

PNAS

www.pnas.org

Supplementary Information for

An Interpreted Atlas of Biosynthetic Gene Clusters and Their Families from 1000 Fungal Genomes.

Matthew T. Robey, Lindsay K. Caesar, Milton T. Drott, Nancy P. Keller, Neil L. Kelleher*

* Neil Kelleher

Email: n-kelleher@northwestern.edu

This PDF file includes:

Supplementary text
Figures S1 to S25
Tables S1 to S3
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1

Supplementary Information Text

Methods

Genome dataset

A set of 1,037 fungal genome assemblies was downloaded from GenBank and the JGI Genome Portal. The subphylum *Saccharomycotina* was excluded, as a large portion of genomes in this taxonomic group are known to lack biosynthetic gene clusters relevant to this study. Genomes were downloaded 11/15/2018. All genomes were processed using the command-line version of antiSMASH 4 with the fungal setting and otherwise default parameters (1). This resulted in 36,399 predicted biosynthetic gene clusters, to which we added 213 quality fungal gene cluster sequences with known metabolite products from the MIBiG database (2).

Gene cluster family network creation

We utilized a workflow similar to that utilized by the recently-developed tool BiG-SCAPE (3). Predicted biosynthetic gene clusters (GenBank files from antiSMASH) were converted to arrays of predicted protein domains using the HMMER suite tool *hmmsearch* using the Pfam 31.0 database. This database of protein domain Hidden Markov Models was supplemented with models corresponding to fungal polyketide product template domains (TIGR04532) and terpene cyclase domain (NCBI CDD cd00687) based on aligned sequences present in NCBI (4).

The Jaccard similarity between all pairs of gene cluster domain arrays was calculated using the Python package scikit-learn, and any pairs with Jaccard similarity less than 0.1 were excluded from further analysis (3, 5). Gene cluster pairs that passed this Jaccard similarity threshold were compared based on sequence identity. Gene clusters were first separated into biosynthetic type based on the presence or absence of protein domains as defined in **Table S3**. This resulted in gene clusters classified as nonribosomal peptide synthase (NRPS), highly-reducing polyketide synthase (HR-PKS), nonreducing polyketide synthase (NR-PKS), hybrid NRPS-PKS, NRPS-like, dimethylallyl transferase (DMAT), or terpene. We chose to use mutually inclusive classifications (e.g. both NRPS and DMAT for a single cluster), which enabled downstream analysis in cases where the antiSMASH-predicted boundaries encompassed multiple gene clusters in close chromosomal proximity. We determined sequence identity between all pairs of gene clusters within a given biosynthetic class by performing sequence-profile alignments of backbone enzyme using the HMMER suite tool *hmmalign* (4). For a given pair of gene clusters, sequence identity was calculated as the mean sequence identity between pairs of backbone enzyme domains. In the case of gene clusters with more than one of a given backbone enzyme domain, the Hungarian matching algorithm was used to identify the configuration of domain pairs with the highest sequence similarity, and only these domain pairs were used (3).

A final sequence similarity score was calculated as:

$$\text{Similarity Score} = \sqrt{0.8 * \text{sequence identity} + 0.2 * \text{Jaccard similarity}}$$

DBSCAN clustering was used on the resulting distance matrix, with an epsilon value of 0.50. This workflow thus resulted in a network or graph structure, in which nodes represent gene clusters and subgraphs represent gene cluster families. All data was output to a PostgreSQL database to interface with the web portal. Comparisons of BGC similarity to compound similarity in **Fig. S10** were performed using structures associated with 213 known fungal BGCs from the MIBiG database (2). Structures were converted to chemical fingerprints using the Morgan Fingerprint method with a radius of 2 and compared via Tanimoto similarity as implemented in RDKit (6).

Web portal creation (Prospect: prospect-fungi.com)

The developed web portal uses the C# framework ASP.NET as a backend REST API and the TypeScript framework Angular as a frontend interface. The main access points to the site include a table of gene cluster families (GCFs) and a table of genomes analyzed in this study. Each GCF page, accessible via a unique GCF ID (e.g. NRPS_42), shows gene clusters within the family, along with corresponding taxonomy information. When a gene cluster is selected, a panel shows various metadata corresponding to the cluster and a Cytoscape network for visualizing the GCF. A tab within the panel for Gene Info shows metadata for a selected gene, along with predicted protein domains. Individual protein domains can be selected to view a description of the domain from the Pfam website. Sequences corresponding to the gene cluster, all proteins, and predicted domains are available for download from the site. An additional Organisms page enables similar access to GCFs through specific genomes of interest.

Phylogenetic trees

All species trees were based on the BUSCO fungal dataset (7). The HMMER suite tool hmmsearch was used to extract protein sequences corresponding to BUSCO Hidden Markov Models (HMMs).(4) Extracted sequences were aligned using MAFFT with the –auto parameter.(8) Positions with >10% gaps were removed from the alignments using TrimAL (9). Trimmed alignments were concatenated, resulting in single aligned sequences for each genome. Phylogenetic trees were constructed in MEGA using Neighbor Joining, 500 bootstrap iterations, and the James-Taylor-Thornton model (10).

Cheminformatics Analyses

A database of known metabolites and their biological sources was downloaded from the Natural Products Atlas on December 10, 2020 (11). Compounds were classified by taxonomy using each metabolite record's genus and species information by referencing the NCBI taxonomy database, downloaded October 22, 2019. Predicted chemical ontology information was determined from each compound's InChi representation using the ClassyFire REST API (12). Compounds were converted to chemical fingerprints using the Morgan Fingerprint method with

a radius of 2, as implemented in RDKit (6). The Tanimoto similarity function in RDKit was used to calculate the pairwise distance between compounds (6). A cutoff of 0.6 Tanimoto similarity was used to create compound families within the network. Bacterial and fungal metabolites were compared to FDA-approved compounds from the DrugBank database (13). All chemical descriptors and properties were calculated using RDKit (6).

Principal Component Analysis (PCA)

Principal Component Analyses were conducted on two datasets and data subsets thereof. To evaluate differences in the frequencies of biosynthetic domains across fungi, each fungal gene cluster was converted into an array representing the frequency of 14,350 protein domains across each taxonomic group. Taxonomic groups with less than ten genomes were removed from analysis to avoid skewing frequency calculations. Differences in fungal and bacterial chemical space were evaluated using the frequencies of chemical ontology descriptors in molecules downloaded from the aforementioned Natural Products Atlas chemical database (11), including only taxonomic groups with 5 or more compounds in the downloaded database. Final chemometric analyses were conducted using Sirius version 10.0 (Pattern Recognition Systems AS).

Taxon	NRPS	HYBRID	HRPKS	TERPENE	NRPSLIKE	NRPKS	DMAT	Genomes analyzed	Per-genome average
Pucciniomycotina	2.0	0.0	0.0	0.6	0.6	0.0	0.0	25	3.2
Ustilaginomycotina	3.0	0.6	0.5	0.0	3.6	0.5	0.3	32	8.5
Agaricomycotina	1.6	0.6	0.5	6.1	4.5	0.5	0.2	173	14.1
Pezizomycotina	9.0	5.4	7.8	4.5	7.9	4.0	1.2	721	39.8
Taphrinomycotina	1.1	0.0	0.0	0.5	1.1	0.0	0.0	12	2.8
Mucoromycota	1.1	0.2	0.0	1.6	2.6	0.0	0.0	36	5.5
Zoopagomycota	3.8	0.1	0.1	0.3	0.7	0.1	0.0	16	5.1
Blastocladiomycota	1.5	0.0	0.0	0.0	0.5	0.0	0.0	2	2
Chytridiomycota	9.1	0.92	1.3	0.9	1.6	0.7	0.0	12	13.7
Microsporidia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2	0
Cryptomycota	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1	1.0

Table S1. Genomes analyzed in this study and the distribution of their gene clusters classified by biosynthetic type.

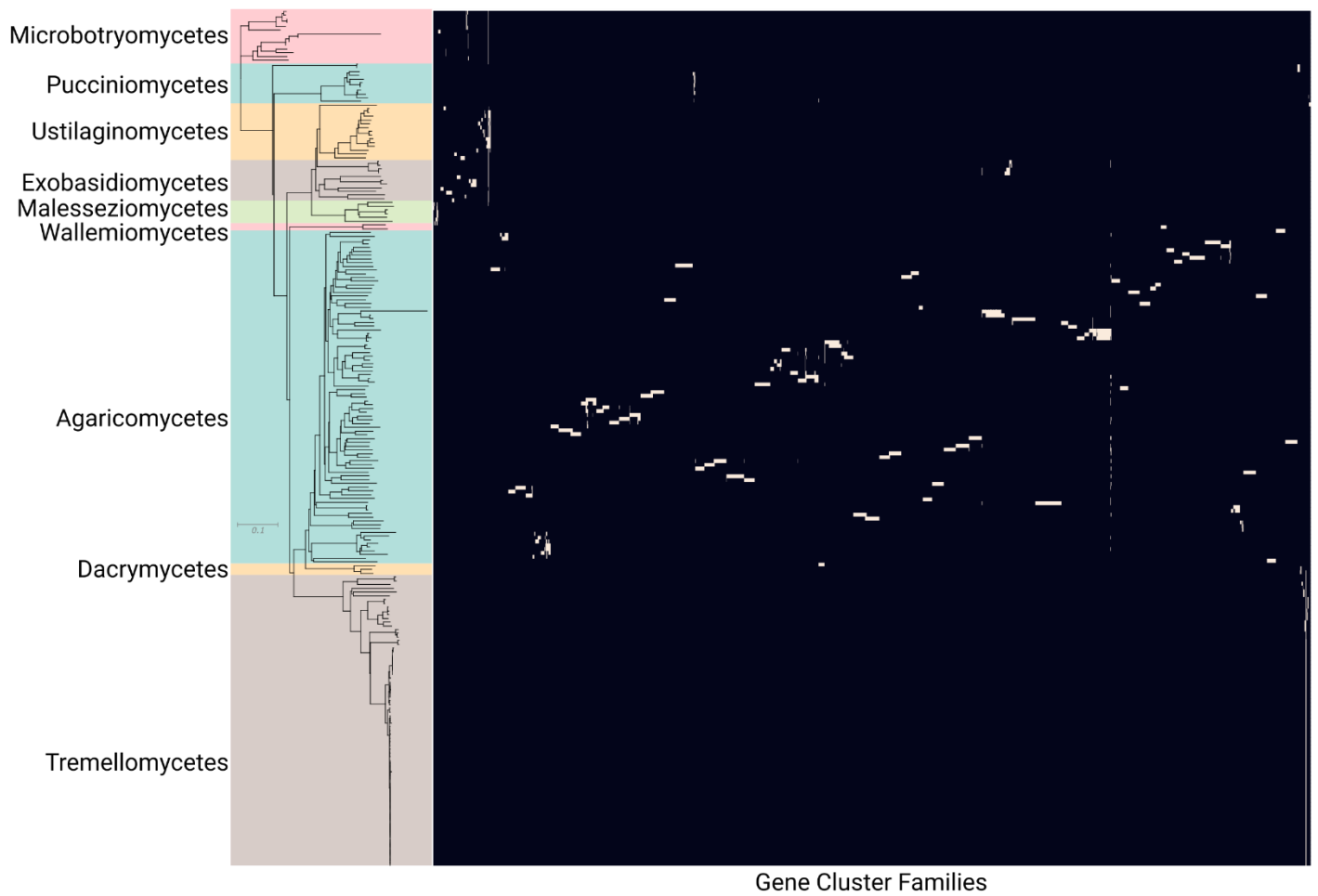


Fig. S1. Distribution of 1933 gene cluster families (GCFs) across Basidiomycota. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes are colored by class, according to NCBI taxonomy information. Genomes within Tremellomycetes are largely composed of subspecies of *Cryptococcus neoformans* and *Cryptococcus gatti* and show little variation in GCF content. Within other classes of Basidiomycota, the majority of GCFs are species- or genus-specific. Several GCFs are distributed across entire classes or shared by organisms within different classes.

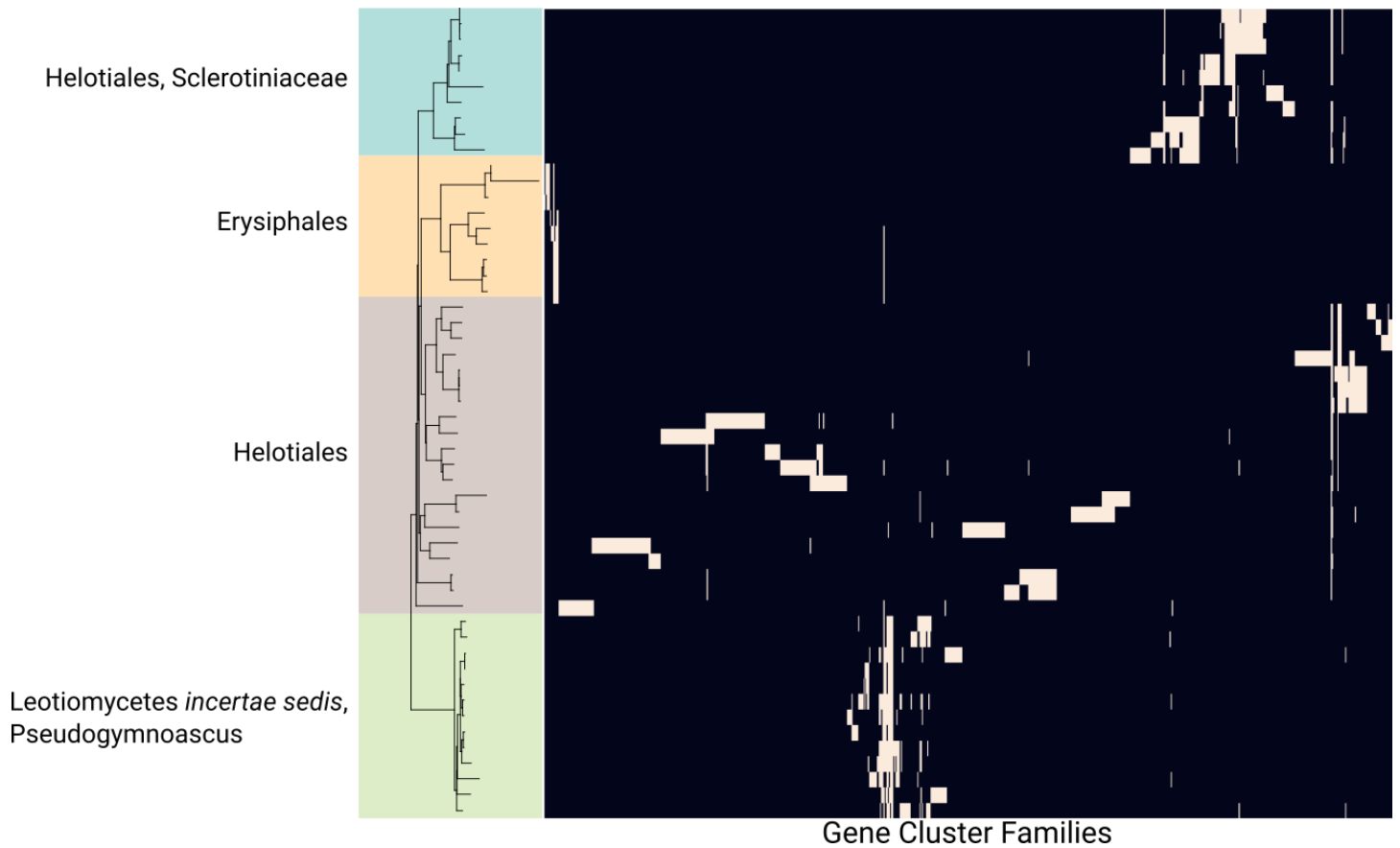


Fig. S2. Distribution of 822 gene cluster families (GCFs) across Leotiomyces. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes are colored by taxonomic order, according to NCBI taxonomy information.

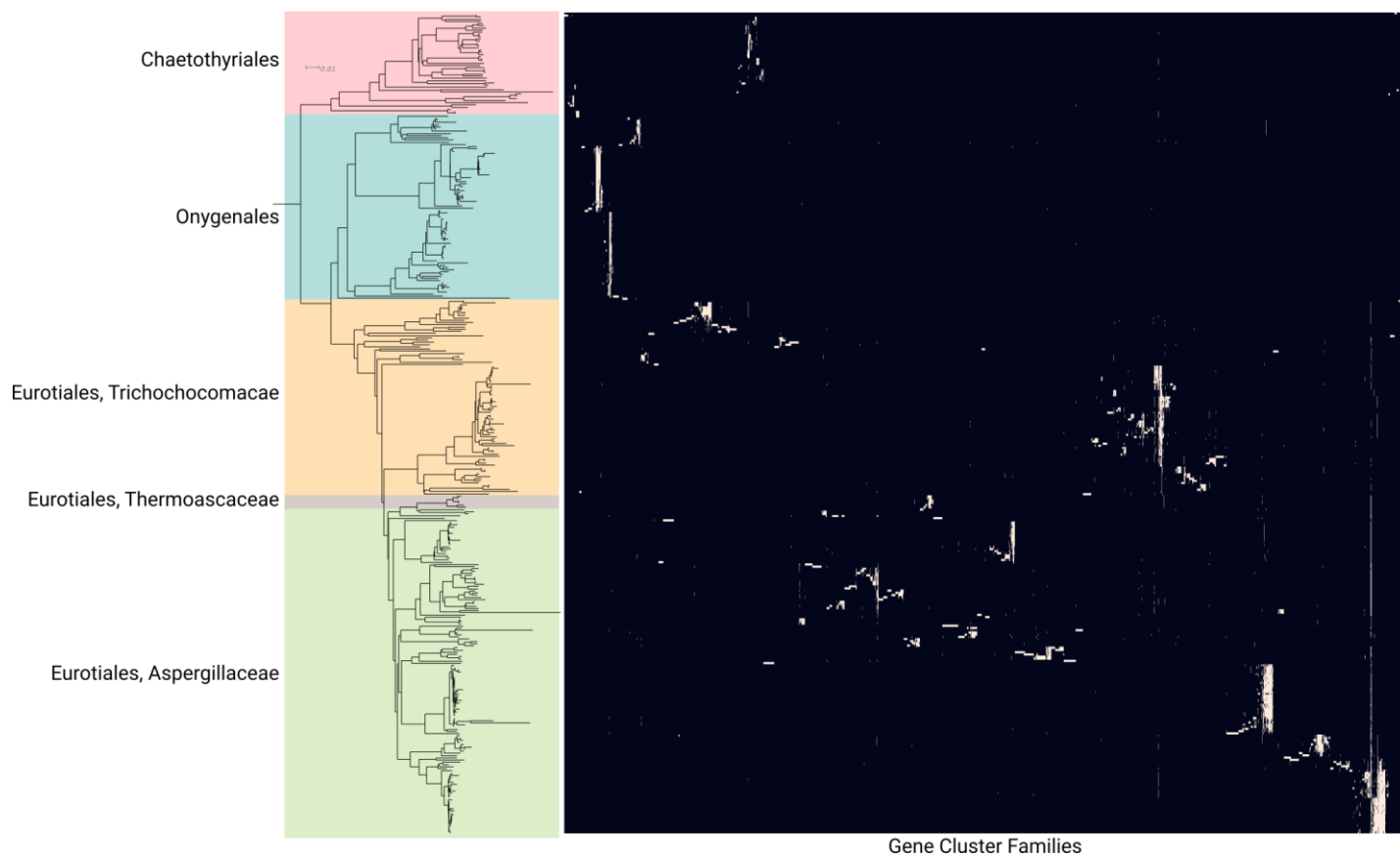


Fig. S3. Distribution of 4926 gene cluster families (GCFs) across Eurotiomycetes. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes are colored by taxonomic order, according to NCBI taxonomy information.

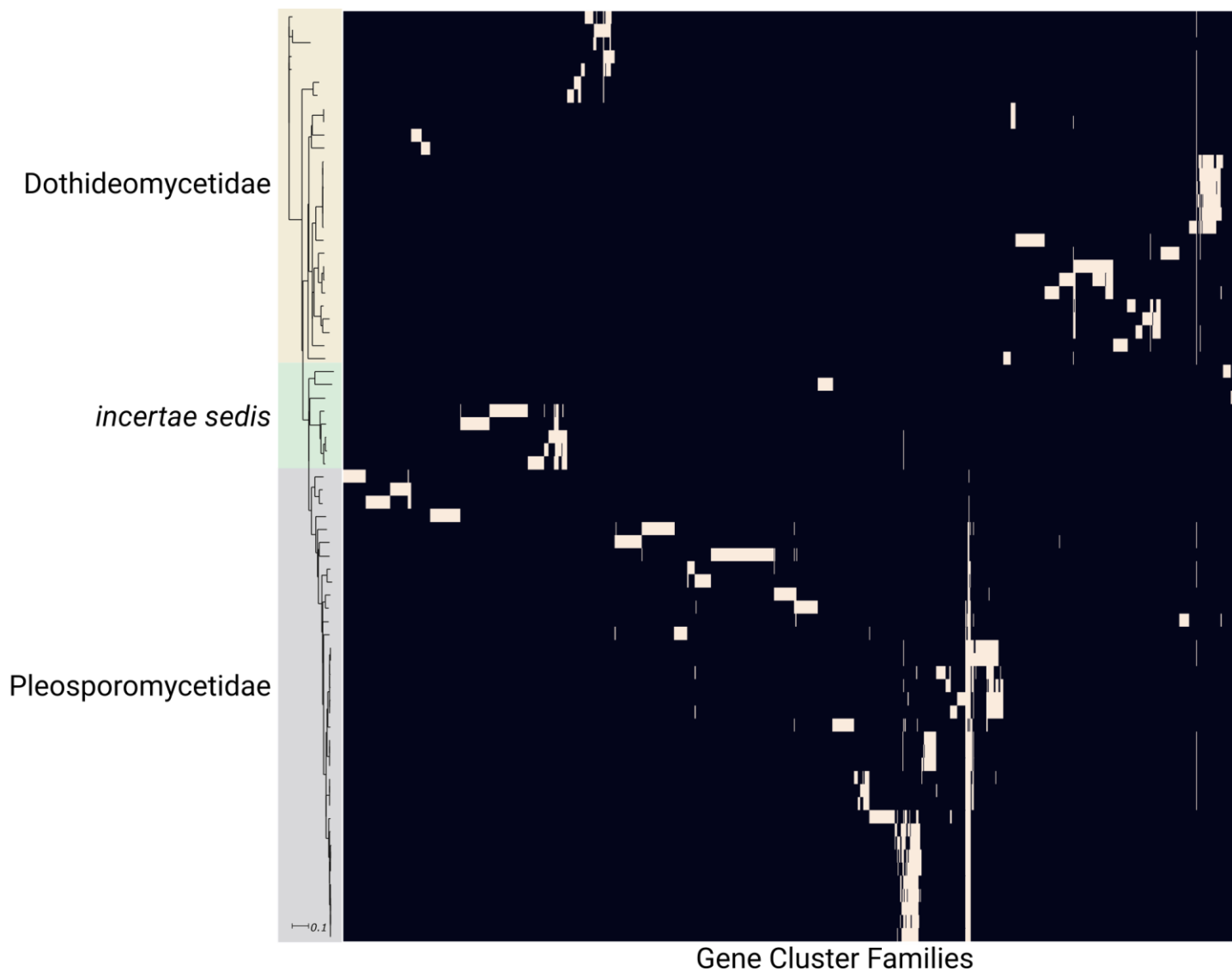


Fig. S4. Distribution of 1176 gene cluster families (GCFs) across Dothideomycetes. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes are colored by taxonomic order, according to NCBI taxonomy information.

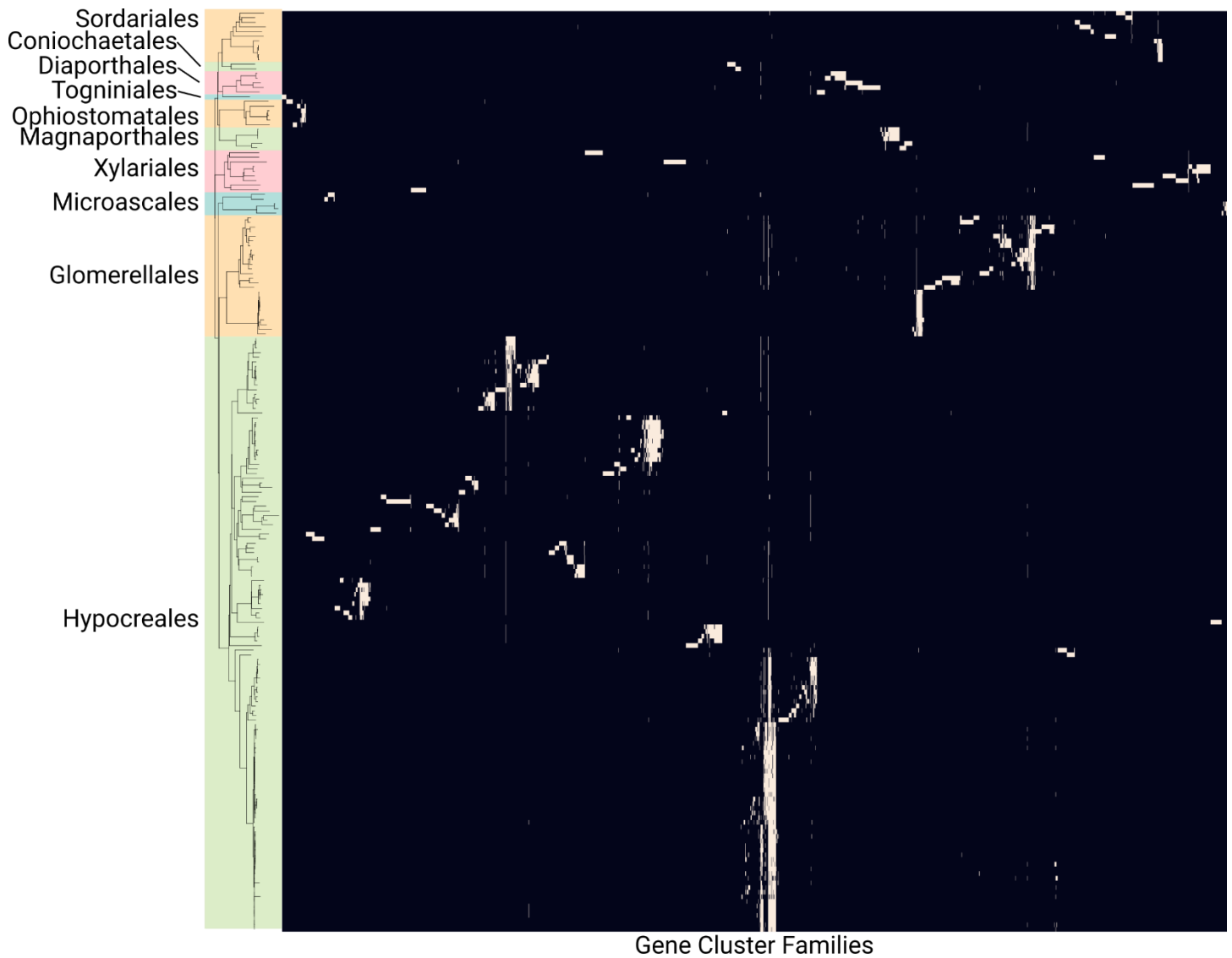


Fig. S5. Distribution of 2884 gene cluster families (GCFs) across Sordariomycetes. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes are colored by taxonomic order, according to NCBI taxonomy information.

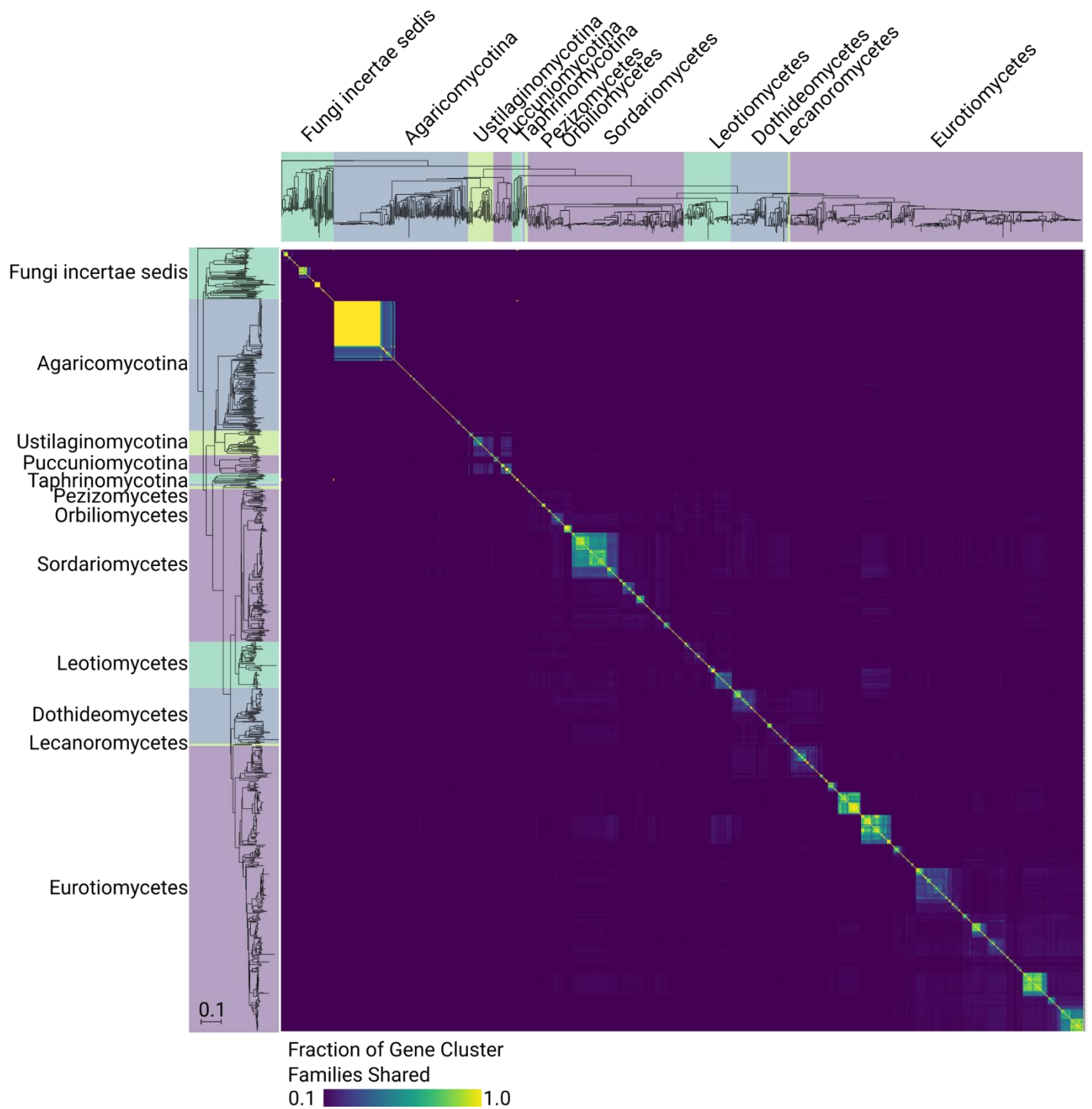


Fig. S6. Relationship between phylogeny and shared gene cluster family (GCF) content. The phylogram to the left shows a Neighbor Joining tree based on 290 orthologous genes, with branches with less than 50% bootstrap support collapsed. Genomes within Pezizomycotina are labeled by taxonomic class, according to NCBI taxonomy information. Other genomes are labeled by subphylum, according to NCBI taxonomy information.

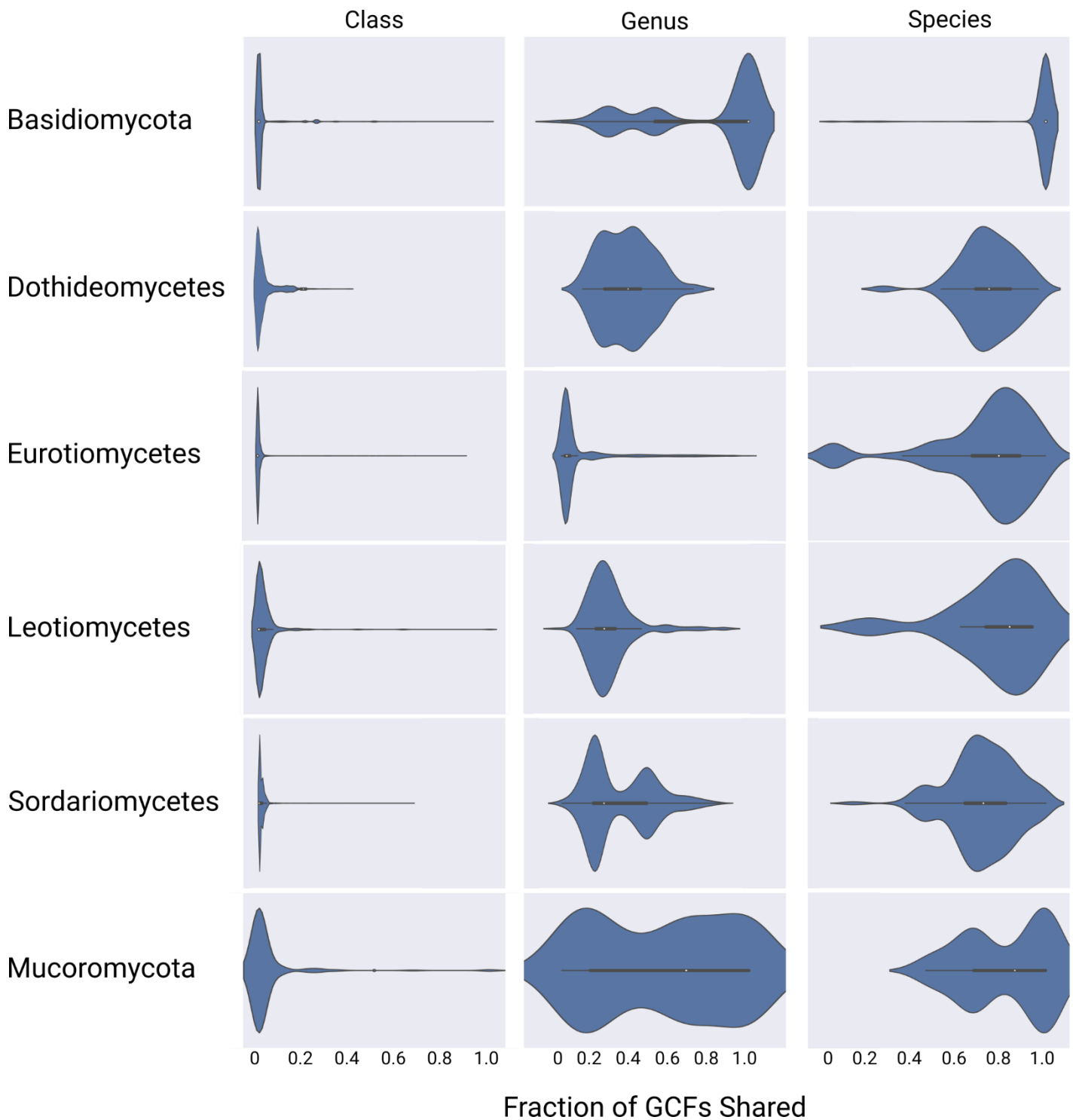
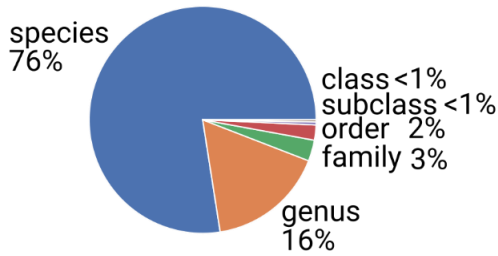
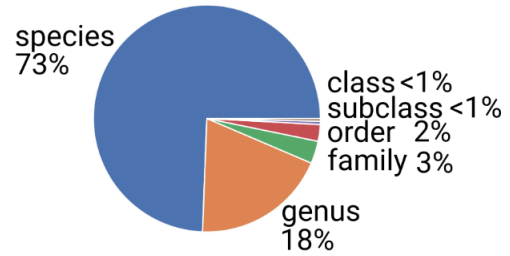


Fig. S7. Relationship between phylogeny and GCFs in six major taxonomic groups. The violin plots represent the fraction of gene cluster families (GCFs) shared by pairs of genomes within the given taxonomic groups. Each genome pair was given a mutually-exclusive classification of same-species, same-genus, or same-class, and the fraction of GCFs shared for each genome pair was determined. Among all taxonomic groups shown, two organisms from the same species will share more GCFs than two organisms from the same genus or class. The fraction of GCFs shared by two organisms within the same genus varies widely by taxonomic group, with organisms from the same genus within Eurotiomycetes averaging ~5% shared GCFs compared to ~95% within Basidiomycota.

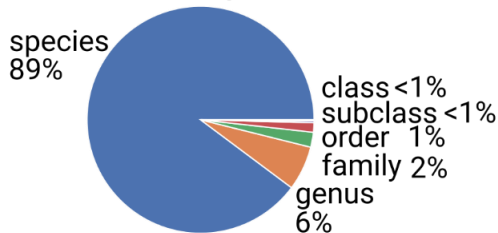
Fungi



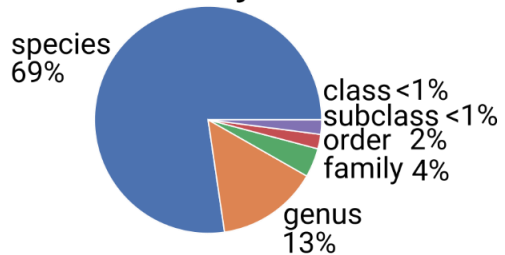
Ascomycota



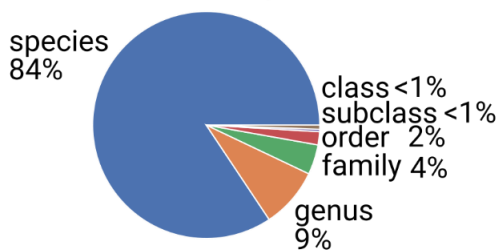
Basidiomycota



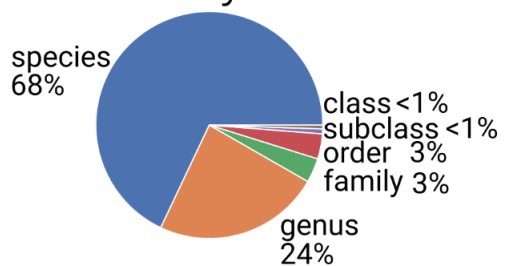
Mucoromycota



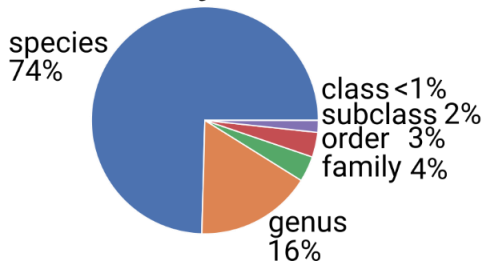
Dothideomycetes



Eurotiomycetes



Leotiomycetes



Sordariomycetes

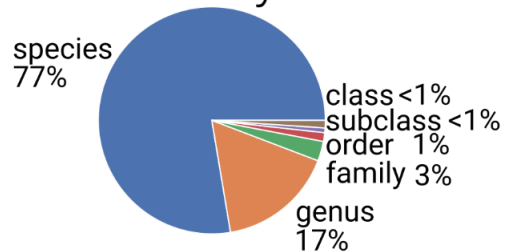


Fig. S8. Fungal gene cluster families (GCFs) are largely species-specific. Each GCF within the given taxonomic group was classified based on highest taxonomic rank shared by organisms with the GCF (i.e. species-specific, genus-specific, family-specific, etc.). Depending on taxonomic group, GCFs are between 68-89% species-specific.

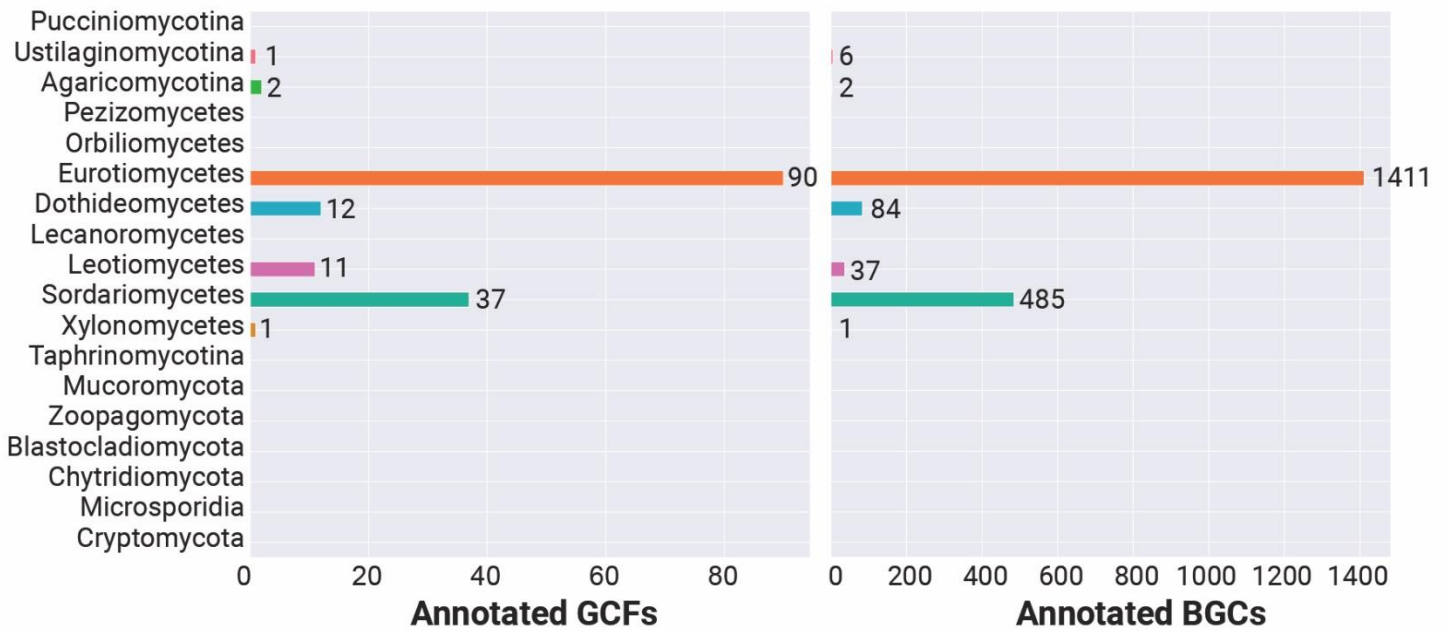


Fig. S9. Using the GCF approach for automated annotation of fungal BGCs with putative metabolite scaffolds. Across the taxonomic groups examined, a total of 154 GCFs contain reference BGCs with known metabolite products. At the level of individual clusters, these amounts to 2,026 BGCs annotated based on their presence in GCFs with known metabolite scaffolds.

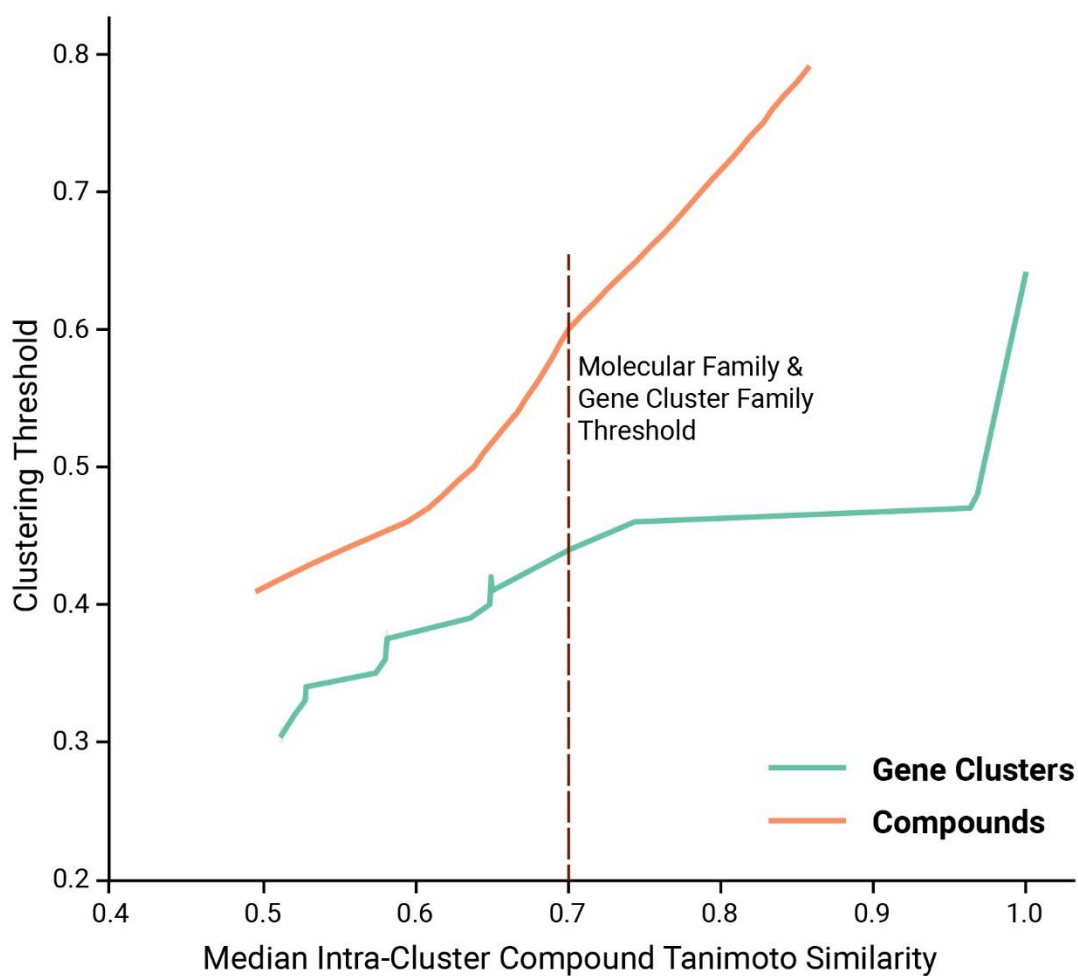


Fig. S10. Comparison of metabolite scaffold chemical space covered by molecular families (MFs) and gene cluster families (GCFs). At each clustering threshold, the median Tanimoto similarity of known compounds within GCFs and MFs was determined. A median intra-cluster Tanimoto similarity of 0.7 was chosen, corresponding to GCF and MF similarity thresholds of 0.45 and 0.6, respectively.

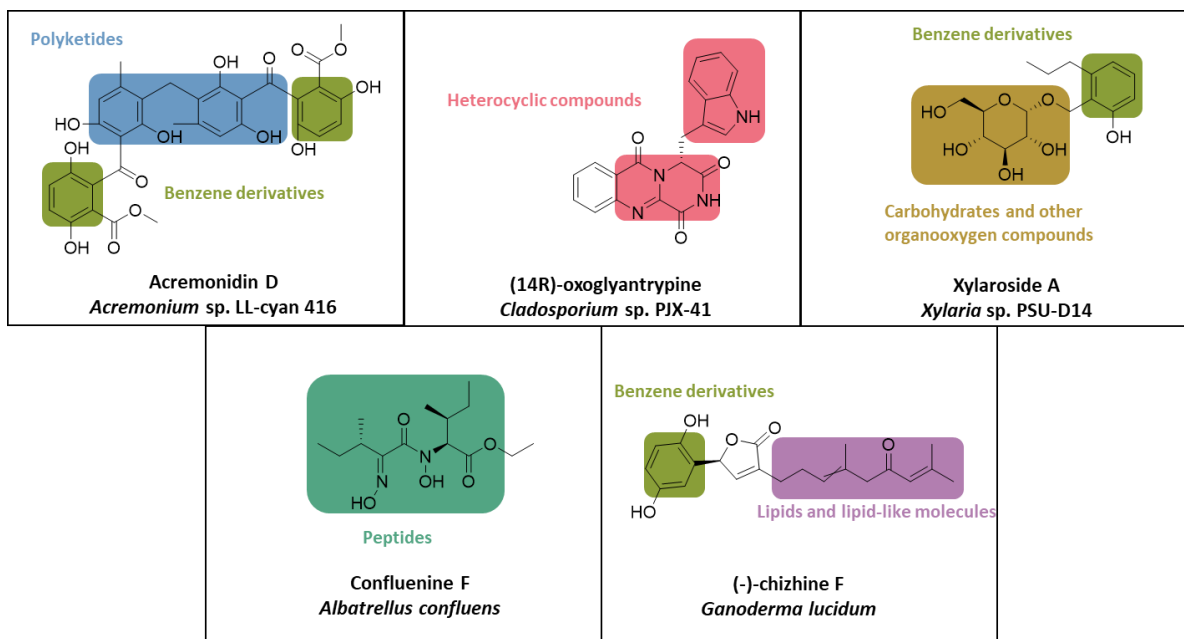


Fig. S11. Chemical ontology-based classification of selected metabolites from diverse fungi. Major ontology superclasses are highlighted for each molecule (14-18).

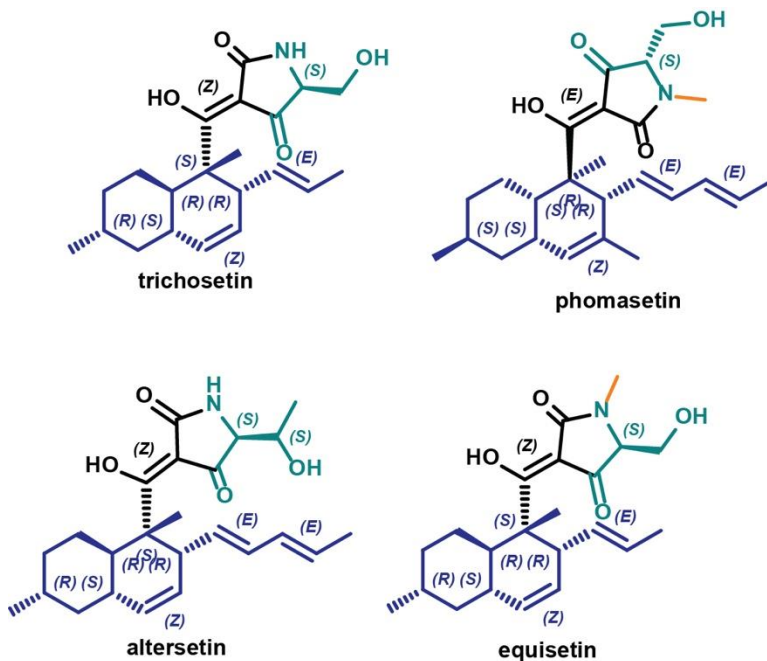


Fig. S12. Compounds from the equisetin structural class that have associated known gene clusters. The scaffold includes a hydrocarbon decalin core varying in methyl and alkenyl substituents and stereochemistry. A tetramic acid moiety derived from serine or threonine is conjugated to the decalin core. N-methylation of the tetramic acid amide is present in equisetin and phomasetin.

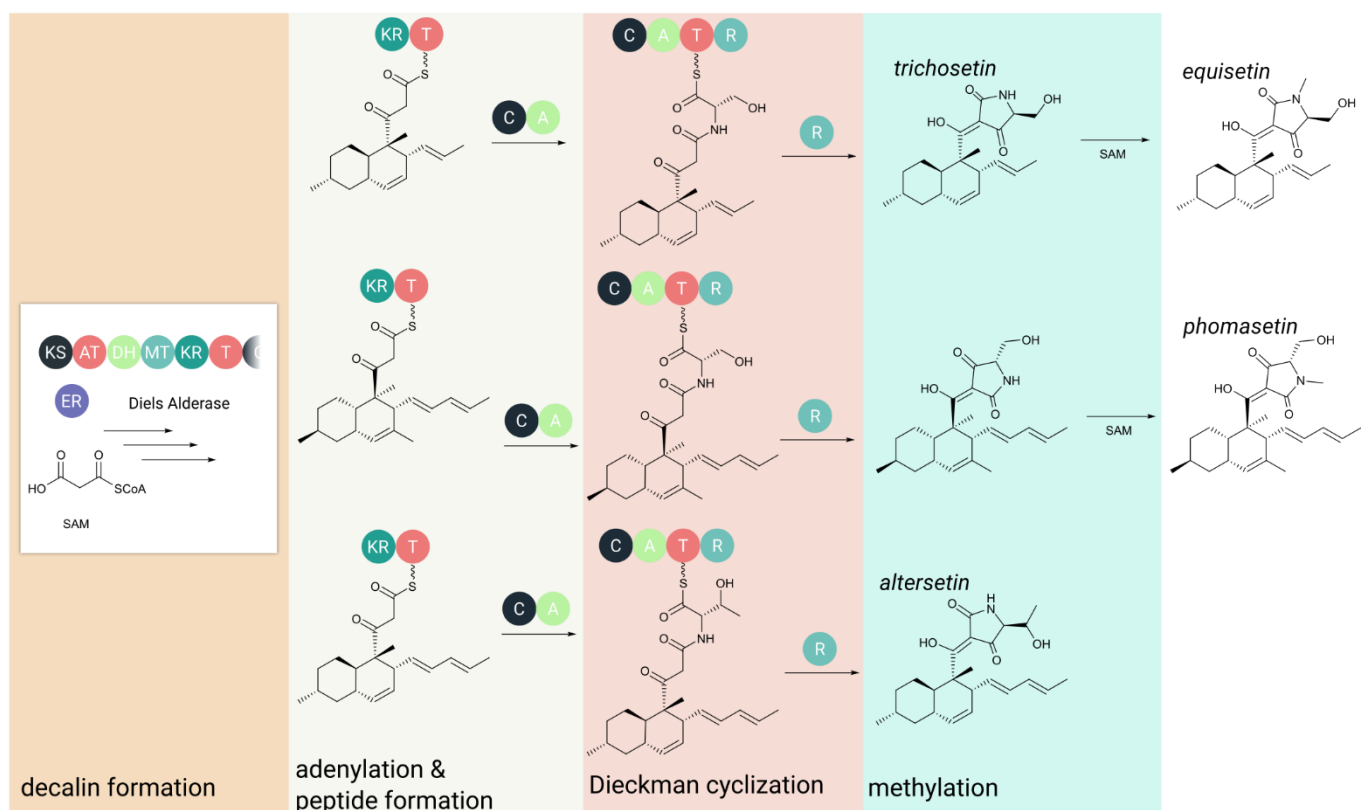


Fig. S13. The biosynthetic pathway for equisetin and related compounds. First the core decalin ring is constructed by a hybrid nonribosomal peptide synthetase-polyketide synthase (NRPS-PKS) enzyme. The PKS domains within the backbone enzyme act in an iterative fashion typical of fungal PKS enzymes, assembling the decalin core from malonyl-CoA monomers. This step is supplemented by the action of a standalone enoyl reductase for ketide monomer reduction and a Diels-Alderase that directs ring closure and controls stereochemistry (19, 20). Second, an NRPS module condenses an amino acid to the decalin core (21). A terminal reductase domain catalyzes Dieckman cyclization to release the intermediate as a tetramic acid, the third step (22). In the final pathway step, a methyltransferase catalyzes N-methylation of the tetramic acid amide (21).

Table S2. The 90 gene cluster families (from total n = 12,067) that are exceptional in that they span multiple taxonomic classes. The Reference column indicates a single GenBank accession number and organism for the backbone enzyme. In cases of multiple backbone enzymes, the provided GenBank reference corresponds to the backbone enzyme in bold text. Abbreviations are as follows: DHONTB, dihydroxy-6-[(3E,5E,7E)-2-oxonona-3,5,7-trienyl]-benzaldehyde; HAS, hexadecahydroastechrome; KS, ketosynthase, AT, acyltransferase; DH, dehydratase; ER, enoyl reductase; KR, ketoreductase; MT, methyltransferase; SAT, starter acyltransferase; PT, product template; A, adenylation; T, thiolation; R, reductase; C, condensation; ICS, isocyanide synthase; DMAT, dimethylallyltransferases; NRPS, nonribosomal peptide synthetase; PKS, polyketide synthase; HRPKS, highly reducing polyketide synthase; NRPKS, nonreducing polyketide synthase; E, Eurotiomycetes; L, Leotiomycetes; S, Sordariomycetes; D, Dothidiomycetes; X, Xylonomycetes; LEC, Lecanoromycetes.

GCF	REFERENCE	BACKBONE	TAXONOMIC CLASSES
HRPKS_30 (DHONTB)	<i>Aspergillus nidulans</i> FGSC A4 (CBF86052)	PKS (KS-AT-DH-ER-KR-T) , PKS (SAT-KS-AT-PT-T-R)	E, S
NRPKS_1343 (BIKAVERIN)	<i>Fusarium fujikuroi</i> (SCO46930.)	PKS (KS-AT-PT-T-TE)	L, S
NRPKS_791 (MELANIN)	<i>Cadophora</i> sp. DSE1049 (PVH73815)	PKS (SAT-KS-AT-PT-T-T-TE)	D, L, S
NRPS_607 (CHAETOGLOBOSIN)	<i>Aspergillus lentulus</i> (GAQ05296)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, E
NRPS_63 (CHRYSOGINE)	<i>Aspergillus arachidicola</i> (PIG85941)	NRPS (C-A-T-C-A-T-C)	E, S
NRPKS_375 (CONIDIAL YELLOW PIGMENT)	<i>Aspergillus sydowii</i> CBS 593.65 (OJJ57401)	PKS (SAT-KS-AT-PT-T)	E, L, LEC, S
NRPS_690 (CYTOCHALASIN)	<i>Aspergillus clavatus</i> NRRL 1 (EAW09117)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, L, S
NRPS_138 (EQUISETIN)	<i>Alternaria alternata</i> (OWY46706)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, E, S
NRPS_123 (FUMITREMORGIN)	<i>Aspergillus fumigatus</i> (OXN23238)	NRPS (A-T-C-A-T-C)	E, L, S
NRPS_1705 (FUMONISIN)	<i>Fusarium verticillioides</i> (RBR13858)	PKS (KS-AT-DH-MT-ER-KR-T)	D, S
NRPS_442 (HAS)	<i>Aspergillus fumigatus</i> (OXN25028)	NRPS (A-T-C-A-T-C-T)	E, S
NRPKS_147 (MELANIN)	<i>Alternaria alternata</i> (OAG24502)	PKS (SAT-KS-AT-PT-T-T-TE)	D, E
NRPS_101 (PHOMASETIN)	<i>Aspergillus clavatus</i> NRRL 1 (EAW07624)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, E, S, X
NRPS_1149 (SERINOCYCLIN)	<i>Metarhizium acridum</i> CQMa 102 (EFY85053)	NRPS (A-T-C-C-A-T-C-A-T-C-A-T-C-C-A-T-C)	L, S
HYBRIDS_151 (SWAINSONINE)	<i>Clohesyomyces aquaticus</i> (ORY11783)	PKS (A-T-KS-AT-KR-T-R)	E, S
NRPS_2042 (UCS1025A)	<i>Oidiodendron maius</i> Zn (KIM94019)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	L, S
DMAT_140	<i>Ophiocordyceps australis</i> (PHH64516)	DMAT	E, S
DMAT_401	<i>Colletotrichum orchidophilum</i> (OHF04557)	PKS (SAT-KS-AT-MT-PT-T-TE), DMAT	L, S
DMAT_411	<i>Cadophora</i> sp. DSE1049 (PVH84683)	DMAT	L, S
HRPKS_1152	<i>Meliniomyces bicolor</i> E (PMD61012)	PKS (KS-AT-DH-MT-ER-KR-T)	L, S
HRPKS_128	<i>Pezoloma ericae</i> (PMD17755)	PKS (KS-AT-DH-MT-ER-KR-T)	E, L
HRPKS_1289	<i>Acremonium chrysogenum</i> ATCC 11550 (KFH46614)	PKS (KS-AT-DH-MT-ER-KR-T-Carnitine_acyltransferase)	L, S
HRPKS_1318	<i>Colletotrichum higginsianum</i> IMI 349063 (OBR06526)	PKS (KS-AT-DH-MT-ER-KR-T-C) , PKS (KS-AT-DH-MT-ER-KR-T)	L, S
HRPKS_159	<i>Penicillium griseofulvum</i> (KXG49005)	PKS (SAT-KS-AT-PT-MT-R) , PKS (KS-AT-DH-MT-ER-KR-T)	E, S
HRPKS_170	<i>Penicillium camemberti</i> (CRL31088)	PKS (KS-AT-DH-MT-ER-KR-T)	E, L
HRPKS_2026	<i>Pseudogymnoascus</i> sp. 05NY08 (OBT71831)	PKS (KS-AT-DH-MT-ER-KR-T-C)	E, L
HRPKS_2048	<i>Coniochaeta pulveracea</i> (RKU46359)	PKS (KS-AT-DH-ER-KR-T)	E, S

HRPKS_2325	<i>Phialocephala scopiformis</i> (KUJ09200)	PKS (SAT-KS-AT-PT-T-T-TE) , PKS (DH-KR)	D, L
HRPKS_2458	<i>Pseudogymnoascus</i> sp. VKM F-103 (KFY80205)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	
HRPKS_3348	<i>Fusarium oxysporum</i> f. sp. <i>cepae</i> (RKK07595)	PKS (KS-AT-DH-MT-KR-T-C-A-T-R)	
HRPKS_3381	<i>Scedosporium apiospermum</i> (KEZ41293)	PKS (KS-AT-DH-MT-KR-T-R)	E, S
HRPKS_3476	<i>Penicillium griseofulvum</i> (KXG49279)	PKS (KS-AT-DH-MT-ER-KR-T-R)	D, E, L
HRPKS_216	<i>Aspergillus sydowii</i> CBS 593.65 (OJJ61536)	NRPS (C-A-T-C-C-A-T-C-A-T-C-A-T-C)	E, L
HRPKS_495	<i>Aspergillus uvarum</i> CBS 121591 (PYH83208)	PKS (KS-AT-DH-ER-KR-T)	E, S
HRPKS_53	<i>Colletotrichum chlorophyti</i> (OLN93260)	PKS (KS-AT-DH-ER-KR-T-R)	E, S
HRPKS_597	<i>Cordyceps</i> sp. RAO-2017 (PHH90746)	PKS (KS-AT-DH-ER-KR-T)	E, S
HRPKS_678	<i>Pseudogymnoascus</i> sp. VKM F-3557 (KFX86927)	PKS (KS-AT-DH-MT-ER-KR-T- Carnitine_acyltransferase)	E, L
HRPKS_694	<i>Phialocephala subalpina</i> (CZR67900)	PKS (KS-AT-DH-MT-ER-KR-T)	E, L
HRPKS_882	<i>Fusarium fujikuroi</i> (SCN83763)	PKS (KS-AT-DH-MT-ER-KR-T)	S, X
HYBRIDS_195	<i>Aspergillus ochraceoroseus</i> IBT 24754 (PTU20620)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, S
HYBRIDS_215	<i>Penicillium camemberti</i> (CRL19370)	PKS (KS-AT-DH-MT-ER-KR-T)	E, S
HYBRIDS_506	<i>Talaromyces stipitatus</i> ATCC 10500 (EED18841)	PKS (KS-AT-DH-MT-ER-KR-T)	E, L
HYBRIDS_9	<i>Penicillium subrubescens</i> (OKP00032)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, E, L
NRPKS_1988	<i>Pseudogymnoascus</i> sp. 23342-1-11 (OBT65120)	PKS (SAT-KS-AT-PT-T)	D, L
NRPKS_250	<i>Aspergillus lentulus</i> (GAQ09994)	PKS (KS-AT-DH-ER-KR-T)	E, L
NRPKS_437	<i>Aspergillus kawachii</i> IFO 4308 (GAA83965)	PKS (A-T-KS-AT-KR-T-R)	
NRPKS_447	<i>Endocarpon pusillum</i> Z07020 (ERF68696)	PKS (A-T-KS-AT-KR-T-R)	
NRPKS_5	<i>Penicillium nalgiovense</i> (OQE96240)	PKS (SAT-KS-AT-PT-T)	E, X
NRPKS_510	<i>Trichoderma asperellum</i> CBS 433.97 (PTB35070)	PKS (SAT-KS-AT-PT-T)	E, S
NRPS_1018	<i>Pseudogymnoascus</i> sp. VKM F-3808 (KFX99775)	NRPS (A-T-C-A-T-R)	D, L
NRPS_1055	<i>Bipolaris victoriae</i> FI3 (EUN25091)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, L
NRPS_1064	<i>Coleophoma cylindrospora</i> (RDW81833)	PKS (SAT-KS-AT-PT-T-T-TE) , NRPS (C-A-T)	E, L
NRPS_111	<i>Aspergillus brasiliensis</i> CBS 101740 (OJJ75537)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R) , NRPS (A-T-C)	E, S
NRPS_1222	<i>Talaromyces stipitatus</i> ATCC 10500 (EED13058)	PKS (KS-AT-DH-MT-ER-KR-T) , NRPS (A-T-C-A-T-C-T)	E, S
NRPS_1295	<i>Penicillium steckii</i> (OQE21884)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, L, S
NRPS_1301	<i>Aspergillus bombycis</i> (OGM48141)	PKS (KS-AT-DH-ER-KR-T-C-A-T-R)	E, S
NRPS_1372	<i>Fusarium avenaceum</i> (KIL86455)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, S
NRPS_1410	<i>Helicocarpus griseus</i> UAMH5409 (PGH19023)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, S
NRPS_1417	<i>Aspergillus heteromorphus</i> CBS 117.55 (PWY81896)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, L
NRPS_151	<i>Madurella mycetomatis</i> (KXX75968)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	
NRPS_1545	<i>Fusarium avenaceum</i> (KIL87829)	PKS (KS-AT-DH-ER-KR-T) , NRPS (T-C-A-T-C-A-T-C-A-T-C-T)	E, L, S
NRPS_1559	<i>Pseudogymnoascus</i> sp. VKM F-3775 (KFY27678)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, L
NRPS_1586	<i>Metarhizium rileyi</i> RCEF 4871 (OAA34246)	NRPS (A-T-C-A-T-C-A-T-R)	E, S

NRPS_2023	<i>Colletotrichum graminicola</i> M1.001 (EFQ35223)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	D, S
NRPS_2636	<i>Bipolaris sorokiniana</i> ND90Pr (EMD59100)	NRPS (A-T-C)	E, S
NRPS_283	<i>Aspergillus bombycis</i> (OGM44044)	PKS (KS-AT-DH-ER-KR-T-C-A-T-R)	E, S
NRPS_353	<i>Beauveria bassiana</i> ARSEF 2860 (EJP61198)	PKS (KS-AT-DH-MT-KR-T-C-A-T-R)	E, S
NRPS_41	<i>Aspergillus steynii</i> IBT 23096 (PLB43453)	NRPS (A-T-C-A-T-C)	E, L
NRPS_457	<i>Coleophoma crateriformis</i> (RDW59260)	PKS (KS-AT-DH-MT-ER-KR-T) , NRPS (A-T-C)	E, L
NRPS_480	<i>Capronia coronata</i> CBS 617.96 (EXJ78804)	NRPS (A-T-C-T-C)	E, L
NRPS_514	<i>Aspergillus mulundensis</i> (RDW86494)	PKS (KS-AT-DH-MT-ER-KR-T) , NRPS (T-C-A-T-C-A-T-C-A-T-C-A-T-C)	E, S, L
NRPS_569	<i>Cladophialophora carrionii</i> (OCT48933)	NRPS (A-T-C-A-T-C-T-C-A-T-C-T-C-T-C)	E, L
NRPS_648	<i>Hypoxyton</i> sp. CO27-5 (OTA94984)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R) , NRPS (A-T-R)	E, S
NRPS_777	<i>Cordyceps fumosorosea</i> ARSEF 2679 (OAA69787)	PKS (KS-AT-DH-MT-ER-KR-T-C-A-T-R)	E, S
NRPS_871	<i>Aspergillus fischeri</i> NRRL 181 (EAW20390)	NRPS (A-T-C-A-T-R) , NRPS (A-T-C-A-T-C)	D, E, L, S
NRPS_932	<i>Aspergillus costaricensis</i> CBS 115574 (RAK83302)	PKS (KS-AT-DH-MT-ER-KR-T) , PKS (KS-AT-DH-MT-ER-KR-T)	D, E, L
NRPSLIKE_10	<i>Aspergillus ochraceoroseus</i> (KKK21469)	NRPS-like (ICS-A-T-Transferase)	D, E, S
NRPSLIKE_1029	<i>Cladophialophora carrionii</i> CBS 160.54 (ETI26263)	NRPS-like (A-T-R)	E, L
NRPSLIKE_11	<i>Aspergillus lentulus</i> (GAQ04120)	NRPS-like (ICS-A-T-Transferase)	E, S
NRPSLIKE_1277	<i>Amorphotheca resinae</i> ATCC 22711 (PSS07172)	NRPS-like (A-T-R)	L, S
NRPSLIKE_128	<i>Exophiala oligosperma</i> (KIW43198)	NRPS-like (A-T-Transferase)	D, S
NRPSLIKE_1465	<i>Neonectria ditissima</i> (KPM46454)	NRPS-like (A-T-R)	D, S
NRPSLIKE_1739	<i>Ophiocordyceps australis</i> (PHH75740)	NRPS-like (A-T-TE)	D, L, S
NRPSLIKE_22	<i>Cladophialophora bantiana</i> CBS 173.52 (KIW93789)	NRPS-like (A-T-R-DH)	
NRPSLIKE_266	<i>Penicillium occitanis</i> (PCG97091)	NRPS-like (A-T-R) , NRPS-like (A-T-R)	E, L
NRPSLIKE_869	<i>Cladophialophora bantiana</i> CBS 173.52 (KIW89508)	NRPS-like (A-T-R)	E, L
NRPSLIKE_873	<i>Cladophialophora carrionii</i> CBS 160.54 (ETI24620)	NRPS-like (A-T-R)	E, L
NRPSLIKE_899	<i>Talaromyces marneffei</i> ATCC 18224 (EEA18553)	NRPS-like (A-T-R)	E, L
TERPENE_1140	<i>Exserohilum turcica</i> Et28A (EOA88708)	trichodiene synthase	D, L
TERPENE_139	<i>Penicillium camemberti</i> (CRL18805)	terpene cyclase	E, S

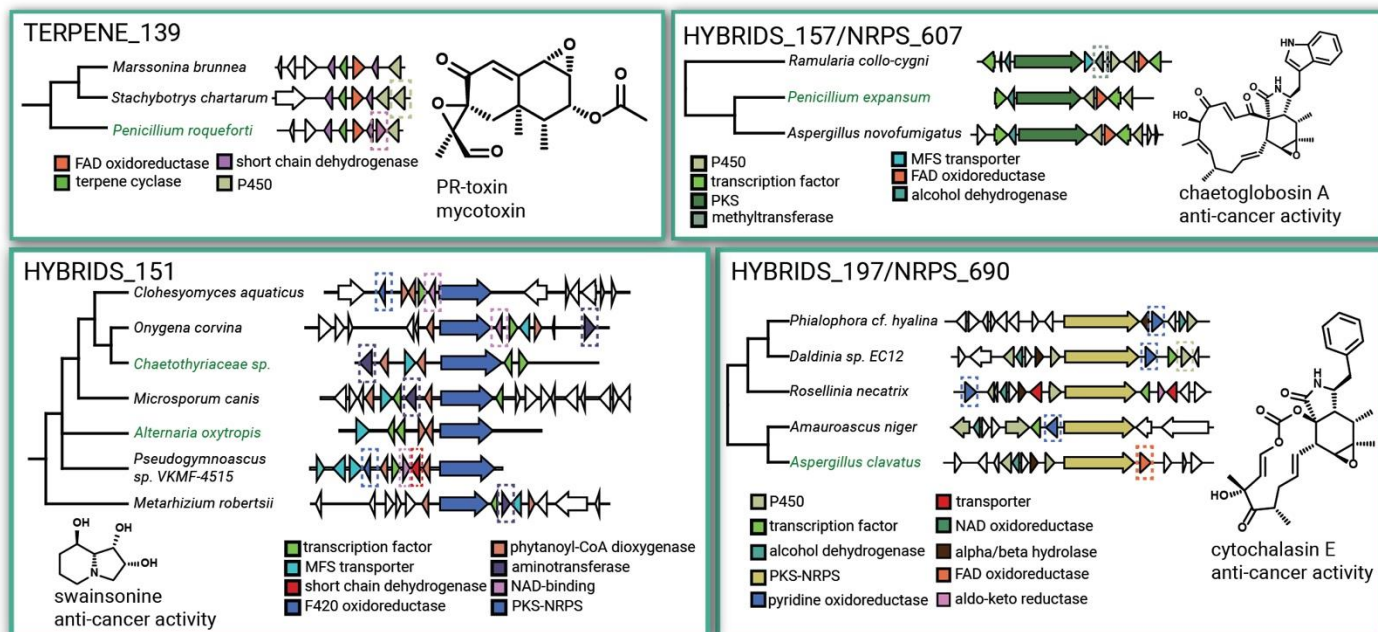


Fig. S14. Diversification of chemical scaffolds across gene cluster families. The GCF for PR-toxin (TERPENE_139), a DNA polymerase mycotoxin produced by *Penicillium roqueforti* (23), contains an additional P450 enzyme in a BGC from the Sordariomycete *Stachybotrys chartarum*. The GCF for chaetoglobosin A, a scaffold with a variety of anti-cancer activities (24), contains a methyltransferase in a BGC from the Dothideomycete *Ramularia collo-cygni* not present in the experimentally-characterized BGC from *Penicillium expansum*. The GCF for swainsonine (HYBRIDS_151), an α -mannosidase inhibitor advanced to clinical trials as a potential anti-cancer therapeutic (25, 26), contains variable F420 oxidoreductase, short chain dehydrogenase, and an NAD oxidoreductase, and aminotransferase enzymes. In the GCF for cytochalasin E (HYBRIDS_197), a compound with anticancer activity, BGCs differ in the presence/absence of a pyridine oxidoreductase and an FAD oxidoreductase present in the experimentally-characterized *Aspergillus clavatus* BGC.

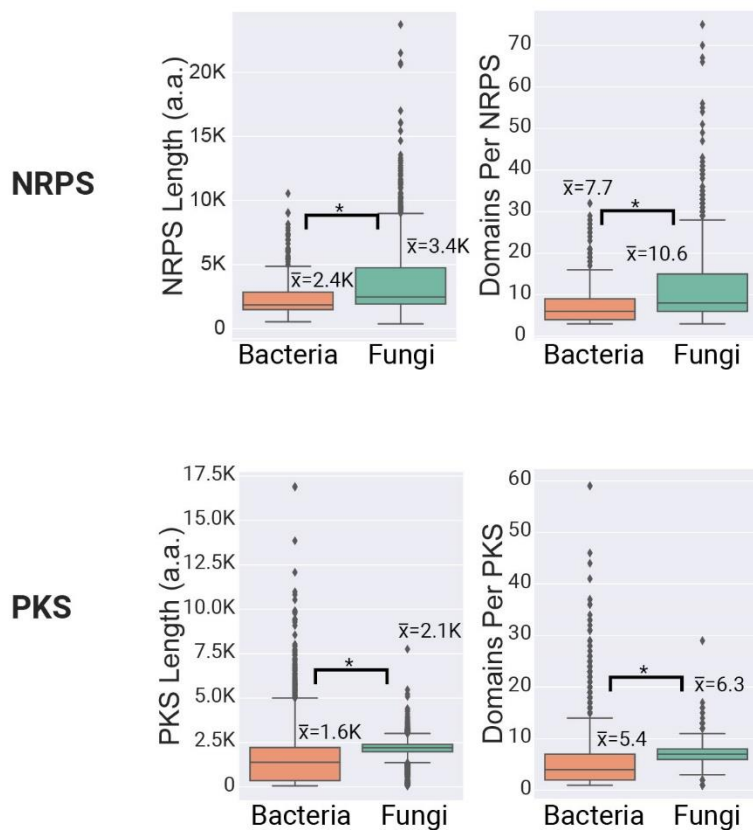
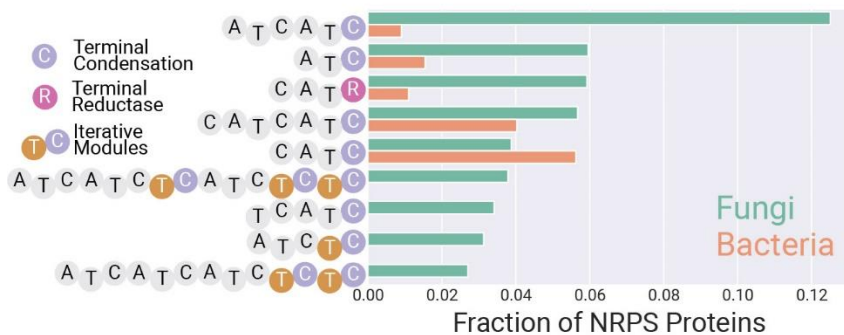


Fig. S15. Comparison of fungal and bacterial NRPS and PKS backbone sizes. For both NRPS and PKS enzymes, fungal backbones are longer both in terms of amino acids and catalytic domains per backbone enzyme.

Top Fungal NRPS Domain Organizations



Top Bacterial NRPS Domain Organizations

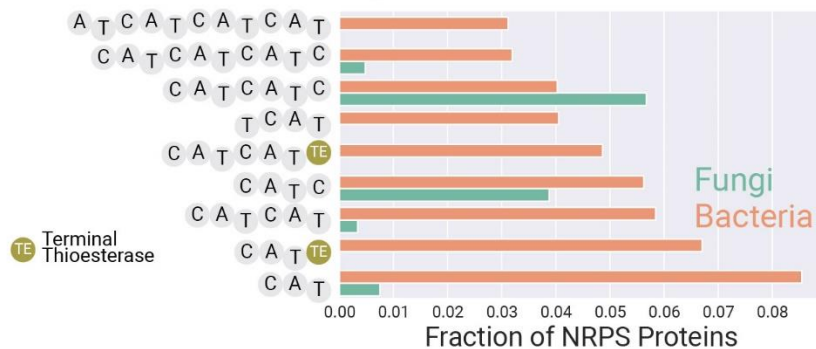


Fig. S16. Comparison of fungal and bacterial NRPS domain organizations. In fungi (top), the most common NRPS domain organizations include terminal condensation or thioester reductase domains. Fungal NRPS enzymes also commonly employ iterative modules. In bacteria, the most common NRPS domain organizations feature terminal thioesterase domains and/or N-terminal condensation domains that interact with an upstream NRPS enzyme catalyze N-acylation.

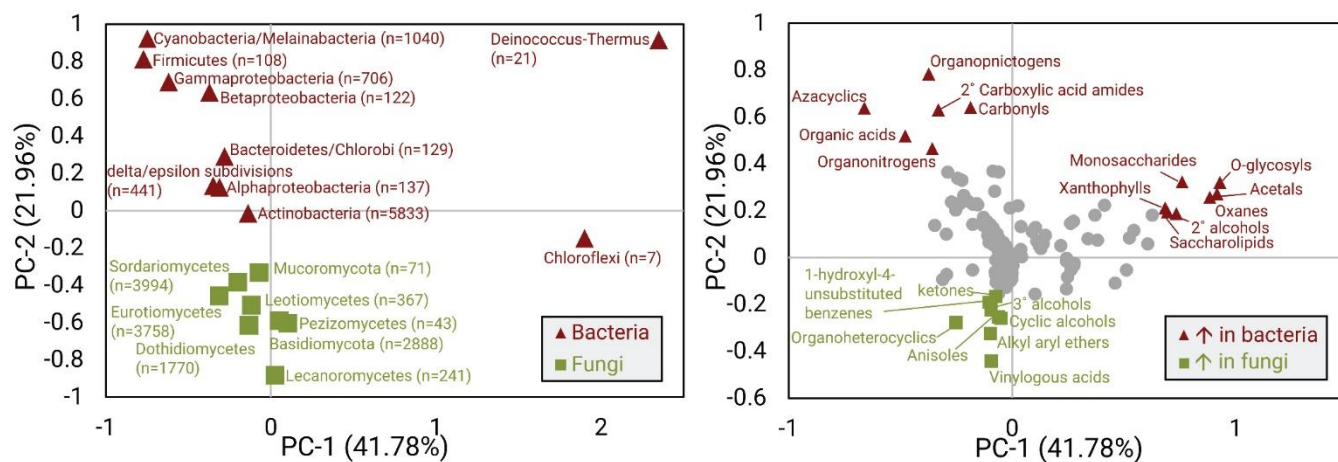


Fig. S17. PCA plot (left) and associated loadings plot (right) of bacterial and fungal chemical space. Fungal and bacterial taxonomic groups represent distinct regions in this space. Fungi are distinguished from bacteria due to an increased frequency of chemical ontology terms associated with aromatic polyketides, such as anisoles, ketones, and alkyl aryl ethers. Bacteria are distinguished largely due to peptide-associated chemical ontology terms (i.e. organic acids, azacyclics, amides).

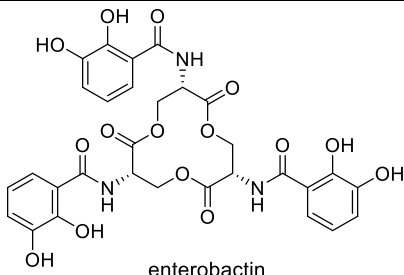
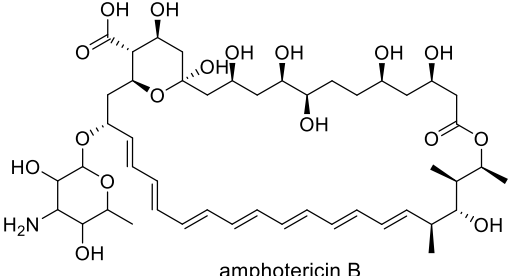
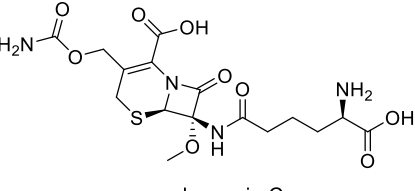
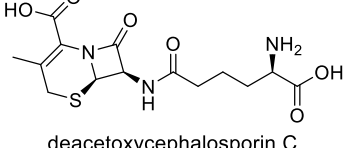
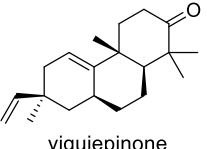
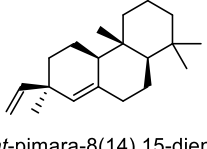
Compound Type	Compound Name and Structure	Fungal Producers	Bacterial Producers	References
Siderophores	 <p>enterobactin</p>	<i>Wickerhamiella versatilis</i>	<i>Escherichia coli</i>	(27)
Polyenes	 <p>amphotericin B</p>	<i>Penicillium nalgiovense</i>	<i>Streptomyces nodosus</i>	(28)
Cephalosporins	 <p>cephamycin C</p>		<i>Streptomyces clavuligerus</i>	(29)
	 <p>deacetoxycephalosporin C</p>	<i>Penicillium chrysogenum</i>		(30)
Diterpenoids	 <p>viguiepinone</p>		<i>Streptomyces</i> sp. KO-3988	(31)
	 <p>ent-pimara-8(14),15-diene</p>	<i>Aspergillus nidulans</i>		(32)

Figure S18. Selected examples of shared chemical space between bacterial and fungal organisms.

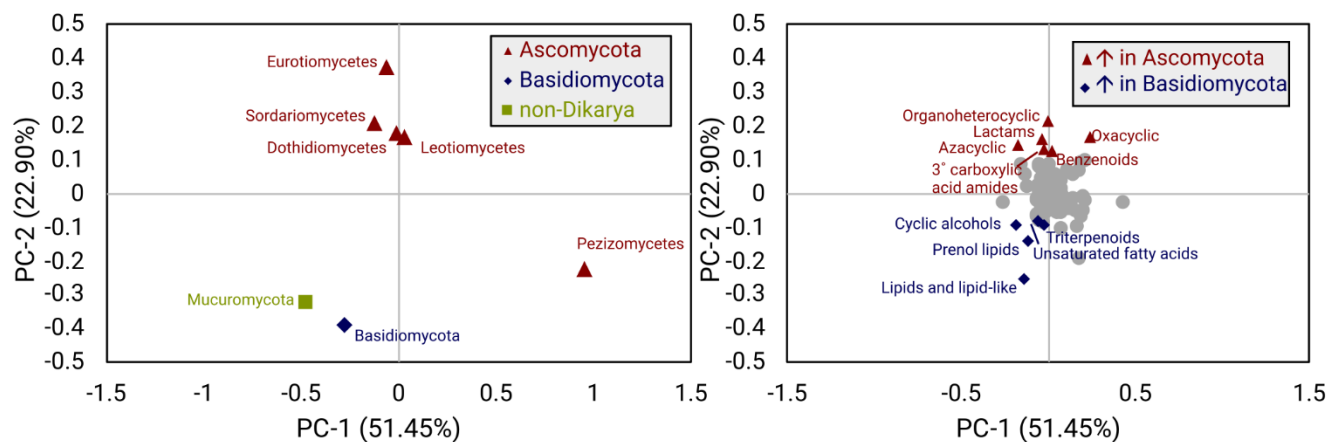


Fig. S19. PCA analysis of fungal chemical space. Eurotiomycetes, Sordariomycetes, Dothideomycetes, and Leotiomycetes (Ascomycota) are distinct largely based on polyketide and peptide-related chemical ontology terms, such as *azacyclic*, *Oxacyclic*, *Benzenoids*, and *Lactams*. Lipid-associated chemical ontology terms are prevalent in Basidiomycota and Mucoromycota.

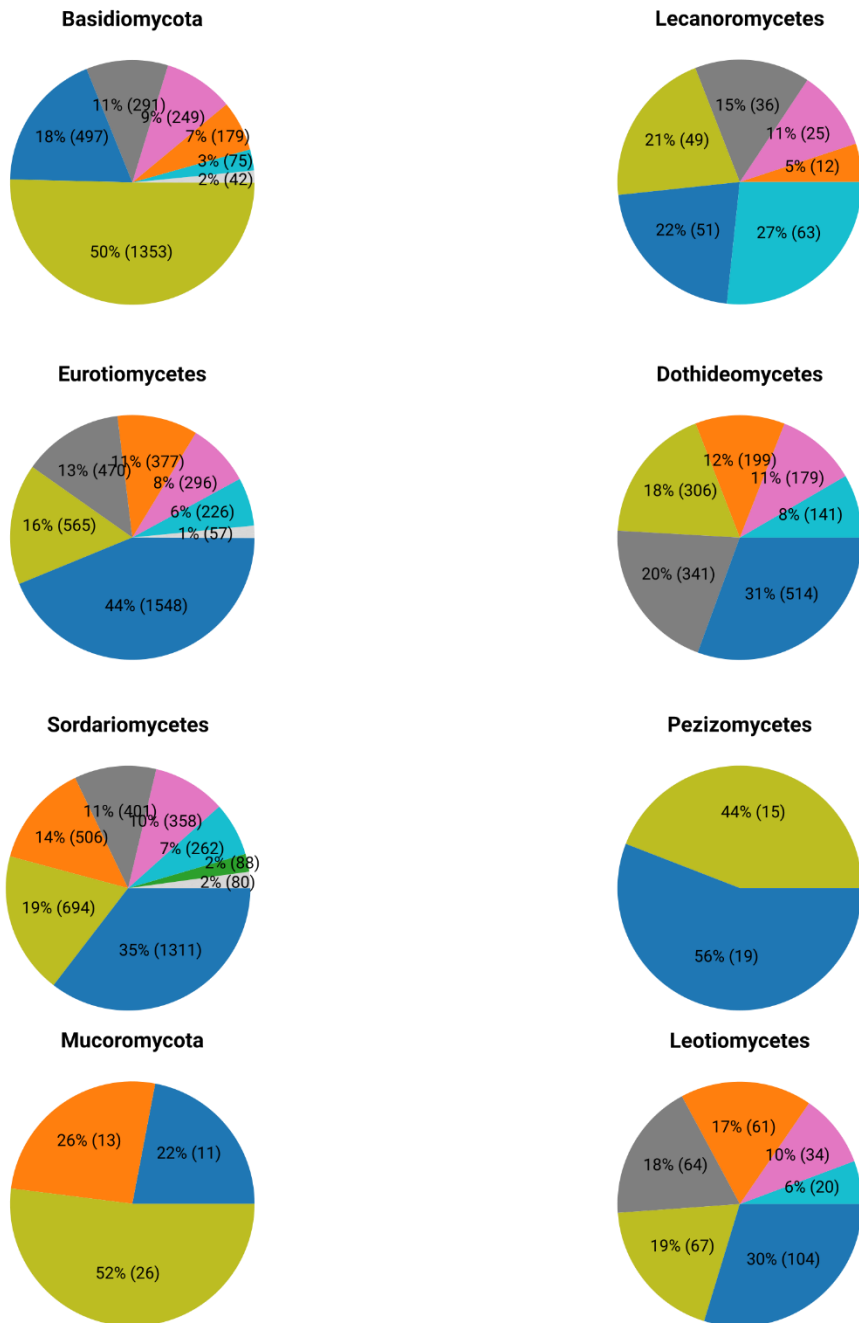


Fig. S20. Breakdown of chemical superclasses in fungal taxa. The chemical space of distinct fungal taxonomic groups varies dramatically. Basidiomycota and Mucoromycota are both ~50% lipids. Other taxonomic groups contain a higher fraction of organoheterocyclic compounds.

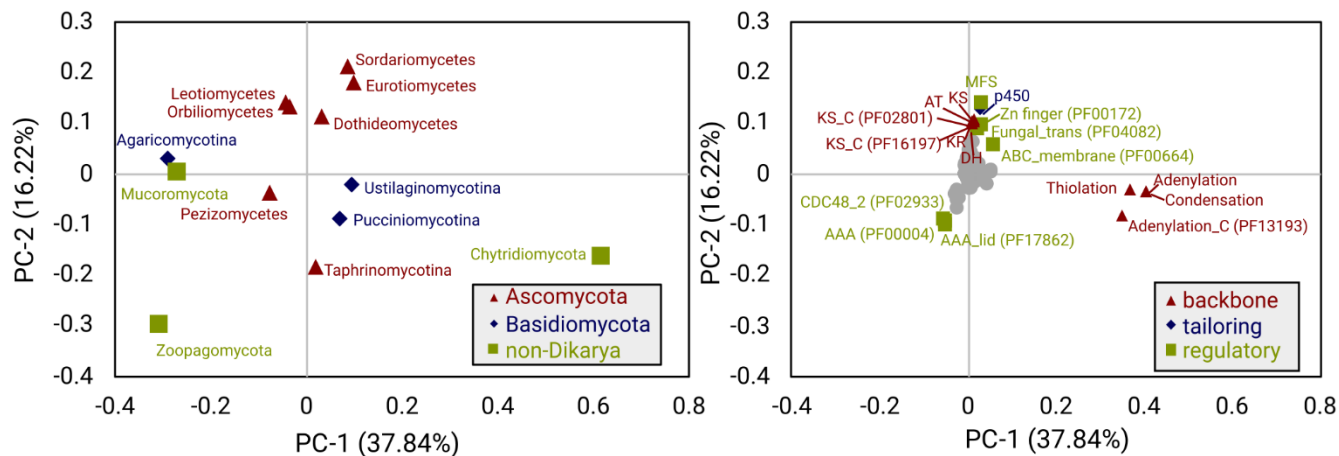


Fig. S21. PCA plot (left) and associated loading plot (right) of biosynthetic domains contained within NRPS-containing biosynthetic gene clusters. Chytridiomycota are pulled in the positive direction on the x-axis due to their high frequency of large NRPS backbone enzymes containing many adenylation, condensation, and thiolation domains, while Pezizomycotina are largely pulled in the “up” direction due to the presence of NRPS-PKS hybrids.

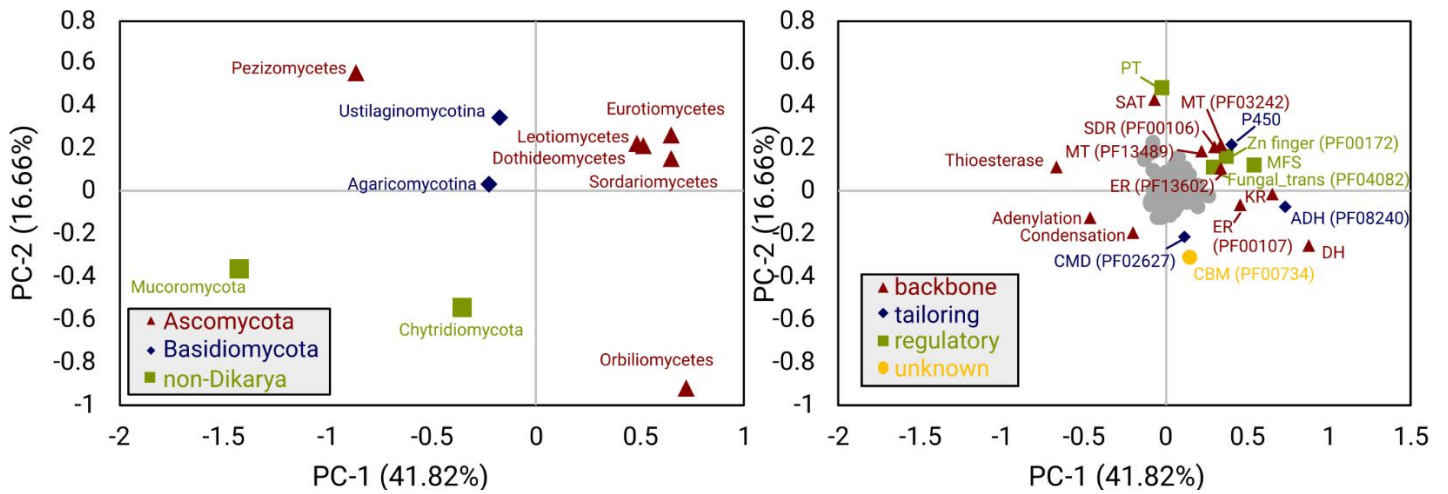


Fig. S22. PCA plot (left) and associated loading plot (right) of biosynthetic domains contained within PKS-containing biosynthetic gene clusters. Eurotiomycetes, Leotiomyces, Dothideomycetes, and Sordariomycetes contain the most PKS backbone enzymes, and are pulled to the right by the corresponding PKS domains. Several regulatory elements are associated with these backbone genes, providing insight into the way fungi regulate PKS biosynthesis.

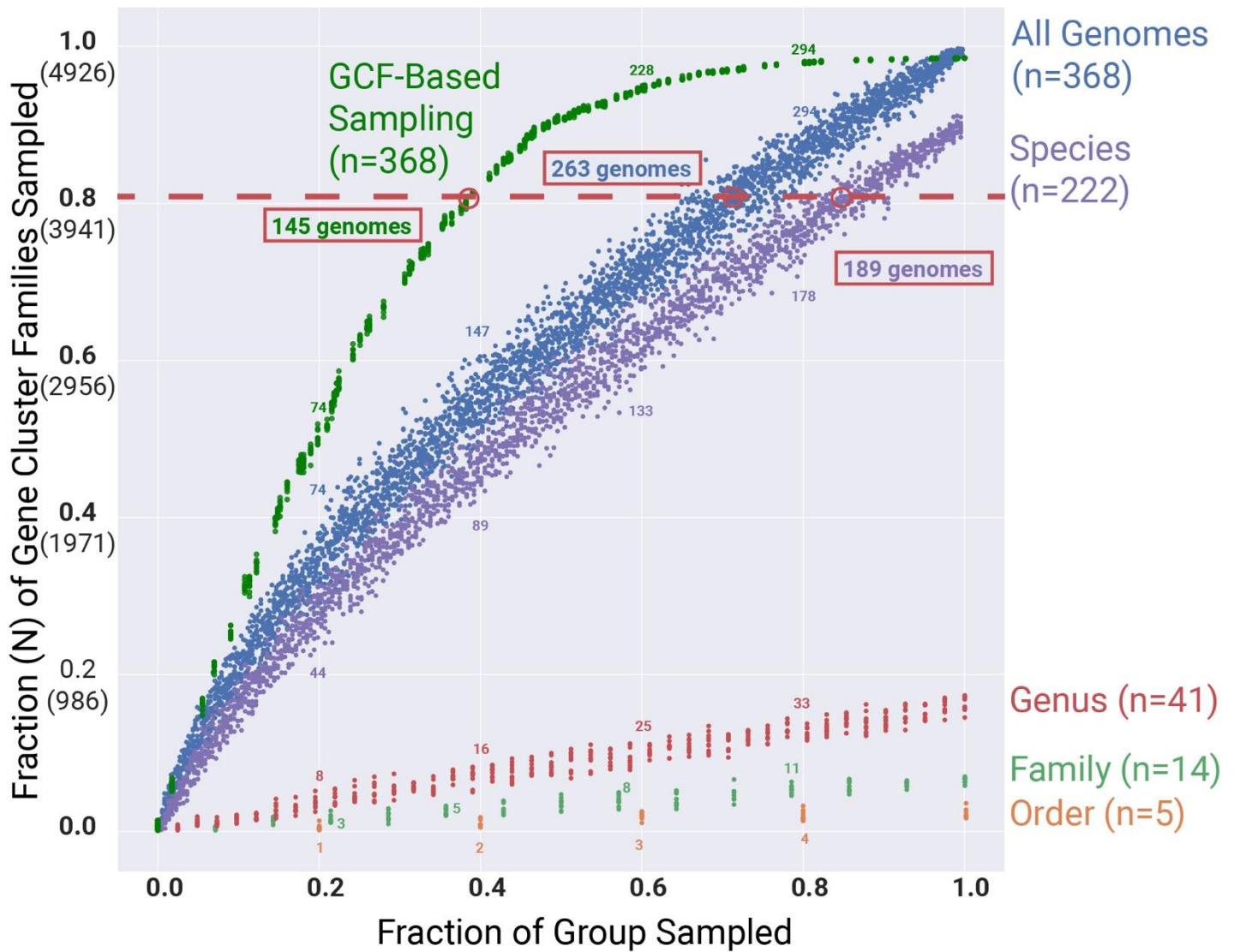


Fig. S23. A roadmap for sampling Eurotiomycetes genomes for natural products discovery based on shared GCFs. Each curve shows the fraction of Eurotiomycetes GCFs that would be present in genomes sampled using different approaches. *All Genomes* shows the results of randomly sampling from all 368 Eurotiomycetes genomes. *Species* and other taxonomic ranks shows the result of randomly sampling unique species, genera, families, or orders. *GCF-Based Sampling* shows the result of sampling clusters of organisms that share GCFs (“clusters” representing the results of density-based clustering, not *biosynthetic gene clusters*). The red boxed numbers indicate the number of genomes required reach 80% GCF coverage, the threshold indicated by the dashed red line. Small numbers along each curve indicate the number of genomes randomly sampled from each group. GCF-based sampling of organisms reaches 80% coverage of GCFs after 145 genomes sampled, species-based sampling of organisms requires 189 genomes, and random sampling of all genomes requires 263 genomes to reach this threshold. This indicates that sampling of organisms for biosynthetic pathway and compound discovery based on GCF overlap can provide a more efficient means of accessing these GCFs. Each random sampling of genomes was performed using 1000 iterations.

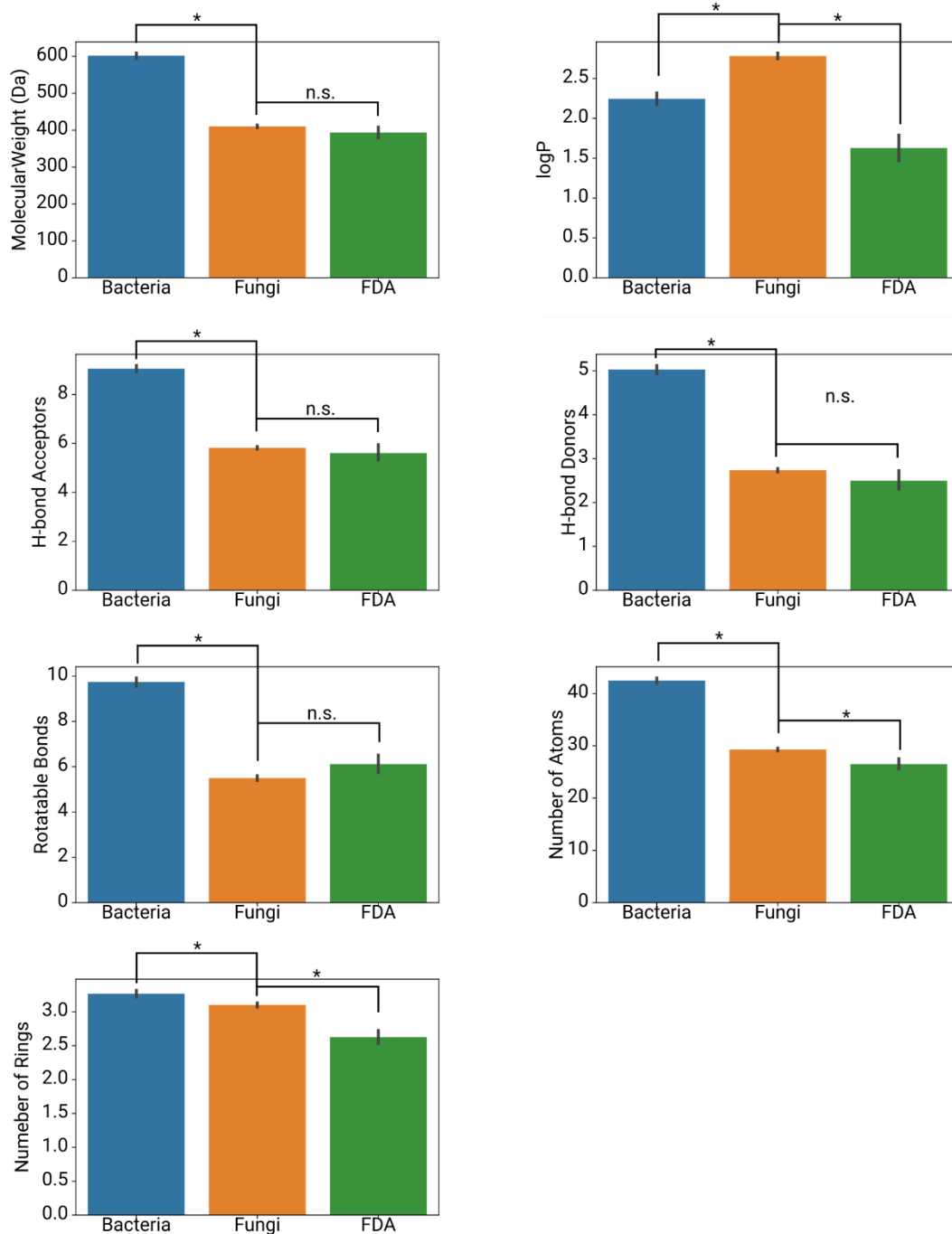


Fig. S24. Comparison of the pharmacological properties of bacterial (n=9,382), fungal (n=15,213), and FDA-approved compounds (n=2884). Error bars represent 95% confidence intervals determined by bootstrap sampling. Asterisks indicate statistically-significant differences between the means ($p < 0.01$, Student's t-test).

BGC TYPE	DOMAINS PRESENT	DOMAINS ABSENT
NRPS	Adenylation, condensation	N/A
HR-PKS	Ketosynthase, dehydratase	N/A
NR-PKS	Ketosynthase and product template or starter acyltransferase	N/A
HYBRID NRPS-PKS	Adenylation, ketosynthase	N/A
NRPS-LIKE	Adenylation	Condensation
DMAT	Dimethylallyl transferase	N/A
TERPENE	Terpene synthase, terpene cyclase, trichodiene synthase, or polyprenyl synthetase	N/A

Table S3. Protein domain rules for classifying gene clusters as nonribosomal peptide synthase (NRPS), highly-reducing polyketide synthase (HR-PKS), nonreducing polyketide synthase (NR-PKS) hybrid NRPS-PKS, NRPS-like, dimethylallyl transferase (DMAT), or terpene.

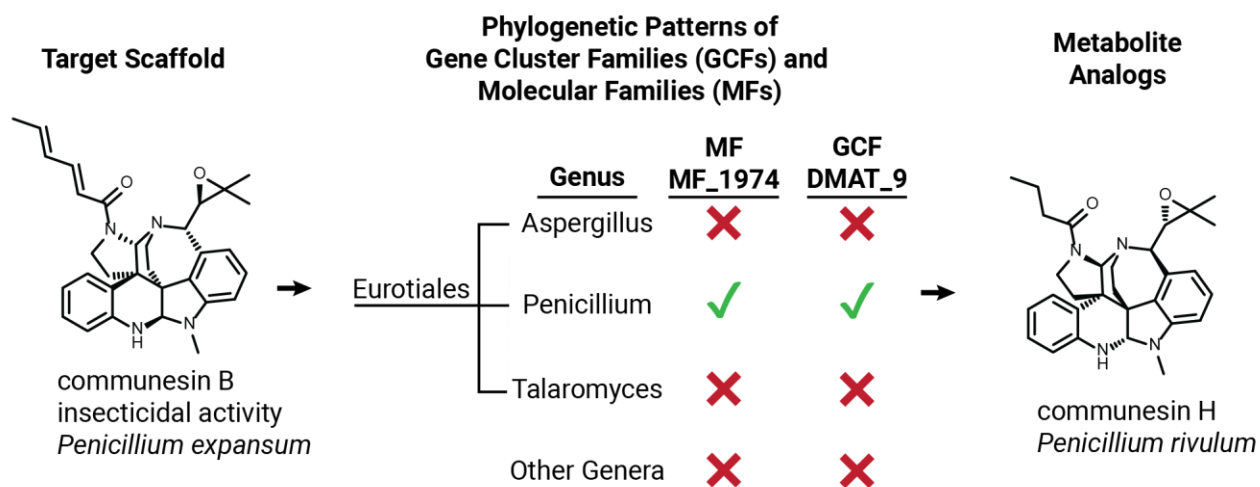


Fig. S25. Genome mining informed by phylogenetic patterns from an interpreted atlas of fungal gene cluster and molecular families. For example, we used *Prospect* to search for the communiesin molecular family (MF_1974), observed that this family as well as the associated GCF (DMAT_9) were only found in the *Penicillium* genus (33, 34). Using these observed patterns could help drive discovery of new analogs informed by their taxonomic distribution.

SI References

1. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**, W81-W87 (2019).
2. M. H. Medema *et al.*, Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625-631 (2015).
3. J. C. Navarro-Muñoz *et al.*, A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60-68 (2020).
4. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-W37 (2011).
5. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).
6. G. Landrum, RDKit: Open-source cheminformatics. (2006).
7. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
8. K. Katoh, G. Asimenos, H. Toh, "Multiple alignment of DNA sequences with MAFFT" in *Bioinformatics for DNA sequence analysis*. (Springer, 2009), pp. 39-64.
9. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
10. S. Kumar, K. Tamura, M. Nei, MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics* **10**, 189-191 (1994).
11. J. A. Van Santen *et al.*, The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Central Science* **5**, 1824-1833 (2019).
12. Y. D. Feunang *et al.*, ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics* **8**, 61 (2016).
13. D. S. Wishart *et al.*, DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901-D906 (2008).
14. H. He, R. Bigelis, E. H. Solum, M. Greenstein, G. T. Carter, Acremonidins, new polyketide-derived antibiotics produced by *Acremonium* sp., LL-Cyan 416. *J Antibiot (Tokyo)* **56**, 923-930 (2003).
15. J. Peng *et al.*, Antiviral alkaloids produced by the mangrove-derived fungus *Cladosporium* sp. PJX-41. *J Nat Prod* **76**, 1133-1140 (2013).
16. W. Pongcharoen *et al.*, Metabolites from the endophytic fungus *Xylaria* sp. PSU-D14. *Phytochemistry* **69**, 1900-1902 (2008).
17. S.-B. Zhang *et al.*, Confluenines A–F, N-oxidized l-isoleucine derivatives from the edible mushroom *Albatrellus confluens*. *Tetrahedron Letters* **59**, 3262-3266 (2018).
18. Q. Luo *et al.*, Isolation and identification of renoprotective substances from the mushroom *Ganoderma lucidum*. *Tetrahedron* **71**, 840-845 (2015).
19. N. Kato *et al.*, Control of the Stereochemical Course of [4+ 2] Cycloaddition during trans-Decalin Formation by Fsa2-Family Enzymes. *Angewandte Chemie* **130**, 9902-9906 (2018).
20. X. Li, Q. Zheng, J. Yin, W. Liu, S. Gao, Chemo-enzymatic synthesis of equisetin. *Chemical Communications* **53**, 4695-4697 (2017).
21. J. W. Sims, J. P. Fillmore, D. D. Warner, E. W. Schmidt, Equisetin biosynthesis in *Fusarium heterosporum*. *Chemical Communications*, 186-188 (2005).
22. J. W. Sims, E. W. Schmidt, Thioesterase-like role for fungal PKS-NRPS hybrid reductive domains. *J Am Chem Soc* **130**, 11149-11155 (2008).
23. Y. Moule, M. Jemmali, N. Rousseau, Mechanism of the inhibition of transcription by PR toxin, a mycotoxin from *Penicillium roqueforti*. *Chemico-biological interactions* **14**, 207-216 (1976).
24. J. Chen *et al.*, Bioactivities and Future Perspectives of Chaetoglobosins. *Evid Based Complement Alternat Med* **2020** (2020).
25. P. E. Goss, J. Baptiste, B. Fernandes, M. Baker, J. W. Dennis, A phase I study of swainsonine in patients with advanced malignancies. *Cancer Res* **54**, 1450-1457 (1994).
26. P. E. Goss, C. L. Reid, D. Bailey, J. W. Dennis, Phase IB clinical trial of the oligosaccharide processing inhibitor swainsonine in patients with advanced malignancies. *Cancer Res* **3**, 1077-1086 (1997).
27. J. Kominek *et al.*, Eukaryotic acquisition of a bacterial operon. *Cell* **176**, 1356-1366. e1310 (2019).
28. K. S. Svahn, E. Chryssanthou, B. Olsen, L. Bohlin, U. Göransson, *Penicillium nalgiovense* Laxa isolated from Antarctica is a new source of the antifungal metabolite amphotericin B. *Fungal biology and biotechnology* **2**, 1-8 (2015).

29. A. Khetan, L.-H. Malmberg, D. H. Sherman, W.-S. Hu, Metabolic engineering of cephalosporin biosynthesis in *Streptomyces clavuligerus*. *Annals of the New York Academy of Sciences* **782**, 17-24 (1996).
30. K. A. ALVI, C. D. REEVES, J. PETERSON, J. LEIN, Isolation and identification of a new cephem compound from *Penicillium chrysogenum* strains expressing deacetoxycephalosporin C synthase activity. *J Antibiot (Tokyo)* **48**, 338-340 (1995).
31. K. Motohashi *et al.*, Studies on terpenoids produced by actinomycetes: oxaloterpins A, B, C, D, and E, diterpenes from *Streptomyces* sp. KO-3988. *J Nat Prod* **70**, 1712-1717 (2007).
32. K. Bromann *et al.*, Identification and characterization of a novel diterpene gene cluster in *Aspergillus nidulans*. *PloS one* **7**, e35450 (2012).
33. H. Hayashi, H. Matsumoto, K. Akiyama, New insecticidal compounds, communesins C, D and E, from *Penicillium expansum* link MK-57. *Bioscience, biotechnology, and biochemistry* **68**, 753-756 (2004).
34. P. W. Dalsgaard, J. W. Blunt, M. H. Munro, J. C. Frisvad, C. Christophersen, Communesins G and H, New Alkaloids from the Psychrotolerant Fungus *Penicillium r ivulum*. *J Nat Prod* **68**, 258-261 (2005).