) were displayed in white and blank squares. Pearson correlation coefficients were calculated using the function cor (the package stats), and the significance test was conducted using the function cor.mtest (the package corrplot). The figure was plotted using the package corrplot with the resulting datasets in Table S6. Values of these 24 functional protein groups and the total number of proteomes are provided in Table S1.

We also counted the number of significantly different protein groups in each pairwise comparison. At the class level (Table S7: sheets subclass-subclass), the most notable subclass is Xylariomycetidae, which has 17 significantly different protein groups with Diaporthomycetidae, 16 with Hypocreomycetidae and Sordariomycetidae. The smallest difference is observed in the pairwise comparison of Diaporthomycetidae and

Sordariomycetidae with 12 significantly different protein groups. In other words, Xylariomycetidae is the easest to distinguish from other subclasses. At the order level (Table S7: sheet order-order), the most notable order is Ophiostomatales, which has 22 significantly different protein groups with Glomerellales and Hypocreales, 21 with Amphisphaeriales, 20 with Diaporthales, 19 with Magnaporthales. The smallest differences are observed in the pairwise comparisons of Magnaporthales-Amphisphaeriales, and Magnaporthales-Diaporthales. Moreover, Magnaporthales has only 2 significantly different protein groups with Glomerellales, 4 with Hypocreales and Xylariales, indicating that it is not easy to distinguish Magnaporthales from the compared orders based on most functional protein groups. At the family level (Table S7: sheet family-family), the largest number of significantly different protein groups is 23, which is observed in three pairwise comparisons of Ceratocystidaceae-Nectriaceae, Glomerellaceae-Ophiostomataceae and Nectriaceae-Ophiostomataceae. Inversely, the smallest number is 1, which is observed in two pairwise comparisons of Bionectriaceae-Nectriaceae and Clavicipitaceae-Ophiocordycipitaceae. For lifestyles (Table S7: sheet lifestyle-lifestyle), plant pathogens are the easiest lifestyle to distinguish from saprotrophs, entomopathogens and mycoparasites, and they have 21, 20, and 17 significantly different protein groups respectively. At the same time, it is also the most difficult to distinguish from endophytes because that they only significant difference in the abundance of PCWDEs and peptidoglycan-degrading enzymes. No significantly different protein group is present in the comparison of endophytes-saprotrophs, indicating that we cannot differentiate them based on the number of functional protein groups.

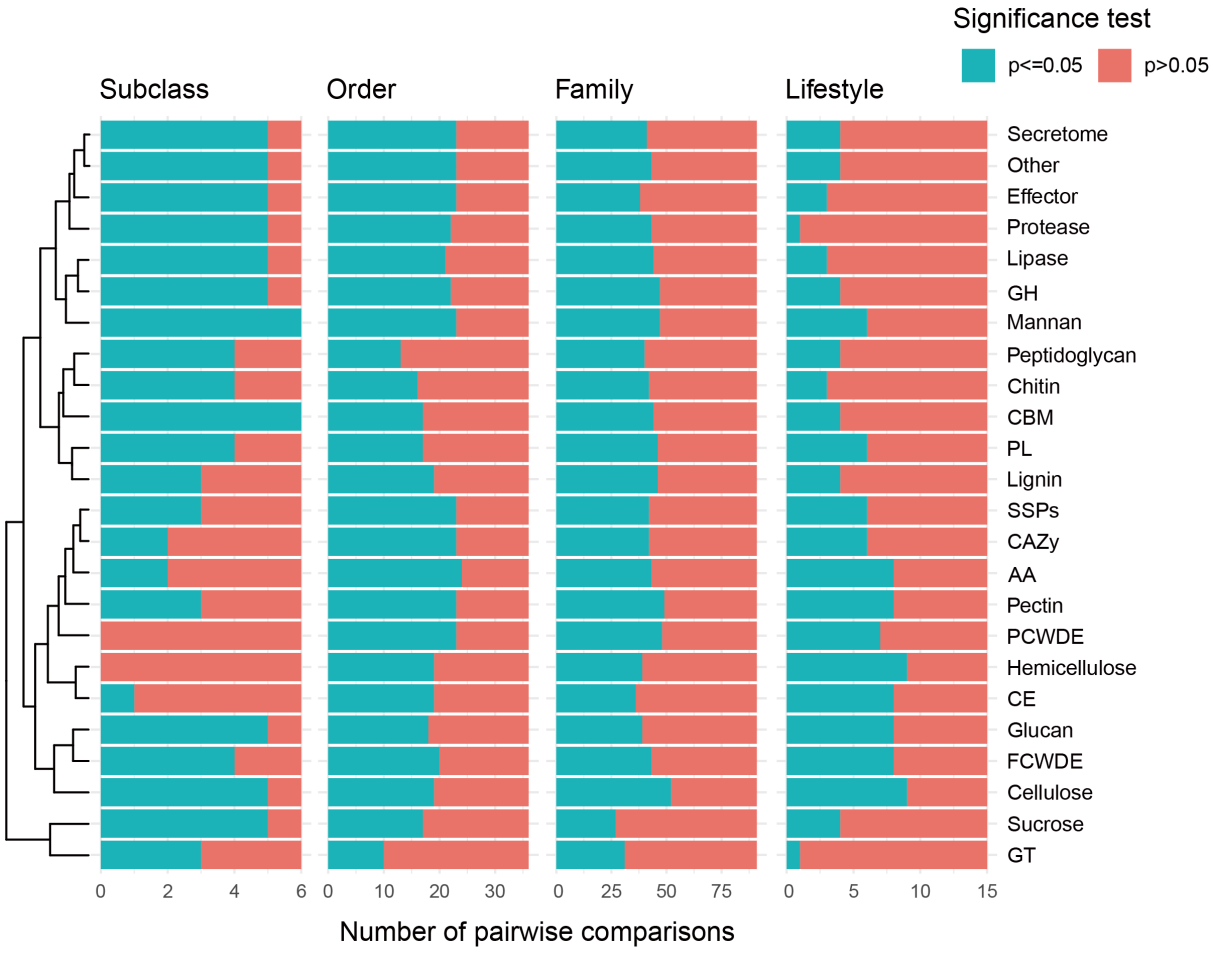

Fig. 6 Contributions of 24 functional protein groups in differentiating different taxonomic groups and lifestyles. Stacked bar plots of the number of significantly (orange; p <= 0.05) and non-significantly (green; p > 0.05) different comparisons across all features based on their taxonomic ranks and lifestyles. The cluster analysis

was performed using the function dist (the package stats) with the dataset in Table S7 sheet: cluster-matrix, to obtain a Euclidean distance matrix, then using the function hclust (the package stats) to cluster these features with the "complete" agglomeration method. All datasets are given in corresponding sheets in Table S7.

Predicting lifestyles using machine learning approaches

Predictive models of six commonly used machine learning algorithms were trained and optimized based on the training subsets of three different datasets, and accuracies in predicting fungal lifestyles were compared and visualized in Fig. 7 (Tables S8). For the dataset of basic genomic features, RF is the best classifier with an average accuracy of 0.7844, followed by KNN (0.7766), SVM (0.7272), DT (0.6675) and Bayes (0.5688); LR is the worst with an average accuracy of 0.5221. For the dataset of the functional protein groups, SVM is the best classifier with an average accuracy of 0.8286, followed by KNN (0.8208), RF (0.8156), DT (0.7065) and Bayes (0.6156); LR is the worst with an average accuracy of 0.5662. For the combined dataset including a total of 49 numerical features, KNN is the best classifier with an average accuracy of 0.8260, followed by RF (0.8234), SVM (0.8026), DT (0.6909) and LR (0.6753); Bayes is the worst with an average accuracy of 0.5766. In terms of machine learning algorithms, KNN, SVM and RF perform better than LR, Bayes and DT in predictive accuracies across the three datasets (Fig. 7 a, c, e). Bayes, DT, RF and SVM obtained the highest-average accuracies based on the functional protein groups, and the other two methods, KNN and LR, were based on the combined datasets. We noticed that all classifiers obtained the worst-average accuracies based on the basic genomic feature alone, and increased accuracies were observed based solely on a functional protein dataset or combined dataset (Fig. S3), indicating that numerical traits of functional protein groups are more useful than basic genomic features for predicting fungal lifestyles.

Based on the test subsets, we tested the performance of the three best classifiers, RF for the dataset of basic genomic features and combined dataset and SVM for the functional protein groups. For the dataset of basic genomic features (Fig. 7b), we noticed that 99.13% of plant pathogens were assigned the correct lifestyles, suggesting that RF is reliable for distinguishing plant pathogens from other lifestyles. However, it performed worse in differentiating endophytes, human pathogens and mycoparasites from other lifestyles. Predictive results of all endophytes, human pathogens and mycoparasites did not match the assigned lifestyles that we determined by a literature survey or the genomic descriptions. About half of endophytes (44.44%) were incorrectly predicted as plant pathogens, and some other genomes were incorrectly recognized as saprotrophs and mycoparasites. Of mycoparasites, 75% were incorrectly predicted as human pathogens and 25% as saprotrophs. Of human pathogens, all of them were incorrectly predicted as plant pathogens. As for the other three lifestyles, RF obtained an increased accuracy. Of entomopathogens, 37.5% were correctly classified, and 50% were incorrectly predicted as plant pathogens and 12.5% as saprotrophs. Of plant pathogens, 99.13% were correctly classified, and the rest were incorrectly predicted as saprotrophs (0.87%). Of saprotrophs, 46.15% were correctly predicted as saprotrophs, 42.13 % were incorrectly predicted as plant pathogens, 7.69% as entomopathogens, 3.85% asmycoparasites. Concerning the dataset of function protein groups (Fig. 7d), we used the SVM algorithm and observed a clear improvement in differentiating entomopathogens (37.50% to 62.50%), human pathogens (0 to 25%), mycoparasites (0 to 50%) and saprotrophs (46.15% to 57.69%) from other lifestyles. Compared with RF predication based on the dataset of genomic features, SVM resulted in the same incorrect results in differentiating human pathogens, with a similar result for endophytes and slightly decreased accuracy in predicting saprotrophs. As for the combined dataset (Fig. 7f), the KNN algorithm was used to predict lifestyles, and we observed a clear improvement in predictive accuracies for endophytes, entomopathogens and human pathogens.

Based on the combined dataset, we obtained the highest average accuracy of 0.7325 for the six algorithms. It was seen that KNN was the best classifier with an average accuracy of 0.8260. Therefore, we used KNN to conduct the prediction of 83 Sordariomycetes genomes with undetermined lifestyles, and the predicted lifestyles with probabilities were listed in Table S9. KNN classified these 83 genomes into 5 lifestyles, 1 endophyte, 4 human pathogens, 4 entomopathogens, 6 saprotrophs, and 68 plant pathogens. We further checked the taxonomic positions of strains, and only one endophyte is distributed in the family Bionectriaceae; 4 human pathogens in Sordariaceae;4 entomopathogens in Ophiocordycipitaceae, Ophiostomataceae, Clavicipitaceae and Cordycipitaceae; 6 saprotrophs in Bionectriaceae, Diatrypaceae, Sordariaceae and Hypoxylaceae; and 68 plant pathogens in 20 families. We traced the lifestyles of phylogenetically closed groups with predicted genomes, and most of the observed lifestyles were consistent with our predictions.

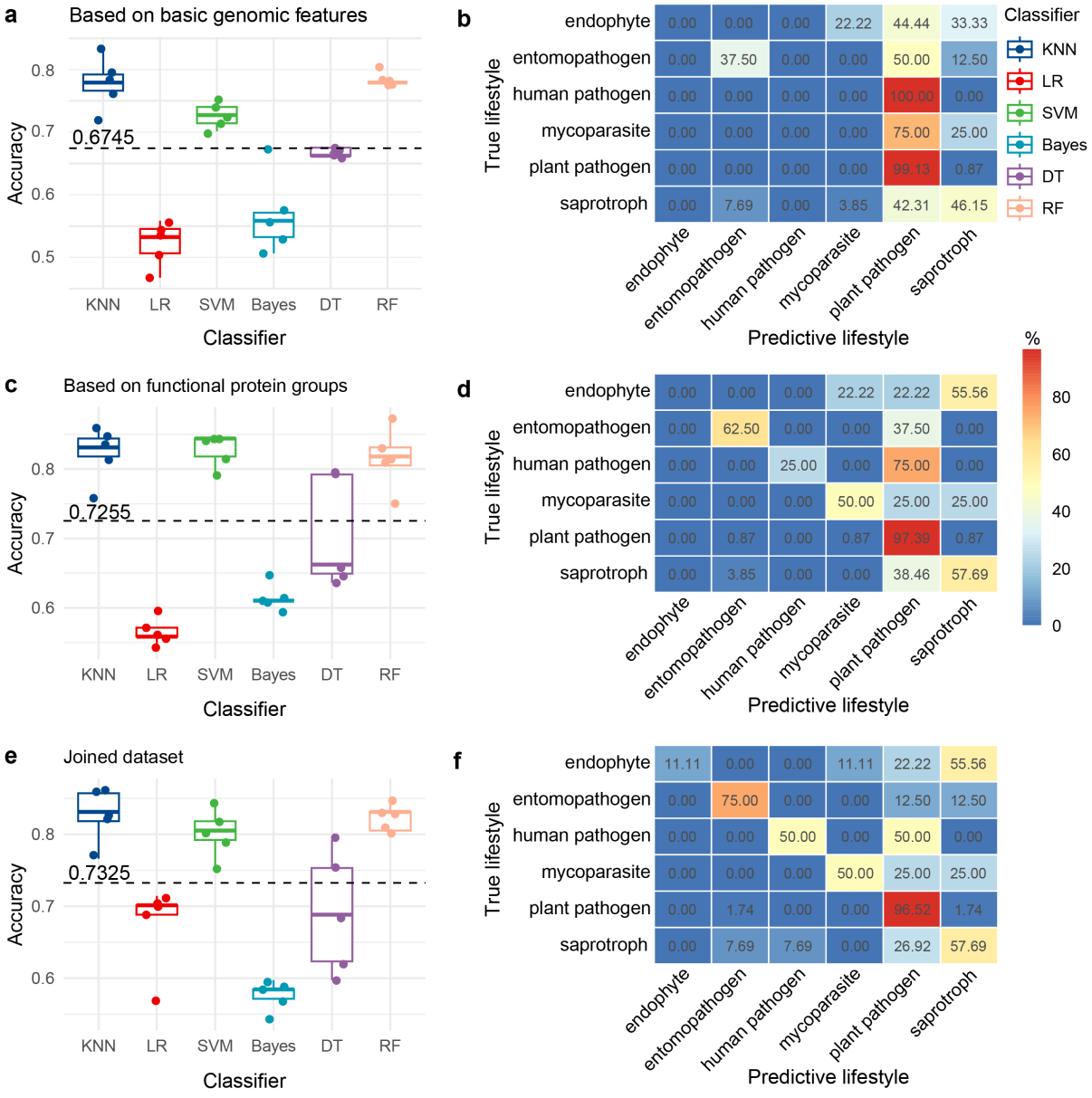

Fig. 7 Lifestyle predictions using machine learning methods. a Boxplots of predictive accuracies usi ng six machine learning algorithms for predicting fungal lifestyles based on the train subset of the basi

c genomic features. **b** Confusion matrix, a performance matrix, to evaluate the performance of the best classifier (RF, average accuracy = 0.7844) in predicting fungal lifestyles based on the test subset of th e basic genomic features. **c** Predictive accuracies of the six commonly used machine learning algorithm s based on the train subset of the functional protein features. **d** Confusion matrix of the best classifier (SVM, average accuracy = 0.8286) in predicting fungal lifestyles based on the test subset of the functi onal protein groups. **e** Predictive accuracies of the six commonly used machine learning algorithms bas ed on the combined datasets. **f** Confusion matrix of the best classifier (KNN, average accuracy = 0.82 60) in predicting fungal lifestyles based on the test subset of the functional protein groups. For the co nfusion matrix, the diagonal elements show the proportion of correctly classified genomes, while the of f-diagonal elements show the number of misclassified genomes.

Discussion

Diverse lifestyles but unbalanced whole genome sequencing

Sordariomycetes has a large number of available genome sequences for an ascomycetes class in public databases; however, many of these genomes are restricted to economically important groups such as plant pathogens (*Fusarium, Diaporthe*, *Calonectria*, *Claviceps, Collectotrichum*), entomopathogens (*Cordyceps, Metarhizium, Ophiocordyceps, Tolypocladium),* mycoparasites (*Clonostachys*), human pathogens (*Sporothrix*, *Sarocladium, Scedosporium)* model organism (*Neurospora*), and biocontrol and secondary metabolites producers (*Trichoderma, Daldinia, Xylaria*). For instance, Hypocreomycetidae includes plant pathogens, entomopathogens, mycoparasites, human pathogens and biocontrol agents and is responsible for 73.20% of the total Sordariomycete genome used in this study. However, the Sordariomycetes include other ecologically important saprotrophs, epiphyllous, hypophyllous, facultative lichenised, fungicolous and extreme inhibiting groups primarily overlooked due to their economically insignificance. Therefore, the current genomic data are largely incomplete and cannot be used to make reliable conclusions about the overall lifestyle of Sordariomycetes fungi. Saprobes are the most common type of fungi, and Sordariomycetes now comprises 195 families, and 171 have a saprobic lifestyle. This is true as many of these fungi can degrade polymers of varying complexity by releasing extracellular enzymes that break down plant and animal debris. We suspect that saprobic Sordariomycete families will likely be more than this as the remaining families are poorly sampled or monotypic. Plant pathogens are the second most abundant lifestyle in Sordariomycetes, distributed over 93 families. The five largest Sordariomycetes orders, Diaporthales, Glomerellales, Hypocreales, Microascales and Ophiostomatales, each contain a large number of highly destructive plant pathogens. These include some of the most important diseases of the cereal (rice, wheat, barley and maize) ornamental, fruit, vegetable and wild crops (Chang et al. 2018; Talhinhas and Baroncelli 2021; Liu et al. 2022; Han et al. 2023). Endophytes are distributed over 40 families of Sordariomycetes. There is publishable evidence that fungal endophytes can switch lifestyles to saprotrophs and pathogens and vice versa (Promputtha et al. 2007; Promputtha et al. 2010). Human pathogens, entomopathogens, mycoparasites and nematophagous fungi are distributed over 17, 11, 5 and 2 families of Sordariomycetes, respectively. The least distributed nematophagous fungi are only present in Hypocreales families Clavicipitaceae and Ophiocordycipitaceae. Their diverse lifestyles and ability to switch to other life modes and inhibit diverse ecological niches that include extreme environmental constraints allow Sordariomycetes to adapt and distributed over all ecosystems on earth and to be the second largest ascomycetes class.

Influence of sequencing technologies on genome assemblies

High-quality genome assemblies are fundamental for genomic studies. Therefore, when we used genomes from public databases, we were meticulous in checking their quality that was inevitably affected by the methods of DNA extraction (Nouws et al. 2020), sequencing technologies (Lang et al. 2020; Murigneux et al. 2020) and assembly algorithms (Miller et al. 2010; Meng et al. 2022). As a user of public genomes, although we cannot improve genome assemblies by optimizing these steps, recognizing the inaccuracies of genome assemblies reduces the possibility of drawing incorrect conclusions. Repetitive DNA sequences present technical challenges for assembly algorithms by bringing in ambiguous alignment during genome assemblies, leading to biases and errors in final assembly results (Treangen and Salzberg 2012; Tørresen et al. 2019). For instance, fungal ribosomal RNA genes (rDNA) as multiple-copy segments organized in tandem arrays exist in genomes (Cooper 2000). Each repeat unit (18S rRNA-internal transcribed spacer 1-5.8S rRNA-internal transcribed spacer 2-28S rRNA-intergenic spacer) is approximately 9kb in length (SONE et al. 2000; Salim et al. 2017), which far exceeds the read length limit of second-generation sequencing, and the reads generated from second-generation sequencers cannot span this kind of long repetitive sequence (Treangen and Salzberg 2012). Assembly algorithms, such as the Greedy strategy, Overlap-Layout-Consensus strategy, and de Bruijn graph strategy, tend to assemble these highly similar or identical sequences into single, collapsed contig (Treangen and Salzberg 2012). Although third-generation sequencing technologies, also called long-read sequencing technologies, can overcome the read length limit by producing 20–200 kb reads (Goodwin et al. 2016), the high cost per genome hinders its widespread application, especially in some fungal species that lack of direct economic interest. Furthermore, our previous study (Chen et al. 2022) found that second-generation sequencing technologies can provide reliable genome assemblies for phylogenomic analyses, which focus on protein-coding genes rather than repetitive sequences. In this study, we included 638 genomes, most of which were generated using second-generation sequencing technologies (n=478, 74.92%). We set the completeness threshold at 80% to remove the unreliable genomes, and confirmed that each group included at least 10 genomes during statistical analyses. Hence, we believe that sequencing strategies did not influence the numerical traits meaningfully.

TEs are mobile genetic elements that are composed of diverse members, including short interspersed nuclear elements (SINEs), Helitrons, Alus, endogenous retroviruses (ERVs), DNA transposons and retrotransposons (Wicker et al. 2007). The ability to move and their repetitive nature make TEs key drivers of genome evolution (Dhillon et al. 2019; Senft and Macfarlan 2021). Many studies have shown that the expansion of TEs resulted in a significantly expanded genome in fungal species, such as in *Cenococcum geophilum* (Peter et al. 2016), *Zymoseptoria tritici* (Oggenfuss et al. 2021) and *Lactarius* species (Lebreton et al. 2022). Large-scale genomic location analysis of TEs has indicated that most TEs are evolutionarily neutral, but animal-related and pathogenic fungi include more TEs inserted in genes compared to fungi with other lifestyles (Muszewska et al. 2019). Kirkland et al. (2018) reported that *hAT* or *Gypsy* TEs located within 1kb of protein-coding genes can decrease the expression of related genes. LTR retrotransposons, a class I transposable element, inserted in the *MFS1* promoter region, resulted in MFS1 overexpression and the presence of multidrug resistance phenotype in the wheat pathogen *Zymoseptoria tritici* (Omrane et al. 2017). TEs are important and biologically functional repetitive sequences, the abundance of which in genomes is inevitably affected by sequencing technologies, especially by second-generation sequencing technologies. In this study, we recognized that TE sizes in the genomes generated from second-generation sequencing technologies are significantly smaller than those from third-generation sequencing technologies. We also discovered the GC content of TEs is significantly lower than other regions in the genomes, and that TE sizes are negatively correlated with the overall GC content of fungal genomes. Hu et al. (2022) showed that GC content is positively correlated with growth temperature in prokaryotes, and Šmarda et al.

(2014) reported that increased GC content helps plants adapt to seasonally cold and/or dry climates. Considering the clear influence of sequencing technologies, the true abundance of TEs in most genomes has been underestimated in previous studies and in this study. Therefore, instead of providing a more in-depth analysis, we only compared the abundance of TEs in multiple groups and displayed their diversity in Table S1 and Fig. S2. We did not observe a significant difference in TE sizes between lifestyles; thus, the underestimated abundance in this study did not affect our statistical and predicted results. However, future studies related to TEs should take into account the influence of sequencing technologies.

Effectors are not a reliable indicator for disease-related fungi but are useful for differentiating specific lifestyles

Effectors, important virulence factors secreted by bacteria (Yu et al. 2020), fungi (Stergiopoulos and Wit 2009), and Oomycetes (Birch et al. 2006), either function in the interaction space between hyphae and host cells or are transferred into host cells to subvert host immunity. A successful fungal infection with significant disease symptoms is a complicated process that depends on the result of the battle between the pathogen and its host (GS 1996). When pathogens start to invade a host, the innate immune system is activated by recognizing microbial invariant molecular patterns (also known as pathogen-associated molecular patterns, PAMPs) (Akira et al. 2006). In fungi, chitin, the important cell wall component, is one of the main PAMPs, which is recognized by pattern-recognition receptors (PRRs) located in the host membrane (Boller 1995), and further activates important chemical pathways and specific gene expressions to eliminate pathogens (Macho and Zipfel 2014). The PAMP-triggered immunity (PTI) is the frontline of the plant host's immune system; if fungi seek to successfully colonize the host, they must avoid inducing PTI or suppress it. Effectors can suppress PTI, but they also can be captured by effector-triggered immunity (ETI). Therefore, linking the disease symptoms and effectors or elucidating their relationships remains a significantly challenging task. We hypothesize that this is why we did not observe a significantly different abundance in the average number of effectors between plant pathogens (the average number = 216) and endophytes (the average number = 207) in our analysis. There is limited capacity to validate the function of effectors in pathogen-host interactions experimentally; accordingly, only a small part of effectors are well studied in model fungi and economically important fungi (Stergiopoulos and Wit 2009), and many effectors have been identified in newly sequenced non-model fungal genomes or not economically important genomes using bioinformatic approaches (Jones et al. 2018). *PgtSR1*, a novel fungal effector identified by Yin et al. (2019) from the wheat rust pathogen *Puccinia graminis*, decreases the abundance of small RNAs by suppressing RNA silencing in plant cells, and further obstructs small RNA-regulated host immune reactions. Czislowski et al. (2021) showed that endophytic *Fusarium oxysporum* strains display different *SIX* gene profiles (a family of effector genes secreted in xylem) with pathogenic strains. However, a larger-scale study of fusarioid fungi did not find a significant difference in the number of effectors (Hill et al. 2022). In this study, we observed that plant pathogens (the average number = 216) include more effectors ($p < 0.05$) than saprotrophs (the average number = 162) and entomopathogens (the average number = 142). The abundance of effectors in endophytes (the average number = 207) is significantly higher ($p < 0.05$) than that found in entomopathogens. To explain these differences, we speculate that the pathogenic *F. oxysporum* isolates and non-pathogenic isolates might have similar numbers of effectors that differ in composition. Compared to results from Hill et al. (2022), we include more genomes with lifestyle information (n = 555 VS n = 61), which provides more numerical information for conducting statistical analysis. Moreover, we believe that our dataset includes more reliable lifestyle information. For instance, entomopathogens are mainly from the families Cordycipitaceae, Clavicipitaceae and Ophiocordycipitaceae; species in these families are more specifically parasitic on insects than species from the family Nectriaceae

(fusarioid fungi) (Simmons et al. 2015; Luangsa-ard et al. 2017; Araújo et al. 2018). We confirmed that the abundance of effectors is significantly different between several compared lifestyles, and future extended studies should focus on the composition to verify whether it is a possible indicator for differentiating lifestyles.

Basic genomic features are generally consistent with higher taxonomic ranks rather than lifestyles

In the genomic era, the rapid development of sequencing technologies and affordable cost of WGS have brought new insights to taxonomy. Genome Taxonomy Database (GTDB) exemplifies the important contribution of genomes in bacterial and archaeal taxonomy (Parks et al. 2018; Rinke et al. 2021). In fungal taxonomy, Gostinčar (2020) first tried to use the genomic distance to delineate fungal species, and obtained a relatively high degree of accuracy in delineating species according to the assumed threshold of genomic distances. However, the proposed criteria have not been widely utilized. Compared with the multilocus phylogenetic taxonomy, huge computational resource requirements, higher sequencing cost, more complicated analytic methods and lower accuracy at higher taxonomic ranks render it useless. In this study, we initially planned to differentiate lifestyles based on the basic numerical features of genomes and exclude the influence of phylogenetic signals. However, we unexpectedly discovered that some basic numerical features, such as genome size, GC content, and gene number, easily accessed from public databases, display powerful resolution for differentiating genomes at the higher levels, especially at the subclass. Inversely, most of these basic genomic features are useless only using the two features tRNA number and genome size without TEs displaying a certain degree of resolving power. To some extent, our discovery agrees with the conclusion of Li et al. (2021), in which fungal genome divergence is broadly consistent with the current taxonomic scheme at higher ranks, even using different genomic information. Fijarczyk et al. (2022) reported that pathogenic fungi include a higher number of protein-coding genes, tRNA genes, and larger genome sizes without repeats than non-pathogenic fungi. Compared with insect-unrelated fungi, they also found that insect-related fungi have smaller genome sizes, gene numbers and exon numbers but increased exon length. In this study, we divided 638 genomes into more specific lifestyles instead of only marking them as pathogenic or non-pathogenic, and our results are partially consistent with the previous discoveries by Fijarczyk et al. (2022). More specifically, we observed that plant pathogens have the largest average gene number of 11858, which is significantly larger than the average gene number of saprotrophs (the average number = 10581) and entomopathogens (the average number = 8821). However, entomopathogens have the smallest average gene number, which is significantly smaller than that of endophytes (the average number = 11577). As for genome size and tRNA number, we observed a similar pattern when we compared both features across lifestyles. In aggregate, although several basic genomic features display a certain degree of discrimination for differentiating lifestyles, we prefer to conclude that differences across these basic genomic features reflect taxonomic ranks rather than lifestyles.

Functional proteins are useful for differentiating lifestyles

Compared with basic genomic features, numerous studies have demonstrated that functional proteins, responsible for degrading substrates, invading host cells and obtaining nutrition are biologically more convincing in differentiating lifestyles (Feldman et al. 2017; Muszewska et al. 2017b; Seong and Krasileva 2023). In the present study, we divided the functional proteins into multiple groups and discovered that these functional proteins generally display relatively high discrimination for differentiating taxonomic groups at different ranks and slightly reduced for distinguishing lifestyles.

Secretome, a collective term representing all secreted proteins of an organism, is assumed to be related to fungal lifestyles. Krijger et al. (2014) reported that plant pathogens and saprotrophs include larger secretomes than animal pathogens, also indicated that differences in fungal secretome size reflects more on the phylogenetic relationships and less on lifestyle differences. Alfaro et al. (2014) believed that lifestyle is correlated to the composition of the secretome rather than its size. Recently, Chang et al. (2022) reported that the secretome size is mainly determined by phylogeny and lifestyle plays an important auxiliary role. Our results (Table S5: sheet pairwise-lifestyle) reveal that plant pathogens have the largest secretomes (the average number = 847), whereas entomopathogens have the smallest secretomes (the average number = 513). Based on the average number, we can clearly differentiate ($p < 0.05$) plant pathogens from entomopathogens, mycoparasites (the average number = 671), saprotrophs (the average number = 667) and entomopathogens, as well as differentiate entophytes (the average number = 828) from entomopathogens. With respect to the main protein groups, including CAZymes, lipases and SSPs, they display similar or higher discrimination than secretome, but lipases display lower discrimination.

PCWDEs play key roles in obtaining nutrients and degrading the main structural components of the plant cell wall, i.e., cellulose, hemicellulose and pectin. Lichenized fungi live as symbionts of green algae or cyanobacteria, obtaining diverse nutrients from their partners; therefore, they have fewer PCWDEs than non-lichenized fungi (Song et al. 2022). The reduction of PCWDEs is a prevailing trend in ectomycorrhizal Russulaceae (Looney et al. 2022), but they retain a certain degree of diversity in components (Kohler et al. 2015). The reduced abundance of PCWDEs in fungi might help in facilitating symbiosis by decreasing the expression of PCWDEs to reduce plant immune responses (Plett and Martin 2011). As for other kinds of lifestyles, the compositions of PCWDEs are different between saprophytic and plant-pathogenic fungi (Zhao et al. 2013; Kubicek et al. 2014). To the best of our knowledge, the present study is the first to conduct a comprehensively comparative analysis on the abundance of PCWDEs across multiple lifestyles. Plant-related fungi including endophytes (the average number = 75), plant pathogens (the average number = 81) and saprotrophs (the average number = 62) have a significantly larger repository of PCWDEs compared with entomopathogens (the average number = 12). For the plant-unrelated fungi, entomopathogens feature the smallest repository of PCWDEs. However, interestingly, human pathogens feature relatively high abundance of PCWDEs (the average number = 72). We investigated the lifestyles of these human pathogens, which belong to *Scedosporium* (Kaur et al. 2019), *Phialemoniopsis* (Alvarez Martinez et al. 2021), *Lomentospora* (Ramirez-Garcia et al. 2018), *Fusarium (Zhang et al. 2020)*, *Sporothrix* (Rodrigues et al. 2016), and *Madurella* (Ahmed et al. 2004), also confirmed that these groups are indeed associated with human diseases. However, we did not receive any clues to help explain the high abundance of PCWDEs in human pathogens. We speculate that these species mainly exist as non-human pathogens, but they rarely infect humans as opportunistic pathogens. Therefore, the contraction of PCWDEs has not yet occurred or is in an early evolutionary stage, while still featuring a large number of PCWDEs. More in-depth studies should be carried out to trace changes of PCWDEs in human pathogens.

FCWDEs are critical for degrading the cell wall of fungal hosts during mycoparasitism. Mycoparasitic species tend to have an expanded repository of FCWDEs (Gruber and Seidl-Seiboth 2012). Our results show that mycoparasites have the largest repository of FCWDEs (the average number = 41), which is significantly larger than entomopathogens (the average number = 28), human pathogens (the average number = 24), and plant pathogens (the average number = 27). To date, there are few studies that investigate the relationship between FCWDEs and fungal lifestyles. Results in the present study represent an important addition to this field.

The promising but limited potential of machine learning for lifestyle prediction

Machine learning algorithms heavily rely on massive amounts of data, the accuracy of which dramatically depends on not only the correctness of the training data and test data but also the quantity of input data (Raudys and Jain 1991; Sordo and Zeng 2005; Read et al. 2011). In classification tasks, inaccurately labeled datasets and inadequate sampling can lead to incorrect predictions. In the present study, there were two main challenges: inadequate sampling in several lifestyles and inaccurate lifestyle labels for some genomes. Unbalanced lifestyle distribution of genomes from public databases is common and unavoidable. Distribution largely depends on economic and medical importance, as well as the availability of samples. In our dataset, we include enough genomes of plant pathogens (n = 372), but fewer genomes of mycoparasites (n = 23), human pathogens (n = 16), and nematophagous fungi (n = 4). We excluded nematophagous fungi during analysis, but the relatively small sample sizes for multiple lifestyles affected the predictive accuracies to some extent, as shown in Fig. 7. Another challenge is assigning lifestyle labels to each genome. We attempted to determine the lifestyle of each genome, but for most genomes, the lifestyle is determined based on published literature or the submitter's description. Moreover, most studies directly characterize fungi isolated from diseased plants as plant pathogens, which does not follow Koch's postulates (van Wyk et al. 2012; Oberti et al. 2020; Telenko et al. 2020). Our predictive models display a high degree of accuracy in differentiating plant pathogens from other lifestyles, and adequate sampling reduced the error caused by inaccurate labeling. In predicting the lifestyle of unlabeled genomes, we further compared the predicted lifestyles and observed lifestyles in phylogenetically closed groups, and most of our predicted lifestyles are consistent with the observed lifestyles. Taken together, we suggest that using machine learning algorithms to predict fungal lifestyles is promising and can be improved with more sequenced genomes in the future.

Predicting potentially adverse fungal lifestyle

Fungi provide food and important medical and industrial secondary metabolites, as well as promote the global carbon cycle (Hyde et al. 2019; Lücking et al. 2021; Maharachchikumbura et al. 2021). However, the past two decades have witnessed the occurrence of new and emerging disease-causing fungi that infect plants, animals and humans (Fisher et al. 2012). Human activities have largely expanded fungal distribution and brought pathogenic fungal species accidentally to new ecosystems (Santini et al. 2013). *Pseudogymnoascus destructans*, an emerging fungal pathogen causing white-nose syndrome in bats, was initially detected in a commercial tourist cave, and it was speculated that the species was brought to external environments by tourist movements and further spread across North America, resulting in widespread mortality of hibernating bats (Blehert et al. 2009; Frick et al. 2015; Langwig et al. 2016). During the long-term interaction between fungal pathogens and hosts, both the fungi and the host have developed mechanisms to counteract each other's actions. Therefore, the hosts do not develop disease symptoms even if the fungi express abundant virulent factors. However, the fungi are introduced to new habitats and colonize new hosts, disease-causing interactions do develop (Parker and Gilbert 2004). *Phytophthora ramorum*, an alien plant pathogen to California and Oregon, causes a disease known as sudden oak death, that led to the death of a large number of trees, seriously threatening the local forest ecosystem (Rizzo and Garbelotto 2003). In addition, some fungal species or strains have multiple lifestyles, including non-pathogenic and pathogenic. Cannon et al. (2012) and Liu et al. (2022) demonstrated that endophytic fungi can switch to pathogenic lifestyle and cause disease symptoms. Due to the lack of effective analytical methods, some potential fungal pathogens were neglected until they caused devastating impacts on human health, food security and ecosystem stability (Anderson et al. 2004; Fisher et al. 2012; McDonald and Stukenbrock 2016). In scientific investigations and daily practices, we only observe one specific lifestyle of a certain fungal isolate under the current condition. Therefore, the experimentally exploring the potential lifestyles is impractical. In the study, our machine learning

model determine the fungal lifestyles according the corresponding probabilities, the highest probability represents the final predictive results, and the secondary high but non-zero probabilities imply that the strain might have other kind of lifestyles. For instance, *Arthrinium puccinioides* CBS 549.89 was predicted as a plant pathogen with a probability of 0.6689, but it may also be an endophyte or saprotroph with a probability of 0.1683 and 0.1628 respectively. Through a literature survey, we did observe endophytic and saprotrophic lifestyles in other species within the genus *Arthrinium* (Wang et al. 2018). With more fungal genomes sequenced and added to the dataset, the accuracy of our predictive model for determining fungal lifestyles using machine learning algorithms will become more reliable. The relatively high probability of harmful lifestyles can be used as an early warning of some devastating fungi. By identifying these harmful fungi early on, appropriate measures can be taken to prevent their spread and minimize their impact.

Supplementary information

Fig. S1. Distribution and proportion (%) of TE families in 638 genome assemblies. The bubble size represents the proportion of the TE in the genome. The bar represents the proportion of total TE size to the genome size.
Fig. S2. Composition and abundance of functional protein groups in 638 genome assemblies. The bubble size represents the number of the protein group. The bar represents the proportion of the secretome size to the total number of proteins per genome.
Fig. S3. Predictive accuracies of six machine learning algorithms based on three data matrices.

Table S1. A summary table containing genome information of 638 genome assemblies, lineage information, and statistics of TE categories, basic genomic features and functional protein groups.
Table S2. Taxonomic and lifestyle coverage of 638 Sordariomycete genomes.
Table S3. Number and proportion of different sequencing technologies.
Table S4. Results of Pearson Correlation of 25 basic genomic features.
Table S5. Comparative analysis results of 25 basic genomic features.
Table S6. Results of Pearson Correlation of 24 functional protein features.
Table S7. Comparative analysis results of 24 functional protein features.
Table S8. Predictive accuracies of six machine learning algorithms
Table S9. Prediction results of 83 undetermined genomes and the observed lifestyles of phylogenetically closed groups.

Code availability

All the scripts used for statistics, visualization and machine learning are written in R or Python. Scripts are available at GitHub (https://github.com/ypchan/Predict-fungal-lifestyles).

Acknowledgments

This research was funded by the Talent Introduction and Cultivation Project, University of Electronic Science and Technology of China, grant number A1098531023601245. Several genomes were produced by the US Department of Energy Joint Genome Institute (JGI) (https://ror.org/04xm1d337; operated under Contract No. DE-AC02-05CH11231) in collaboration with the user community and we acknowledge Professor J.W. Spatafora for allowing us to use these genomes submitted in JGI. K.D. Hyde acknowledges the National Research Council of Thailand (NRCT) grant “Total fungal diversity in a given forest area with implications towards species numbers, chemical

diversity and biotechnology" (grant no. N42A650547).

Author contributions

SSNM and YPC designed the study. YPC collected genome data and performed all bioinformatic analyses. SSNM, PWS, WHT and YPC checked the tables and performed lifestyle assignments. YPC and SSNM wrote the first draft of the manuscript. RX checked all R codes and Python codes, and provided a portion of computing resources. SSNM, HKD and MS help in revision. All authors provided valuable comments on the manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by Talent Introduction and Cultivation Project, University of Electronic Science and Technology of China, grant number A1098531023601245.

Declarations

The authors declare that there is no conflict of interest related to this study.

Reference

Ahmed AOA, van Leeuwen W, Fahal A, van de Sande W, Verbrugh H, van Belkum A (2004) Mycetoma caused by *Madurella mycetomatis*: a neglected infectious burden. Lancet Infect Dis 4 (9):566-574. https://doi.org/10.1016/S1473-3099(04)01131-4

Akira S, Uematsu S, Takeuchi O (2006) Pathogen recognition and innate immunity. Cell 124 (4):783-801. https://doi.org/10.1016/j.cell.2006.02.015

Alfaro M, Oguiza JA, Ramírez L, Pisabarro AG (2014) Comparative analysis of secretomes in basidiomycete fungi. J Proteomics 102:28-43. https://doi.org/10.1016/j.jprot.2014.03.001

Alvarez Martinez D, Alberto C, Riat A, Schuhler C, Valladares P, Ninet B, Kraak B, Crous PW, Hou LW, Toutous Trellu L (2021) *Phialemoniopsis limonesiae* sp. nov. causing cutaneous phaeohyphomycosis in an immunosuppressed woman. Emerging Microbes Infect 10 (1):400-406. https://doi.org/10.1080/22221751.2021.1892458

Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P (2004) Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol 19 (10):535-544. https://doi.org/10.1016/j.tree.2004.07.021

Araújo JPM, Evans HC, Kepler R, Hughes DP (2018) Zombie-ant fungi across continents: 15 new species and new combinations within *Ophiocordyceps*. I. Myrmecophilous hirsutelloidÂ species. Stud Mycol 90:119-160. https://doi.org/10.1016/j.simyco.2017.12.002

Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6 (1):11. https://doi.org/10.1186/s13100-015-0041-9

Barros MBdL, Paes RdA, Schubach AO (2011) *Sporothrix schenckii* and Sporotrichosis. Clin Microbiol Rev 24 (4):633-654. https://doi.org/10.1128/CMR.00007-11

Bartlett P, Eberhardt U, Schütz N, Beker HJ (2022) Species determination using AI machine-learning algorithms: *Hebeloma* as a case study. IMA Fungus 13 (1):13. https://doi.org/10.1186/s43008-022-00099-x

Birch PRJ, Rehmany AP, Pritchard L, Kamoun S, Beynon JL (2006) Trafficking arms: oomycete effectors enter host plant cells. Trends Microbiol 14 (1):8-11. https://doi.org/10.1016/j.tim.2005.11.007

Blehert DS, Hicks AC, Behr M, Meteyer CU, Berlowski-Zier BM, Buckles EL, Coleman JTH, Darling SR, Gargas A, Niver R, Okoniewski JC, Rudd RJ, Stone WB (2009) Bat white-nose syndrome: an emerging fungalpathogen? Science 323 (5911), 227-227. https://doi.org/10.1126/science.1163874

Boddy L (2016) Chapter 9 - Interactions with humans and other animals. In: Watkinson SC, Boddy L, Money NP (eds) The Fungi (Third Edition). Academic Press, Boston, pp 293-336. https://doi.org/10.1016/B978-0-12-382034-1.00009-8

Boller T (1995) Chemoperception of microbial signals in plant cells. Annu Rev Plant Phys 46 (1):189-214. https://doi.org/10.1146/annurev.pp.46.060195.001201

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinf 3 (1). https://doi.org/10.1093/nargab/lqaa108

Brůna T, Lomsadze A, Borodovsky M (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinf 2 (2). https://doi.org/10.1093/nargab/lqaa026

Bzdok D, Krzywinski M, Altman N (2018) Machine learning: supervised methods. Nat Methods 15 (1):5-6. https://doi.org/10.1038/nmeth.4551

Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ (2018) Next-generation machine learning for biological networks. Cell 173 (7):1581-1592. https://doi.org/10.1016/j.cell.2018.05.015

Cannon PF, Damm U, Johnston PR, Weir BS (2012) *Colletotrichum*: current status and future directions. Stud. Mycol 73 (1):181-213. https://doi.org/10.3114/sim0014

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38 (12):5825-5829. https://doi.org/10.1093/molbev/msab293

Chan Patricia P, Lin Brian Y, Mak Allysia J, Lowe Todd M (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res 49 (16):9077-9096. https://doi.org/10.1093/nar/gkab688

Chang TH, Hassan O, Lee YS (2018) First Report of Anthracnose of Japanese Plum (*Prunus salicina*) Caused by *Colletotrichum nymphaeae* in Korea. Plant Disease 102 (7):1461-1461. https://doi.org/10.1094/pdis-01-18-0018-pdn

Chang Y, Wang Y, Mondo S, Ahrendt S et al (2022) Evolution of zygomycete secretomes and the origins of terrestrial fungal ecologies. iScience 25 (8). https://doi.org/10.1016/j.isci.2022.104840

Chen YP, Wu T, Tian WH, Ilyukhin F, Hyde KD, Maharachchikumbura SSN (2022) Comparative genomics provides new insights into the evolution of *Colletotrichum*. Mycosphere 13 (2):56. https://doi.org/10.5943/mycosphere/si/1f/5

Consortium TU (2020) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49 (D1):D480-D489. https://doi.org/10.1093/nar/gkaa1100

Cooper, G M. (2000). The Cell: A Molecular Approach. 2nd edition. Sinauer Associates.

Crawford K, Heatley NG, Boyd PF, Hale CW, Kelly BK, Miller GA, Smith N (1952) Antibiotic production by a species of *Cephalosporium*. J Gen Microbiol 6 (1-2):47-59. https://doi.org/10.1099/00221287-6-1-2-47

Crous PW, Lombard L, Sandoval-Denis M, Seifert KA et al (2021) *Fusarium*: more than a node or a foot-shaped basal cell. Stud Mycol 98:100116. https://doi.org/10.1016/j.simyco.2021.100116

Czislowski E, Zeil-Rolfe I, Aitken EAB (2021) Effector profiles of endophytic *Fusarium* associated with asymptomatic banana (*Musa* sp.) hosts. Int J Mol Sci 22 (5):2508. https://doi.org/10.3390/ijms22052508

Dasari P, Shopova IA, Stroe M, Wartenberg D, Martin-Dahse H, Beyersdorf N, Hortschansky P, Dietrich S, Cseresnyés Z, Figge MT, Westermann M, Skerka C, Brakhage AA, Zipfel PF (2018) Aspf2 from *Aspergillus fumigatus* recruits human immune regulators for immune evasion and cell damage. Front Immunol 9. https://doi.org/10.3389/fimmu.2018.01635

de Jonge R, Bolton MD, Thomma BPHJ (2011) How filamentous pathogens co-opt plants: the ins and outs of fungal effectors. Curr Opin Plant Biol 14 (4):400-406. https://doi.org/10.1016/j.pbi.2011.03.005

Dean R, Van Kan JL, Pretorius ZA, Hammond-kosack KE, Dipietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J, Foster GD (2012) The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol 13 (4):414-430. https://doi.org/10.1111/j.1364-3703.2011.00783.x

Delaye L, García-Guzmán G, Heil M (2013) Endophytes versus biotrophic and necrotrophic pathogens—are fungal lifestyles evolutionarily stable traits? Fungal Divers 60 (1):125-135. https://doi.org/10.1007/s13225-013-0240-y

Deo RC (2015) Machine learning in medicine. Circulation 132 (20):1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Derkarabetian S, Castillo S, Koo PK, Ovchinnikov S, Hedin M (2019) A demonstration of unsupervised machine learning in species delimitation. Mol Phylogenet Evol 139:106562. https://doi.org/10.1016/j.ympev.2019.106562

Eastwood DC, Floudas D, Binder M, Majcherczyk A et al (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science 333 (6043):762-765.

https://doi.org/10.1126/science.1205411
Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2 (4):953-971. https://doi.org/10.1038/nprot.2007.131
Eriksson OE WK (1997) Supraordinal taxa of Ascomycota. Myconet 1 (1):1–16
Feldman D, Kowbel DJ, Glass NL, Yarden O, Hadar Y (2017) A role for small secreted proteins (SSPs) in a saprophytic fungal lifestyle: ligninolytic enzyme regulation in *Pleurotus ostreatus*. Sci Rep 7 (1):14553. https://doi.org/10.1038/s41598-017-15112-2
Fijarczyk A, Hessenauer P, Hamelin RC, Landry CR (2022) Lifestyles shape genome size and gene content in fungal pathogens. bioRxiv:2022.2008.2024.505148. https://doi.org/10.1101/2022.08.24.505148
Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ (2012) Emerging fungal threats to animal, plant and ecosystem health. Nature 484 (7393), 186-194. https://doi.org/10.1038/nature10947
Fontana DC, de Paula S, Torres AG, de Souza VHM, Pascholati SF, Schmidt D, Dourado Neto D (2021) Endophytic fungi: biological control and induced resistance to phytopathogens and abiotic stresses. Pathogens 10 (5):570. https://doi.org/10.3390/pathogens10050570
Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D (2019) Stress-driven transposable element de-repression dynamics and virulence evolution in a fungal pathogen. Mol Biol Evol 37 (1):221-239. https://doi.org/10.1093/molbev/msz216
Frey-Klett P, Burlinson P, Deveau A, Barret M, Tarkka M, Sarniguet A (2011) Bacterial-fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists. Microbiol Mol Biol Rev 75 (4):583-609. https://doi.org/10.1128/MMBR.00020-11
Frick WF, Puechmaille SJ, Hoyt JR, Nickel BA, Langwig KE, Foster JT, Barlow KE, Bartonička T, Feller D, Haarsma A-J, Herzog C, Horáček I, van der Kooij J, Mulkens B, Petrov B, Reynolds R, Rodrigues L, Stihler CW, Turner GG, Kilpatrick AM (2015) Disease alters macroecological patterns of North American bats. Glob Ecol Biogeogr 24 (7), 741-749. https://doi.org/10.1111/geb.12290
Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet 38 (8):953-956. https://doi.org/10.1038/ng1839
Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28 (23):3150-3152. https://doi.org/10.1093/bioinformatics/bts565
Gíslason MH, Nielsen H, Almagro Armenteros JJ, Johansen AR (2021) Prediction of GPI-anchored proteins with pointer neural networks. Curr Res Biotechnol 3:6-13. https://doi.org/10.1016/j.crbiot.2021.01.001
Glazebrook J (2005) Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. Annu Rev Phytopathol 43 (1):205-227. https://doi.org/10.1146/annurev.phyto.43.040204.135923
Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17 (6):333-351. https://doi.org/10.1038/nrg.2016.49
Gostinčar C (2020) Towards genomic criteria for delineating fungal species. J Fungi 6 (4):246. https://doi.org/10.3390/jof6040246
Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2013) MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res 42 (D1):D699-D704. https://doi.org/10.1093/nar/gkt1183
Gruber S, Seidl-Seiboth V (2012) Self versus non-self: fungal cell wall degradation in *Trichoderma*. Microbology 158 (1):26-34. https://doi.org/10.1099/mic.0.052613-0
GS K (1996) Disease mechanisms of fungi. In: Baron S (ed) Medical Microbiology. vol 4th edition. University of

Texas Medical Branch at Galveston, Galveston (TX).
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59 (3):307-321. https://doi.org/10.1093/sysbio/syq010
Han S, Wang M, Ma Z, Raza M, Zhao P, Liang J, Gao M, Li Y, Wang J, Hu D (2023) *Fusarium* diversity associated with diseased cereals in China, with an updated phylogenomic assessment of the genus. Stud Mycol 104:87-148. https://doi.org/10.3114/sim.2022.104.02
Haridas S, Albert R, Binder M, Bloem J et al (2020) 101 Dothideomycetes genomes: a test case for predicting lifestyles and emergence of pathogens. Stud Mycol 96:141-153. https://doi.org/10.1016/j.simyco.2020.01.003
Hill R, Buggs RJA, Vu DT, Gaya E (2022) Lifestyle transitions in fusarioid fungi are frequent and lack clear genomic signatures. Mol Biol Evol 39 (4). https://doi.org/10.1093/molbev/msac085
Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2017) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35 (2):518-522. https://doi.org/10.1093/molbev/msx281
Hobbie EA, Horton TR (2007) Evidence that saprotrophic fungi mobilise carbon and mycorrhizal fungi mobilise nitrogen during litter decomposition. New Phytol 173 (3):447-449. https://doi.org/10.1111/j.1469-8137.2007.01984.x
Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35 (suppl_2):W585-W587. https://doi.org/10.1093/nar/gkm259
Hu E-Z, Lan X-R, Liu Z-L, Gao J, Niu D-K (2022) A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics 23 (1):110. https://doi.org/10.1186/s12864-022-08353-7
Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ (2015) The Dfam database of repetitive DNA families. Nucleic Acids Res 44 (D1):D81-D89. https://doi.org/10.1093/nar/gkv1272
Hyde KD, Norphanphoun C, Maharachchikumbura SSN, Bhat DJ et al (2020) Refined families of Sordariomycetes. Mycosphere 11 (1):305-1059. https://doi.org/10.5943/mycosphere/11/1/7
Hyde KD, Xu J, Rapior S, Jeewon R et al (2019) The amazing potential of fungi: 50 ways we can exploit fungi industrially. Fungal Divers 97 (1), 1-136. https://doi.org/10.1007/s13225-019-00430-9
Ismaiel AA, Papenbrock J (2015) Mycotoxins: producing fungi and mechanisms of phytotoxicity. Agriculture 5 (3):492-537. https://doi.org/10.3390/agriculture5030492
Jenks JD, Reed SL, Seidel D, Koehler P, Cornely OA, Mehta SR, Hoenigl M (2018) Rare mould infections caused by Mucorales, *Lomentospora prolificans* and *Fusarium*, in San Diego, CA: the role of antifungal combination therapy. Int J Antimicrob Agents 52 (5):706-712. https://doi.org/10.1016/j.ijantimicag.2018.08.005
Jia M, Chen L, Xin H-L, Zheng C-J, Rahman K, Han T, Qin L-P (2016) A friendly relationship between endophytic fungi and medicinal plants: a systematic review. Front Microbiol 7. https://doi.org/10.3389/fmicb.2016.00906
Jones DAB, Bertazzoni S, Turo CJ, Syme RA, Hane JK (2018) Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. Curr Opin Microbiol 46:43-49. https://doi.org/10.1016/j.mib.2018.01.017
Kaewchai S, Soytong K, Hyde KD (2009) Mycofungicides and fungal biofertilizers. Fungal Divers 38:25-50
Kale SD, Tyler BM (2011) Entry of oomycete and fungal effectors into plant and animal host cells. Cell Microbiol 13 (12):1839-1848. https://doi.org/10.1111/j.1462-5822.2011.01659.x
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14 (6):587-589. https://doi.org/10.1038/nmeth.4285
Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment

based on fast Fourier transform. Nucleic Acids Res 30 (14):3059-3066. https://doi.org/10.1093/nar/gkf436

Kaur J, Kautto L, Penesyan A, Meyer W, Elbourne LDH, Paulsen IT, Nevalainen H (2019) Interactions of an emerging fungal pathogen *Scedosporium aurantiacum* with human lung epithelial cells. Sci Rep 9 (1):5035. https://doi.org/10.1038/s41598-019-41435-3

Kim K-T, Jeon J, Choi J, Cheong K, Song H, Choi G, Kang S, Lee Y-H (2016) Kingdom-wide analysis of fungal small secreted proteins (SSPs) reveals their potential role in host association. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00186

Kirkland TN, Muszewska A, Stajich JE (2018) Analysis of transposable elements in *Coccidioides* species. J Fungi 4 (1):13. https://doi.org/10.3390/jof4010013

Knapp DG, Németh JB, Barry K, Hainaut M, Henrissat B, Johnson J, Kuo A, Lim JHP, Lipzen A, Nolan M, Ohm RA, Tamás L, Grigoriev IV, Spatafora JW, Nagy LG, Kovács GM (2018) Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. Sci Rep 8 (1):6321. https://doi.org/10.1038/s41598-018-24686-4

Kohler A, Kuo A, Nagy LG, Morin E et al (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet 47 (4):410-415. https://doi.org/10.1038/ng.3223

Krijger J-J, Thon MR, Deising HB, Wirsel SGR (2014) Compositions of fungal secretomes indicate a greater impact of phylogenetic history than lifestyle adaptation. BMC Genomics 15 (1):722. https://doi.org/10.1186/1471-2164-15-722

Kubicek CP, Starr TL, Glass NL (2014) Plant cell wall–degrading enzymes and their secretion in plant-pathogenic fungi. Annu Rev Phytopathol 52 (1):427-451. https://doi.org/10.1146/annurev-phyto-102313-045831

Kwon SL, Park MS, Jang S, Lee YM, Heo YM, Hong J-H, Lee H, Jang Y, Park J-H, Kim C, Kim G-H, Lim YW, Kim J-J (2021) The genus *Arthrinium* (Ascomycota, Sordariomycetes, Apiosporaceae) from marine habitats from Korea, with eight new species. IMA Fungus 12 (1):13. https://doi.org/10.1186/s43008-021-00065-z

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L, Wang D, Hu F, Wang W, Liu S (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. GigaScience 9 (12). https://doi.org/10.1093/gigascience/giaa123

Langwig KE, Frick WF, Hoyt JR, Parise KL, Drees KP, Kunz TH, Foster JT, Kilpatrick AM (2016) Drivers of variation in species impacts for a multi-host fungal disease of bats. Philos Trans R Soc B 371 (1709): 20150456. https://doi.org/10.1098/rstb.2015.0456

Lebreton A, Tang N, Kuo A, LaButti K, Andreopoulos W, Drula E, Miyauchi S, Barry K, Clum A, Lipzen A, Mousain D, Ng V, Wang R, Dai Y, Henrissat B, Grigoriev IV, Guerin-Laguette A, Yu F, Martin FM (2022) Comparative genomics reveals a dynamic genome evolution in the ectomycorrhizal milk-cap (*Lactarius*) mushrooms. New Phytol 235 (1):306-319. https://doi.org/10.1111/nph.18143

Lee S-J, Rose JKC (2010) Mediation of the transition from biotrophy to necrotrophy in hemibiotrophic plant pathogens by secreted effector proteins. Plant Signaling Behav 5 (6):769-772. https://doi.org/10.4161/psb.5.6.11778

Li J, Cornelissen B, Rep M (2020) Host-specificity factors in plant pathogenic fungi. Fungal Genet Biol 144:103447. https://doi.org/10.1016/j.fgb.2020.103447

Li Y, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, Spatafora JW, Groenewald M, Dunn CW, Hittinger CT, Shen X-X, Rokas A (2021) A genome-scale phylogeny of the kingdom fungi. Curr Biol 31 (8):1653-1665.e1655. https://doi.org/10.1016/j.cub.2021.01.074

Liu F, Ma ZY, Hou LW, Diao YZ, Wu WP, Damm U, Song S, Cai L (2022) Updating species diversity of *Colletotrichum*, with a phylogenomic overview. Stud Mycol 101 (1):1-56. https://doi.org/10.3114/sim.2022.101.01

Looney B, Miyauchi S, Morin E, Drula E, Courty PE, Kohler A, Kuo A, LaButti K, Pangilinan J, Lipzen A, Riley R, Andreopoulos W, He G, Johnson J, Nolan M, Tritt A, Barry KW, Grigoriev IV, Nagy LG, Hibbett D, Henrissat B, Matheny PB, Labbé J, Martin FM (2022) Evolutionary transition to the ectomycorrhizal habit in the genomes of a hyperdiverse lineage of mushroom-forming fungi. New Phytol 233 (5):2294-2309. https://doi.org/10.1111/nph.17892

Lorrain C, Feurtey A, Möller M, Haueisen J, Stukenbrock E (2021) Dynamics of transposable elements in recently diverged fungal pathogens: lineage-specific transposable element content and efficiency of genome defenses. G3-Genes Genom Genet 11 (4). https://doi.org/10.1093/g3journal/jkab068

Lu S, Edwards MC (2016) Genome-wide analysis of small secreted cysteine-rich proteins identifies candidate effector proteins potentially involved in *Fusarium graminearum*−wheat interactions. Phytopathology 106 (2):166-176. https://doi.org/10.1094/phyto-09-15-0215-r

Luangsa-ard JJ, Mongkolsamrit S, Thanakitpipattana D, Khonsanit A, Tasanathai K, Noisripoom W, Humber RA (2017) Clavicipitaceous entomopathogens: new species in *Metarhizium* and a new genus *Nigelia*. Mycol Prog 16 (4):369-391. https://doi.org/10.1007/s11557-017-1277-1

Lücking R, Aime MC, Robbertse B, Miller AN, Aoki T, Ariyawansa HA, Cardinali G, Crous PW, Druzhinina IS, Geiser DM, Hawksworth DL, Hyde KD, Irinyi L, Jeewon R, Johnston PR, Kirk PM, Malosso E, May TW, Meyer W, Nilsson HR, Öpik M, Robert V, Stadler M, Thines M, Vu D, Yurkov AM, Zhang N, Schoch CL (2021) Fungal taxonomy and sequence-based nomenclature. Nat Microbiol 6 (5), 540-548. https://doi.org/10.1038/s41564-021-00888-x

Luo Z-L, Hyde KD, Liu J-K, Maharachchikumbura SSN, Jeewon R, Bao D-F, Bhat DJ, Lin C-G, Li W-L, Yang J, Liu N-G, Lu Y-Z, Jayawardena RS, Li J-F, Su H-Y (2019) Freshwater Sordariomycetes. Fungal Divers 99 (1):451-660. https://doi.org/10.1007/s13225-019-00438-1

Ma C, Zhang HH, Wang X (2014) Machine learning for Big Data analytics in plants. Trends Plant Sci 19 (12):798-808. https://doi.org/10.1016/j.tplants.2014.08.004

Macho Alberto P, Zipfel C (2014) Plant PRRs and the activation of innate immune signaling. Mol Cell 54 (2):263-272. https://doi.org/10.1016/j.molcel.2014.03.028

Magyar D, Tartally A, Merényi Z (2022) *Hagnosa longicapillata*, gen. nov., sp. nov., a new sordariaceous Ascomycete in the indoor environment, and the proposal of Hagnosaceae fam. nov. Pathogens 11 (5):593. https://doi.org/10.3390/pathogens11050593

Maharachchikumbura SSN, Hyde KD, Jones EBG, McKenzie EHC et al (2015) Towards a natural classification and backbone tree for Sordariomycetes. Fungal Divers 72 (1):199-301. https://doi.org/10.1007/s13225-015-0331-z

Maharachchikumbura SSN, Wanasinghe DN, Cheewangkoon R, Al-Sadi AM (2021) Uncovering the hidden taxonomic diversity of fungi in Oman. Fungal Divers 106 (1):229-268. https://doi.org/10.1007/s13225-020-00467-1

Maharachchikumbura SSN, Chen Y, Ariyawansa HA, Hyde KD, Haelewaters D, Perera RH, Samarakoon MC, Wanasinghe DN, Bustamante DE, Liu J-K, Lawrence DP, Cheewangkoon R, Stadler M (2021) Integrative approaches for species delimitation in Ascomycota. Fungal Divers 109 (1):155-179. https://doi.org/10.1007/s13225-021-00486-6

Mäkelä MR, Donofrio N, de Vries RP (2014) Plant biomass degradation by fungi. Fungal Genet Biol 72:2-9. https://doi.org/10.1016/j.fgb.2014.08.010

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38 (10):4647-4654. https://doi.org/10.1093/molbev/msab199

Mapuranga J, Zhang N, Zhang L, Chang J, Yang W (2022) Infection strategies and pathogenicity of biotrophic plant fungal pathogens. Front Microbiol 13. https://doi.org/10.3389/fmicb.2022.799396

McCotter SW, Horianopoulos LC, Kronstad JW (2016) Regulation of the fungal secretome. Curr Genet 62 (3):533-545. https://doi.org/10.1007/s00294-016-0578-2

McDonald BA, Stukenbrock EH (2016) Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. Philos Trans R Soc B 371 (1709):9. https://doi.org/10.1098/rstb.2016.0026

Melén K, Krogh A, von Heijne G (2003) Reliability measures for membrane protein topology prediction algorithms. J Mol Biol 327 (3):735-744. https://doi.org/10.1016/S0022-2836(03)00182-7

Mendgen K, Hahn M (2002) Plant infection and the establishment of fungal biotrophy. Trends Plant Sci 7 (8):352-356. https://doi.org/10.1016/S1360-1385(02)02297-5

Meng Y, Lei Y, Gao J, Liu Y, Ma E, Ding Y, Bian Y, Zu H, Dong Y, Zhu X (2022) Genome sequence assembly algorithms and misassembly identification methods. Mol Biol Rep 49 (11):11133-11148. https://doi.org/10.1007/s11033-022-07919-8

Mengiste T (2012) Plant immunity to necrotrophs. Annu Rev Phytopathol 50 (1):267-294. https://doi.org/10.1146/annurev-phyto-081211-172955

Mesny F, Miyauchi S, Thiergart T, Pickel B et al (2021) Genetic determinants of endophytism in the *Arabidopsis* root mycobiome. Nat Commun 12 (1):7227. https://doi.org/10.1038/s41467-021-27479-y

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95 (6):315-327. https://doi.org/10.1016/j.ygeno.2010.03.001

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37 (5):1530-1534. https://doi.org/10.1093/molbev/msaa015

Miyauchi S, Kiss E, Kuo A, Drula E et al (2020) Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. Nat Commun 11 (1):5125. https://doi.org/10.1038/s41467-020-18795-w

Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, Harliwong I, Wei H, Yang B, Ye Q, Anderson E, Mao Q, Drmanac R, Wang O, Peters BA, Xu M, Wu P, Topp B, Coin LJM, Henry RJ (2020) Comparison of long-read methods for sequencing and assembly of a plant genome. GigaScience 9 (12). https://doi.org/10.1093/gigascience/giaa146

Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K (2017a) Cut-and-paste transposons in fungi with diverse lifestyles. Genome Biol Evol 9 (12):3463-3477. https://doi.org/10.1093/gbe/evx261

Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K (2019) Transposable elements contribute to fungal genes and impact fungal lifestyle. Sci Rep 9 (1):4307. https://doi.org/10.1038/s41598-019-40965-0

Muszewska A, Stepniewska-Dziubinska MM, Steczkiewicz K, Pawlowska J, Dziedzic A, Ginalski K (2017b) Fungal lifestyle reflected in serine protease repertoire. Sci Rep 7 (1):9147. https://doi.org/10.1038/s41598-017-09644-w

Newman TE, Derbyshire MC (2020) The evolutionary and molecular features of broad host-range necrotrophy in plant pathogenic fungi. Front Plant Sci 11. https://doi.org/10.3389/fpls.2020.591733

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal

peptides and prediction of their cleavage sites. Protein Eng Des Sel 10 (1):1-6. https://doi.org/10.1093/protein/10.1.1

Nouws S, Bogaerts B, Verhaegen B, Denayer S, Piérard D, Marchal K, Roosens NHC, Vanneste K, De Keersmaecker SCJ (2020) Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. Sci Rep 10 (1):14649. https://doi.org/10.1038/s41598-020-71207-3

O'Connell RJ, Thon MR, Hacquard S, Amyotte SG et al (2012) Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. Nat. Genet. 44 (9):1060-1065. https://doi.org/10.1038/ng.2372

Oberti H, Dalla Rizza M, Reyno R, Murchio S, Altier N, Abreo E (2020) Diversity of *Claviceps paspali* reveals unknown lineages and unique alkaloid genotypes. Mycologia 112 (2):230-243. https://doi.org/10.1080/00275514.2019.1694827

Oggenfuss U, Badet T, Wicker T, Hartmann FE, Singh NK, Abraham L, Karisto P, Vonlanthen T, Mundt C, McDonald BA, Croll D (2021) A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. eLife 10:e69249. https://doi.org/10.7554/eLife.69249

Omrane S, Audéon C, Ignace A, Duplaix C, Aouini L, Kema G, Walker A-S, Fillinger S (2017) Plasticity of the *MFS1* promoter leads to multidrug resistance in the wheat pathogen *Zymoseptoria tritici*. mSphere 2 (5):e00393-00317. https://doi.org/10.1128/mSphere.00393-17

Parker IM, Gilbert GS (2004) The evolutionary ecology of novel plant-pathogen interactions. Annu Rev Ecol Evol Syst 35 (1): 675-700. https://doi.org/10.1146/annurev.ecolsys.34.011802.132339

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36 (10):996-1004. https://doi.org/10.1038/nbt.4229

Pellegrin C, Morin E, Martin FM, Veneault-Fourrey C (2015) Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins. Front Microbiol 6. https://doi.org/10.3389/fmicb.2015.01278

Peter M, Kohler A, Ohm RA, Kuo A, Krützmann J, Morin E, Arend M, Barry KW, Binder M, Choi C, Clum A, Copeland A, Grisel N, Haridas S, Kipfer T, LaButti K, Lindquist E, Lipzen A, Maire R, Meier B, Mihaltcheva S, Molinier V, Murat C, Pöggeler S, Quandt CA, Sperisen C, Tritt A, Tisserant E, Crous PW, Henrissat B, Nehls U, Egli S, Spatafora JW, Grigoriev IV, Martin FM (2016) Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus *Cenococcum geophilum*. Nat Commun 7 (1):12662. https://doi.org/10.1038/ncomms12662

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8 (10):785-786. https://doi.org/10.1038/nmeth.1701

Phurailatpam L, Mishra S (2020) Role of plant endophytes in conferring abiotic stress tolerance. In: Hasanuzzaman M (ed) Plant Ecophysiology and Adaptation under Climate Change: Mechanisms and Perspectives II: Mechanisms of Adaptation and Stress Amelioration. Springer Singapore, Singapore, pp 603-628. https://doi.org/10.1007/978-981-15-2172-0_22

Plett JM, Martin F (2011) Blurred boundaries: lifestyle lessons from ectomycorrhizal fungal genomes. Trends Genet 27 (1):14-22. https://doi.org/10.1016/j.tig.2010.10.005

Presti LL, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, Zuccaro A, Reissmann S, Kahmann R (2015) Fungal effectors and plant susceptibility. Annu Rev Plant Biol 66 (1):513-545. https://doi.org/10.1146/annurev-arplant-043014-114623

Promputtha I, Hyde KD, McKenzie EHC, Peberdy JF, Lumyong S (2010) Can leaf degrading enzymes provide

evidence that endophytic fungi becoming saprobes? Fungal Divers 41 (1):89-99. https://doi.org/10.1007/s13225-010-0024-6

Promputtha I, Lumyong S, Dhanasekaran V, McKenzie EHC, Hyde KD, Jeewon R (2007) A phylogenetic evaluation of whether endophytes become saprotrophs at host senescence. Microb Ecol 53 (4):579-590. https://doi.org/10.1007/s00248-006-9117-x

R Core Team (2022) R: A language and environment for statistical computing. in R Foundation for Statistical Computing. (2020).

Rai M, Agarkar G (2016) Plant–fungal interactions: What triggers the fungi to switch among lifestyles? Crit Rev Microbiol 42 (3):428-438. https://doi.org/10.3109/1040841X.2014.958052

Ramirez-Garcia A, Pellon A, Rementeria A, Buldain I et al (2018) *Scedosporium* and *Lomentospora*: an updated overview of underrated opportunists. Med Mycol 56 (suppl_1):S102-S125. https://doi.org/10.1093/mmy/myx113

Raudys SJ, Jain AK (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13 (3):252-264. https://doi.org/10.1109/34.75512

Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD (2017) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 46 (D1):D624-D632. https://doi.org/10.1093/nar/gkx1134

Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. Mach Learn 85 (3):333-359. https://doi.org/10.1007/s10994-011-5256-5

Réblová M, Miller AN, Rossman AY, Seifert KA, Crous PW et al (2016) Recommendations for competing sexual-asexually typified generic names in Sordariomycetes (except Diaporthales, Hypocreales, and Magnaporthales). IMA Fungus 7 (1):131-153. https://doi.org/10.5598/imafungus.2016.07.01.08

Řehulka J, Kubátová A, Hubka V (2016) *Cephalotheca sulfurea* (Ascomycota, Sordariomycetes), a new fungal pathogen of the farmed rainbow trout *Oncorhynchus mykiss*. J Fish Dis 39 (12):1413-1419. https://doi.org/10.1111/jfd.12477

Rinke C, Chuvochina M, Mussig AJ, Chaumeil P-A, Davín AA, Waite DW, Whitman WB, Parks DH, Hugenholtz P (2021) A standardized archaeal taxonomy for the genome taxonomy database. Nat Microbiol 6 (7):946-959. https://doi.org/10.1038/s41564-021-00918-8

Rizzo DM, Garbelotto M (2003) Sudden oak death: endangering California and Oregon forest ecosystems. Front Ecol Environ 1 (4):197-204. https://doi.org/10.1890/1540-9295(2003)001[0197:SODECA]2.0.CO;2

Rodrigues AM, de Hoog GS, de Camargo ZP (2016) *Sporothrix* species causing outbreaks in animals and humans driven by animal–animal transmission. PLoS Pathog 12 (7):e1005638. https://doi.org/10.1371/journal.ppat.1005638

Salim D, Bradford WD, Freeland A, Cady G, Wang J, Pruitt SC, Gerton JL (2017) DNA replication stress restricts ribosomal DNA copy number. PLos Genet 13 (9):e1007006. https://doi.org/10.1371/journal.pgen.1007006

Santini A, Ghelardini L, De Pace C, Desprez-Loustau ML et al (2013) Biogeographical patterns and determinants of invasion by forest pathogens in Europe. New Phytol 197 (1), 238-250. https://doi.org/10.1111/j.1469-8137.2012.04364.x

Senft AD, Macfarlan TS (2021) Transposable elements shape the evolution of mammalian development. Nat Rev Genet 22 (11):691-711. https://doi.org/10.1038/s41576-021-00385-1

Seong K, Krasileva KV (2023) Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. Nat Microbiol 8 (1):174-187. https://doi.org/10.1038/s41564-022-01287-6

Shang Y, Feng P, Wang C (2015) Fungi that infect insects: altering host behavior and beyond. PLoS Pathog 11

(8):e1005037. https://doi.org/10.1371/journal.ppat.1005037
Shen X, Opulente DAx, Kominek J, Zhou X et al (2018) Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175 (6):1533-1545.e1520. https://doi.org/10.1016/j.cell.2018.10.023
Shen X, Steenwyk JL, LaBella AL, Opulente DA, Zhou X, Kominek J, Li Y, Groenewald M, Hittinger CT, Rokas A (2020) Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. Sci Adv 6 (45). https://doi.org/10.1126/sciadv.abd0079
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19):3210-3212. https://doi.org/10.1093/bioinformatics/btv351
Simmons DR, Kepler RM, Renner SA, Groden E (2015) Phylogeny of *Hirsutella* species (Ophiocordycipitaceae) from the USA: remedying the paucity of *Hirsutella* sequence data. IMA Fungus 6 (2):345-356. https://doi.org/10.5598/imafungus.2015.06.02.06
Singh VK, Meena M, Zehra A, Tiwari A, Dubey MK, Upadhyay RS (2014) Fungal toxins and their impact on living systems. In: Kharwar RN, Upadhyay RS, Dubey NK, Raghuwanshi R (eds) Microbial Diversity and Biotechnology in Food Security. Springer India, New Delhi, pp 513-530. https://doi.org/10.1007/978-81-322-1801-2_47
Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová O (2014) Ecological and evolutionary significance of genomic GC content diversity in monocots. PNAS 111 (39):E4096-E4102. https://doi.org/https://doi.org/10.1073/pnas.1321152111
Smits THM (2019) The importance of genome sequence quality to microbial comparative genomics. BMC Genomics 20 (1):662. https://doi.org/10.1186/s12864-019-6014-5
Solla A, Bohnens J, Collin E, Diamandis S, Franke A, Gil L, Burón M, Santini A, Mittempergher L, Pinon J, Broeck AV (2005) Screening european elms for resistance to *Ophiostoma novo-ulmi*. For Sci 51 (2):134-141.
Sone T, Fukiya S, Kodama M, Tomita F (2000) Molecular structure of rDNA repeat unit in *Magnaporthe grisea*. Biosci Biotechnol Biochem 64 (8):1733-1736. https://doi.org/10.1271/bbb.64.1733
Song H, Kim K-T, Park S-Y, Lee G-W, Choi J, Jeon J, Cheong K, Choi G, Hur J-S, Lee Y-H (2022) A comparative genomic analysis of lichen-forming fungi reveals new insights into fungal lifestyles. Sci Rep 12 (1):10724. https://doi.org/10.1038/s41598-022-14340-5
Sordo M, Zeng Q On sample size and classification accuracy: a performance comparison. In, Berlin, Heidelberg, 2005. Biological and Medical Data Analysis. Springer Berlin Heidelberg, pp 193-201
Spanu PD, Abbott JC, Amselem J, Burgis TA et al (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Sciences 330 (6010):1543-1546. https://doi.org/10.1126/science.1194573
Sperschneider J, Dodds PN (2022) EffectorP 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. Mol Plant-Microbe Interact 35 (2):146-156. https://doi.org/10.1094/mpmi-08-21-0201-r
Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24 (5):637-644. https://doi.org/10.1093/bioinformatics/btn013
Stergiopoulos I, Wit PJGMd (2009) Fungal effector proteins. Annu Rev Phytopathol 47 (1):233-263. https://doi.org/10.1146/annurev.phyto.112408.132637
Sugita R, Tanaka K (2022) *Thyridium* revised: Synonymisation of *Phialemoniopsis* under *Thyridium* and establishment of a new order, Thyridiales. MycoKeys 86:147-176. https://10.3897/mycokeys.86.78989
Sun Y, Liu N, Samarakoon MC, Jayawardena RS, Hyde KD, Wang Y (2021) Morphology and phylogeny reveal

Vamsapriyaceae fam. nov. (Xylariales, Sordariomycetes) with two novel *Vamsapriya* species. J Fungi 7 (11):891. https://doi.org/10.3390/jof7110891

Tongcham P, Supa P, Pornwongthong P, Prasitmeeboon P (2020) Mushroom spawn quality classification with machine learning. Comput Electron Agric 179:105865. https://doi.org/10.1016/j.compag.2020.105865

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 (21):10994-11006. https://doi.org/10.1093/nar/gkz841

Tortorano AM, Prigitano A, Esposto MC, Arsic Arsenijevic V et al (2014) European Confederation of Medical Mycology (ECMM) epidemiological survey on invasive infections due to *Fusarium* species in europe. Eur J Clin Microbiol Infect Dis 33 (9):1623-1630. https://doi.org/10.1007/s10096-014-2111-1

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13 (1):36-46. https://doi.org/10.1038/nrg3117

Troy GC, Panciera DL, Pickett JP, Sutton DA, Gene J, Cano JF, Guarro J, Thompson EH, Wickes BL (2013) Mixed infection caused by *Lecythophora canina* sp. nov. and *Plectosphaerella cucumerina* in a German shepherd dog. Med Mycol 51 (5):455-460. https://doi.org/10.3109/13693786.2012.754998

van Kan JAL (2006) Licensed to kill: the lifestyle of a necrotrophic plant pathogen. Trends Plant Sci 11 (5):247-253. https://doi.org/10.1016/j.tplants.2006.03.005

Wang D, Tian L, Zhang DD, Song J, Song SS, Yin CM, Zhou L, Liu Y, Wang B-L, Kong Z-Q, Klosterman SJ, Li J-J, Wang J, Li T-G, Adamu S, Subbarao KV, Chen J-Y, Dai X-F (2020) Functional analyses of small secreted cysteine-rich proteins identified candidate effectors in *Verticillium dahliae*. Mol Plant Pathol 21 (5):667-685. https://doi.org/10.1111/mpp.12921

Wang M, Tan X-M, Liu F, Cai L (2018) Eight new *Arthrinium* species from China. MycoKeys 34. https://doi.org/10.3897/mycokeys.34.24221

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8 (12):973-982. https://doi.org/10.1038/nrg2165

Wijayawardene NN, Hyde KD, Dai DQ, Sanchez-Garcia M, et al (2022) Outline of fungi and fungus-like taxa-2021. Mycosphere 13 (1):53-453. https://doi.org/10.5943/mycosphere/13/1/2

Wingfield MJ, De Beer ZW, Slippers B, Wingfield BD, Groenewald JZ, Lombard L, Crous PW (2012) One fungus, one name promotes progressive plant pathology. Mol Plant Pathol 13 (6):604-613. https://doi.org/10.1111/j.1364-3703.2011.00768.x

Wu W, Chen W, Liu S, Wu J, Zhu Y, Qin L, Zhu B (2021) Beneficial relationships between endophytic bacteria and medicinal plants. Front. Plant Sci 12. https://doi.org/10.3389/fpls.2021.646146

Xu C, Jackson SA (2019) Machine learning and complex biological data. Genome Biol 20 (1):76. https://doi.org/10.1186/s13059-019-1689-0

Xu J, Yang X, Lin Q (2014) Chemistry and biology of *Pestalotiopsis*-derived natural products. Fungal Divers 66 (1):37-68. https://doi.org/10.1007/s13225-014-0288-3

Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, Wu T, Hu E, Jiang Y, Bo X, Yu G (2021) ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol Biol Evol 38 (9):4039-4042. https://doi.org/10.1093/molbev/msab166

Yin C, Ramachandran SR, Zhai Y, Bu C, Pappu HR, Hulbert SH (2019) A novel fungal effector from *Puccinia graminis* suppressing RNA silencing and plant defense responses. New Phytol 222 (3):1561-1572. https://doi.org/10.1111/nph.15676

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y (2017) GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8 (1):28-36. https://doi.org/10.1111/2041-210X.12628

Yu G, Xian L, Xue H, Yu W, Rufian JS, Sang Y, Morcillo RJL, Wang Y, Macho AP (2020) A bacterial effector protein prevents MAPK-mediated phosphorylation of SGT1 to suppress plant immunity. PLoS Pathog 16 (9):e1008933. https://doi.org/10.1371/journal.ppat.1008933

Zeng T, Holmer R, Hontelez J, te Lintel-Hekkert B, Marufu L, de Zeeuw T, Wu F, Schijlen E, Bisseling T, Limpens E (2018) Host- and stage-dependent secretome of the arbuscular mycorrhizal fungus *Rhizophagus irregularis*. Plant J 94 (3):411-425. https://doi.org/10.1111/tpj.13908

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 (W1):W95-W101. https://doi.org/10.1093/nar/gky418

Zhang Y, Yang H, Turra D, Zhou S, Ayhan DH, DeIulio GA, Guo L, Broz K, Wiederhold N, Coleman JJ, Donnell KO, Youngster I, McAdam AJ, Savinov S, Shea T, Young S, Zeng Q, Rep M, Pearlman E, Schwartz DC, Di Pietro A, Kistler HC, Ma L-J (2020) The genome of opportunistic fungal pathogen *Fusarium oxysporum* carries a unique set of lineage-specific chromosomes. Commun Biol 3 (1):50. https://doi.org/10.1038/s42003-020-0770-2

Zhao Z, Liu H, Wang C, Xu JR (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14 (1):274. https://doi.org/10.1186/1471-2164-14-274

Zieliński B, Sroka-Oleksiak A, Rymarczyk D, Piekarczyk A, Brzychczy-Włoch M (2020) Deep learning approach to describe and classify fungi microscopic images. PLoS One 15 (6):e0234806. https://doi.org/10.1371/journal.pone.0234806