New Phytologist

Research

Ecological generalism drives hyperdiversity of secondary metabolite gene clusters in xylarialean endophytes

Mario E. E. Franco[1], Jennifer H. Wisecaver[2], A. Elizabeth Arnold[3], Yu-Ming Ju[4], Jason C. Slot[5], Steven Ahrendt[6], Lillian P. Moore[1], Katharine E. Eastman[2], Kelsey Scott[5], Zachary Konkel[5], Stephen J. Mondo[6], Alan Kuo[6], Richard D. Hayes[6], Sajeet Haridas[6], Bill Andreopoulos[6], Robert Riley[6], Kurt LaButti[6], Jasmyn Pangilinan[6], Anna Lipzen[6], Mojgan Amirebrahimi[6], Juying Yan[6], Catherine Adam[6], Keykhosrow Keymanesh[6], Vivian Ng[6], Katherine Louie[6], Trent Northen[6], Elodie Drula[7,8], Bernard Henrissat[9,10], Huei-Mei Hsieh[4], Ken Youens-Clark[1], François Lutzoni[11], Jolanta Miadlikowska[11], Daniel C. Eastwood[12], Richard C. Hamelin[13], Igor V. Grigoriev[6,14] and Jana M. U'Ren[1]

[1]BIO5 Institute and Department of Biosystems Engineering, The University of Arizona, Tucson, AZ 85721, USA; [2]Center for Plant Biology and Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA; [3]School of Plant Sciences and Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ 85721, USA; [4]Institute of Plant and Microbial Biology, Academia Sinica, Taipei 11529, Taiwan; [5]Department of Plant Pathology, The Ohio State University, Columbus, OH 43210, USA; [6]Department of Energy, The Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [7]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille 13288, France; [8]INRAE, Marseille 13288, France; [9]Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby DK-2800, Denmark; [10]Department of Biological Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia; [11]Department of Biology, Duke University, Durham, NC 27708, USA; [12]Department of Biosciences, Swansea University, Swansea, SA2 8PP, UK; [13]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [14]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

Author for correspondence:
Jana M. U'Ren
Email: juren@email.arizona.edu

Received: *16 August 2021*
Accepted: *7 November 2021*

New Phytologist (2022) **233:** 1317–1330
doi: 10.1111/nph.17873

Key words: Ascomycota, endophyte, plant–fungal interactions, saprotroph, specialised metabolism, symbiosis, trophic mode, Xylariales.

Summary

- Although secondary metabolites are typically associated with competitive or pathogenic interactions, the high bioactivity of endophytic fungi in the Xylariales, coupled with their abundance and broad host ranges spanning all lineages of land plants and lichens, suggests that enhanced secondary metabolism might facilitate symbioses with phylogenetically diverse hosts.
- Here, we examined secondary metabolite gene clusters (SMGCs) across 96 Xylariales genomes in two clades (Xylariaceae s.l. and Hypoxylaceae), including 88 newly sequenced genomes of endophytes and closely related saprotrophs and pathogens. We paired genomic data with extensive metadata on endophyte hosts and substrates, enabling us to examine genomic factors related to the breadth of symbiotic interactions and ecological roles.
- All genomes contain hyperabundant SMGCs; however, Xylariaceae have increased numbers of gene duplications, horizontal gene transfers (HGTs) and SMGCs. Enhanced metabolic diversity of endophytes is associated with a greater diversity of hosts and increased capacity for lignocellulose decomposition.
- Our results suggest that, as host and substrate generalists, Xylariaceae endophytes experience greater selection to diversify SMGCs compared with more ecologically specialised Hypoxylaceae species. Overall, our results provide new evidence that SMGCs may facilitate symbiosis with phylogenetically diverse hosts, highlighting the importance of microbial symbioses to drive fungal metabolic diversity.

Introduction

Fungal endophytes inhabit asymptomatic, living photosynthetic tissues of all major lineages of plants and lichens to form one of Earth's most prevalent groups of symbionts (Arnold *et al.*, 2009; Peay *et al.*, 2016). Known from a wide range of biomes and agroecosystems (e.g. U'Ren *et al.*, 2012, 2019), endophytes impact plant health, productivity and evolution (Rodriguez *et al.*, 2009). Although classified together due to ecologically similar patterns of colonisation, transmission and *in planta* biodiversity (Rodriguez *et al.*, 2009), foliar fungal endophytes represent a diversity of evolutionary histories, life history strategies and functional traits (Porras-Alfaro & Bayman, 2011). Despite the recent surge of interest in

plant microbiome research (Trivedi *et al.*, 2020) the genomic and molecular mechanisms foliar fungal endophytes employ to establish symbiotic host associations remain largely unknown.

Global, large-scale surveys of phylogenetically diverse plant and lichen hosts have revealed that many foliar endophyte species preferentially associate with particular host species and lineages, resulting in host structured endophyte communities at local to global scales (e.g. U'Ren *et al.*, 2019). However, endophytic fungi in the Xylariales (Sordariomycetes, Pezizomycotina, Ascomycota) appear unique in that they frequently have broad host ranges that span multiple lineages of land plants (e.g. angiosperms, conifers, lycophytes, ferns and mosses) as well as green algae and cyanobacteria within lichen thalli (e.g. Arnold *et al.*, 2009; U'Ren *et al.*, 2016). By contrast, described Xylariales species associate primarily with angiosperms as wood- or litter-degrading saprotrophs or woody pathogens (Hsieh *et al.*, 2005, 2010). Although the genetic factors that determine foliar endophyte host range are currently unknown, research on fungal pathogens has shown that host specificity is often determined by the presence of avirulence proteins (i.e. effectors), proteinaceous host-specific toxins and secondary metabolites (SMs) (Li *et al.*, 2020). Horizontal gene transfer (HGT) of these host-determining genes frequently alters and/or expands pathogen host range (Li *et al.*, 2020).

Xylariales genomes sequenced to date have revealed a rich repertoire of secondary metabolite gene clusters (SMGCs) (Wibberg *et al.*, 2021), often exceeding the numbers reported for saprotrophic fungi well known for their SM production (*Aspergillus*, *Penicillium*) (Nielsen *et al.*, 2017; Drott *et al.*, 2021). Previously, it was postulated that intense competition with diverse communities of soil organisms increases selection to maintain and diversify SMGCs (Slot, 2017). However, the high bioactivity of Xylariales fungi (> 500 SMs reported to date; Becker & Stadler, 2021), their broad host ranges as endophytes and their ability to persist in leaf litter as saprotrophs that decompose lignocellulose (U'Ren & Arnold, 2016; U'Ren *et al.*, 2016), led us to hypothesise that enhanced secondary metabolism might play a role in facilitating ecological generalism in both substrate use and the phylogenetic breadth of their symbiotic associations with plants and lichens.

To test this hypothesis, we examined the genomic factors associated with endophyte host range and ecological roles (i.e. endophytic, pathogenic and saprotrophic) across 96 genomes of Xylariales, including 88 newly sequenced genomes of endophytes, saprotrophs and plant pathogens within two major clades of Xylariales [Hypoxylaceae and Xylariaceae *sensu lato* (s.l.)]. We paired genomic data with extensive metadata on endophyte host associations, geographic distributions and substrate usage gleaned from a collection of > 6000 xylarialean endophytes isolated from phylogenetically diverse plants and lichens across North America (U'Ren *et al.*, 2016), enabling us to examine for the first time the genomic factors related to the breadth of symbiotic interactions and ecological roles in this dynamic and ecologically important fungal clade.

Materials and Methods

Fungal strain selection

We sequenced genomes of 44 endophytic taxa (U'Ren *et al.*, 2012; U'Ren & Arnold, 2016) and 44 named taxa of Xylariaceae s.l. and Hypoxylaceae representing *c.* 24 genera and 80 species, as well as two undescribed species of endophytic Xylariales (*Pestalotiopsis* sp. NC0098 and Xylariales sp. AK1849) included in the outgroup (Supporting Information Table S1). Isolates were selected based on their phylogenetic position and ecological mode from U'Ren *et al.* (2016). Although classifying fungal ecological modes broadly as 'endophytic' or 'saprotrophic' based on the condition of the tissue from which they are cultured is often insufficient to adequately define their ecological roles, for the purposes of this study, isolates cultured from living host tissues (either plant or lichen) are referred to as endophytes even if other isolates in the same fungal operational taxonomic unit (OTU) were found in nonliving tissues as well. Isolates were defined as saprotrophs only if all isolates in the OTU were cultured from nonliving plant tissues such as senescent leaves or leaf litter (U'Ren *et al.*, 2016). To minimise the effect of phylogeny when assessing the impact of ecological mode on genome evolution, we also selected 15 pairs of closely related sister taxa with contrasting ecological modes (i.e. endophyte vs nonendophyte) (U'Ren *et al.*, 2016). For reference species that lacked host and substrate metadata, ecological modes were estimated based on information for that species in the literature as described in U'Ren *et al.* (2016).

DNA and RNA purification

We used two different mycelial growth and cultivation techniques to obtain DNA for either Illumina or PacBio Single-Molecule Real-Time (SMRT) sequencing. DNA isolations were performed using modified phenol/chloroform extractions (Methods S1). RNA was extracted for each isolate with the Ambion Purelink RNA Kit (Thermo Fisher Scientific, Waltham, MA, USA). DNA and RNA were quantified with a Qubit fluorometer (Invitrogen) and sample purity was assessed using NanoDrop spectrophotometer (BioNordika, Herlev, Denmark). RNA was treated with DNase (Thermo Fisher Scientific) following the manufacturer's instructions and RNA integrity was assessed on a BioAnalyser at the University of Arizona Genomics Core Facility.

Genome and transcriptome sequencing and assembly

Genomes were generated at the Department of Energy Joint Genome Institute using Illumina and PacBio technologies (Table S1). For 66 isolates, Illumina standard shotgun libraries (insert sizes of 300 bp or 600 bp) were constructed and sequenced using the NovaSeq platform. Raw reads were filtered using the JGI QC pipeline. An assembly of the target genome was generated using the resulting nonorganelle reads with SPAdes (Bankevich *et al.*, 2012). PacBio SMRT sequencing was

performed for 22 isolates of Xylariaceae s.l. and Hypoxylaceae, as well as Xylariales spp. NC0098 and AK1849 on a PacBio SEQUEL. Library preparation was performed using either the PacBio low input 10 kb or PacBio > 10 kb with AMPure bead size selection. Filtered subread data were processed with the JGI QC pipeline and *de novo* assembled using Falcon (SEQUEL) or Flye (SEQUEL II) systems. Stranded RNA-seq libraries were created and quantified using qPCR and transcriptome sequencing was performed on an Illumina NovaSeq S4 instrument. For both Hypoxylaceae and Xylariaceae s.l., *c.* 25% of genomes were sequenced with PacBio, although a higher proportion of endophyte genomes were sequenced with PacBio rather than Illumina (43% vs 28% overall). Genome completeness was assessed by benchmarking universal single-copy orthologues (BUSCO) v.2.0 using the 'eukaryota_odb9' (2016-11-02) dataset (10.1093/bioinformatics/btv351).

Genome annotation

Gene prediction and annotation was performed using the JGI pipeline (Grigoriev *et al.*, 2014; Kuo *et al.*, 2014) (Methods S1). Predicted genes were annotated using functional information from InterPro (Mitchell *et al.*, 2019), PFAM (El-Gebali *et al.*, 2019), gene ontology (GO) (The Gene Ontology Consortium, 2019), kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa *et al.*, 2006), eukaryotic orthologous groups of proteins (KOG) (Tatusov *et al.*, 2003), the carbohydrate-active enzymes database (CAZy) (Lombard *et al.*, 2014), MEROPS database (Rawlings *et al.*, 2016), the transporter classification database (TCDB) (Saier *et al.*, 2016), SIGNALP v.3.0a (Nielsen, 2017) and EFFECTORP 2.0 (Sperschneider *et al.*, 2018). CAZymes involved in the degradation of the plant cell wall were classified by substrate (Kameshwar *et al.*, 2019). We examined repetitive elements using REPEATSCOUT (Price *et al.*, 2005), which identifies novel repeats in the genomes, and REPEATMASKER (http://repeatmasker.org), which identifies known repeats based on the Repbase library (Bao *et al.*, 2015).

Orthogroup prediction, functional annotation and ancestral state reconstruction

For comparative analyses, data from an additional eight taxa in Xylariaceae s.l. (Wu *et al.*, 2017) and 23 additional genomes of Sordariomycetes were obtained from MycoCosm (Grigoriev *et al.*, 2014) (Table S1). Orthologous gene families (i.e. orthogroups) for all 121 genomes (ingroup and outgroup) were inferred by ORTHOFINDER v.2.3.3 (Emms & Kelly, 2019), which was executed using DIAMOND v.0.9.22 (Buchfink *et al.*, 2015) for the all-versus-all sequence similarity search and MAFFT v.7.427 (Katoh & Standley, 2013) for sequence alignment. Orthogroups were assigned functional annotations with KINFIN v.1.0 (Laetsch & Blaxter, 2017), which performs a representative functional annotation of the orthogroups based on both the proportion of proteins in the group carrying a specific annotation, as well as the proportion of taxa in the cluster with such annotation. KINFIN was also used to perform network analysis of orthogroups, classify orthogroups and SMGCs into isolate-specific, clade-specific (Hypoxylaceae and Xylariaceae s.l.) and universal (i.e. orthogroups present in all taxa) categories, and to identify orthogroups that were significantly enriched or depleted in the Xylariaceae s.l. or Hypoxylaceae using the Mann–Whitney *U*-test. We used COUNT v.10.04 (Csurös, 2010) with the unweighted Wagner parsimony method (gain and loss penalties both set to 1) to assess changes in the size of orthologous gene families over evolutionary time. Orthogroup annotations were also used to reconstruct the ancestral gene content for subsets of orthologous gene families corresponding to different functional categories.

Phylogenomic analysis

Protein sequences of 1526 single-copy orthogroups defined by ORTHOFINDER were aligned using MAFFT v.7.427 (Katoh & Standley, 2013), concatenated and analysed using maximum likelihood in IQ-TREE multicore v.1.6.11 (Nguyen *et al.*, 2015) with the Le Gascuel (LG) substitution model. Node support was calculated with 1000 ultrafast bootstrap replicates. Additional phylogenomic analyses with different models of evolution, gene sets and outgroup taxa resulted in nearly identical topologies (Methods S1).

Metabolic gene cluster prediction

Secondary metabolite gene clusters were predicted using ANTISMASH v.5.1.0 (Blin *et al.*, 2019) setting the strictness to 'relaxed' and enabling 'KnownClusterBlast', 'ClusterBlast', 'SubClusterBlast', 'ActiveSiteFinder', 'Cluster Pfam analysis' and 'Pfam-based GO term annotation'. CLINKER and CLUSTERMAP.JS were used to visualise and compare SMGCs (Gilchrist & Chooi, 2021). Sequence similarity network analysis of the SMGCs was performed using BIG-SCAPE v.1.0.1 (Navarro-Muñoz *et al.*, 2020). BIG-SCAPE was executed under the hybrid mode, enabling the inclusion of singletons and the SMGCs from the MIBiG repository v.1.4 (Medema *et al.*, 2015). The output from BIG-SCAPE was incorporated into KINFIN (Laetsch & Blaxter, 2017) to visualise gene content similarity as network graphs and examine SMGC distribution across clades.

We used a custom pipeline (https://github.com/egluckthaler/cluster_retrieve) to examine fungal metabolic gene clusters involved in the degradation of a broad array of plant phenylpropanoids (Gluck-Thaler *et al.*, 2018) (from this point forwards, catabolic gene clusters: CGCs). Cluster_retrieve searches for multiple 'cluster models' containing one of 13 anchor genes (Gluck-Thaler *et al.*, 2018). Homologous genes in each locus were defined by a minimum BLASTP (v.2.2.25+) bitscore of 50, 30% amino acid identity, and target sequence alignment 50–150% of the query sequence length. Homologues of query genes were considered clustered if separated by < 7 intervening genes. However, CGCs often share many gene families among classes, resulting in overlapping and adjacent clusters detected by different cluster profile searches. As the majority of CGCs have not been functionally characterised, rather than splitting loci by functional

annotation alone, we empirically assessed the spatial distribution of genes in 25 contigs that contained multiple consolidated cluster predictions. Based on these results, we selected a gap size of 30 kb to define discrete clusters (i.e. clusters on the same contig were consolidated if separated by < 30 kb). Homologous cluster families across genomes were inferred using a modified version of BIG-SCAPE (Navarro-Muñoz *et al.*, 2020) (i.e. adding catabolic anchor genes to 'anchor_domains.txt' and manually tuning the 'Others' cluster type model parameters until known related clusters, such as quinate dehydrogenase clusters, merged into families). Tuning resulted in the values 0.35 for the Jaccard dissimilarity of cluster Pfams, 0.63 for Pfam sequence similarity, 0.02 adjacency index and 2.0 anchor boost.

Detection of HGT events

We used the ALIEN INDEX (AI) pipeline (https://github.itap.purdue.edu/jwisecav/phylowise) (Wisecaver *et al.*, 2016; Verster *et al.*, 2019) to identify HGT candidate genes. Each predicted protein sequence was queried against a custom protein database using DIAMOND v.0.9.22.123 (Buchfink *et al.*, 2015). The custom database consisted of protein sequences from NCBI RefSeq (release 98) (O'Leary *et al.*, 2016), the marine microbial eukaryotic transcriptome sequencing project (MMETSP) (Keeling *et al.*, 2014) and the 1000 plants transcriptome sequencing project (OneKP) (Matasci *et al.*, 2014). DIAMOND results were sorted based on the normalised bitscore (*nbs*), where *nbs* was calculated as the bitscore of the single best high scoring segment pair (HSP) in the hit sequence divided by the best bitscore possible for the query sequence (i.e. the bitscore of the query aligned to itself).

To identify HGT candidates, an ancestral lineage is first specified and the AI score calculated using the formula: AI = *nbsO* – *nbsA*, where *nbsO* is the normalised bit score of the best hit to a species outside of the ancestral lineage and *nbsA* is the normalised bit score of the best hit to a species within the ancestral lineage. AI scores range from −1 to 1, being greater than zero if the predicted protein sequence had a better hit to species outside of the ancestral lineage and can be suggestive of either HGT or contamination (Wisecaver *et al.*, 2016). To identify HGTs present in multiple species, a recipient sublineage within the larger ancestral lineage may also be specified to identify their shared HGT candidates (Fig. S1). All hits to the recipient lineage are skipped so as not to be included in the *nbsA* calculation. To identify candidate HGTs acquired from distant gene donors (e.g. viruses, bacteria, or plants) we performed a first AI screen using Ascomycota (NCBI: txid4890) and Xylariomycetidae (NCBI: txid 222545) as the ancestral and recipient lineages, respectively (Fig. S1). To identify candidate horizontal transfers of genes predicted by antiSMASH to be in a SMGC from more closely related donors (e.g. other filamentous fungi), we ran the AI pipeline for a second time using Xylariales (NCBI: txid 37989) as the ancestral lineage and manually curated subclades (see Table S1) as recipient lineages (see Fig. S1). Genes from both the first (i.e. all genes, distant donors) and second (i.e. SMGC genes, closely related donors) were considered putative HGT candidates if they passed the following filters: (1) AI score of > 0; (2) significant hits to at least 25 sequences in the custom database; and (3) at least 50% of top hits to sequences outside of the ancestral lineage.

Candidates from the first AI screen were further validated using phylogenetic analyses (described below) and designated as either high or low confidence HGT. Full-length proteins corresponding to the top < 200 hits (*E*-value < 1×10^{-3}) to each AI screen 1 candidate were extracted from the custom database using esl-sfetch (Eddy, 2009). As our initial query-based trees often lacked sufficient taxon sampling to assess HGT, we combined all orthogroup sequences with all extracted top hits to each AI candidate. Sequences were aligned using MAFFT v.7.407 using --auto (Katoh & Standley, 2013) and the number of well aligned columns was determined with TRIMAL v.1.4. rev15 using its gappyout strategy (Capella-Gutiérrez *et al.*, 2009). Only alignments with ≥ 50 retained columns after TRIMAL were retained for phylogenetic analysis. Phylogenetic trees were constructed with IQ-TREE v.1.6.10 (Nguyen *et al.*, 2015) in a single run with MODELFINDER (Kalyaanamoorthy *et al.*, 2017) and SH-ALRT combined with ultrafast bootstrapping analyses (1000 replicates each). Phylogenies were visualised using iTOL v.4 (Letunic & Bork, 2019). Each phylogenetic tree was manually curated to verify HGT with either high or low confidence. High-confidence HGT events had to meet the following criteria: (1) the association between donor and recipient clades was supported by ultrafast bootstrap ≥ 95; (2) recipient clade consisted of sequences from two or more species. If the candidate met one of the two criteria, HGT was considered lower confidence.

Statistical analyses

To assess whether genes within different functional categories are associated with endophytic ecological mode we performed phylogenetically independent contrasts (PICs) (Felsenstein, 1985) with the function 'brunch' of the package CAPER v.1.0.1 (Orme *et al.*, 2012) in R v.3.6.1. All other statistical analyses were done in R v.3.6.1 or JMP v.15.1 (SAS Institute Inc., Cary, NC, USA).

Results and Discussion

Genomes of 96 Xylariales taxa correspond to the previously recognised family Xylariaceae (Ju & Rogers, 1996) that was recently split into multiple families (Hypoxylaceae, Xylariaceae, Graphostromataceae, Barrmaeliaceae) (Voglmayr *et al.*, 2018; Wendt *et al.*, 2018) (Figs 1a, S2). Here, we use the term xylarialean to refer to this monophyletic clade within the Xylariales. In addition, because our analyses revealed seven undescribed endophytic isolates in five distinct clades (i.e. clades E2, E4, E5, E6 and E6; Fig. S2) nested between the Graphostromataceae and Xylariaceae *sensu stricto* (s.s), we refer to the sister clade to Hypoxylaceae as Xylariaceae s.l. (from this point forwards, Xylariaceae) following Voglmayr *et al.* (2018) (Fig. 1).

Genome sequencing yielded eukaryotic BUSCO values ≥ 95% (Table S1). Xylarialean genomes ranged in size from 33.7–60.3 Mbp (average 43.5 Mbp; Fig. S3; Table S1) and contained *c.* 8000–15 000 predicted genes (mean 11 871; Fig. S3), congruent with average genome and proteome sizes of other Pezizomycotina

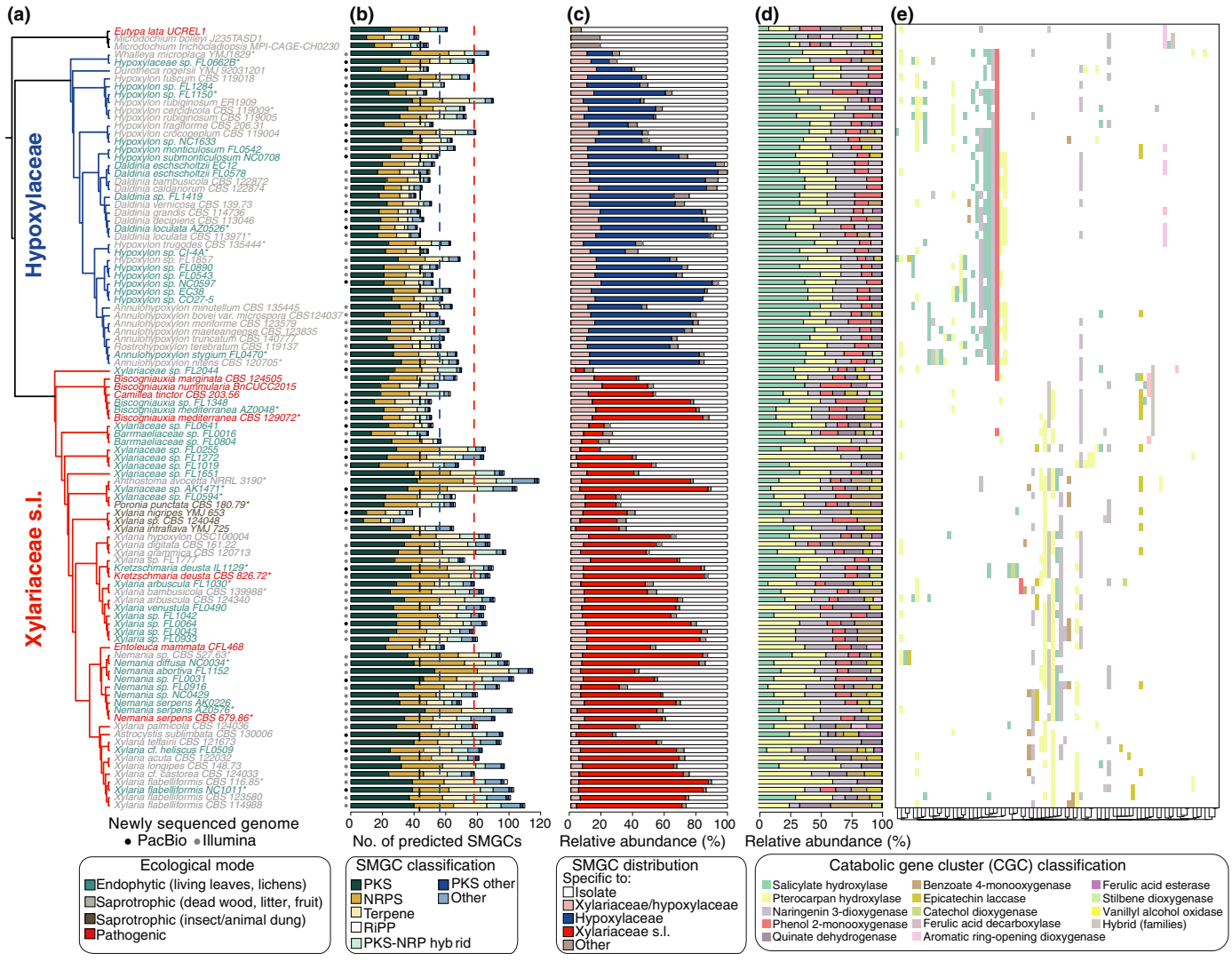

Fig. 1 Xylariaceae s.l. and Hypoxylaceae genomes are characterised by hyperdiverse and dynamic metabolic gene clusters. (a) Maximum likelihood phylogenetic analyses of 1526 universal, single-copy orthogroups support the sister relationship of the Xylariaceae s.l. (containing Xylariaceae *sensu stricto*, Graphostromataceae and Barrmaeliaceae) and the Hypoxylaceae, as well as previously denoted relationships among genera (see Supporting Information Fig. S2). Phylogenetic analyses included genomes of 25 outgroup taxa representing five other families of Xylariales and eight orders of Sordariomycetes (total 121 genomes; Fig. S2). Taxon names are coloured by ecological mode and branches coloured by major clade (red: Xylariaceae s.l.; blue: Hypoxylaceae). Taxa with asterisks (*) represent 15 pairs of endophyte/nonendophyte sister taxa used to assess differences in genomic content due to ecological mode (see Fig. 4). Within this phylogenetic framework, we compared the: (b) abundance of different secondary metabolite gene cluster (SMGC) families per genome. Dotted lines indicate the averages for Pezizomycotina (black), Xylariaceae s.l. (red) and Hypoxylaceae (blue); (c) relative abundance of family-specific, clade-specific and isolate-specific SMGCs; (d) relative abundance and (e) presence/absence of catabolic gene clusters (CGCs), coloured by anchor gene identity (Gluck-Thaler & Slot, 2018). Hierarchical clustering of CGCs (see bottom) was performed with the unweighted pair group method with arithmetic mean.

(Shen *et al.*, 2020). The percentage of repetitive elements per genome ranged from < 1–24% (average 1.6%; Table S2), but unlike mycorrhizal fungi (Miyauchi *et al.*, 2020), repeat content was not corrected with ecological mode (Fig. S3).

Xylariaceae and Hypoxylaceae genomes contain hyperdiverse metabolic gene clusters

To investigate the diversity and composition of metabolic gene clusters in xylarialean genomes, we used antiSMASH (Blin *et al.*, 2019) to mine genomes for SMGCs, as well as a custom pipeline to examine catabolic gene clusters (CGCs) involved in fungal degradation of a broad array of plant phenylpropanoids (Gluck-Thaler *et al.*, 2018). Across 96 xylarialean genomes we predicted a total of 6879 putative SMGCs (belonging to 3313 cluster families) and 973 putative CGCs (belonging to 190 cluster families) (Tables S3, S4). In comparison, recent large-scale analyses predicted 3399 SMGCs (in 719 cluster families) across 101 Dothideomycetes genomes (Gluck-Thaler *et al.*, 2020) and 1110 CGCs across 341 fungal genomes (Gluck-Thaler & Slot, 2018). Only 25% of predicted SMGCs ($n = 1711$ belonging to 816 cluster families)

had BLAST hits to 168 unique MIBiG (Medema *et al.*, 2015) accession numbers (Table S3).

Total SMGCs diversity in the Xylariaceae and Hypoxylaceae is reflected in a high number of SMGCs per genome: the average number of SMGCs per genome was 71.2 (median 68), which is significantly higher than the average for other fungi in the Pezizomycotina (mean 42.8; Fig. 1b). At least eight xylarialean genomes contained more than 100 predicted SMGCs, with a maximum of 119 in *Anthostoma avocetta* NRRL 3190 (Fig. 1b; Table S3). In comparison, a recent study of 24 species of *Penicillium* found an average of 54.9 SMGCs per genome, with a maximum number of 78 SMGCs observed in *P. polonicum* (Nielsen *et al.*, 2017). Genomes of Xylariaceae and Hypoxylaceae contained on average 3.3× more CGCs per genome (average 10.1; Table S4) compared with other genomes of Pezizomycotina (average 3.0 (Gluck-Thaler *et al.*, 2018)).

Every xylarialean genome contained SMGCs for the production of polyketides (PK; 2871 total), nonribosomal peptides (NRP; 2482 total) and terpenes (1322 total; Fig. 1b; Table S3). SMGCs for ribosomally synthesised and post-translationally modified peptides (RiPPs) and hybrid NRP-PK compounds occurred less frequently (Fig. 1b). The most widely distributed and abundant CGCs were pterocarpan hydroxylases ($n = 93$), putatively involved in isoflavonoid metabolism (Fig. 1d,e; Table S5). CGCs involved in the breakdown of plant salicylic acid (Ambrose *et al.*, 2015) ($n = 251$ salicylate hydroxylases) and plant flavonoids ($n = 170$ naringenin 3-dioxygenases) were also abundant (Fig. 1d,e). CGCs classified into nine other categories (e.g. phenol 2-monooxygenase, quinate dehydrogenase) (Gluck-Thaler *et al.*, 2018) occurred more rarely (Table S4). Vanillyl alcohol oxidases, which were previously shown to be enriched in genomes of soil saprotrophs (Gluck-Thaler *et al.*, 2018), were absent in xylarialean genomes.

Consistent with the hyperdiversity of SMGCs in the Hypoxylaceae and Xylariaceae, we observed that only *c.* 10% of SMGCs were shared among genomes from both Xylariaceae and Hypoxylaceae (Fig. 1c), and no SMGCs were universally present in both clades (Table S3). On average, 21.4% and 28.2% of SMGCs per genome were unique to either a taxon in the Hypoxylaceae or the Xylariaceae, respectively (range 0–82%; Fig. 1c; Table S4), but no SMGCs were universally present within either clade. For most isolates, the majority of SMGCs were unique (i.e. 'isolate specific'; Fig. 1c). Isolate-specific SMGCs represented an average of 36.6% (SD ± 21.1) of the clusters per genome (range 0–85.7%; Fig. 1c). Even when multiple isolates of the same species were compared (e.g. *Nemania serpens* clade) 30–41% of the SMGCs appeared specific to a single isolate (Fig. 1b; Table S3), similar to intraspecific SMGC variation in *Aspergillus flavus* (Drott *et al.*, 2021).

Impact of HGT on xylarialean genome evolution

To assess the role of HGT in shaping the genome evolution of Xylariaceae and Hypoxylaceae we performed two AI analyses (Alexander *et al.*, 2016; Wisecaver *et al.*, 2016; Gonçalves *et al.*, 2018). The first AI screen – designed to detect candidate HGTs from more distantly related donor lineages (e.g. bacteria, plants) – flagged 4262 genes representing 647 orthogroups (Table S5). Using a custom phylogenetic pipeline (see the Materials and Methods section) we manually validated 168 of these genes as likely to be HGT events to Xylariaceae and Hypoxylaceae. Based on branch support and the presence of multiple xylarialean taxa in the recipient clade, we deemed 92 of these genes as high-confidence HGTs and the remaining 76 as lower confidence HGTs (Fig. 2; Table S5). Similar to previous studies (Marcet-Houben & Gabaldón, 2010; Lawrence *et al.*, 2011), the majority of high-confidence HGTs were predicted to have been acquired from bacteria ($n = 86$) (Fig. 2). Overall, 66% of genes identified as HGT from bacterial donors did not contain introns (compared with 6% of genes across 121 genomes). Other donor lineages include viruses ($n = 3$), Basidiomycota ($n = 2$) and plants ($n = 1$) (Fig. 2; Table S5). On average, xylarialean genomes had 16.2 high-confidence HGT events per genome (range: 7–30; Table S5). The highest number of high-confidence HGT events per genome occurred in the genome of *Xylaria flabelliformis* CBS 123 580 ($n = 30$).

Horizontal gene transfer candidate genes were typically distributed across taxa in numerous diverse clades ($n = 85$ of 92 genes) rather than in monophyletic clades (Fig. 2). For example, an enoyl-acyl carrier protein reductase protein (EC 1.3.1.9) – a key enzyme of the type II fatty acid synthesis (FAS) system (Massengo-Tiassé & Cronan, 2009) – occurred in bacteria (putative donor) and four distantly related recipient taxa: Xylariales sp. PMI 506, *Hypoxylon rubiginosum* ER1909; *H. cercidicola* CBS 119 009; *H. fuscum* CBS 119 018 (HGT0001; Table S5). Multiple evolutionary scenarios could result in patchy taxonomic distributions. For example, multiple fungi could have independently acquired the same gene from closely related bacterial donors (Marcet-Houben & Gabaldón, 2010). Alternatively, an initial HGT from bacteria to fungi may have been followed by fungal–fungal HGTs. In total, 38 HGT candidate genes occurred in genomes of both Sordariomycetes outgroup and Xylariales genomes, 28 were found in only Xylariales genomes, and 26 were only observed in genomes of Xylariaceae and Hypoxylaceae (Fig. 2; Table S5).

Functional annotation revealed that most candidate HGT genes were associated with at least one type of annotation (i.e. 95% of the highly confident and 82% of the ambiguous events; Table S5). Six high-confidence HGT candidate genes were annotated as CAZymes, including three predicted plant cell wall-degrading enzymes (PCWDEs) transferred from bacteria to diverse Xylariales (Fig. 2). No genes predicted in CGCs were identified as candidate HGTs, consistent with convergent evolution to result in similar clustering of fungal phenolic metabolism genes (Gluck-Thaler *et al.*, 2018). However, 43% of candidate HGT genes were predicted to be part of an SMGC (i.e. 40 of 92) (Fig. 2; Tables S3, S5). These include 13 genes predicted to have a biosynthetic function, such as a putative FsC-acetyl coenzyme A–N^2-transacetylase (HGT076; Table S5), which is part of the siderophore biosynthetic pathway in *Aspergillus* implicated in fungal virulence (Blatzer *et al.*, 2011).

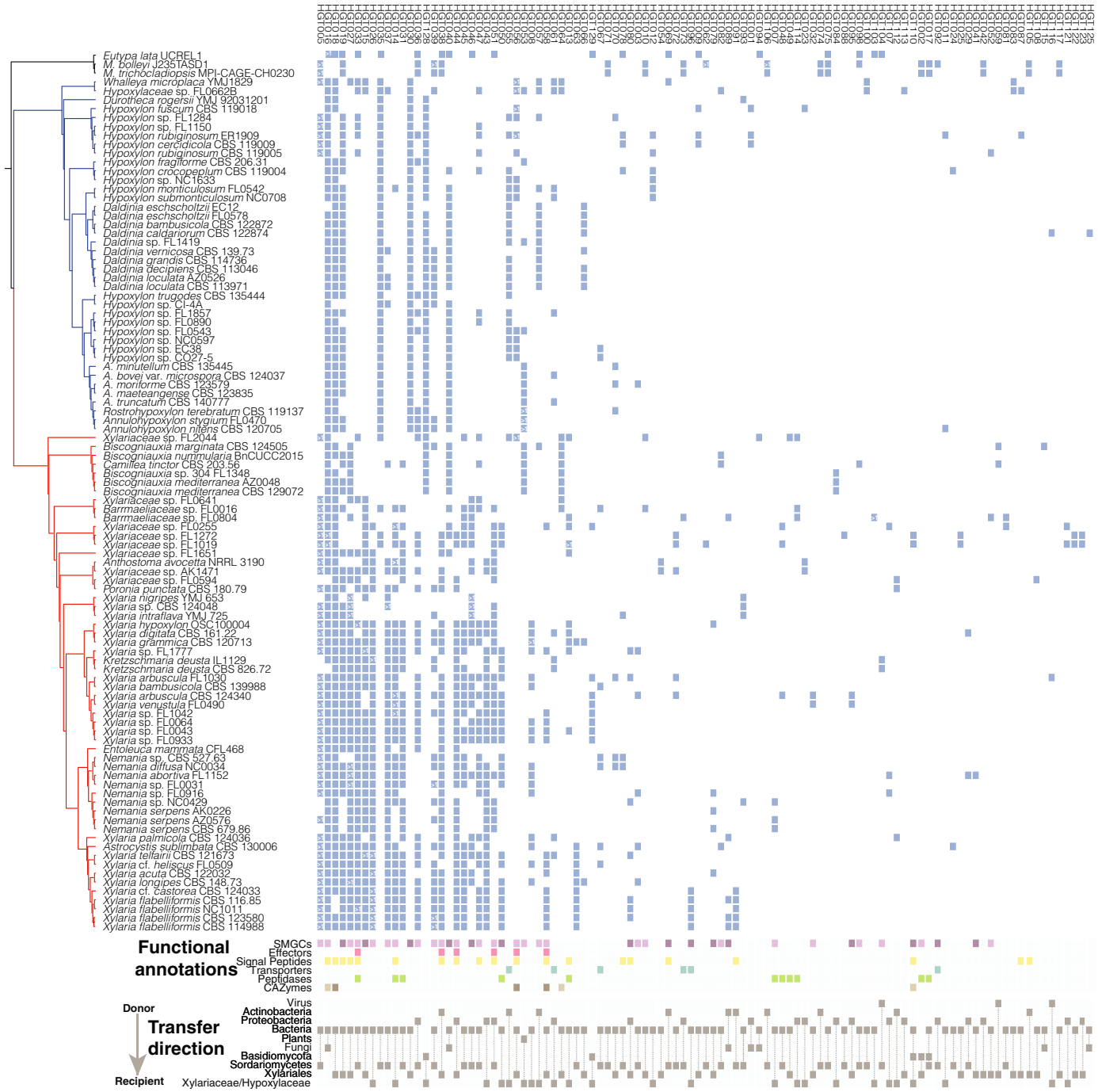

Fig. 2 Phylogenetic distribution and functional annotation of high-confidence horizontal gene transfers (HGTs) to genomes of Xylariaceae s.l. and Hypoxylaceae. Phylogeny matches Fig. 1(a). Blue boxes represent genes predicted to be high-confidence HGT events (detected with the first round of Alien Index analyses; Supporting Information Table S5). HGT events are ordered from left to right based on their abundance. Transfers with more than one gene copy per genome are indicated with > 1. Coloured boxes indicate putative functional annotations of HGTs: secondary metabolite gene clusters (SMGCs), effectors, signalling peptides, transporters, peptidases and Carbohydrate-Active enZYmes (CAZymes). SMGCs predicted as 'biosynthetic core' and 'biosynthetic additional' are shown with darker purple, whereas other genes in SMGCs are shown with light purple. For CAZyme predictions, a dark brown colour indicates plant cell wall-degrading enzymes. The bottom panel (transfer direction) indicates the taxonomic identity of putative donor and recipient lineage(s) inferred from phylogenetic analyses.

Due to the high prevalence of HGT among genes predicted to be part of SMGCs, we performed a second AI screen to detect intrafungal HGT events of genes within the boundaries of SMGCs ($n = 93\,066$ genes) (see the Materials and Methods section; Fig. S1). The second AI screen identified 1148 genes in 660 SMGCs (belonging to 594 cluster families) that were putatively transferred from other fungi to members of the Xylariales (Table S5). Candidate HGT genes were primarily for polyketide

and nonribosomal peptide production (518 PK, 270 NRP and 180 PK–NRP hybrid clusters). In addition, > 75% of hits to MIBiG contained genes identified by AI analyses as putative HGTs (see Fig. S4, bottom). SMGCs with HGT candidate genes included those with 100% similarity to MIBiG accessions from *Aspergillus*, *Fusarium* and *Parastagonospora* involved in mycotoxin (e.g. cyclopiazonic acid, alternariol, fusarin) and antimicrobial compound (asperlactone, koraiol) production, and clusters from *Alternaria* that produce host-selective toxins (e.g. ACT-Toxin II) (Tables S3, S5). Although the second AI analysis did not flag every gene in these clusters as potential HGTs (e.g. only four of the 19 genes in the alternariol cluster from *Hypoxylon cercidicola* CBS 119 009 were HGT candidates based on AI; Table S5) and we were not able to further validate candidates based on the same criteria used for high-confidence HGT, the phylogenetic distribution of many of these SMGCs across Xylariales is consistent with the acquisition of SMGCs via HGT (Fig. S4).

In addition to the AI screen for HGT candidates, we identified additional putative HGTs of SMGCs to Xylariaceae and Hypoxylaceae based on their (1) high similarity to fungal MIBiG accessions from distantly related fungi; and (2) discontinuous phylogenetic distributions (Fig. S4). Putative HGT of SMGCs included xylarialean SMGCs with > 70% similarity to clusters for ergoline alkaloids and their precursors (e.g. loline, ergovaline and lysergic acid production) produced by Clavicipitaceae endophytes, as well as the phytotoxin cichorine cluster from *Aspergillus* (Fig. S4; Table S3). The griseofulvin cluster from *Penicillium aethiopicum*, which produces a potent antifungal compound (Chooi *et al.*, 2010), also appears horizontally transferred to the clade containing *X. castorea* and *X. flabelliformis* isolates (Figs S4, S5). Although the discontinuous phylogenetic distributions of SMGCs observed here may represent unequal gene loss across taxa (Slot, 2017; Rokas *et al.*, 2018), the presence of entire clusters known from Eurotiomycetes and Sordariomycetes in multiple endophytic and nonendophytic taxa provides additional support for HGTs. Overall, our first AI analysis provides the highest support for HGTs primarily from distantly related hosts such as bacteria (Fig. 2) (see also Marcet-Houben & Gabaldón, 2010), yet our second AI screen and comparisons of SMGCs to MIBiG within our phylogenomic framework also support fungal–fungal HGT as an important mechanism of metabolic innovation in the Xylariales, similar to pathogenic fungi (Qiu *et al.*, 2016).

Expansion of Xylariaceae genomes due to increased gene duplication and HGTs

Despite the close evolutionary relationship and similar ecological niches of taxa in the Xylariaceae and Hypoxylaceae, genomes of Xylariaceae were on average *c.* 7.2 Mbp larger than genomes of Hypoxylaceae (Fig. 3a; Table S6). Larger genome size was associated with higher repeat content: Xylariaceae genomes contained an average of two-fold more repetitive elements (Fig. 3b; Table S6) and had a higher density of repetitive elements surrounding genes (including effectors and genes identified as HGT candidates) compared with Hypoxylaceae genomes (Fig. S6).

In addition to greater repeat content, Xylariaceae genomes also contained on average 750 more protein-coding genes compared with Hypoxylaceae ($P < 0.0001$; Table S6). Ancestral state reconstructions reveal that Xylariaceae genomes have experienced significantly more gene gains ($n = 472$), gene duplication events ($n = 136$), orthogroup gains ($n = 313$) and orthogroup expansion events ($n = 90$) compared with the Hypoxylaceae clade since the radiation from their last common ancestor (Fig. 3c,d), although both clades underwent similar numbers of gene losses ($t_{95} = 0.51$, $P = 0.61$; Table S6). Xylariaceae genomes also experienced on average *c.* two-fold more HGT events compared with Hypoxylaceae genomes (Fig. 3e).

Horizontal gene transfer events were positively associated with increased numbers of SMGCs across both clades (Fig. 3f), reflecting the fact that clustered metabolite genes in fungi are more likely to undergo HGT compared with unclustered genes (Wisecaver *et al.*, 2014). Genomes of Xylariaceae contained on average *c.* 20 more SMGCs than Hypoxylaceae genomes (Table S6) and *c.* two-fold greater cumulative richness of SMGCs compared with the Hypoxylaceae clade (2336 vs 1075 total; 587 vs 282 nonsingleton). Rarefaction analysis revealed that the richness of SMGCs per genome sampled also increased at a greater rate in the Xylariaceae clade (Fig. S7). Genomes of Xylariaceae also contained a greater fraction of isolate-specific SMGCs compared with Hypoxylaceae, regardless of SMGC type (Xylariaceae: 31.2 ± 16.1; Hypoxylaceae: 19.8 ± 15.3; $P = 0.0007$; Figs 1c, S8). Yet despite the high variation of SMGCs among taxa, network analysis illustrated that the composition of SMGCs was more similar among isolates from the same clade, regardless of ecological mode (Fig. S9).

In contrast with the pattern observed for SMGCs, genomes of Hypoxylaceae contained a greater number of CGCs than Xylariaceae genomes (Xylariaceae: 9.5 ± 0.4; Hypoxylaceae: 11.0 ± 0.4; $P = 0.0068$; Table S4) and different classes of CGCs dominated the two clades (Fig. 1d,e). For example, salicylate hydroxylases were the most abundant CGCs among Hypoxylaceae, but were absent from 25% of Xylariaceae genomes (Fig. 1d). Four types of CGCs were universally present across Hypoxylaceae: salicylate hydroxylase, pterocarpan hydroxylase, naringenin 3-dioxygenase, phenol 2-monooxygenase (Fig. 1d). CGCs classified as pterocarpan hydroxylases were the most abundant CGC type in genomes of Xylariaceae (Fig. 1d), but were not found in all Xylariaceae genomes. Only CGCs classified as naringenin 3-dioxygenases were found across all Xylariaceae genomes.

In addition to distinct metabolic gene cluster content and prevalence of HGT between clades, comparison of GO terms for shared orthogroups significantly enriched in either Xylariaceae or Hypoxylaceae (i.e. 74 and 26, respectively) revealed that Hypoxylaceae genomes had a significant increase in the number of GO terms associated with membrane transport, whereas Xylariaceae genomes had a significant increase in the number of GO terms for catalytic activities and binding (Fig. S10). Xylariaceae genomes also contained greater numbers of genes with signalling peptides, as well as genes annotated as effectors, membrane transport proteins, transcription factors, peptidases and CAZymes compared with Hypoxylaceae, even after accounting for differences in genome size

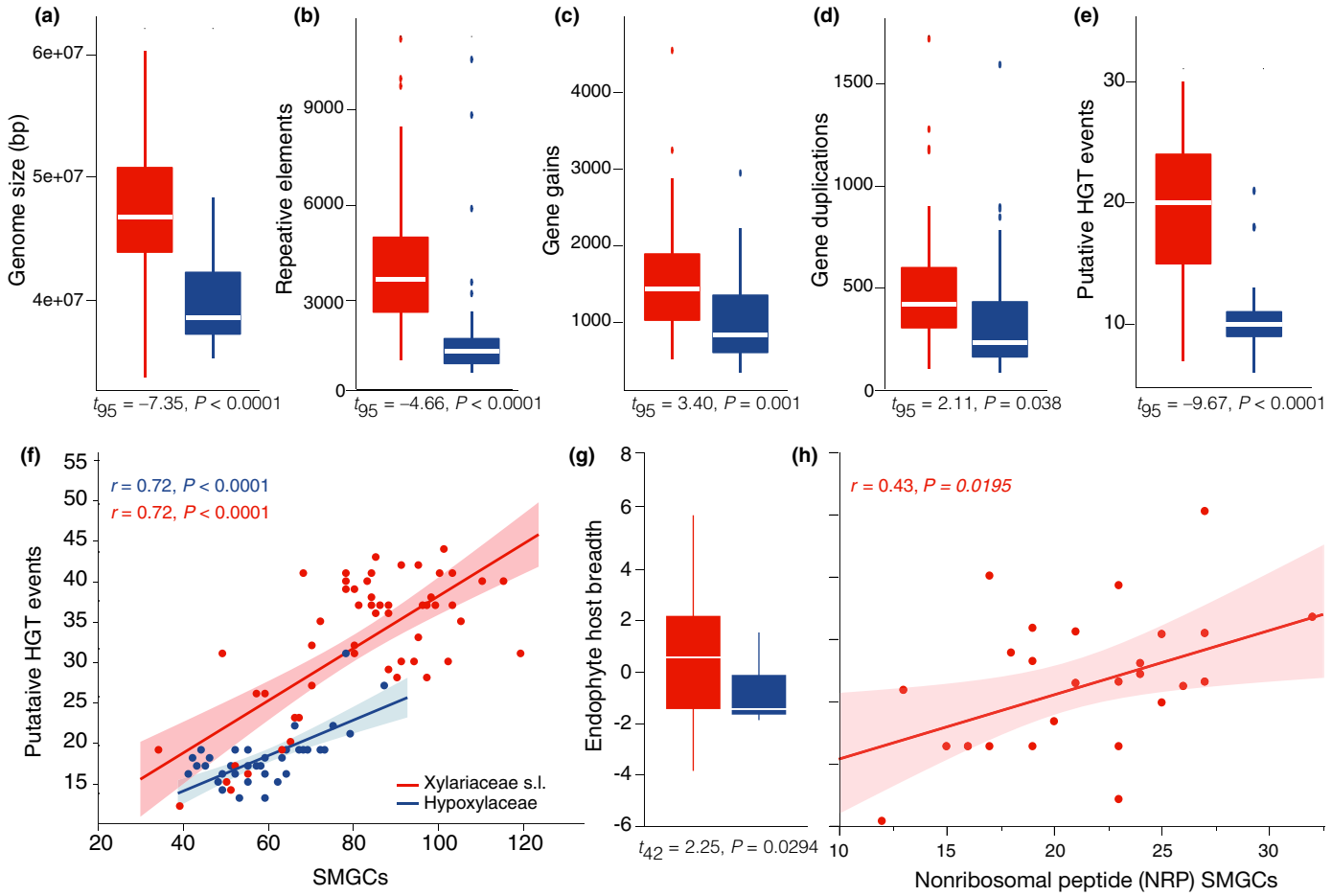

Fig. 3 Larger genomes in the Xylariaceae s.l. clade reflect increased repetitive regions, gene gains and duplications, and horizontal gene transfers (HGTs). Median (a) genome size, (b) repetitive element content, (c) gene gains, (d) gene duplications and (e) number of putative HGT events (high confidence only) for genomes of Xylariaceae s.l. (red) and Hypoxylaceae (blue). Box plot boundaries reflect the interquartile range. Summary statistics (mean, standard deviation and sample size) are reported in Supporting Information Table S6. Gene gains/losses were inferred with Wagner parsimony under a gain penalty = loss penalty = 1. (f) Relationship between the number of HGT events and secondary metabolite gene clusters (SMGCs) as a function of clade. (g) A quantile box plot showing the interquartile range and median of endophyte host breadth [measured as total number of plant families and lichen orders with which a fungal operational taxonomic unit (OTU) was cultured (U'Ren *et al*., 2016)] as a function of major clade (colour). A similar pattern was observed when only the number of plant families are compared (Wilcoxon: $\chi^2 = 4.14$, $P = 0.0413$), but not lichen orders (Wilcoxon: $\chi^2 = 1.77$, $P = 0.1834$). (h) Relationship of Xylariaceae endophyte host breadth and the number of SMGCs classified as nonribosomal peptides (NRPs) per genome. For panels f and h, the shaded region indicates the 95% confidence interval of the linear fit and statistics represent Pearson's correlation coefficient (*r*) and *P*-value.

(Table S6). On average, genomes of Xylariaceae contained *c.* 50 more CAZymes than Hypoxylaceae (Xylariaceae 579.9 ± 7.7; Hypoxylaceae 529.6 ± 9.1, $P < 0.0001$), including a significant increase in PCWDEs involved in the degradation of cellulose, hemicellulose, lignin, pectin and starch (Table S6).

As genomes of fungi with saprotrophic lifestyles typically encode more CAZymes and PCWDEs compared with plant pathogens and mycorrhizal symbionts (Knapp *et al*., 2018; Haridas *et al*., 2020; Miyauchi *et al*., 2020), our genomic results are consistent with the potential for Xylariaceae fungi (including endophytes) to have greater saprotrophic abilities compared with Hypoxylaceae fungi (Osono, 2006). To test this prediction, we compared the abilities of 20 isolates to degrade leaves of *Pinus* and *Quercus*. Regardless of trophic mode, isolates of Xylariaceae with expanded CAZyme and PCWDE repertoires caused greater mass loss compared with taxa with fewer genes predicted to degrade lignocellulose (i.e. Hypoxylaceae and Xylariaceae from animal-dung clade; Fig. S11). In addition to increased capacity for lignocellulose degradation, Xylariaceae endophyte species associated with a greater phylogenetic diversity of plant and lichen hosts compared with species of Hypoxylaceae endophytes ($t_{42} = 2.25$; $P = 0.0294$; Fig. 3g). Host breadth of Xylariaceae endophytes also was positively associated with the number of total HGT events ($r = 0.43$, $P = 0.0193$), as well as the number of peptidases ($r = 0.37$, $P = 0.0444$) and NRP SMGCs (Fig. 3h).

Genomic differences between endophytic and nonendophytic fungi

Both culture-based and culture-free studies of healthy photosynthetic tissues of plants and lichens demonstrate the abundance and novel diversity represented by xylarialean endophytes

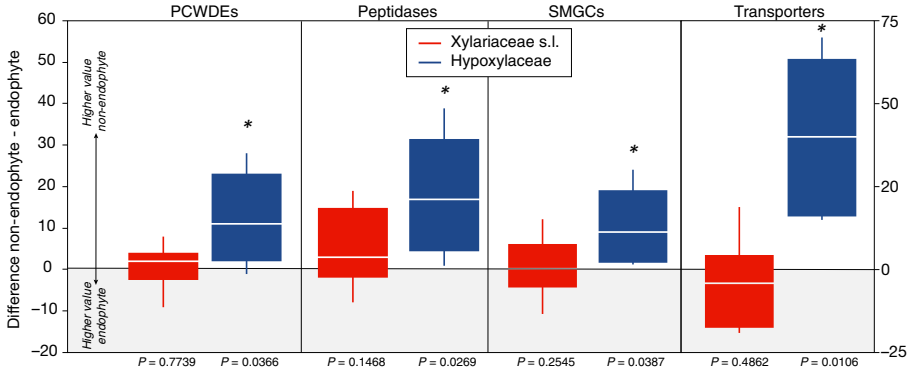

Fig. 4 Pairwise comparisons of sister taxa illustrate ecological modes are more distinct in the Hypoxylaceae. Box plots of the median and interquartile difference in gene counts of plant cell wall-degrading enzymes (PCWDEs), peptidases, secondary metabolite gene clusters (SMGCs) (*y*-axis on left) and transporters (*y*-axis on right) between 15 pairs of sister taxa with contrasting ecological modes for Xylariaceae s.l. and Hypoxylaceae (sister taxa are indicated with asterisks in Fig. 1a). Values greater than zero indicate higher gene counts in nonendophytic taxa, whereas differences less than zero indicate higher gene counts in endophytes. Statistical differences were assessed with least squares means contrast under the null hypothesis: nonendophyte value – endophyte value = 0 (see Supporting Information Table S6 for summary statistics). *, $P < 0.05$.

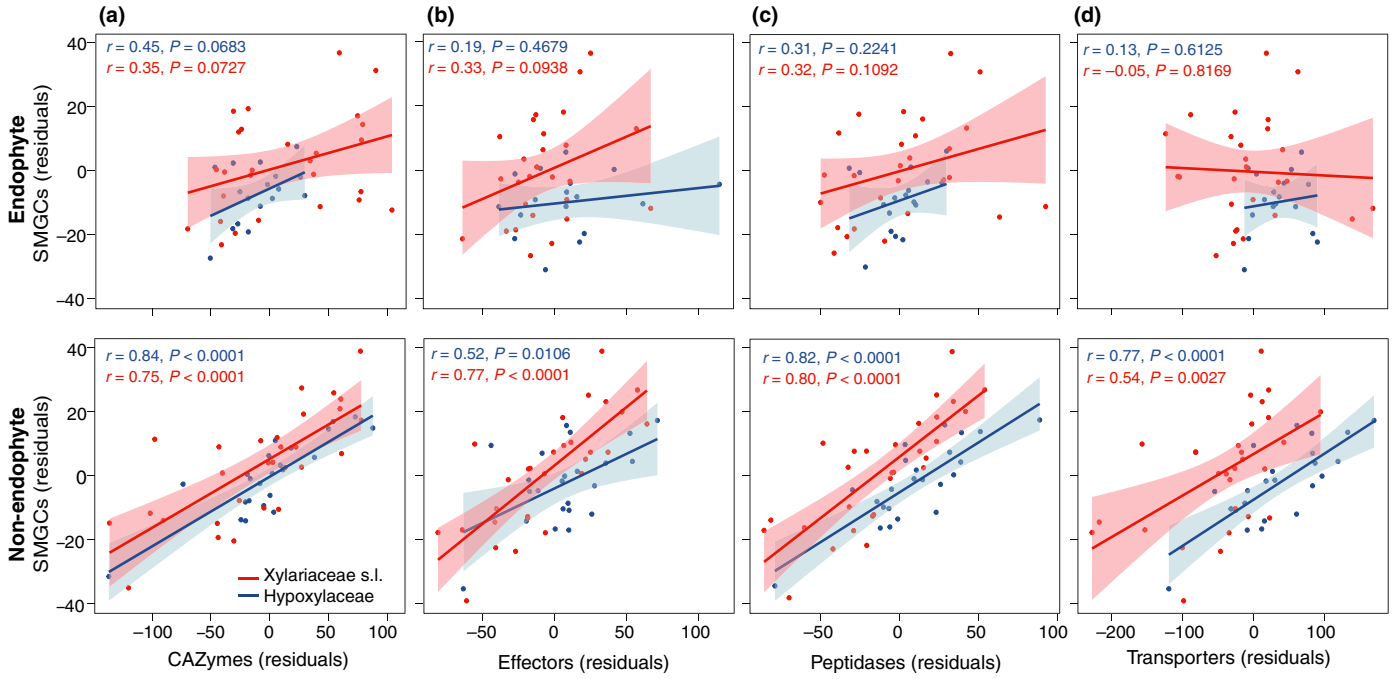

Fig. 5 Correlation of secondary metabolite gene cluster (SMGC) content and genes involved in saprotrophy and/or pathogenicity. Relationship between SMGC abundance and number of genes annotated as (a) Carbohydrate-Active enZYmes (CAZymes), (b) effectors, (c) peptidases and (d) transporters for endophytes (top row) and nonendophytes (bottom row). Values for each genome represent the residuals after accounting for genome size. Points, linear regression lines and shaded 95% confidence intervals of fit are colour-coded by clade (red, Xylariaceae; blue Hypoxylaceae). Statistical values represent Pearson's correlation coefficient (*r*) and *P*-value. See Supporting Information Table S6 for additional details.

(U'Ren *et al.*, 2016). However, some endophytes can occur in both living host tissues as well as decomposing plant materials (Okane *et al.*, 2008; U'Ren & Arnold, 2016; U'Ren *et al.*, 2016) and are often closely related to described species of saprotrophs and pathogens (U'Ren *et al.*, 2016). This suggests that, for some species, endophytism may represent only part of a complex life cycle that blurs the lines between distinct ecological modes (U'Ren *et al.*, 2016; Chen *et al.*, 2018) and few genomic signatures may be associated with the evolution of endophytism in the Xylariaceae and Hypoxylaceae.

Overall, when we analysed all ingroup genomes we observed no clear distinctions in genome size or content due to different ecological modes, even after taking phylogeny into account (Table S6). One exception was the reduced genomes and CAZyme content of termite-associated *Xylaria* spp. (i.e. *X. nigripes* YMJ 653, *X.* sp. CBS 124 048 and *X. intraflava* YMJ725; Figs S3, S12) that reflects a single evolutionary transition to specialisation on termite nest substrates decomposed by a basidiomycete fungus (Hsieh *et al.*, 2010). However, as evolutionary distances among taxa can impede detection of finer-scale genomic

differences due to ecological mode (e.g. Harrington *et al.*, 2019), we restricted our analyses to comparisons of 15 pairs of sister taxa across both clades with contrasting ecological modes. These pairwise comparisons revealed that endophytic Hypoxylaceae genomes contained significantly fewer genes with signalling peptides, protein-coding genes, transporters, peptidases, PCWDEs (especially those involved in decomposition of cellulose and lignin), SMGCs and CGCs compared with nonendophytes (Fig. 4). Yet, similar to the lack of reduced genome repertoires in some root endophytes (Lahrmann *et al.*, 2015; Xu *et al.*, 2015), no significant differences in genomic content were observed between paired endophytes and nonendophytes in the Xylariaceae clade (Fig. 4; Table S6).

These results suggest that, compared with endophytes and saprotrophs in the Hypoxylaceae, Xylariaceae taxa have less distinct ecological modes and their increased metabolic versatility may be the result of selection maintaining diverse genes for both endophytism and saprotrophy. As saprotrophs, fungi experience strong selection to maintain highly diverse SMGCs that increase competitive abilities in diverse microbial communities (Richards & Talbot, 2013; Rokas *et al.*, 2018; Naranjo-Ortiz & Gabaldón, 2020), as well as large gene repertoires to degrade lignocellulosic compounds (Haridas *et al.*, 2020). Accordingly, we observed that in genomes of nonendophytic Xylariaceae and Hypoxylaceae, SMGC abundance was positively correlated with the number of genes important for saprotrophy (e.g. CAZymes, transporters) and putative pathogenicity (e.g. signalling peptides, effectors, peptidases), even after accounting for differences among clades and genome sizes (Fig. 5; Table S6). By contrast, we found that endophyte SMGC abundance was decoupled from the majority of other genomic factors (Fig. 5), due in part to fewer numbers of CAZymes, transporters and peptidases annotated in SMGCs (Table S6). These results are consistent with different selection pressures and ecological roles of SMGCs in endophytic and nonendophytic fungi and highlight the importance of phylogenetically informed comparisons to detect genomic differences associated with endophytism, as well as the complexity of linking genotype to phenotype for complex traits, especially in dynamic genomes undergoing frequent HGT.

Conclusions

Our analysis of 96 phylogenetically and ecologically diverse Xylariaceae and Hypoxylaceae genomes reveals that gene duplication, gene family expansion and HGT of SMGCs, effectors and peptidases from putative bacterial and fungal donors drives metabolic versatility in the Xylariaceae. Expanded metabolic diversity and secondary metabolism of Xylariaceae taxa is associated with greater ecological generalism in both substrate usage and the phylogenetic breadth of symbiotic associations compared with Hypoxylaceae taxa. Correlations between endophyte host breadth, HGT and abundance of NRPs also indicate that SMGCs may play a key role in facilitating xylarialean endophyte colonisation of diverse hosts. For example, although NRPs are known for their role as virulence factors of phytopathogenic fungi (e.g. host-selective toxins or siderophores) (Oide & Turgeon, 2020), previous research has shown that an NRP is essential for the endophyte *Neotyphodium*/*Epichloë* to establish symbiosis with its host (Johnson *et al.*, 2007). Overall, our results highlight the importance of plant–fungal symbioses to drive not only fungal speciation and ecological diversification (Joy, 2013), but vast chemical biodiversity that can be leveraged for novel pharmaceuticals and agrochemicals (Becker & Stadler, 2021; Robey *et al.*, 2021).

Acknowledgements

Funding for the project was provided by the DOE JGI Large-scale Community Science Project (Grant no. 503506 to JMU, JHW, AEA). MEEF was funded by the Office for Research, Innovation and Impact at the University of Arizona and the University of Arizona BIO5 Postdoctoral Fellowship Program. FL and JM received financial support from NSF DEB-1541548 and DEB-1046065, and AEA received support from NSF DEB-1541496 and DEB-1045766. These awards and NSF DEB-0640996 to AEA and DEB-1010675 to AEA and JMU supported the initial collections of endophytes. We thank F. Martin, P. Gladieux, J. Spatafora, R. Vilgalys, and K. O'Donnell for permission to use unpublished JGI F1000 genomes; D. Bellomo, Y. Sanchez-Rosario, and S. Valdez for laboratory assistance; and the Genomics Analysis and Sequencing Core (GATC), the Arizona Genomics Institute (AGI), and the High-Performance Computer (HPC) at the University of Arizona for technical support. The authors declare no competing interests.

Author contributions

JMU, JHW, AEA and MEEF designed research; JMU, LPM, YMJ, HMH, AEA, FL and JM performed field or laboratory research; YMJ, HMH, AEA, DCE and RCH contributed fungal isolates; SA, SJM, AK, RH, SH, BA, RR, K. LaButti, JP, AL, MA, JY, CA, KK, VN and IVG involved in genome and transcriptome sequencing, assembly and annotation; K. Louie and TN performed metabolomics; JHW, JCS and KYC contributed analytic tools; MEEF, JMU, JHW, SA, KEE, KS, ZK, ED and BH analyzed data; MEEF, JMU and JHW wrote the manuscript, with contributions from AEA, JCS, FL, JM, IVG and SA.

ORCID

Steven Ahrendt https://orcid.org/0000-0001-8492-4830
A. Elizabeth Arnold https://orcid.org/0000-0002-7013-4026
Elodie Drula https://orcid.org/0000-0002-9168-5214
Katharine E. Eastman https://orcid.org/0000-0002-4438-3854
Daniel C. Eastwood https://orcid.org/0000-0002-7015-0739
Mario E. E. Franco https://orcid.org/0000-0002-4959-1257
Sajeet Haridas https://orcid.org/0000-0002-0229-0975
Richard D. Hayes https://orcid.org/0000-0002-5236-7918
Bernard Henrissat https://orcid.org/0000-0002-3434-8588
Huei-Mei Hsieh https://orcid.org/0000-0002-7142-3209
Yu-Ming Ju https://orcid.org/0000-0002-8202-6145

Kurt LaButti https://orcid.org/0000-0002-5838-1972
Anna Lipzen https://orcid.org/0000-0003-2293-9329
François Lutzoni https://orcid.org/0000-0003-4849-7143
Jolanta Miadlikowska https://orcid.org/0000-0002-5545-2130
Vivian Ng https://orcid.org/0000-0001-8941-6931
Kelsey Scott https://orcid.org/0000-0003-1378-5348
Jason C. Slot https://orcid.org/0000-0001-6731-3405
Jana M. U'Ren https://orcid.org/0000-0001-7608-5029
Jennifer H. Wisecaver https://orcid.org/0000-0001-6843-5906
Ken Youens-Clark https://orcid.org/0000-0001-9961-144X

Data availability

Raw sequence data, assembled sequences, and genome annotations are available through the corresponding MycoCosm portal (https://mycocosm.jgi.doe.gov/). NCBI accession numbers are listed in Table S1. All other data, including Notes S1, can be found in FigShare Repository (10.6084/m9.figshare.c.5314025).

References

Alexander WG, Wisecaver JH, Rokas A, Hittinger CT. 2016. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proceedings of the National Academy of Sciences, USA* **113**: 4116–4121.

Ambrose KV, Tian Z, Wang Y, Smith J, Zylstra G, Huang B, Belanger FC. 2015. Functional characterization of salicylate hydroxylase from the fungal endophyte *Epichloë festucae*. *Scientific Reports* **5**: doi: 10.1038/srep10939.

Arnold AE, Miadlikowska J, Higgins KL, Sarvate SD, Gugger P, Way A, Hofstetter V, Kauff F, Lutzoni F. 2009. A phylogenetic estimation of trophic transition networks for ascomycetous fungi: are lichens cradles of symbiotrophic fungal diversification? *Systematic Biology* **58**: 283–297.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al.* 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**: 455–477.

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 11.

Becker K, Stadler M. 2021. Recent progress in biodiversity research on the Xylariales and their secondary metabolism. *Journal of Antibiotics* **74**: 1–23.

Blatzer M, Schrettl M, Sarg B, Lindner HH, Pfaller K, Haas H. 2011. SidL, an *Aspergillus fumigatus* transacetylase involved in biosynthesis of the siderophores ferricrocin and hydroxyferricrocin. *Applied and Environmental Microbiology* **77**: 4959–4966.

Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* **47**: W81–W87.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59–60.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Chen K-H, Liao H-L, Arnold AE, Bonito G, Lutzoni F. 2018. RNA-based analyses reveal fungal communities structured by a senescence gradient in the moss *Dicranum scoparium* and the presence of putative multi-trophic fungi. *New Phytologist* **218**: 1597–1611.

Chooi Y-H, Cacho R, Tang Y. 2010. Identification of the viridicatumtoxin and griseofulvin gene clusters from *Penicillium aethiopicum*. *Chemistry & Biology* **17**: 483–494.

Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**: 1910–1912.

Drott MT, Rush TA, Satterlee TR, Giannone RJ, Abraham PE, Greco C, Venkatesh N, Skerker JM, Glass NL, Labbé JL *et al.* 2021. Microevolution in the pansecondary metabolome of *Aspergillus flavus* and its potential macroevolutionary implications for filamentous fungi. *Proceedings of the National Academy of Sciences, USA* **118**. doi: 10.1073/pnas.2021683118.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Informatics* **23**: 205–211.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al.* 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**: D427–D432.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**: 238.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* **125**: 1–15.

Gilchrist CLM, Chooi Y-H. 2021. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Cold Spring Harbor Laboratory* **37**: 2473–2475.

Gluck-Thaler E, Haridas S, Binder M, Grigoriev IV, Crous PW, Spatafora JW, Bushley K, Slot JC. 2020. The architecture of metabolism maximizes biosynthetic diversity in the largest class of fungi. *Molecular Biology and Evolution* **37**: 2838–2856.

Gluck-Thaler E, Slot JC. 2018. Specialized plant biochemistry drives gene clustering in fungi. *ISME Journal* **12**: 1694–1705.

Gluck-Thaler E, Vijayakumar V, Slot JC. 2018. Fungal adaptation to plant defences through convergent assembly of metabolic modules. *Molecular Ecology* **27**: 5120–5136.

Gonçalves C, Wisecaver JH, Kominek J, Oom MS, Leandro MJ, Shen X-X, Opulente DA, Zhou X, Peris D, Kurtzman CP *et al.* 2018. Evidence for loss and reacquisition of alcoholic fermentation in a fructophilic yeast lineage. *eLife* **7**: e33034.

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

Haridas S, Albert R, Binder M, Bloem J, LaButti K, Salamov A, Andreopoulos B, Baker SE, Barry K, Bills G *et al.* 2020. 101 Dothideomycetes genomes: a test case for predicting lifestyles and emergence of pathogens. *Studies in Mycology* **96**: 141–153.

Harrington AH, del Olmo-Ruiz M, U'Ren JM, Garcia K, Pignatta D, Wespe N, Sandberg DC, Huang Y-L, Hoffman MT, Arnold AE. 2019. *Coniochaeta endophytica* sp. nov., a foliar endophyte associated with healthy photosynthetic tissue of *Platycladus orientalis* (Cupressaceae). *Plant and Fungal Systematics* **64**: 65–79.

Hsieh H-M, Ju Y-M, Rogers JD. 2005. Molecular phylogeny of *Hypoxylon* and closely related genera. *Mycologia* **97**: 844–865.

Hsieh H-M, Lin C-R, Fang M-J, Rogers JD, Fournier J, Lechat C, Ju Y-M. 2010. Phylogenetic status of *Xylaria* subgenus *Pseudoxylaria* among taxa of the subfamily Xylarioideae (Xylariaceae) and phylogeny of the taxa involved in the subfamily. *Molecular Phylogenetics and Evolution* **54**: 957–969.

Johnson R, Voisey C, Johnson L, Pratt J, Fleetwood D, Khan A, Bryan G. 2007. Distribution of NRPS gene families within the *Neotyphodium/Epichloë* complex. *Fungal Genetics and Biology* **44**: 1180–1190.

Joy JB. 2013. Symbiosis catalyses niche expansion and diversification. *Proceedings of the Royal Society B: Biological Sciences* **280**. doi: 10.1098/rspb.2012.2820.

Ju YM, Rogers JD. 1996. *A revision of the genus Hypoxylon. Mycologia memoir no. 20*. St. Paul, MN, USA: APS Press.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Kameshwar AKS, Ramos LP, Qin W. 2019. CAZymes-based ranking of fungi (CBRF): an interactive web database for identifying fungi with extrinsic plant biomass degrading abilities. *Bioresources and Bioprocessing* **6**: 51.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* **34**: D354–D357.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software v.7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ *et al.* 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology* **12**: e1001889.

Knapp DG, Németh JB, Barry K, Hainaut M, Henrissat B, Johnson J, Kuo A, Lim JHP, Lipzen A, Nolan M *et al.* 2018. Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. *Scientific Reports* **8**: 6321.

Kuo A, Bushnell B, Grigoriev IV. 2014. Fungal genomics: sequencing and annotation. *Advances in Botanical Research* **70**: 1–52.

Laetsch DR, Blaxter ML. 2017. KinFin: software for taxon-aware analysis of clustered protein sequences. *G3: Genes|Genomes|Genetics* **7**: 3349–3357.

Lahrmann U, Strehmel N, Langen G, Frerigmann H, Leson L, Ding Y, Scheel D, Herklotz S, Hilbert M, Zuccaro A. 2015. Mutualistic root endophytism is not associated with the reduction of saprotrophic traits and requires a noncompromised plant innate immunity. *New Phytologist* **207**: 841–857.

Lawrence DP, Kroken S, Pryor BM, Arnold AE. 2011. Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. *PLoS ONE* **6**: e28231.

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v.4: recent updates and new developments. *Nature Acids Research* **47**: W256–W259.

Li J, Cornelissen B, Rep M. 2020. Host-specificity factors in plant pathogenic fungi. *Fungal Genetics and Biology* **144**. doi: 10.1016/j.fgb.2020.103447.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **42**: D490–D495.

Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics* **26**: 5–8.

Massengo-Tiassé RP, Cronan JE. 2009. Diversity in enoyl-acyl carrier protein reductases. *Cellular and Molecular Life Sciences* **66**: 1507–1517.

Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al.* 2014. Data access for the 1,000 plants (1KP) project. *GigaScience* **3**: 17.

Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC *et al.* 2015. Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology* **11**: 625–631.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI *et al.* 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* **47**: D351–D360.

Miyauchi S, Kiss E, Kuo A, Drula E, Kohler A, Sánchez-García M, Morin E, Andreopoulos B, Barry KW, Bonito G *et al.* 2020. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nature Communications* **11**: 5125.

Naranjo-Ortiz MA, Gabaldón T. 2020. Fungal evolution: cellular, genomic and metabolic complexity. *Biological Reviews* **95**: 1198–1232.

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S *et al.* 2020. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* **16**: 60–68.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Nielsen H. 2017. Predicting secretory proteins with SignalP. *Methods in Molecular Biology* **1611**: 59–73.

Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, Frisvad JC, Workman M, Nielsen J. 2017. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nature Microbiology* **2**: 17044.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al.* 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**: D733–D745.

Oide S, Turgeon BG. 2020. Natural roles of nonribosomal peptide metabolites in fungi. *Mycoscience* **61**: 101–110.

Okane I, Toyama K, Nakagiri A, Suzuki K-I, Srikitikulchai P, Sivichai S, Hywel-Jones N, Potacharoen W, Læssøe T. 2008. Study of endophytic Xylariaceae in Thailand: diversity and taxonomy inferred from rDNA sequence analyses with saprobes forming fruit bodies in the field. *Mycoscience* **49**: 359–372.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012. *caper: comparative analyses of phylogenetics and evolution in R.* R package v.1.0.1. [WWW document] URL https://CRAN.R-project.org/package=caper.

Osono T. 2006. Role of phyllosphere fungi of forest trees in the development of decomposer fungal communities and decomposition processes of leaf litter. *Canadian Journal of Microbiology* **52**: 701–716.

Peay KG, Kennedy PG, Talbot JM. 2016. Dimensions of biodiversity in the earth mycobiome. *Nature Reviews Microbiology* **14**: 434–447.

Porras-Alfaro A, Bayman P. 2011. Hidden fungi, emergent properties: endophytes and microbiomes. *Annual Review of Phytopathology* **49**: 291–315.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**: i351–i358.

Qiu H, Cai G, Luo J, Bhattacharya D, Zhang N. 2016. Extensive horizontal gene transfers between plant pathogenic fungi. *BMC Biology* **14**: 41.

Rawlings ND, Barrett AJ, Finn R. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research* **44**: D343–D350.

Richards TA, Talbot NJ. 2013. Horizontal gene transfer in osmotrophs: playing with public goods. *Nature Reviews Microbiology* **11**: 720–727.

Robey MT, Caesar LK, Drott MT, Keller NP, Kelleher NL. 2021. An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes. *Proceedings of the National Academy of Sciences, USA* **118**: doi: 10.1073/pnas.2020230118.

Rodriguez RJ, White JF Jr, Arnold AE, Redman RS. 2009. Fungal endophytes: diversity and functional roles: tansley review. *New Phytologist* **182**: 314–330.

Rokas A, Wisecaver JH, Lind AL. 2018. The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology* **16**: 731–744.

Saier MH Jr, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. 2016. The transporter classification database (TCDB): recent advances. *Nature Acids Research* **44**: D372–D379.

Shen X-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. *Science Advances* **6**: eabd0079.

Slot JC. 2017. Fungal gene cluster diversity and evolution. *Advances in Genetics* **100**: 141–178.

Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. 2018. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology* **19**: 2094–2110.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

The Gene Ontology Consortium. 2019. The gene ontology resource: 20 years and still GOing strong. *Nature Acids Research* **47**: D330–D338.

Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 2020. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**: 607–621.

U'Ren JM, Arnold AE. 2016. Diversity, taxonomic composition, and functional aspects of fungal communities in living, senesced, and fallen leaves at five sites across North America. *PeerJ* **4**: e2768.

U'Ren JM, Lutzoni F, Miadlikowska J, Laetsch AD, Arnold AE. 2012. Host and geographic structure of endophytic and endolichenic fungi at a continental scale. *American Journal of Botany* **99**: 898–914.

U'Ren JM, Lutzoni F, Miadlikowska J, Zimmerman NB, Carbone I, May G, Arnold AE. 2019. Host availability drives distributions of fungal endophytes in the imperiled boreal realm. *Nature Ecology and Evolution* **3**: 1430–1437.

U'Ren JM, Miadlikowska J, Zimmerman NB, Lutzoni F, Stajich JE, Arnold AE. 2016. Contributions of North American endophytes to the phylogeny, ecology,

and taxonomy of Xylariaceae (Sordariomycetes, Ascomycota). *Molecular Phylogenetics and Evolution* **98**: 210–232.

Verster KI, Wisecaver JH, Karageorgi M, Duncan RP, Gloss AD, Armstrong EE, Price DK, Menon AR, Ali ZM, Whiteman NK. 2019. Horizontal transfer of bacterial cytolethal distending toxin B genes to insects. *Molecular Biology and Evolution* **36**: 2105–2110.

Voglmayr H, Friebes G, Gardiennet A, Jaklitsch WM. 2018. *Barrmaelia* and *Entosordaria* in Barrmaeliaceae (fam. nov., Xylariales) and critical notes on *Anthostomella*-like genera based on multigene phylogenies. *Mycological Progress* **17**: 155–177.

Wendt L, Sir EB, Kuhnert E, Heitkämper S, Lambert C, Hladki AI, Romero AI, Luangsa-ard JJ, Srikitikulchai P, Peršoh D *et al.* 2018. Resurrection and emendation of the Hypoxylaceae, recognised from a multigene phylogeny of the Xylariales. *Mycological Progress* **17**: 115–154.

Wibberg D, Stadler M, Lambert C, Bunk B, Spröer C, Rückert C, Kalinowski J, Cox RJ, Kuhnert E. 2021. High quality genome sequences of thirteen Hypoxylaceae (Ascomycota) strengthen the phylogenetic family backbone and enable the discovery of new taxa. *Fungal Diversity* **106**: 7–28.

Wisecaver JH, Alexander WG, King SB, Hittinger CT, Rokas A. 2016. Dynamic evolution of nitric oxide detoxifying flavohemoglobins, a family of single-protein metabolic modules in Bacteria and Eukaryotes. *Molecular Biology and Evolution* **33**: 1979–1987.

Wisecaver JH, Slot JC, Rokas A. 2014. The evolution of fungal metabolic pathways. *PLoS Genetics* **10**: e1004816.

Wu W, Davis RW, Tran-Gyamfi MB, Kuo A, LaButti K, Mihaltcheva S, Hundley H, Chovatia M, Lindquist E, Barry K *et al.* 2017. Characterization of four endophytic fungi as potential consolidated bioprocessing hosts for conversion of lignocellulose into advanced biofuels. *Applied Microbiology and Biotechnology* **101**: 2603–2618.

Xu X-H, Su Z-Z, Wang C, Kubicek CP, Feng X-X, Mao L-J, Wang J-Y, Chen C, Lin F-C, Zhang C-L. 2015. The rice endophyte *Harpophora oryzæ* genome reveals evolution from a pathogen to a mutualistic endophyte. *Scientific Reports* **4**: 5783.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Overview of Alien Index (AI) calculations to identify horizontal gene transfers (HGTs).

Fig. S2 Results of phylogenomic and network analysis of 1526 universal single-copy orthologous protein sequences.

Fig. S3 Phylogenomic reconstruction of Xylariaceae s.l. and Hypoxylaceae and genome statistics.

Fig. S4 Dynamic distribution of 168 Xylariaceae and Hypoxylaceae secondary metabolite gene clusters (SMGCs) with hits to known metabolites in the MIBiG repository.

Fig. S5 Similarity of the griseofulvin SMGC in *Penicillium* and *Xylaria* supports HGT.

Fig. S6 The density of repetitive elements surrounding genes was higher for Xylariaceae s.l. than for Hypoxylaceae genomes.

Fig. S7 Rarefaction analysis illustrates higher SMGC diversity in Xylariaceae compared with Hypoxylaceae.

Fig. S8 Most SMGCs are specific to Hypoxylaceae or Xylariaceae s.l. clades or individual isolates regardless of SMGC type.

Fig. S9 Network analysis illustrates the importance of clade rather than ecological mode for SMGC content.

Fig. S10 Orthogroup enrichment suggests functional differences for Xylariaceae s.l. and Hypoxylaceae.

Fig. S11 Xylariaceae s.l. taxa demonstrate increased decomposition abilities (estimated via mass loss) on leaf litter compared with fungi with reduced genomes (i.e. Hypoxylaceae and animal dung Xylariaceae s.l. in the *Poronia* clade).

Fig. S12 Relative abundance of functional gene categories across Xylariaceae s.l. and Hypoxylaceae.

Methods S1 Additional information on strain selection, fungal growth and nucleic acid extraction, genome and transcriptome sequencing, gene prediction and genome assembly, phylogenetic analyses, comparative genomic analyses, litter decomposition assays and metabolomics.

Notes S1 List and description of appendices S1–S10 available on FigShare Repository, 10.6084/m9.figshare.c.5314025.

Table S1 Sequencing and assembly statistics for the 121 genomes included in this study.

Table S2 RepeatMasker, RepeatScout and RepBase Update classification of repetitive elements for 96 genomes of Xylariaceae s.l. and Hypoxylaceae.

Table S3 Secondary metabolite gene clusters and cluster families for the 96 Hypoxylaceae and Xylariaceae s.l. genomes included in this study.

Table S4 Catabolic gene clusters and cluster families for the 96 Hypoxylaceae and Xylariaceae s.l. genomes included in this study.

Table S5 Taxonomic, phylogenetic, and functional annotation information for HGT candidate genes identified by Alien Index analyses.

Table S6 Counts and statistical comparisons of genome content as a function of major clade (Xylariaceae s.l. vs Hypoxylaceae) and ecological mode (endophyte vs nonendophyte).

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.