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[7,8,9], Huei-Mei Hsieh[4], Ken Youens-Clark[1], François Lutzoni[10], Jolanta Miadlikowska[10], Daniel C. Eastwood[11], Richard C. Hamelin[12], Igor V. Grigoriev[6,13], Jana M. U'Ren[1*]

[1]BIO5 Institute and Department of Biosystems Engineering, The University of Arizona, Tucson, Arizona, United States of America; [2]Department of Biochemistry, Purdue University, West Lafayette, Indiana, United States of America; [3]School of Plant Sciences and Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, Arizona, United States of America; [4]Institute of Plant and Microbial Biology, Academic Sinica, Taipei, Taiwan; [5]Department of Plant Pathology, The Ohio State University, Columbus, Ohio, United States of America; [6]Department of Energy, The Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America; [7]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille, France; [8]INRAE, Marseille, France; [9]Department of Biological Sciences, King Abdulaziz University, Saudi Arabia; [10]Department of Biology, Duke University, Durham, North Carolina, United States of

America; [11]Department of Biosciences, Swansea University, Swansea, Wales, United Kingdom; [12]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada; [13]Department of Plant and Microbial Biology, University of California, Berkeley, California, United States of America.

* Corresponding author

Email: juren@email.arizona.edu, Phone: (520) 626-0426

Keywords: Ascomycota, endophyte, plant-fungal interactions, saprotroph, specialized metabolism, trophic mode, symbiosis, 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 to more ecologically specialized 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 (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 colonization, 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 (U'Ren *et al.*, 2019). In contrast, endophytic fungi in the Xylariales (Sordariomycetes, Pezizomycotina, Ascomycota) appear unique in that they typically 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 (Arnold *et al.*, 2009; U'Ren *et al.*, 2016). Although the genetic factors that determine foliar endophyte host range are 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.*, 2020), 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 xylarialean fungi (>500 SMs reported to date; (Becker & Stadler, 2021)), their broad host ranges as endophytes, and ability to persist in leaf litter as saprotrophs that decompose lignocellulose (U'Ren *et al.*, 2016; U'Ren & Arnold, 2016) led us to hypothesize that enhanced secondary metabolism might play a role in facilitating ecological generalism in both substrate usage 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 s.l., hereafter Xylariaceae). We paired genomic data with extensive metadata on endophyte host associations, geographic distributions, and substrate usage gleaned from a collection of >6,000 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 and Hypoxylaceae representing ca. 24 genera and 80 species, as well as an additional two undescribed species of endophytic Xylariales (*Pestalotiopsis* sp. NC0098 and Xylariales sp. AK1849) included in the outgroup (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 non-living tissues as well. Isolates were defined as saprotrophs only if all isolates in the OTU were cultured from non-living plant tissues such as senescent leaves or leaf litter (U'Ren *et al.*, 2016). To minimize 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. non-endophyte) (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 (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 (see Supporting Information). DNA isolations were performed using modified phenol:chloroform extractions (U'Ren & Moore). RNA was extracted for each isolate with the

Ambion Purelink RNA Kit (Thermo Fisher Scientific, Waltham, MA). DNA and RNA were quantified with a Qubit fluorometer (Invitrogen) and sample purity was assessed with a NanoDrop (BioNordika). RNA was treated with DNase (Thermo Fisher Scientific) following the manufacturer's instructions and RNA integrity was assessed on a BioAnalyzer 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 300bp or 600bp) 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 non-organelle reads with SPAdes (Bankevich *et al.*, 2012). PacBio SMRT sequencing was performed for 22 isolates of Xylariaceae and Hypoxylaceae, as well as Xylariales spp. NC0098, and AK1849 on a PacBio Sequel. Library preparation was performed either using the PacBio Low Input 10kb or PacBio >10kb with AMPure Bead Size Selection. Filtered sub-read data were processed with the JGI QC pipeline and *de novo* assembled using Falcon (SEQUEL) or Flye (SEQUEL II). Stranded RNASeq libraries were created and quantified by qPCR and transcriptome sequencing was performed on an Illumina NovaSeq S4. For both Hypoxylaceae and Xylariaceae ~25% of genomes were sequenced with PacBio, although a higher proportion of endophyte genomes were sequenced with PacBio than Illumina (43% vs. 28% overall). Genome completeness was assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0" using the "eukaryota_odb9" (2016-11-02) dataset (https://doi.org/10.1093/bioinformatics/btv351).

Genome annotation. Gene prediction and annotation was performed using the JGI pipeline (Kuo *et al.*, 2014; Grigoriev *et al.*, 2014) (see Supporting Information). 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 v3.0a (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). Candidate effectors were predicted using EffectorP v2.0 (Sperschneider *et al.*, 2018).

Orthogroup prediction, functional annotation, and ancestral state reconstruction. For comparative analyses, data from an additional eight taxa in Xylariaceae *sensu lato* (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 v2.3.3 (Emms & Kelly, 2019), which was executed using DIAMOND v0.9.22 (Buchfink *et al.*, 2015) for the all-versus-all sequence similarity search and MAFFT v7.427 (Katoh & Standley, 2013) for sequence alignment. Orthogroups were assigned functional annotations with KinFin v1.0 (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 also was used to perform network analysis of orthogroups, classify orthogroups and SMGCs into isolate-specific, clade-specific (Hypoxylaceae and Xylariaceae), and universal (i.e., orthogroups present in all taxa) categories, and to identify orthogroups that were significantly enriched or depleted in the Xylariaceae or Hypoxylaceae using the Mann-Whitney U test. KOG annotations were used to compare putative functions of universal ("core") vs. isolate-specific ("dispensable") orthogroups (see Supporting Information). We used Count v10.04 (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 1,526 single-copy orthogroups defined by OrthoFinder were aligned using MAFFT v7.427 (Katoh & Standley, 2013), concatenated, and analyzed using maximum-likelihood in IQ-TREE multicore v1.6.11 (Nguyen *et al.*, 2015) with the Le Gascuel (LG) substitution model. Node support was calculated with 1,000 ultrafast bootstrap replicates. Additional phylogenomic analyses with different models of evolution, gene sets, and outgroup taxa resulted in nearly identical topologies (see Supporting Information).

Metabolic gene cluster prediction. SMGCs were predicted using antiSMASH version 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 visualize and compare SMGCs (Gilchrist & Chooi, 2020). Sequence similarity network analysis of the SMGCs was performed using BiG-SCAPE v1.0.1 (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 version 1.4 (Medema *et al.*, 2015). The output from BiG-SCAPE was incorporated into KinFin (Laetsch & Blaxter, 2017) to visualize gene content similarity as network graphs as well as 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) (hereafter, 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 (v2.2.25+) bitscore of 50, 30% amino acid identity, and target sequence alignment 50-150% of the query sequence length. Homologs 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 characterized, 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 30kb to define discrete clusters (i.e., clusters on the same contig were consolidated if separated by less than 30kb). 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/wise) (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 v0.9.22.123 (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 normalized 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 normalized bit score of the best hit to a species outside of the ancestral lineage and *nbsA* is the normalized 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 sub-lineage 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 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) Genes were considered putative HGT candidates if they passed the following filters: (i) AI score of > 0, (ii) significant hits to at least 25 sequences in the custom database, and (iii) 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 v7.407 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 v1.6.10 (Nguyen *et al.*, 2015) in a single run with ModelFinder (Kalyaanamoorthy *et al.*, 2017) and SH-aLRT combined with ultrafast bootstrapping analyses (1,000 replicates each). Phylogenies were visualized using iTOL

v4 (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: (i) the association between donor and recipient clades was supported by ultrafast bootstrap >= 95 and (ii) 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' version 1.0.1 (Orme *et al.*, 2012) in R version 3.6.1. All other statistical analyses were done in R version 3.6.1 or JMP version 15.1 (SAS Institute Inc., Cary, NC).

RESULTS AND DISCUSSION

Genomes of 96 xylarialean taxa correspond to the previously recognized family Xylariaceae (Ju & Rogers, 1996) that was recently split into multiple families (Hypoxylaceae, Graphostromataceae, Barrmaeliaceae (Voglmayr *et al.*, 2018; Wendt *et al.*, 2018); Fig. 1a; Fig. S2). 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 ca. 8,000-15,000 predicted genes (average 11,871; Fig. S3), congruent with average genome and proteome sizes of Pezizomycotina (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 6,879 putative SMGCs (belonging to 3,313 cluster families) and 973 putative CGCs (belonging to 190 cluster families) (Tables S3 and S4). In comparison, recent large-scale analyses predicted 3,399 SMGCs (in 719 cluster families) across 101 Dothideomycetes genomes (Gluck-Thaler *et al.*, 2020) and 1,110 CGCs across 341 fungal genomes (Gluck-Thaler & Slot, 2018). Only 25% of predicted SMGCs (n = 1,711 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 fungi in the Pezizomycotina (average 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 3.3X more CGCs per genome (average 10.1; Table S4) compared to genomes of Pezizomycotina (average 3.0 (Gluck-Thaler *et al.*, 2018)).

Every xylarialean genome contained SMGCs for the production of polyketides (PK; 2,871 total), nonribosomal peptides (NRP; 2,482 total), and terpenes (1,322 total; Fig. 1b; Table

S3). SMGCs for ribosomally synthesized 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 (Gluck-Thaler *et al.*, 2018) (n = 170 naringenin 3−dioxygenases) also were 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 ca. 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; see also 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 Alien Index (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 4,262 genes representing 647 orthogroups (Table S5). Using a custom phylogenetic pipeline (see Methods) we manually validated 168 of these genes as likely 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 HGT (Fig. 2; Table S5). Similar to previous studies (Marcet-Houben & Gabaldón, 2010; Lawrence *et al.*, 2011), the majority of high-confidence HGTs are predicted to have been acquired from bacteria (n = 86) (Fig. 2). Overall, 66% of genes identified as HGT from bacteria do not contain introns (compared to 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 123580 (n = 30).

HGT 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 119009; *H. fuscum* CBS 119018 (HGT0001; Table S5). Multiple evolutionary scenarios could result in patchy taxonomic distributions. For example, multiple fungi could have

353 independently acquired the same gene from closely related bacterial donors (Marcet-Houben &
354 Gabaldón, 2010). Alternatively, an initial HGT from bacteria to fungi may have been followed
355 by fungal-fungal HGTs. In total, 38 HGT candidate genes occurred in genomes of both
356 Sordariomycetes outgroup and Xylariales genomes, 28 were found in only Xylariales genomes,
357 and 26 were only observed in genomes of Xylariaceae and Hypoxylaceae (Fig. 2; Table S5).
358 Functional annotation revealed the majority of candidate HGT genes were associated
359 with at least one type of annotation (i.e., 95% of the highly confident and 82% of the ambiguous
360 events; Table S5). Six high-confidence HGT candidate genes were annotated as CAZymes,
361 including three predicted plant cell wall degrading enzymes (PCWDEs) transferred from bacteria
362 to diverse Xylariales (Fig. 2). No genes predicted in CGCs were identified as candidate HGTs,
363 consistent with convergent evolution to result in similar clustering of fungal phenolic metabolism
364 genes (Gluck-Thaler *et al.*, 2018). However, 43% of candidate HGT genes were predicted to be
365 part of a SMGC (i.e., 40 of 92) (Fig. 2; Tables S3 and S5). These include 13 genes predicted to
366 have a biosynthetic function, such as a putative FsC-acetyl coenzyme A-N^2-transacetylase
367 (HGT076; Table S5), which is part of the siderophore biosynthetic pathway in *Aspergillus*
368 implicated in fungal virulence (Blatzer *et al.*, 2011).
369 Due to the high prevalence of HGT among genes predicted to be part of SMGCs, we
370 performed a second AI screen to detect intra-fungal HGT events of genes within the boundaries
371 of SMGCs (n = 93,066 genes) (see Methods; Fig. S1). AI identified 1,148 genes in 660 SMGCs
372 (belonging to 594 cluster families) that were putatively transferred from other fungi to members
373 of the Xylariales (Table S5). Candidate HGT genes were primarily for polyketide and
374 nonribosomal peptide production (518 PKSs, 270 NRPSs, and 180 PKS-NRPS hybrid clusters).
375 In addition, >75% of hits to MIBiG contain genes identified by AI analyses as putative HGTs

(see Fig. S4, bottom). SMGCs with HGT candidate genes include 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 and S5). Although the second AI analysis did not identify each gene in these clusters as potential HGTs (e.g., 4 of the 19 genes in the alternariol cluster from *Hypoxylon cercidicola* CBS 119009 were predicted to be HGT; 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).

Although AI is a robust high-throughput method to identify candidate HGT events, we identified additional SMGCs not flagged by the second AI analysis with high similarity to fungal MIBiG accessions (Medema *et al.*, 2011) and phylogenetic distributions that support potential HGT to Xylariaceae and Hypoxylaceae (Fig. S4). For example, 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 phytotoxins 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 non-endophytic 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 a 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 ca. 7.2 Mbp larger than genomes of Hypoxylaceae (Fig. 3a; Table S6). Larger genome size was associated with higher repeat content: Xylariaceae contained an average of 2-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 to Hypoxylaceae genomes (Fig. S6).

In addition to greater repeat content, Xylariaceae genomes also contained on average 750 more protein-coding genes compared to 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 to 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 ca. 2-fold more HGTs events compared to Hypoxylaceae genomes (Fig. 3e).

Increased genome sizes resulting from HGTs 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 to unclustered genes (Wisecaver *et al.*, 2014). Genomes of Xylariaceae contained on average ca. 20 more SMGCs than Hypoxylaceae genomes (Table S6) and ca. 2-fold greater cumulative richness of SMGCs compared to Hypoxylaceae clade (2,336 vs. 1,075 total; 587 vs. 282 non-singleton). Rarefaction analysis reveals the richness of SMGCs increases at a greater rate in the Xylariaceae clade (Fig. S7). Genomes of Xylariaceae also contained a greater fraction of isolate specific SMGCs compared to Hypoxylaceae, regardless of SMGC type (Xylariaceae: 31.2 ± 16.1; Hypoxylaceae: 19.8 ± 15.3; P = 0.0007; Fig. 1c; Fig. S8). Yet despite the high variation of SMGCs among taxa, network analysis illustrates that the composition of SMGCs is more similar among isolates from the same clade, regardless of ecological mode (Fig. S9).

In contrast to 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 CGC dominated the two clades (Fig. 1d,e). For example, salicylate hydroxylases were the most abundant CGCs among Hypoxylaceae, but were absent from 25% 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 Xylariacaee 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, comparison of gene ontology (GO) terms for shared orthogroups significantly enriched in either Xylariaceae or Hypoxylaceae (i.e., 74 and 26, respectively) revealed that the Hypoxylaceae had a significant

445 increase in the number of GO terms associated with membrane transport, whereas Xylariaceae
446 had a significant increase in the number of GO terms for catalytic activities and binding (Fig.
447 S10). Xylariaceae genomes also contained greater numbers of genes with signaling peptides, as
448 well as genes annotated as effectors, membrane transport proteins, transcription factors,
449 peptidases, and CAZymes compared to Hypoxylaceae, even after accounting for differences in
450 genome size (Table S6). On average genomes of Xylariaceae contained ca. 50 more CAZymes
451 than Hypoxylaceae (Xylariaceae 579.9 ± 7.7; Hypoxylaceae 529.6 ± 9.1, P <0.0001), including a
452 significant increase in PCWDEs involved in the degradation of cellulose, hemicellulose, lignin,
453 pectin, and starch (Table S6).
454 As genomes of fungi with saprotrophic lifestyles typically contain more CAZymes and
455 PCWDEs compared to plant pathogens and mycorrhizal symbionts (Knapp *et al.*, 2018; Haridas
456 *et al.*, 2020; Miyauchi *et al.*, 2020), our genomic results are consistent with the potential for
457 Xylariaceae fungi (including endophytes) to have greater saprotrophic abilities compared to
458 Hypoxylaceae fungi (Osono, 2006). To test this prediction, we compared the abilities of 20
459 isolates to degrade leaves of *Pinus* and *Quercus*. Regardless of trophic mode, isolates of
460 Xylariaceae with expanded CAZymes and PCWDEs repertoires caused greater mass loss
461 compared to taxa with fewer genes predicted to degrade lignocellulose (i.e., Hypoxylaceae and
462 Xylariaceae from animal-dung clade; Fig. S11). In addition to increased capacity for
463 lignocellulose degradation, Xylariaceae endophyte species associate with a greater phylogenetic
464 diversity of plant and lichen hosts compared to species of Hypoxylaceae endophytes (t_{42} = 2.25;
465 P = 0.0294; Fig. 3g). Host breadth of Xylariaceae endophytes is positively associated with the
466 number of total HGT events (r = 0.43, P = 0.0193), as well as the number of peptidases (r = 0.37,
467 P = 0.0444) and nonribosomal peptide (NRP) SMGCs (Fig. 3h).

Genomic differences between endophytic and non-endophytic fungi. The majority of described Xylariaceae and Hypoxylaceae species are wood- or litter-degrading saprotrophs or woody pathogens (Hsieh *et al.*, 2005, 2010), although 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 (U'Ren *et al.*, 2016). Previous studies have identified isolates with highly similar sequences of the fungal internal transcribed spacer nuclear ribosomal DNA (ITS nrDNA) barcode locus occurring in both living host tissues as well as decomposing plant materials (Okane *et al.*, 2008; U'Ren *et al.*, 2016). This suggests that 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 analyzed 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 124048, and *X. intraflava* YMJ725; Figs. S3 and S12) that reflects a single evolutionary transition to specialization on termite nest substrates decomposed by a basidiomycete fungus (Hsieh *et al.*, 2010). However, as evolutionary distance among taxa can impede detection of finer-scale genomic differences due to ecological mode (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 contain significantly fewer genes with signaling peptides, protein coding genes, transporters, peptidases, PCWDEs (especially those

involved in decomposition of cellulose and lignin), SMGCs, and CGCs compared to non-endophytes (Fig. 4). Yet, similar to the lack of reduced genome repertoires in some root endophytes (Xu *et al.,* 2015; Lahrmann *et al.*, 2015), no significant differences in genomic content were observed between paired endophytes and non-endophytes in the Xylariaceae clade (Fig. 4; Table S6).

These results suggest that compared to endophytes and saprotrophs in the Hypoxylaceae, Xylariaceae taxa may 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 non-endophytic Xylariaceae and Hypoxylaceae SMGC abundance is positively correlated with the number of genes important for saprotrophy (e.g., CAZymes, transporters) and putative pathogenicity (e.g., signaling peptides, effectors, peptidases), even after accounting for differences among clades and genome sizes (Fig. 5; Table S6). In contrast, we found that endophyte SMGC abundance was decoupled from the majority of genomic factors involved in plant-fungal interactions (Fig. 5), due in part to fewer numbers of CAZymes, transports, and peptidases annotated in SMGCs (Table S6). These results are consistent with different selection pressures and ecological roles of SMGCs in endophytic and non-endophytic fungi and highlight the importance of phylogenetically informed comparisons to detect genomic differences associated with endophytism, as well as 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 to 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 colonization 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 NRPS 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 number 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. 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.

AUTHOR CONTRIBUTIONS

Designed research: JMU, JHW, AEA, MEEF; Performed field or laboratory research: JMU, LPM, YMJ, AEA, FL, JM; Contributed fungal isolates: YMJ, AEA, DCE, RCH; Contributed analytic tools: JHW, JCS, KYC, JGI authors; Analyzed data: MEEF, JMU, JHW, SA, KEE, KS, ZK; Wrote the paper: MEEF, JMU, JHW, with contributions from AEA, JCS, FL, JM, IVG, SA.

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 for raw reads and assemblies are listed in Table S1. All other data can be found in FigShare Repository (DOI 10.6084/m9.figshare.c.5314025; https://figshare.com/s/1684aefe7896295e8fb9)

COMPETING INTERESTS

The authors declare no competing interests.

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 of the United States of America* **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**: 10939.

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: a journal of computational molecular cell 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.

Bastian M, Heymann S, Jacomy M. **2009**. Gephi: an open source software for exploring and manipulating networks. In: Third international AAAI conference on weblogs and social media. aaai.org.

Becker K, Stadler M. **2021**. Recent progress in biodiversity research on the Xylariales and their

secondary metabolism. *The 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. *The 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 of the United States of America* **118**.

Eddy SR. **2009**. A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on 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. *The American naturalist* **125**: 1–15.

Gilchrist CLM, Chooi Y-H. **2020**. clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Cold Spring Harbor Laboratory*: 2020.11.08.370650.

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. *The ISME journal* **12**: 1694–1705.

Gluck-Thaler E, Vijayakumar V, Slot JC. **2018**. Fungal adaptation to plant defences through

618 convergent assembly of metabolic modules. *Molecular ecology* **27**: 5120–5136.

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

622 **Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X,**
623 **Korzeniewski F, *et al.* 2014**. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic*
624 *acids research* **42**: D699–704.

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

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

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

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

638 **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. Biological sciences / The Royal Society* **280**: 20122820.

Ju YM, Rogers JD. **1996**. A revision of the genus Hypoxylon. Mycologia Memoir no. 20. *St. Paul (MN): 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**: D354–7.

Katoh K, Standley DM. **2013**. MAFFT multiple sequence alignment software version 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* **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. *The 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) v4: recent updates and new developments. *Nucleic acids research* **47**: W256–W259.

Li J, Cornelissen B, Rep M. **2020**. Host-specificity factors in plant pathogenic fungi. *Fungal genetics and biology: FG & B* **144**: 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**: D490–5.

Marcet-Houben M, Gabaldón T. **2010**. Acquisition of prokaryotic genes by fungal genomes.

Trends in genetics: TIG **26**: 5–8.

Massengo-Tiassé RP, Cronan JE. **2009**. Diversity in enoyl-acyl carrier protein reductases. *Cellular and molecular life sciences: CMLS* **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, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. **2011**. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39**: W339–46.

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 of the Cambridge Philosophical Society*.

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.

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.

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: D733–45.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W, Orme MD. **2012**. Package 'caper'. *Reference manual, available at* **467**.

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 Suppl 1**: i351–8.

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**: D343–50.

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 of the United States of America **118**.

Rodriguez RJ, White JF Jr, Arnold AE, Redman RS. **2009**. Fungal endophytes: diversity and functional roles: Tansley review. *The 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. *Nucleic acids research* **44**: D372–9.

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**.

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

762 strong. *Nucleic acids research* **47**: D330–D338.

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

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

768 **U'Ren JM, Lutzoni F, Miadlikowska J, Laetsch AD, Arnold AE**. **2012**. Host and geographic
769 structure of endophytic and endolichenic fungi at a continental scale. *American journal of botany*
770 **99**: 898–914.

771 **U'Ren JM, Lutzoni F, Miadlikowska J, Zimmerman NB, Carbone I, May G, Arnold AE**.
772 **2019**. Host availability drives distributions of fungal endophytes in the imperiled boreal realm.
773 *Nature Ecology and Evolution* **3**: 1430–1437.

774 **U'Ren JM, Miadlikowska J, Zimmerman NB, Lutzoni F, Stajich JE, Arnold AE**. **2016**.
775 Contributions of North American endophytes to the phylogeny, ecology, and taxonomy of
776 Xylariaceae (Sordariomycetes, Ascomycota). *Molecular phylogenetics and evolution* **98**: 210–
777 232.

778 **U'Ren JM, Moore LP**. Large Volume Fungal Genomic DNA Extraction Protocol for PacBio.

779 **Verster KI, Wisecaver JH, Karageorgi M, Duncan RP, Gloss AD, Armstrong EE, Price**
780 **DK, Menon AR, Ali ZM, Whiteman NK**. **2019**. Horizontal Transfer of Bacterial Cytolethal
781 Distending Toxin B Genes to Insects. *Molecular biology and evolution* **36**: 2105–2110.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al.* 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**: 261–272.

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. **2020**. High quality genome sequences of thirteen Hypoxylaceae (Ascomycota) strengthen the phylogenetic family backbone and enable the discovery of new taxa. *Fungal diversity*.

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 oryzae genome reveals evolution from a pathogen to a mutualistic endophyte. *Scientific Reports* **4**.

SUPPORTING INFORMATION

Table S1. (a) Information for the 121 genomes included in this study; (b) Genome and assembly information for 96 Xylariaceae s.l. and Hypoxylaceae 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. (a) Secondary metabolite gene cluster (SMGC) annotations for the 121 genomes included in this study (according to antiSMASH) and grouped into families with BiG-SCAPE; (b) Distribution and percent similarity of Xylariaceae s.l. and Hypoxylacae SMGCs to 168 MIBiG accessions; (c) Count and percentage of all SMGCs and SMGC families per category (A-J); (d-j) Count and percentage of SMGCs per type (e.g., NRPS, Terpene, Other PKS, PKS-NRP Hybrids, Other, RiPP) per category (A-J).

Table S4. (a) Count of catabolic gene clusters (CGCs) by anchor gene; (b) Presence/Absence of CGC families per genome; (c) Composition of the CGC families; (d) Genomic position and annotation of CGCs.

826

827 **Table S5**. (a) Taxonomic and phylogenetic information for 4,262 putative HGT candidate genes
828 identified by Alien Index (AI); (b) Manual curation of phylogenetic trees reveals 168 HGT
829 candidates (each row is a unique transfer event; orthogroups may appear more than once); (c)
830 Distribution of HGT counts per genome (HGT001-HGT-129 are high confidence transfers and
831 HGT130-HGT290 are ambiguous transfers); (d) Functional annotation of 1,148 SMGC genes
832 identified by the second Alien Index as candidate HGTs.

833

834 **Table S6**. (a) Number of genes annotated as MEROPS, CAZymes, PCWDCs, SMGCs, CGCs,
835 and putative HGTs for genomes of 96 Xylariaceae s.l. and Hypoxylaceae; (b) Statistical
836 comparison between Xylariaceae s.l. and Hypoxylaceae genomes; (c) Statistical comparison
837 between endophytic and non-endophytic genomes with phylogenetic independent contrasts
838 (PICS); (d) Statistical analysis of genomic features for paired endophyte/non-endophyte sister
839 taxa using least-squares means contrasts; (e) Pearson correlation of genomic features as a
840 function of ecological mode and clade.

841

842 **Table S7**. (a) Orthogroup summary statistics; (b) Orthogroup annotations; (c) Count and
843 percentage of orthogroups and proteins per orthogroup category (A-J). (d) Orthogroups that
844 comprise each category (A-J).

845

846 **Fig. S1. Overview of Alien Index (AI) calculations to identify HGT**. In this example, *Xylaria*
847 *flabelliformis* CBS 116.85 is the query genome. (**a**) AI screen to identify HGT candidates from
848 more distant gene donors (grey box); candidates must have a better hit to sequences outside the

ancestral lineage (Ascomycota; green box). By skipping all sequences to other Xylariales (orange box), HGT candidates could have been acquired at any point back to their last common ancestor (red branches) (**b**) AI screen to identify more recently acquired HGT candidates from other filamentous fungi (grey box). For this screen, candidates must have a better hit to sequences outside the Xylariales (green box). All sequences to other *Xylaria* "PO" clade were skipped (orange box) to identify shared HGT candidates acquired at any point back to the last common ancestor of the clade (red branches).

Fig. S2. Phylogenomic tree inferred by maximum likelihood based on a combination of 1,526 universal single-copy orthologous protein sequences. Twenty-five Sordariomycetes species outside Xylariales were used as the outgroup (Table S1a). Isolates sequenced in this study are highlighted in bold. Endophytes (i.e., fungi isolated from living, photosynthetic tissues of plants and lichens (U'Ren *et al.*, 2016)) are indicated in green. Clade information is based on previously published studies (see (Hsieh *et al.*, 2005, 2010; U'Ren *et al.*, 2016; Voglmayr *et al.*, 2018; Wendt *et al.*, 2018)). Numbers at nodes indicate ultrafast bootstrap support values from IQ-TREE (Nguyen *et al.*, 2015). The scale bar corresponds to the number of substitutions per site.

Fig. S3. **Phylogenomic reconstruction of Xylariaceae s.l. and Hypoxylaceae and genome statistics.** (**a**) The maximum likelihood phylogram is based on 1,526 single-copy orthologous genes present in all genomes. Bootstrap values are shown in Fig. S2. The scale bar indicates the number of substitutions per site. Names of reference taxa are colored according to their clade affiliation (dark blue: Hypoxylaceae; red: Xylariaceae s.l.). Undescribed endophyte species,

putatively named based on phylogenetic analyses (U'Ren *et al.*, 2016), are shown in teal blue; (**b**) genome size; (**c**) predicted protein coding genes; and (**d**) percent transposable element (TE) content (bar colors correspond to ecological mode; see legend). Averages per major clade are shown with dotted lines in panels a-d; (**e**) relative abundance of core, family-specific, clade-specific, and isolate-specific orthogroups (see legend; Table S3d).

Fig S4. Dynamic distribution of 168 Xylariaceae and Hypoxylaceae SMGCs with hits to known metabolites in the MIBiG repository. Rows are sorted by the taxonomic identity (class and species) of the best MIBiG hit (top). Shading indicates the similarity of predicted SMGCs to reference metabolites, defined as the percentage of genes in an SMGC with significant BLAST hits to a known SMGC in the MIBiG database (Medema *et al.*, 2011). Black boxes (bottom) indicate SMGCs predicted by Alien Index (Wisecaver *et al.*, 2016; Verster *et al.*, 2019) to contain at least one gene putatively transferred via HGT (Table S5). For MIBiG clusters that occurred more than once per genome, only the hit with the highest similarity is shown (Table S3).

Fig. S5. Similarity of the griseofulvin SMGC in *Penicillium* and *Xylaria* supports HGT. (a) Comparison of the griseofulvin cluster from *Penicillium aethiopicum* IBT 5753 (top) to five newly sequenced *Xylaria* genomes. Homologous genes are colored by PFAM domain. Connecting ribbons indicate percent amino acid identity to genes in the *Penicillium* cluster; (**b**) Metabolomic analysis of pairwise comparisons of *X. flabelliformis* NC1011, *Xylaria arbuscula* FL1030, and *Daldinia* sp. FL1419 illustrates production of griseofulvin by NC1011 during the interaction with FL1419, but not when grown alone or with isolate FL1030.

Fig. S6. The density of repetitive elements surrounding genes was higher for Xylariaceae s.l. than for Hypoxylaceae genomes. Overlapped density plot of all genomes in each clade (red: Xylariaceae s.l.; blue: Hypoxylaceae), illustrating the distance of the nearest repetitive elements from genes in the following categories: (a, b) effectors, (c, d) non-effector genes; (e, f) high confidence HGT candidate genes, and (g, h) non-HGT genes. Negative distances indicate that repetitive elements are located upstream of genes, while positive distances indicate repetitive elements downstream. Repetitive elements were identified by RepeatScout and RepeatMasker. Effector genes were predicted by EffectorP 2.0. High confidence HGT candidates were predicted using the first Alien Index analysis. Distances were computed using BEDTools v2.29.2.

Fig. S7. Rarefaction analysis illustrates higher SMGC diversity in Xylariaceae compared to Hypoxylaceae. Rarefaction curves of (a) all SMGCs and (b) non-singleton SMGCs by clade (Xylariaceae s.l. Hypoxylaceae, and Sordariomycetes outgroup). Comparison of rarefaction curves for all SMGCs vs. non-singleton SMGCs illustrates the high number of singleton SMGCs present in the outgroup, which is consistent with the phylogenetic diversity of outgroup genomes that span 13 orders of Sordariomycetes. In contrast, richness of non-singleton SMGCs is ca. 4-7X greater for Xylariaceae and Hypoxylaceae genomes compared to outgroup genomes (n = 71 SMGCs).

Fig. S8. The majority of SMGCs are specific to Hypoxylaceae or Xylariaceae s.l. clades or individual isolates regardless of SMGC type. Phylogenomic tree of Xylariaceae s.l. and Hypoxylaceae and outgroup taxa with bar plots illustrating the number of SMGC families per

genome, as well as the percentage of clade-specific and isolate-specific SMGC families for (**a**) PKSI; (**b**) NRPS; (**c**) Terpene; (**d**) PKS-Other; (**e**) PKS-NRP Hybrid; and (**f**) Other.

Fig. S9. Network analysis illustrates the importance of clade rather than ecological mode for SMGC content. Network representation of SMGCs clustering from BiG-SCAPE. Each node represents the SMGC content per genome for (**a**) all SMGCs and SMGC sub-types:(**b**) PKSI; (**c**) NRPS; (d) PKS other; (e) PKS-NRPS Hybrids; (f) terpenes; (g) other; and (h) RiPPs. Networks edited with Gephi v0.9.1 (Bastian *et al.*, 2009), where nodes were scaled by the count of gene clusters and positioned by a force-directed layout algorithm (as described by (Laetsch & Blaxter, 2017)). Edges between two nodes are weighted by the number of shared clusters. Node color corresponds to clade. Nodes representing endophytic isolates are shown with blue borders. To compare the distribution of all SMGC families (**a**), BiG-SCAPE families representing different SMGC types were combined into a single dataset and SMGCs assigned to multiple families were arbitrarily assigned to the largest family.

Fig. S10. Orthogroup enrichment suggests functional differences for Xylariaceae and Hypoxylaceae. Twenty-six orthogroups were significantly enriched in the Hypoxylaceae clade, while 74 orthogroups were significantly expanded in the Xylariaceae s.l. clade. (a) Volcano plot of the protein count representation tests for orthogroups shared between the Hypoxylaceae and Xylariaceae s.l. clades. Orthogroups significantly enriched in Xylariaceae s.l. taxa are colored in red, while orthogroups significantly enriched in Hypoxylaceae taxa are colored in blue. Two-sided Mann-Whitney U-tests, p-value ≤ 0.01 and $|\log2FC| \geq 1$. (b) Comparison of enriched GO terms (level 2) of genes from orthogroups significantly enriched in Hypoxylaceae taxa (blue) vs.

941 Xylariaceae s.l. taxa (red). GO terms were analyzed and visualized using Web Gene Ontology
942 Annotation Plot 2.0 (WEGO). See also Table S3f for KOG annotation of enriched orthologs. The
943 two-sided Mann-Whitney U-test was performed using SciPy (Virtanen *et al.*, 2020) through
944 KinFin v1.0 (Laetsch & Blaxter, 2017).
945
946 **Fig. S11. Relative abundance of functional gene categories across Xylariaceae s.l. and**
947 **Hypoxylaceae.** Phylogenomic tree and bar plot showing the abundance and identity of (**a**)
948 carbohydrate-active enzymes (CAZyme); (**b**) peptidases and their inhibitors (MEROPs); (**c**)
949 transporters (TCDB); (**d**) secreted proteins (SignalP); and (**e**) effectors (EffectorP). Colors refer
950 to different classifications within each database (see legends).
951
952 **Fig. S12. Xylariaceae s.l. taxa demonstrate increased decomposition abilities (estimated via**
953 **mass loss) on leaf litter compared to fungi with reduced genomes (i.e., Hypoxylaceae and**
954 **animal dung Xylariaceae s.l. in the *Poronia* clade).** Interquartile box plots showing median
955 and interquartile range. We observed significant differences among means of each clade on both
956 *Pinus* and *Quercus* leaves (ANOVA). Letters indicate significant differences after post-hoc
957 Tukey's HSD. See Table S1 for a list of isolates included in the mass loss experiment.
958
959 **Fig. S13. Phylogenetic tree topology was similar regardless of methodology, except for**
960 **relationships among taxa in the *Xylaria* HY and E9 clades.** Subclade topology for three
961 phylogenetic analyses: concatenated analysis of 1,526 single-copy orthogroups with the (**a**) LG
962 model of evolution (i.e., analysis 1; see also Fig. S2) or (**b**) JTT + F + I + G4 model of evolution
963 (i.e., analysis 2); and (**c**) ASTRAL coalescent analysis of gene trees.

Fig. S14. Network analysis of individual proteomes illustrates the importance of major clade affiliation. Proteomes are represented by nodes, scaled by the count of proteins, colored by clade (fill) and ecological mode (border), and positioned by a force-directed layout algorithm. Edges between two nodes are weighted by the number of shared orthogroups. The node with a star represents Xylariaceae sp. FL2044.

Fig S15. Comparison of functional annotations for core and dispensable orthogroups. Bar graphs showing the relative abundance of different functional categories represented by "core" vs. "dispensable" orthogroups. Orthogroups were annotated with euKaryotic Orthologous Groups (KOGs; see Table S7f).

Fig S16. **The number of predicted SMGCs is not related to genome assembly.** Relationship of predicted SMGC content (residuals after accounting for genome size) and the number of scaffolds for 121 genomes. Points are colored by clade and their size is proportional to the raw number of SMGC per genome (range 16-119).

MAIN FIGURE LEGENDS

Fig 1. **Xylariaceae and Hypoxylaceae genomes are characterized by hyperdiverse and dynamic metabolic gene clusters.** (**a**) Maximum likelihood phylogenetic analyses of 1,526 universal, single-copy orthogroups support the sister relationship of the Xylariaceae (Voglmayr *et al.*, 2018) (containing Xylariaceae *sensu stricto* and Graphostromataceae) and the

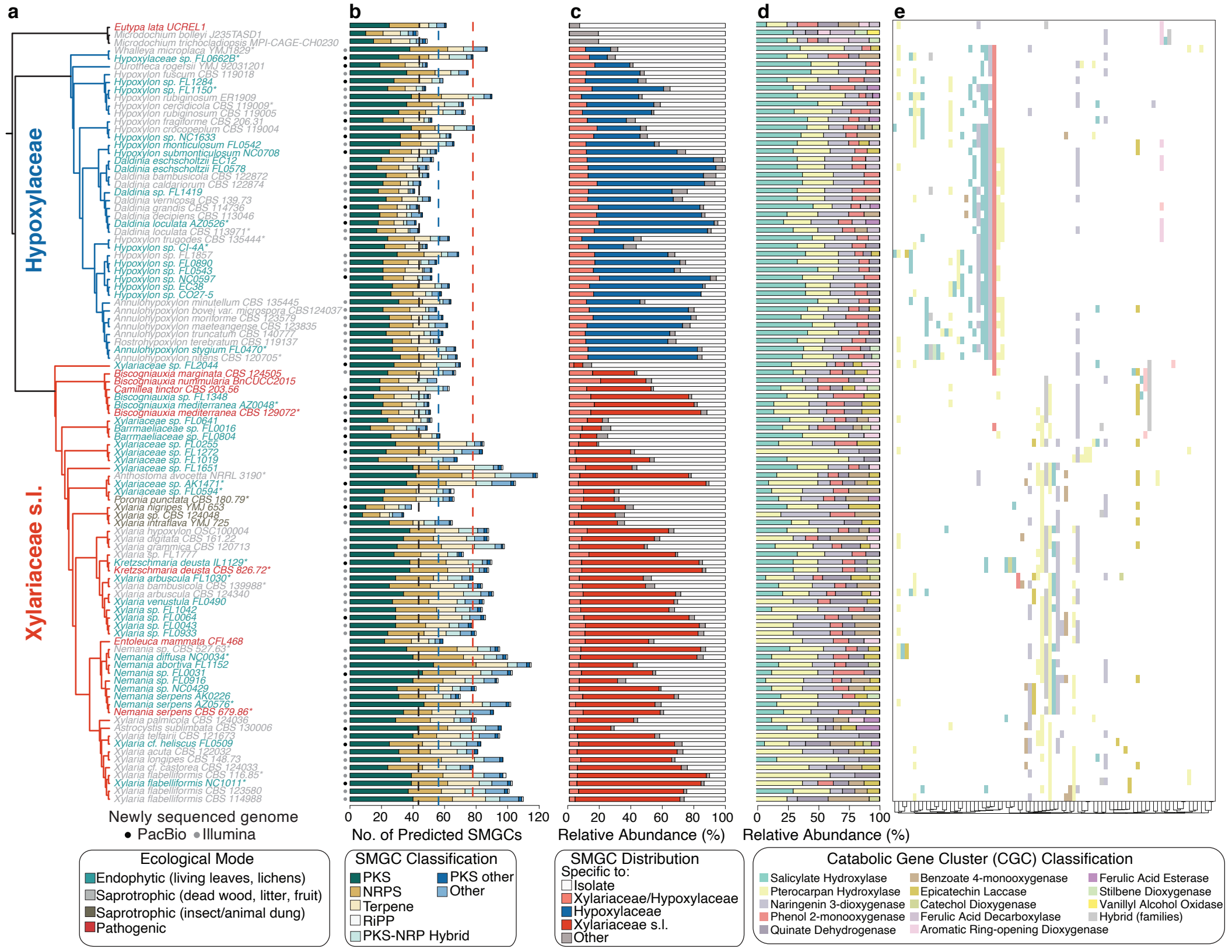
Hypoxylaceae (Wendt *et al.*, 2018) (Fig. 1a; Figs. S2 and S3), as well as previously denoted relationships among genera (U'Ren *et al.*, 2016). Phylogenetic analyses included genomes of 25 outgroup taxa representing five other families of Xylariales and eight orders of Sordariomycetes (total 121 genomes; Fig. S2). Taxon names are colored by ecological mode and branches colored by major clade (red: Xylariaceae; blue: Hypoxylaceae). Taxa with asterisks (*) represent 15 pairs of endophyte/non-endophyte 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 SMGC families per genome. Dotted lines indicate the averages for Pezizomycotina (black), Xylariaceae (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), colored by anchor gene identity (sensu (Gluck-Thaler & Slot, 2018)). Hierarchical clustering of CGCs (see bottom) was performed with the unweighted pair group method with arithmetic mean (UPGMA).

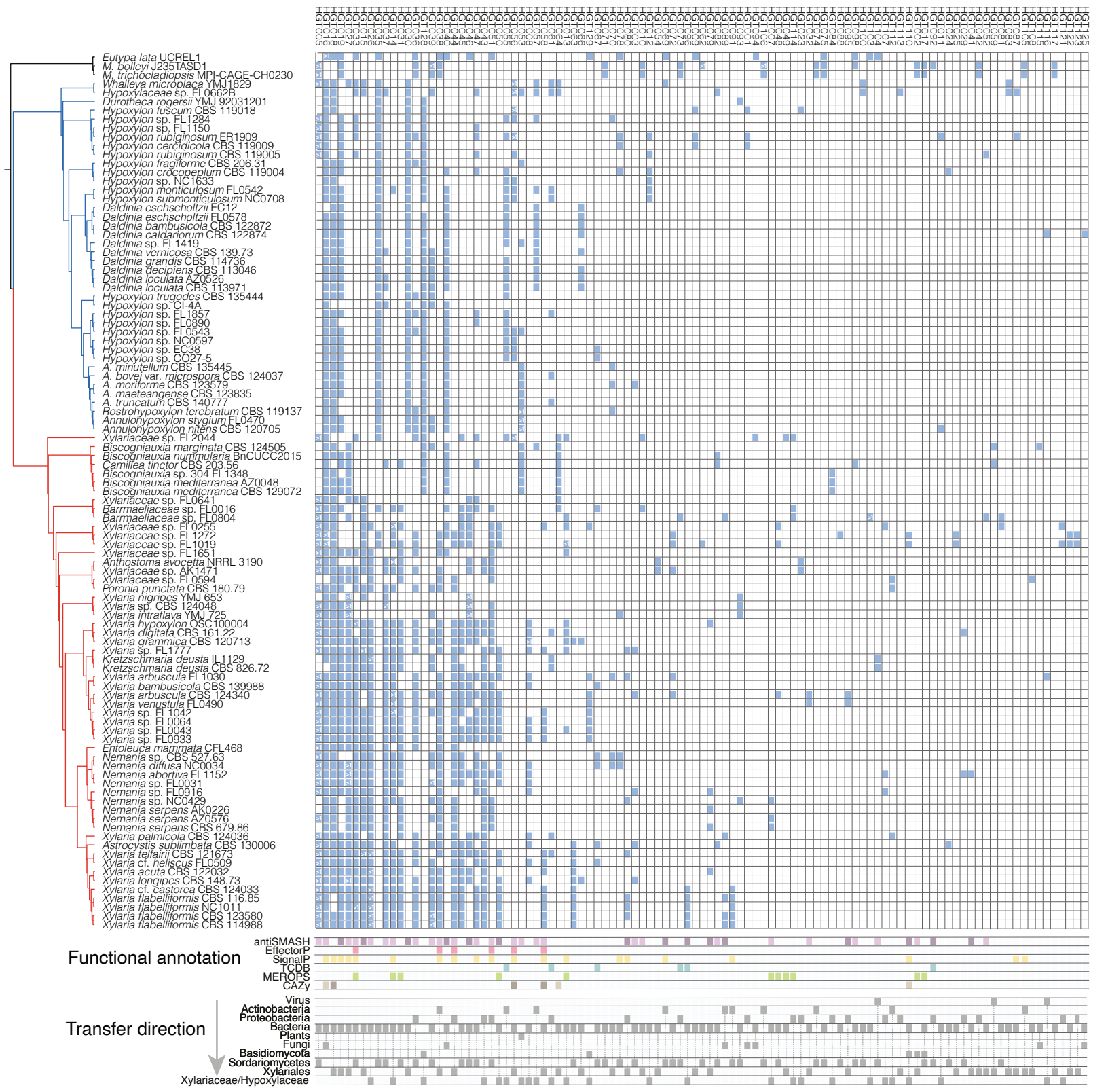
Fig 2. Phylogenetic distribution and functional annotation of high confidence HGTs to genomes of Xylariaceae and Hypoxylaceae. Phylogeny matches Fig. 1a. Blue boxes represent genes predicted to be high-confidence HGT events (detected with the first round of Alien Index analyses; Table S5). HGT events are ordered from left to right based on their abundance. Transfers with more than one gene copy per genome are indicated with >1. Functional annotations (bottom) are based on antiSMASH, EffectorP, SignalP, TCDB, MEROPS, and CAZyme. 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, dark brown color indicates plant cell wall-degrading carbohydrate-active enzyme

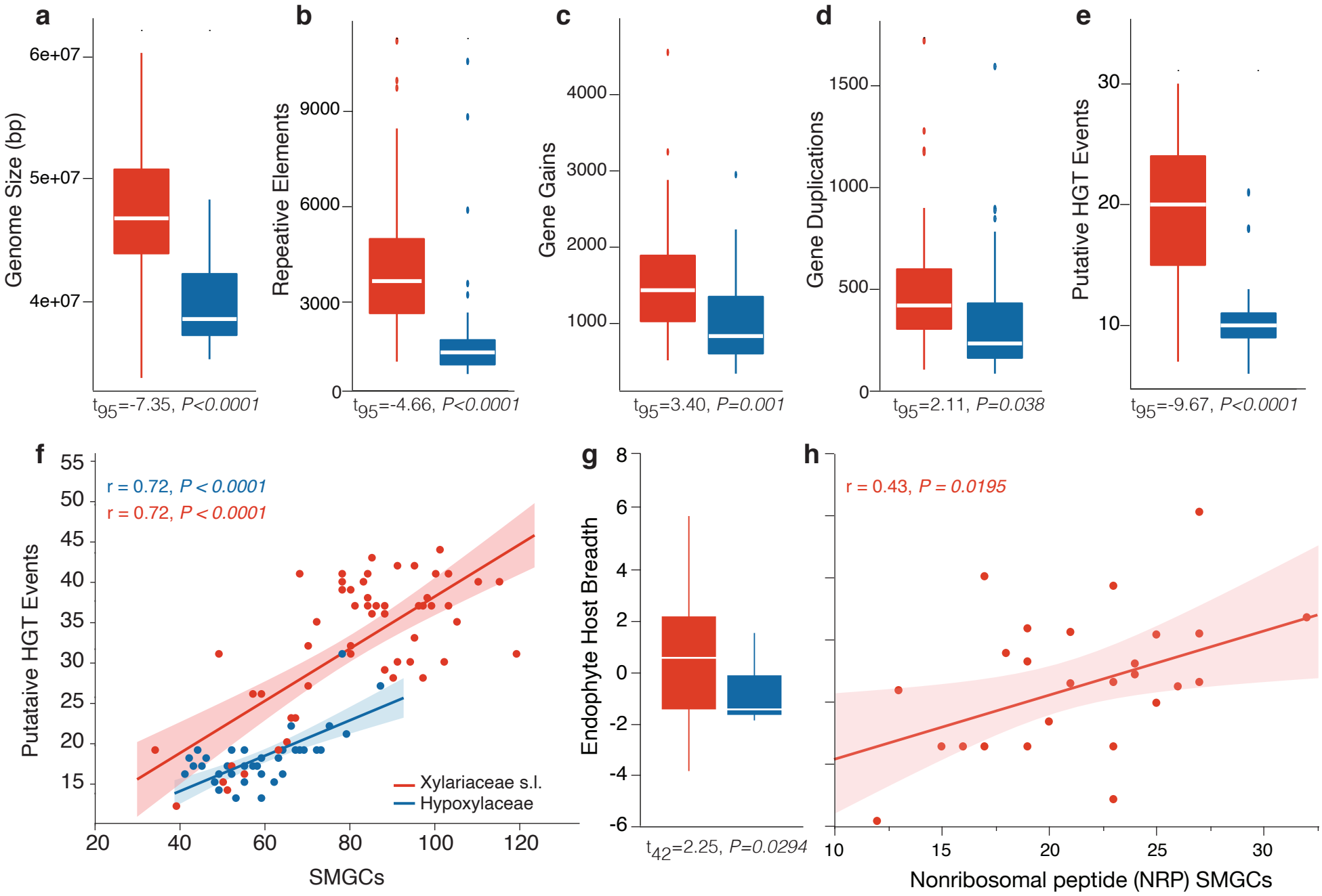
1010 domains (PCWDs). The bottom panel (Transfer Direction) indicates the taxonomic identity of
1011 putative donor and recipient lineages inferred from phylogenetic analyses.
1012
1013 **Fig 3. Larger genomes in the Xylariaceae clade reflect increased repetitive regions, gene**
1014 **gains and duplications, and HGTs.** Median (**a**) genome size, (**b**) repetitive element content, (**c**)
1015 gene gains, (**d**) gene duplications, and (**e**) number of putative HGT events (high confidence only)
1016 for genomes of Xylariaceae (red) and Hypoxylaceae (blue). Box plot boundaries reflect the
1017 interquartile range. Summary statistics (averages, standard deviations, and sample sizes) are
1018 reported in Table S6. Gene gains/losses were inferred with Wagner Parsimony under a gain
1019 penalty=loss penalty=1; (**f**) Relationship between the number of HGT events and SMGCs as a
1020 function of clade (Pearson correlation for each clade was the same; r = 0.72, P<0.0001); (**g**) A
1021 quantile box plot showing the interquartile range and median of endophyte host breadth
1022 (measured as total number of plant families and lichen orders with which a fungal OTU was
1023 cultured; see (U'Ren *et al.*, 2016)) as a function of major clade (color). A similar pattern was
1024 observed when only the number of plant families are compared (Wilcoxon: χ^2 = 4.14, P=0.0413),
1025 but not lichen orders (Wilcoxon: χ^2 = 1.77, P=0.1834). **(h)** Relationship of Xylariaceae
1026 endophyte host breadth and the number of SMGCs classified as NRPS.
1027
1028 **Fig 4. Pairwise comparisons of sister taxa illustrate ecological modes are more distinct in**
1029 **the Hypoxylaceae.** Box plots of the median and interquartile difference in gene counts of
1030 PCWDEs, peptidases, SMGCs (y-axis on left), and transporters (y-axis on right) between 15
1031 pairs of sister taxa with contrasting ecological modes for Xylariaceae and Hypoxylaceae (sister
1032 taxa are indicated with asterisks in Fig. 1a). Values greater than zero indicate higher gene counts

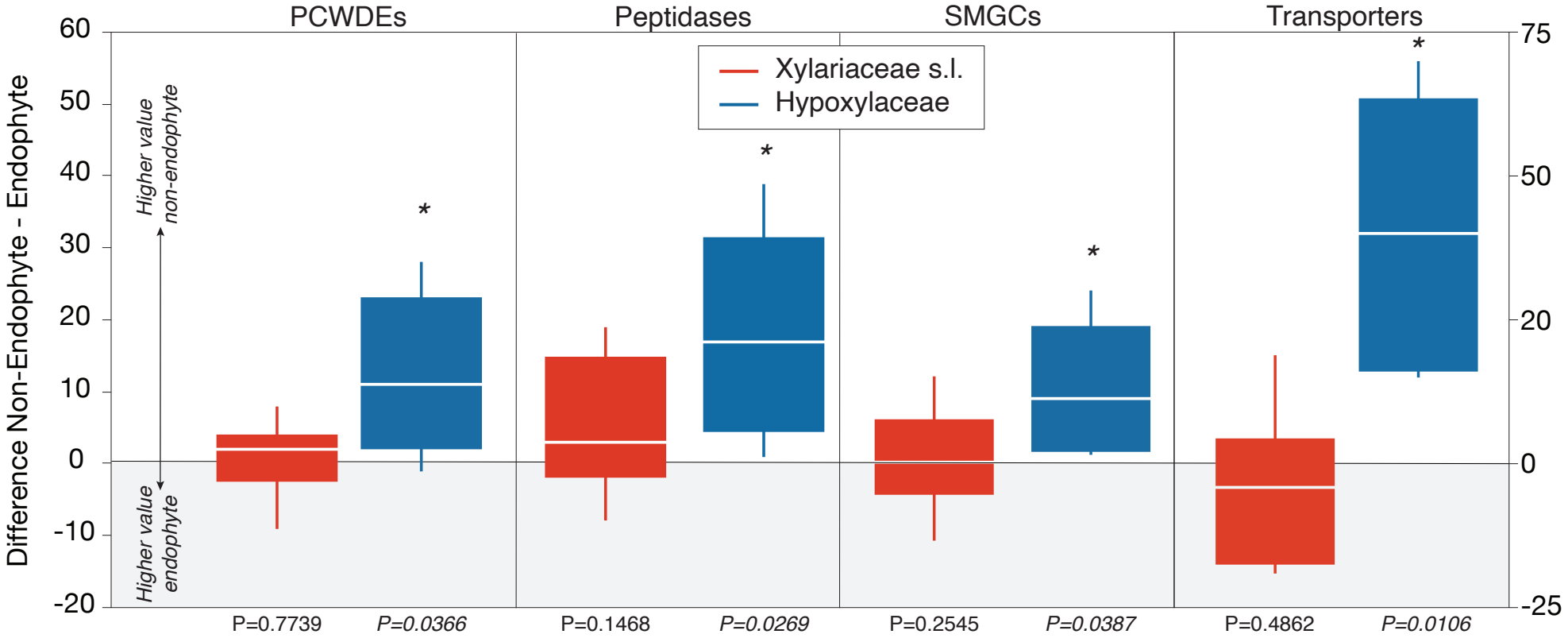
in non-endophytic taxa, whereas differences less than zero indicate higher gene counts in endophytes. Statistical differences were assessed with least squares means contrast under the null hypothesis: non-endophyte value - endophyte value = 0 (see Table S6 for summary statistics). P-values <0.05 are indicated with an asterisk (*).

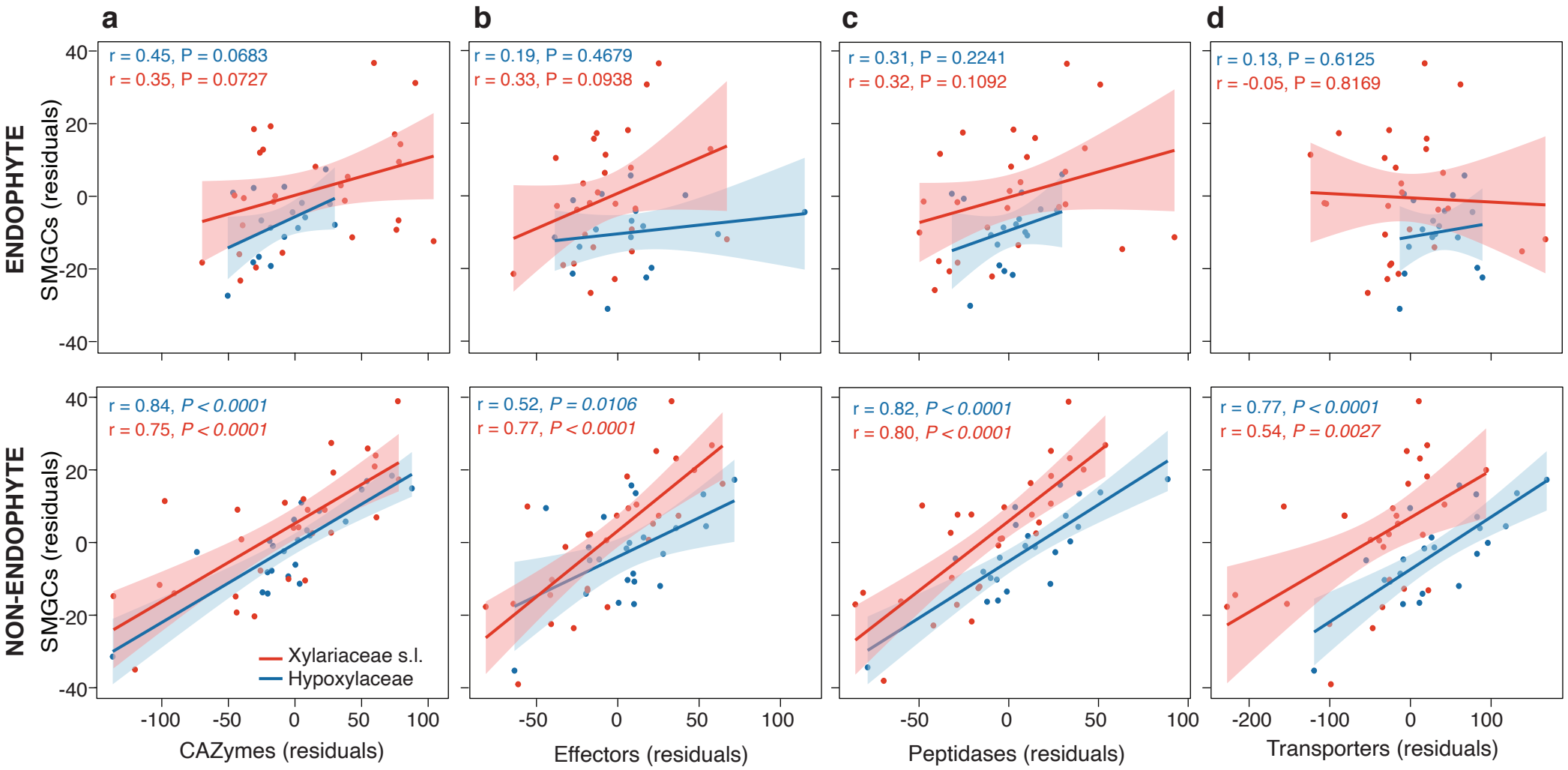
Fig. 5. Non-endophyte genomes display correlation of SMGC content and genes involved in saprotrophy and/or pathogenicity. Relationship between SMGC abundance and number of genes annotated as (**a**) CAZymes, (**b**) effectors, (**c**) peptidases, and (**d**) transporters for endophytes (top row) and non-endophytes (bottom row). Values for each genome represent the residuals after accounting for genome size. Points, linear regression line, and shaded 95% confidence intervals of fit are color-coded by clade (red, Xylariaceae; blue Hypoxylaceae). Statistical values represent Pearson correlation coefficient. See Table S6 for additional details.


a
b
c
d
e
Hypoxylaceae
Xylariaceae s.l.
Eutypa lata UCREL1
Microdochium bolleyi J235TASD1
Microdochium trichocladiopsis MPI-CAGE-CH0230
Whalleya microplaca YMJ829*
Hypoxylaceae sp. FL0662B*
Durotheca rogersii YMJ 92031201
Hypoxylon fuscum CBS 119018
Hypoxylon sp. FL1284
Hypoxylon sp. FL1150*
Hypoxylon rubiginosum ER1909
Hypoxylon cercidicola CBS 119009*
Hypoxylon rubiginosum CBS 119005
Hypoxylon fragiforme CBS 206.31
Hypoxylon crocopeplum CBS 119004
Hypoxylon sp. NC1633
Hypoxylon monticulosum FL0542
Hypoxylon submonticulosum NC0708
Daldinia eschscholtzii EC12
Daldinia eschscholtzii FL0578
Daldinia bambusicola CBS 122872
Daldinia caldariorum CBS 122874
Daldinia sp. FL1419
Daldinia vernicosa CBS 139.73
Daldinia grandis CBS 114736
Daldinia decipiens CBS 113046
Daldinia loculata AZ0526*
Daldinia loculata CBS 113971*
Hypoxylon trugodes CBS 135444*
Hypoxylon sp. CI-4A*
Hypoxylon sp. FL1857
Hypoxylon sp. FL0890
Hypoxylon sp. FL0543
Hypoxylon sp. NC0597
Hypoxylon sp. EC38
Hypoxylon sp. CO27-5
Annulohypoxylon minutellum CBS 135445
Annulohypoxylon bovei var. microspora CBS124037*
Annulohypoxylon moriforme CBS 123579
Annulohypoxylon maeteangense CBS 123835
Annulohypoxylon truncatum CBS 140777
Rostrohypoxylon terebratum CBS 119137
Annulohypoxylon stygium FL0470*
Annulohypoxylon nitens CBS 120705*
Xylariaceae sp. FL2044
Biscogniauxia marginata CBS 124505
Biscogniauxia nummularia BnCUCC2015
Camillea tinctor CBS 203.56
Biscogniauxia sp. FL1348
Biscogniauxia mediterranea AZ0048*
Biscogniauxia mediterranea CBS 129072*
Xylariaceae sp. FL0641
Barrmaeliaceae sp. FL0016
Barrmaeliaceae sp. FL0804
Xylariaceae sp. FL0255
Xylariaceae sp. FL1272
Xylariaceae sp. FL1019
Xylariaceae sp. FL1651
Anthostoma avocetta NRRL 3190*
Xylariaceae sp. AK1471*
Xylariaceae sp. FL0594*
Poronia punctata CBS 180.79*
Xylaria nigripes YMJ 653
Xylaria sp. CBS 124048
Xylaria intraflava YMJ 725
Xylaria hypoxylon OSC100004
Xylaria digitata CBS 161.22
Xylaria grammica CBS 120713
Xylaria sp. L177
Kretzschmaria deusta IL1129*
Kretzschmaria deusta CBS 826.72*
Xylaria arbuscula FL1030*
Xylaria bambusicola CBS 139988*
Xylaria arbuscula CBS 124340
Xylaria venustula FL0490
Xylaria sp. FL1042
Xylaria sp. FL0064
Xylaria sp. FL0043
Xylaria sp. FL0933
Entoleuca mammata CFL468
Nemania sp. CBS 527.63*
Nemania diffusa NC0034*
Nemania abortiva FL1152
Nemania sp. FL0031
Nemania sp. FL0916
Nemania sp. NC0429
Nemania serpens AK0226
Nemania serpens AZ0576*
Nemania serpens CBS 679.86*
Xylaria palmicola CBS 124036
Astrocystis sublimbata CBS 130006
Xylaria telfairii CBS 121673
Xylaria cf. heliscus FL0509
Xylaria acuta CBS 122032
Xylaria longipes CBS 148.73
Xylaria cf. castorea CBS 124033
Xylaria flabelliformis CBS 116.85*
Xylaria flabelliformis NC1011*
Xylaria flabelliformis CBS 123580
Xylaria flabelliformis CBS 114988
Newly sequenced genome
• PacBio • Illumina
0 20 40 60 80 100 120
No. of Predicted SMGCs
0 20 40 60 80 100
Relative Abundance (%)
0 25 50 75 100
Relative Abundance (%)
Ecological Mode
Endophytic (living leaves, lichens)
Saprotrophic (dead wood, litter, fruit)
Saprotrophic (insect/animal dung)
Pathogenic
SMGC Classification
PKS
NRPS
Terpene
RiPP
PKS-NRP Hybrid
PKS other
Other
SMGC Distribution
Specific to:
Isolate
Xylariaceae/Hypoxylaceae
Hypoxylaceae
Xylariaceae s.l.
Other
Catabolic Gene Cluster (CGC) Classification
Salicylate Hydroxylase
Pterocarpan Hydroxylase
Naringenin 3-dioxygenase
Phenol 2-monooxygenase
Quinate Dehydrogenase
Benzoate 4-monooxygenase
Epicatechin Laccase
Catechol Dioxygenase
Ferulic Acid Decarboxylase
Aromatic Ring-opening Dioxygenase
Ferulic Acid Esterase
Stilbene Dioxygenase
Vanillyl Alcohol Oxidase
Hybrid (families)


Eutypa lata UCREL1
M. bolleyi J235TASD1
M. trichocladiopsis MPI-CAGE-CH0230
Whalleya microplaca YMJ829
Hypoxylaceae sp. FL0662B
Durotheca rogersii YMJ 92031201
Hypoxylon fuscum CBS 119018
Hypoxylon sp. FL1284
Hypoxylon sp. FL1150
Hypoxylon rubiginosum ER1909
Hypoxylon cercidicola CBS 119009
Hypoxylon rubiginosum CBS 119005
Hypoxylon fragiforme CBS 206.31
Hypoxylon crocopeplum CBS 119004
Hypoxylon sp. NC1633
Hypoxylon monticulosum FL0542
Hypoxylon submonticulosum NC0708
Daldinia eschscholtzii EC12
Daldinia eschscholtzii FL0578
Daldinia bambusicola CBS 122872
Daldinia caldariorum CBS 122874
Daldinia sp. FL1419
Daldinia vernicosa CBS 139.73
Daldinia grandis CBS 114736
Daldinia decipiens CBS 113046
Daldinia loculata AZ0526
Daldinia loculata CBS 113971
Hypoxylon trugodes CBS 135444
Hypoxylon sp. CI-4A
Hypoxylon sp. FL1857
Hypoxylon sp. FL0890
Hypoxylon sp. FL0543
Hypoxylon sp. NC0597
Hypoxylon sp. EC38
Hypoxylon sp. CO27-5
A. minutellum CBS 135445
A. bovei var. microspora CBS 124037
A. moriforme CBS 123579
A. maeteangense CBS 123835
A. truncatum CBS 140777
Rostrohypoxylon terebratum CBS 119137
Annulohypoxylon stygium FL0470
Annulohypoxylon nitens CBS 120705
Xylariaceae sp. FL2044
Biscogniauxia marginata CBS 124505
Biscogniauxia nummularia BnCUCC2015
Camillea tinctor CBS 203.56
Biscogniauxia sp. 304 FL1348
Biscogniauxia mediterranea AZ0048
Biscogniauxia mediterranea CBS 129072
Xylariaceae sp. FL0641
Barrmaeliaceae sp. FL0016
Barrmaeliaceae sp. FL0804
Xylariaceae sp. FL0255
Xylariaceae sp. FL1272
Xylariaceae sp. FL1019
Xylariaceae sp. FL1651
Anthostoma avocetta NRRL 3190
Xylariaceae sp. AK1471
Xylariaceae sp. FL0594
Poronia punctata CBS 180.79
Xylaria nigripes YMJ 653
Xylaria sp. CBS 124048
Xylaria intraflava YMJ 725
Xylaria hypoxylon OSC100004
Xylaria digitata CBS 161.22
Xylaria grammica CBS 120713
Xylaria sp. FL1777
Kretzschmaria deusta IL1129
Kretzschmaria deusta CBS 826.72
Xylaria arbuscula FL1030
Xylaria bambusicola CBS 139988
Xylaria arbuscula CBS 124340
Xylaria venustula FL0490
Xylaria sp. FL1042
Xylaria sp. FL0064
Xylaria sp. FL0043
Xylaria sp. FL0933
Entoleuca mammata CFL468
Nemania sp. CBS 527.63
Nemania diffusa NC0034
Nemania abortiva FL1152
Nemania sp. FL0031
Nemania sp. FL0916
Nemania sp. NC0429
Nemania serpens AK0226
Nemania serpens AZ0576
Nemania serpens CBS 679.86
Xylaria palmicola CBS 124036
Astrocystis sublimbata CBS 130006
Xylaria telfairii CBS 121673
Xylaria cf. heliscus FL0509
Xylaria acuta CBS 122032
Xylaria longipes CBS 148.73
Xylaria cf. castorea CBS 124033
Xylaria flabelliformis CBS 116.85
Xylaria flabelliformis NC1011
Xylaria flabelliformis CBS 123580
Xylaria flabelliformis CBS 114988
Functional annotation
antiSMASH
EffectorP
SignalP
TCDB
MEROPS
CAZy
Transfer direction
Virus
Actinobacteria
Proteobacteria
Bacteria
Plants
Fungi
Basidiomycota
Sordariomycetes
Xylariales
Xylariaceae/Hypoxylaceae


a
Genome Size (bp)
6e+07
5e+07
4e+07
t95=-7.35, P<0.0001
b
Repeative Elements
9000
6000
3000
0
t95=-4.66, P<0.0001
c
Gene Gains
4000
3000
2000
1000
t95=3.40, P=0.001
d
Gene Duplications
1500
1000
500
0
t95=2.11, P=0.038
e
Putative HGT Events
30
20
10
t95=-9.67, P<0.0001
f
Putataive HGT Events
r = 0.72, P < 0.0001
r = 0.72, P < 0.0001
Xylariaceae s.l.
Hypoxylaceae
SMGCs
g
Endophyte Host Breadth
t42=2.25, P=0.0294
h
r = 0.43, P = 0.0195
Nonribosomal peptide (NRP) SMGCs


PCWDEs
Peptidases
SMGCs
Transporters
Xylariaceae s.l.
Hypoxylaceae
Difference Non-Endophyte - Endophyte
Higher value non-endophyte
Higher value endophyte
60
50
40
30
20
10
0
-10
-20
75
50
20
0
-25
*
P=0.7739
P=0.0366
P=0.1468
P=0.0269
P=0.2545
P=0.0387
P=0.4862
P=0.0106


a
b
c
d
ENDOPHYTE
NON-ENDOPHYTE
SMGCs (residuals)
r = 0.45, P = 0.0683
r = 0.35, P = 0.0727
r = 0.19, P = 0.4679
r = 0.33, P = 0.0938
r = 0.31, P = 0.2241
r = 0.32, P = 0.1092
r = 0.13, P = 0.6125
r = -0.05, P = 0.8169
r = 0.84, P < 0.0001
r = 0.75, P < 0.0001
r = 0.52, P = 0.0106
r = 0.77, P < 0.0001
r = 0.82, P < 0.0001
r = 0.80, P < 0.0001
r = 0.77, P < 0.0001
r = 0.54, P = 0.0027
Xylariaceae s.l.
Hypoxylaceae
CAZymes (residuals)
Effectors (residuals)
Peptidases (residuals)
Transporters (residuals)