Mining for a new class of fungal natural products: the evolution, diversity, and distribution of isocyanide synthase biosynthetic gene clusters

Grant R. Nickles [1], Brandon Oestereicher [2], Nancy P. Keller [1,3,*] and Milton T. Drott [4,*]

[1]Department of Medical Microbiology and Immunology, University of Wisconsin—Madison, Madison, WI 53706, USA, [2]Madison, WI 53703, USA, [3]Department of Plant Pathology, University of Wisconsin—Madison, Madison, WI 53706, USA and [4]USDA-ARS Cereal Disease Lab (CDL), St. Paul, MN 55108, USA

ABSTRACT

The products of non-canonical isocyanide synthase (ICS) biosynthetic gene clusters (BGCs) mediate pathogenesis, microbial competition, and metal-homeostasis through metal-associated chemistry. We sought to enable research into this class of compounds by characterizing the biosynthetic potential and evolutionary history of these BGCs across the Fungal Kingdom. We amalgamated a pipeline of tools to predict BGCs based on shared promoter motifs and located 3800 ICS BGCs in 3300 genomes, making ICS BGCs the fifth largest class of specialized metabolites compared to canonical classes found by antiSMASH. ICS BGCs are not evenly distributed across fungi, with evidence of gene-family expansions in several Ascomycete families. We show that the ICS *dit1/2* gene cluster family (GCF), which was prior only studied in yeast, is present in ∼30% of all Ascomycetes. The *dit* variety ICS exhibits greater similarity to bacterial ICS than other fungal ICS, suggesting a potential convergence of the ICS backbone domain. The evolutionary origins of the *dit* GCF in Ascomycota are ancient and these genes are diversifying in some lineages. Our results create a roadmap for future research into ICS BGCs. We developed a website (https://isocyanides.fungi.wisc.edu/) that facilitates the exploration and downloading of all identified fungal ICS BGCs and GCFs.

GRAPHICAL ABSTRACT
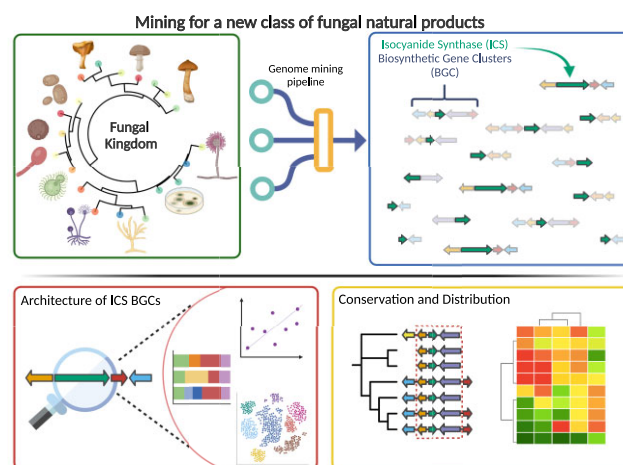
INTRODUCTION

Fungal specialized metabolites (SMs; also called secondary metabolites or natural products) are essential sources of antimicrobial (e.g. penicillin, griseofulvin) and therapeutic (e.g. cyclosporine, mycophenolate) compounds (1–3). In nature, SMs enable niche adaptation by conferring protection from abiotic and biotic stressors, as well as aiding in nutrient acquisition and competitive interactions (4). The bioactive properties of these molecules have propelled the research community to identify strategies to find new fungal SMs. Current genome-mining algorithms that predict SM-producing biosynthetic gene clusters (BGCs) in fungal genomes utilize the contiguous physical arrangement of genes to enable BGC predictions (5–7). This approach has successfully revealed millions of putative BGCs, many of which remain uncharacterized (8,9). However, these algorithms rely on prior knowledge of chemical-class defining 'backbone' synthases/synthetases (e.g. nonribosomal

*To whom correspondence should be addressed. Tel: +1 612 625 3780; Email: Milton.Drott@usda.gov
Correspondence may also be addressed to Nancy P. Keller. Tel: +1 608 262 9795; Email: npkeller@wisc.edu

synthetases, polyketide synthases, terpene synthases/cyclases), biasing results towards well-characterized classes of BGCs. This approach cannot find BGC architectures with uncharacterized backbone genes (i.e. non-canonical). While there have been several recent efforts to predict non-canonical BGC architectures (10,11), the uncharacterized nature of non-canonical BGCs has made them challenging to incorporate into genome-mining software, obfuscating our broader understanding of their prevalence and function. Our inability to mine 'unknown' BGCs (12) remains one of the greatest challenges hindering the discovery of novel compounds with bioactive properties (13).

Isocyanides (also called isonitriles) are a chemical class of SM produced by bacteria and fungi (14). These compounds are produced by non-canonical BGCs that are not detected by current genome-mining software. They are characterized by the presence of the highly reactive isocyanide functional group ($R \equiv N^+ \text{-} C^-$), which is formed by the conversion of the amino group on select amino acids (15). Isocyanide SMs can participate in unique chemical reactions (16,17) and possess potent antifungal, antibacterial, antitumor, and antiprotozoal bioactivity (18–20). The isocyanide functional group is renowned for its ability to chelate to metals (21,22), and compounds with this group can preferentially bind to specific transition metals and metalloproteins (16,19,23).

The first isocyanide backbone enzyme, isocyanide synthase (ICS), was discovered by (24) while investigating the metabolome of an unculturable soil bacterium. Since this initial finding, a small number of ICSs have been shown to serve as backbone synthases in bacterial and fungal BGCs (23,25–27). The only three characterized fungal ICS BGCS are: (i) *xan* (19), (ii) *crm* (16) and (iii) *dit* (28–30). While the *dit1* (ICS) *dit2* (p450) cluster produces dityrosine, a core component in the cell wall of yeast ascospores that has no known metal association (28–30), the *xan* and *crm* BGCs are both strongly up-regulated under copper starvation (31). Despite the mounting evidence that ICS BGCs are a rich source of bioactive microbial natural products, many of which may have metal-associated properties (17,19,23), their exclusion from existing genome mining software has limited research on them relative to other canonical classes of SMs. Lim et al. (2018) reported a diversification of ICS proteins across both the fungal and bacterial kingdoms. However, the chemical diversity of compounds that result from these synthases depends on the extent to which ICSs function as backbones in BGCs, as opposed to existing as standalone genes (32), and has yet to be fully elucidated.

In the present study, we addressed the hypothesis that ICS genes are core members in diverse fungal BGCs. We present the first genome-mining pipeline, assembled from preexisting tools to identify fungal ICS BGCs. We discovered 3800 ICS BGCs in 3300 fungal genomes, making ICSs the fifth-largest class of SM compared to canonical classes found by antiSMASH (6). In exploring this dataset, we focused on two unanswered questions: (i) What is the prevalence of ICS BGCs across different taxonomic groups? And (ii) what are the structural variations and commonalities among ICS BGCs in fungi (e.g. size, protein domain content)? Finally, we conducted additional analysis to examine the distinctive genomic and evolutionary signatures of the *dit1/2* gene cluster family (GCF) that was previously only described in yeast (referred to as the *dit* superfamily in this study).

MATERIALS AND METHODS

Dataset and genomic-database annotation

All publicly available annotated fungal genomes were downloaded from the NCBI database on 09/09/2021 using the NCBI's Dataset tool, version 11.32.1. We generated protein domain predictions using HMMER v3.1b2 (e-value < = 1e-4) (33) with the Pfam database v34 (34). ICS proteins were identified based on the presence of the ICS-specific domain PF05141.1 (named 'DIT1_PvcA' in the Pfam database in reference to the *dit* cluster) (31). All canonical BGCs were predicted using the default settings within the fungal version of antiSMASH v5 (6). See 'Step 1′ in the Reproducible Script for details on the code and specific parameters. See Supplemental Table S1 for information on the taxonomy and metadata of each genome.

Locating putative ICS BGCs from regulatory-motif conservation

Genes within a BGC often share cluster-specific promoter motifs (i.e. genetic regulatory elements) that allow for the co-regulation of a cluster by a single transcription factor (35). To generate our BGC predictions, we examined the genes surrounding ICSs for shared promoter motifs using the eukaryotic BGC-detection algorithm, CASSIS v1 (36). All CASSIS predictions that contained one or more neighboring genes next to the ICS gene were considered BGCs and included in the subsequent analysis in accordance with past definitions of BGCs (7). Scripts detailing the implementation of CASSIS can be found in 'Step 2' of the Reproducible Script.

Guidance for interpreting promoter-motif based BGC predictions

Resulting predictions from CASSIS assume co-regulation of all genes within a BGC and that co-regulation is mediated by a single, consistent regulatory motif (37). A major strength of this approach is that differences in regulation may be biologically relevant (e.g. divergence of regulatory motifs might suggest a change in or loss of a cluster's functionality). However, we cannot be certain that all co-regulatory motifs across all genomes and/or genes are called correctly. We urge caution when interpreting individual predictions (or lack thereof) in individual species of interest. The focus of this study is to identify kingdom-wide patterns in ICS BGCs. At this scale, BGCs can be grouped into GCFs. This amalgamation offers additional data on the gene content of specific clusters. We have implemented strategies to harness patterns across GCF to strengthen our inferences and offer fundamental insights into ICS GCFs across the fungal kingdom.

Converting BGC predictions into standard file formats

Raw CASSIS predictions were run through a custom Python pipeline that converted each prediction into a

GenBank (gbk), FASTA (fna) and General Feature Format (gff) file, enabling us to leverage existing BGC-processing software such as BiG-SCAPE (38), and cblaster (39). The raw gbk, fna, and gff files for all 3800 ICS BGCs can be downloaded from the Supplementary Repository. Coding regions (CDSes) encoding SM-related biosynthetic protein domains (6) were flagged as 'biosynthetic' because BiG-SCAPE places more emphasis on these proteins when determining the relatedness of BGCs. The scripts used in this step can be found in 'Step 3′ of the Reproducible Script. Refer to Supplemental Tables 2–4 for detailed summaries of the number of ICS BGCs in each species/genome and the protein domains present in each BGC.

Grouping the ICS BGCs into GCFs

BGCs that belong to the same GCF are thought to produce identical or closely related natural products based on gene content and protein identity thresholds. ICS BGC gbk files were clustered into GCFs using BiG-SCAPE v1.0.1 ('Step 4' in Reproducible Script) (38). We created a modified 'anchor file' to ensure BiG-SCAPE treated each ICS gene as a chemical-class defining synthase (i.e. a backbone gene). Parameters for ICS GCF classification were tested using 13 different cutoffs ranging from 0.2 to 0.8 in increments of 0.05. Cutoff values above 0.45 were found to be too relaxed, leading to the merging of major ICS GCFs that were separate at lower cutoffs. Conversely, cutoffs below 0.25 were too strict, forming only small GCFs comprising BGCs from closely related species. A value of 0.3 (which is also the default value) was confirmed as the optimal cutoff. A network visualization of the BiG-SCAPE output was created with Cytoscape v3.9.1 (40) and can be found in Supplementary Figure S1. For more information on the dereplication pipeline run on the BiG-SCAPE output, refer to the Supplementary Methods.

Identifying core genes in highly conserved ICS GCFs

CASSIS assumes that all the genes within a BGC share a cluster-specific promoter motif (36,37). Although promoter-based BGC mining is an effective method for identifying fungal BGCs (6,37), this approach fails to locate BGCs with differentially regulated genes (41). Therefore, we employed a second GCF-verification approach to identify genes that were common across a GCF but may not have conserved regulatory motifs. Briefly, we searched for co-localized orthologous genes in each GCF to identify genes conserved across evolutionary time using the program cblaster (v1.3.15) (39). The set of highly conserved genes within a GCF is referred to as a GCF's biosynthetic core and represents the subset of genes essential for the biosynthesis of a given SM pathway (25,42). Accessory genes, which are defined by their variability within GCFs, are thought to modify the SM produced by the biosynthetic core (4). Therefore, GCFs with the same core genes often produce related families of SMs (38). Several GCFs were identified as sharing a common biosynthetic core with other GCFs suggesting that these GCFs were not entirely resolved (9). In these instances, we collapsed GCFs with identical cores into a composite GCF that we call a 'superfamily' and refer to original GCFs as 'Clans' to clarify their

distinct accessory gene complements. It was not feasible to perform manual steps of this analysis on all BGC predictions, so we focused on clusters present in the 31 ICS GCFs that contained at least five distinct species (core regions are also more easily identified among species). A detailed explanation of this methodology can be found in the Supplemental Methods. In this study we use cblaster predictions to identify core and accessory genes within our predictions from CASSIS (Supplementary Table S11). However, a small subset of clusters that were predicted only with cblaster are included in Supplementary Table S14 to facilitate future research.

Estimating the extent of linkage disequilibrium in GCFs

We validated GCFs by confirming that the co-occurrence of genes found in BGCs within a GCF (i.e. linkage disequilibrium) could not be explained by random chance (i.e. genetic drift) (10). Whole-genome levels of linkage disequilibrium (LD) for species represented in a GCF were calculated around BUSCOs (e.g. 'Benchmarking Single Copy Orthologs') (v3.0.3) (43) present across all species ($n = 287$). Around each BUSCO, we identified orthologous genes using Orthofinder v2.5.2 (default settings) (43) and calculated the percent conservation of orthologous genes across all species in a GCF as a proxy for LD. An identical process was used to estimate LD at ICS loci. The distribution of LD estimates around each BUSCO served as a null distribution to determine the significance of LD acting on ICS loci. The significance of LD was calculated using a percentile rank for each GCF (SciPy v1.5.2) (44). Visualizations of the genes detected under significant LD can be found in Supplementary Figure S11. Plots showing estimations of LD across each GCF can be found in Supplementary Figure S13. Additionally, the co-occurrence rate of the four most common protein domains in ICS gene clusters was compared to a null distribution using a binomial expression to determine if such domains were preferentially occurring next to ICS backbone genes (refer to the Supplemental Methods for details).

Phylogenetic inference by maximum likelihood (ML)

All maximum likelihood gene and protein trees were made using the following pipeline. First, sequences were aligned using MAFFT v7.475 with the '-auto' parameter (45). Next, the alignments were trimmed using trimAl v1.2 with either the '-gappyout' or '-gt' parameters (46). Lastly, we made phylogenomic trees from the trimmed alignments using IQTree v2.0.3 (47) and ran 1000 ultrafast bootstrap replicates (48). The optimal model of sequence evolution was selected using ModelFinder (49). All tree visualizations were constructed using either iTOL v5 (50) ggtree v3.1.1 (51) or ETE 3 v3.1.2 (52). All trees created for this publication can be found in the Supplementary Repository, along with the specific parameters used in their construction.

Reconstruction of the fungal kingdom's species tree

We chose a coalescent-based approach to construct a kingdom-level species tree because of this method's

computational scalability (53), compatibility with incomplete lineage sorting (54) and ability to handle missing data (55). We identified a set of 287 BUSCOs that were conserved across all 3300 genomes (lineage database: 'fungi_odb9') (43). Unrooted maximum likelihood trees were generated for each BUSCO as defined above and used as the input for ASTRAL v5.7.8 (56). A species map file was generated and passed into ASTRAL using the '-a' parameter to force isolates within the same species to be monophyletic. We verified that our reconstructed tree closely aligns with Li et al.'s (2021) established fungal tree of life. The tree was rooted at *Rozella allomycis*, the only publicly available genome from the earliest diverging lineage of the Fungal Kingdom, Cryptomycota (57,58). BUSCO gene trees and the final species tree are available for download from the publication's Supplemental Repository.

Constructing a phylogeny of ICS backbone enzyme sequences

The phylogenetic relationships between GCFs were inferred from ICS backbone protein sequences using two datasets: (i) the ICS-domain sequence alone and (ii) the entire protein sequence. Because both alignments produced topologically identical phylogenies, only the complete ICS protein alignment is presented in subsequent analyses.

Creating the phylogeny of p450 proteins to gain insight into the relatedness of dit2 orthologs

We investigated evolutionary relationships between Dit2 and other fungal p450 proteins to identify additional Dit2 orthologs. Constructing a phylogeny from every species' set of p450 proteins was computationally prohibitive, as a single Ascomycete genome can contain hundreds of p450 genes. A subset of ten species was selected to represent the total diversity of the *dit* cluster. The species included: *Aspergillus clavatus*, *Candida tropicalis*, *Scheffersomyces stipitis*, *Saccharomyces cerevisiae, Monosporascus cannonballus*, *Kluyveromyces lactis*, *Rhizodiscina lignyota*, *Ophiobolus disseminans*, *Zygotorulaspora mrakii* and *Macrophomina phaseolina*. Candidate p450s were identified in these genomes based on the presence of p450 protein domains. The resulting dataset of 845 p450 protein sequences was constructed into a maximum likelihood phylogeny as defined above.

Topological comparison of the dit1/2 phylogeny and the reconstructed species tree

Tests of phylogenetic incongruence were used as they are considered the most reliable way to infer historical events from genomic and proteomic sequences (59,60). Before creating a concatenated phylogeny from the Dit1 (ICS) and Dit2 (p450) protein sequences, we verified that phylogenies inferred from these proteins had similar topologies, suggesting that they had co-evolved. The individual protein trees were topologically aligned using the R packages ape v5.6–2 (61), phangorn v2.8.1 (62), and TreeDist v2.4.1 (63). The optimal model of sequence evolution was estimated independently for both genes using ModelFinder and found to be LG + I + G4 in both cases (49). After verifying the co-evolution of the Dit1 and Dit2 proteins, we concatenated these sequences using FASconCAT-G v.1.05.1 (64) and repeated the process, performing a topological comparison between the concatenated Dit1/2 tree and the species tree to determine if this GCF's evolutionary history could be explained by vertical transmission.

RESULTS AND DISCUSSION

ICS BGCs are not evenly distributed across the kingdom fungi

Our ICS BGC-detection pipeline (see Methods) identified 4341 ICS genes within 3300 genomes representing 1329 unique species (Figure 1, Supplementary Table S1). Cluster-specific promoter motifs were found in genes surrounding 3800 (87.5%) of the 4341 ICS genes. After selecting a single representative for each species, the ratio of ICS proteins in BGCs remained very similar, with 1190 BGCs found around 1394 detected ICS proteins (85.4%). This data suggests that most ICSs function as backbone enzymes in fungal BGCs (Supplementary Tables S2–S4).

The genomes of species within the Ascomycete classes Sordariomycetes, Dothideomycetes, and Eurotiomycetes contained the most ICS BGCs (Fisher's exact test: $p = 1.99^{-42}$; Supplementary Figure S2; Supplementary Methods). However, the genomes of some individual species outside of these groups were also found to contain several ICS clusters; for example, the genome of the saprobic Basidiomycete species *Tetrapyrgos nigripes* contained 6 ICS BGCs and the coral mushroom *Clavulina* sp.'s genome contained five ICS BGCs (Figure 2; Supplementary Table S2). A preliminary examination of the expansion of ICS BGCs across the phylogeny of Ascomycota identified 17 branches where sister lineages showed significant differences in the number of clusters (Supplementary Figure S3; Supplementary Methods), suggesting a gene-family expansion and/or contraction in these lineages. For example, genomes in one lineage of the genus *Colletotrichum* contained an average of four ICS BGCs, while its sister lineage had an average of one ICS BGC. Plant pathogenic taxa, such as *Penicillium*, *Bipolaris* (*Cochliobolus*), *Colletotrichum*, and *Fusarium* (65,66), were found to harbor the greatest number of ICS BGCs. However, the overrepresentation of plant pathogenic genomes relative to fungi of other ecologies could skew this inference. Better annotation of fungal ecology is needed on a kingdom level to clarify the importance of this trend. Future research should focus on determining the ecological drivers of ICS gene cluster expansions and contractions in these lineages.

ICS BGCs accounted for 3% of the Fungal Kingdom's predictable biosynthetic potential (i.e. not including cryptic clusters or undescribed chemistries), with an additional 84 657 canonical BGCs identified by antiSMASH (Table 1; Supplementary Figure S4) (6). Proportionally, this makes ICSs the fifth-largest class of predictable fungal BGCs. While siderophores are the most prominent group of SMs that facilitate microbial interactions with metals (67,68), we discovered three times more ICS BGCs than siderophore BGCs (Table 1). Much like the metal-regulation of siderophore synthesis (69), both characterized isocyanide natural products in *Aspergillus fumigatus* were recently shown to be tightly regulated by external copper
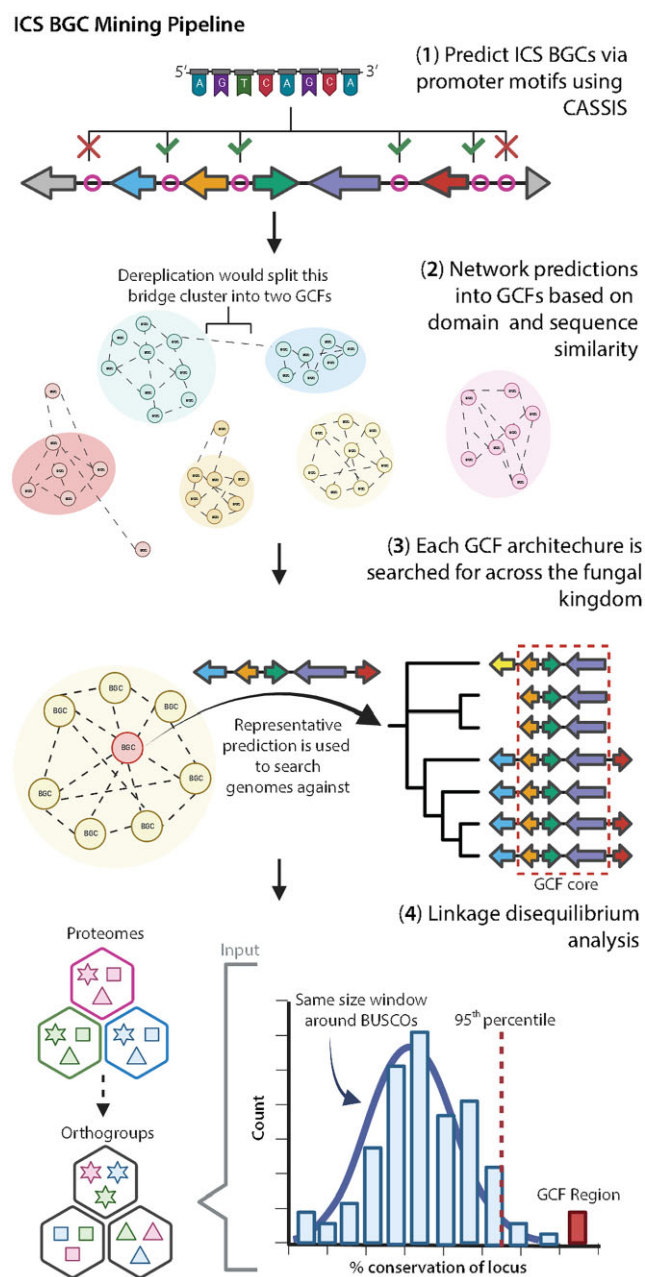
Figure 1. Full isocyanide synthase (ICS) biosynthetic gene cluster (BGC) genome-mining pipeline. First, clusters of co-regulated genes around ICSs are predicted. These BGCs are then grouped into gene cluster families (GCFs) based on sequence similarity and shared protein domains. To identify the core genes in GCFs, we searched for co-localized orthologous genes across every genome in the dataset. Finally, a subset of the GCFs were validated by confirming that the co-occurrence of genes found in the GCFs (i.e. conservation of genes across BGCs in a given GCF) exhibited significant linkage disequilibrium (LD) relative to whole-genome levels of LD (i.e. conservation of orthologous genes around BUSCOs).

concentrations. The isocyanide, xanthocillin, synthesized from the *xan* BGC not only provides chalkophore-like properties for *A. fumigatus* but also inhibits growth of bacteria and other fungi by withholding copper from these microbes [19]. Similar metal-associations have been found with bacteria isocyanides [2,23]. A bacterial isocyanide, rhabduscin,

decorates the cell surface of some gram-negative bacteria and enables infection of insects by inhibiting the activity of central immune system copper-dependent metalloenzymes [17]. The prevalence of fungal ICS BGCs, along with evidence of their metal associations, raises questions about the potential importance of this class of SM in mediating metal-associated fungal ecology and pathogenicity, emphasizing a new opportunity to expand our understanding of how fungi interact with and utilize metals.

While most taxonomic families contained 0–2 ICS BGCs, the right tail of the ICS BGC distribution among taxa comprised 12 families with 4–7 ICS gene clusters (Supplementary Figure S5a, b). In this group, ICSs accounted for 7–15% of the species' total SM BGCs (Supplementary Table S5). Overall, the number of ICS BGCs within a genome had a strong positive correlation with the number of BGCs (canonical & ICS, Pearson's coefficient: 0.39; $P = 2.09 \times 10^{-75}$). Part of this association may reflect that larger genomes (as defined by gene counts) have more BGCs (Pearson's coefficient: 0.30; $P = 1.86 \times 10^{-26}$) (Figure 3A). We also investigated the impact of genome sequence quality on the number of BGCs and observed a strong positive correlation (Supplementary Figure S5c; Pearson's coefficient: 0.22; $P = 1.14 \times 10^{-14}$). While several studies have posited ecological drivers for the number of clusters in genomes [10,70], they have not counted for genome size. Our finding suggests that the number of BGCs in a genome may, to some degree, correlate with size of the genome or genomic sequence.

The structural diversity of ICS proteins and ICS BGCs

All ICS enzymes in our dataset matched one of three protein domain variations summarized by Lim et al. (2018) (Figure 4B): monodomain ICSs [29], two domain ICS-dioxygenases (ICS-TauD) [19], and the multidomain ICS-nonribosomal peptide synthase-like (ICS-NRPS-like) variants [16] (refer to Lim *et al.* (2018) for diagrams detailing each backbone variant). The relative abundance of each backbone-gene variant across species (with only a single genome representing each species) was approximately equal. ICS-TauD, ICS-NRPS-like and ICS monodomain accounted for 36.2%, 32.8%, and 31% of the total ICS proteins, respectively. Outside of the backbones, 55.5% of clustered genes encoded protein domains associated with SM biosynthesis, transportation, or regulation, 22.7% encoded protein domains not typically associated with SM biosynthesis (e.g. sugar transporters, membrane proteins, RNA-related enzymes), and 21.8% lacked a predictable domain altogether.

ICS BGCs are significantly smaller than all other classes of SM BGCs except for siderophores (refer to Supplementary Methods; Supplementary Table S6). ICS BGCs have a mean length of 20 589 bp and contain an average of 8.42 genes per cluster (Figure 3B; Supplementary Figure S6). The small size of ICS BGCs may suggest that resulting compounds are subject to less chemical modification by tailoring enzymes than compounds resulting from other classes of BGC, although we cannot rule out that differences in BGC size between SM classes could result from biases in BGC-border prediction defined by genome-mining algorithms. The observation that ICS BGCs tend
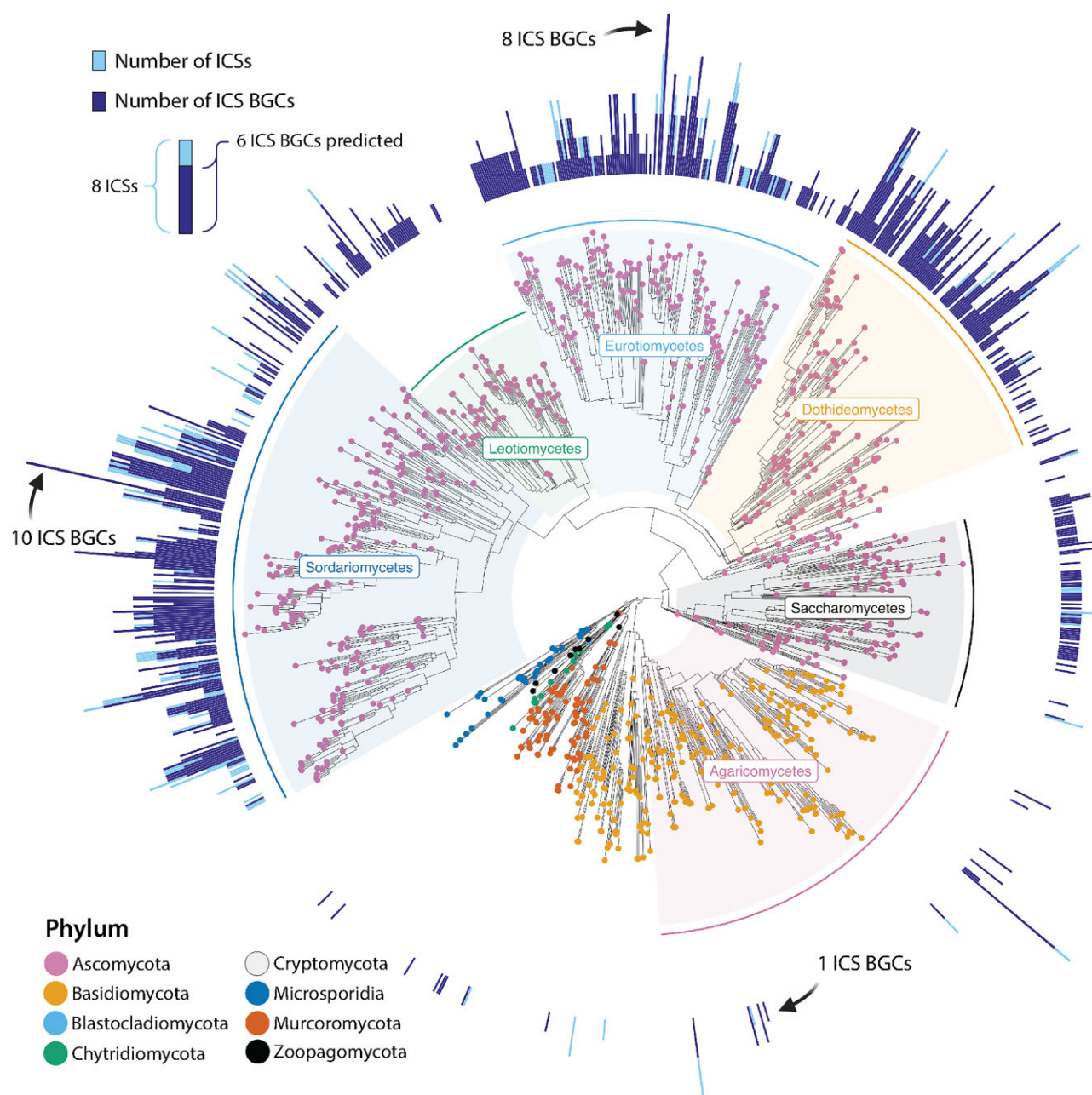
Figure 2. Distribution of isocyanide synthases (ICSs) and associated clusters biosynthetic gene clusters (BGCs) across the fungal kingdom. The species tree was estimated using a multi-species coalescent model, and is rooted at the early-divergent subphylum, Cryptomycota. Tips are colored based on phylum, and major classes within the Ascomycota and Basidiomycota are labelled. The overlay bar chart on the outer ring depicts the number of ICS genes found (dark blue), and number of ICS BGCs (light blue). Fully dark blue bars denote that all ICS genes act as backbones in ICS BGCs, while fully light blue bars indicate that none of the predicted ICS genes were found in co-regulated gene clusters.

to be small is consistent with what has been found in characterized fungal ICS BGCs, with dityrosine production in *S. cerevisiae* requiring only two genes (29) and the *xan* BGC in *A. fumigatus* containing only five genes (19). However, small BGCs can generate outsized chemical diversity through trans-BGC interactions. We speculate that the highly reactive isocyanide backbone may be more likely to have such interactions. Indeed, the ICS product of the *crm* BGC can react with an ergot alkaloid precursor produced by another BGC, resulting in a novel class of alkaloids known as the fumivalines (16). Such reactions could dramatically increase the chemical diversity generated by ICS BGCs despite the relatively small size of these clusters. The nucleophilic, electrophilic, and metal-ligand properties of isocyanides (18) that have made them widely utilized in synthetic (71,72) and medicinal chemistry (22,73), suggests exciting prospects for exploring their potential biological activities.

Table 1. The proportion of each biosynthetic gene cluster (BGC) class across all fungal genomes. Isocyanide synthase (ICS) BGCs are bolded. Multiple isolates within the same species were filtered to avoid overrepresentation bias of estimates stemming from differences in the number of genomes available for a given species

BGC classes	Percent
Nonribosomal peptide synthase (NRPS)	34.60%
Type 1 polyketide synthase (T1PKS)	22.70%
Terpene	22.40%
NRPS,T1PKS	6.40%
Isocyanide synthase (ICS)	**3.00%**
Indole	2.80%
Type 3 polyketide synthase (T3PKS)	1.40%
Betalactone	1.10%
Siderophore[a]	0.90%
Fungal ribosomally synthesized, post-translationally modified peptide (fungal-RiPP)	0.40%
[b]Other	4.30%

[a] The siderophore BGC class is composed of both NRPS and PKS siderophore pathways (per antiSMASH classifications).
[b] The 'Other' row includes all fungal BGC classes present at <0.4%.

ICS biosynthetic diversity across taxonomic ranks

We dereplicated ICS BGCs into GCFs (Supplementary Figure S1) (8–10) to clarify the biosynthetic potential that these clusters represent. Of the 3800 ICS BGCs, 92.4% could be categorized into GCFs (Supplementary Table S7), while the remaining clusters were found only in a single genome. Of the 200 GCFs we identified, 47.3% ($n = 90$) were detected in multiple species, while 52.7% ($n = 110$) were found only within isolates of the same species (Supplementary Figure S7; Supplementary Table S8). In rare instances, we observed multiple copies of the same GCF in individual genomes (Supplementary Table S9). While on average the ICS GCFs were highly conserved within species (98.6%), we located 1963 instances where intraspecies isolates shared less than 50% of the same ICS GCFs. While intraspecific differences in SMs have historically been treated as random variation, recent work has shown that such differences can help to explain niche adaptation and macroevolutionary processes (74). More work is needed to clarify if intraspecific variations in ICS BGCs are associated with specific lineages within species. The vast majority of GCFs found across multiple species were unique to a single genus (70%). Although a small portion of the GCFs we identified ($n = 22$ or 11.5%) was found in two or more taxonomic families, these GCFs were highly conserved across the Fungal Kingdom and accounted for a disproportionate 40.7% (1547/3800) of all ICS BGC predictions.

The distribution and evolution of ICS GCFs in fungi

Understanding the evolutionary history of fungal GCFs presents a challenge because different GCFs frequently lack a shared set of orthologous genes (75,76). However, by definition, backbone genes (and or enzymes) are conserved across GCFs of a given biosynthetic class and can thus be used to infer the evolution of these clusters (77). We investigated the phylogenetic relationships of ICS backbone proteins from a subset of 10 ICS GCFs (see Materials and Methods), with the goal of determining whether

the gene combinations that define BGCs within a GCF shared a single origin or whether they evolved convergently. Most GCFs could be mapped to monophyletic clades of backbone enzymes, indicating that these GCFs had a single common ancestor (Figure 4B). The ICS enzymes from GCFs 9 and *xan* were nested within GCFs 6 and 7, respectively. This suggests the ICS proteins in GCFs 9 and *xan* may have evolved from ICS within GCFs 6 and 7, respectively, sometime after the latter GCFs had already evolved. However, these BGCs do not share any genes beyond their ICSs. This pattern could be explained by the translocation of an ICS backbone gene from a parent GCF into a new genomic context. However, despite strong bootstrap support for the topology of these branches (Supplementary Figure S8), we cannot rule out that these relationships are not fully resolved in our phylogeny, as single-protein trees offer relatively little phylogenetic resolution. The two ICS backbone enzymes found within GCF 1 formed two distinct clades stemming from a single common ancestor (Figure 4B). A small clade of ICSs intercedes the two aforementioned clades in the full ICS phylogeny (see full ICS phylogeny in Supplementary Repository), however, none of the species represented in this clade have GCF 1. This observation suggests that the two enzymes in GCF 1 arose from an ancestral duplication event, which has been conserved over evolutionary time.

The three ICS protein-domain variations (see structural diversity section above) did not form monophyletic groupings, raising questions about the possibility of multiple emergences of the two variants with discontinuous distributions: the ICS-TauD and monodomain ICS variations (outer ring on Figure 4B). A parsimonious explanation is that an ICS-TauD ICS was the common ancestor of the ICSs found in GCFs 6, 7, and 9 and that multiple loss-of-function mutations led to the evolution of the monodomain ICS found in GCFs 1, 4, 5, and 8. However, we cannot rule out the possibility that the ICS-TauD could have independently emerged once in the ancestor of GCF 7 & *xan* and again in the ancestor of GCFs 6 & 9. It is also possible that the observed discontinuous patterns of ICS domain variations result from hemiplasy caused by incomplete lineage sorting, rather than gene loss or independent emergences (78). The *crm* GCF is the only example of an ICS-NRPS-like backbone gene (16). However, the GCF with an ICS that is the closest relative of *crm* (GCF 8) contained a monodomain ICS backbone that is co-localized with an NRPS gene identical to the NRPS-like region in *crm* ICS-NRPS-like backbones (Supplementary Figure S11). A parsimonious interpretation is that a mono domain ICS fused with a nearby NRPS in some lineages, but that these genes remained separate in an ancestral *Fusarium* species.

We observed three main patterns of distribution among GCFs. GCFs 5, 8 and 9 had narrow distributions and were primarily present in 1–2 closely related genera (Figure 4C). Similar distributional patterns of canonical GCFs are thought to reflect the conserved functionality of a GCF that may help define the ecological niche of closely related taxa (1,79–81). GCFs *xan*, 1, 4 and 7 had discontinuous distributions (Figure 4C) but were more broadly spread across taxonomic groups. Three GCFs—*dit*, *crm*, and 6—were widely distributed across Ascomycota (Figure 4C). While these
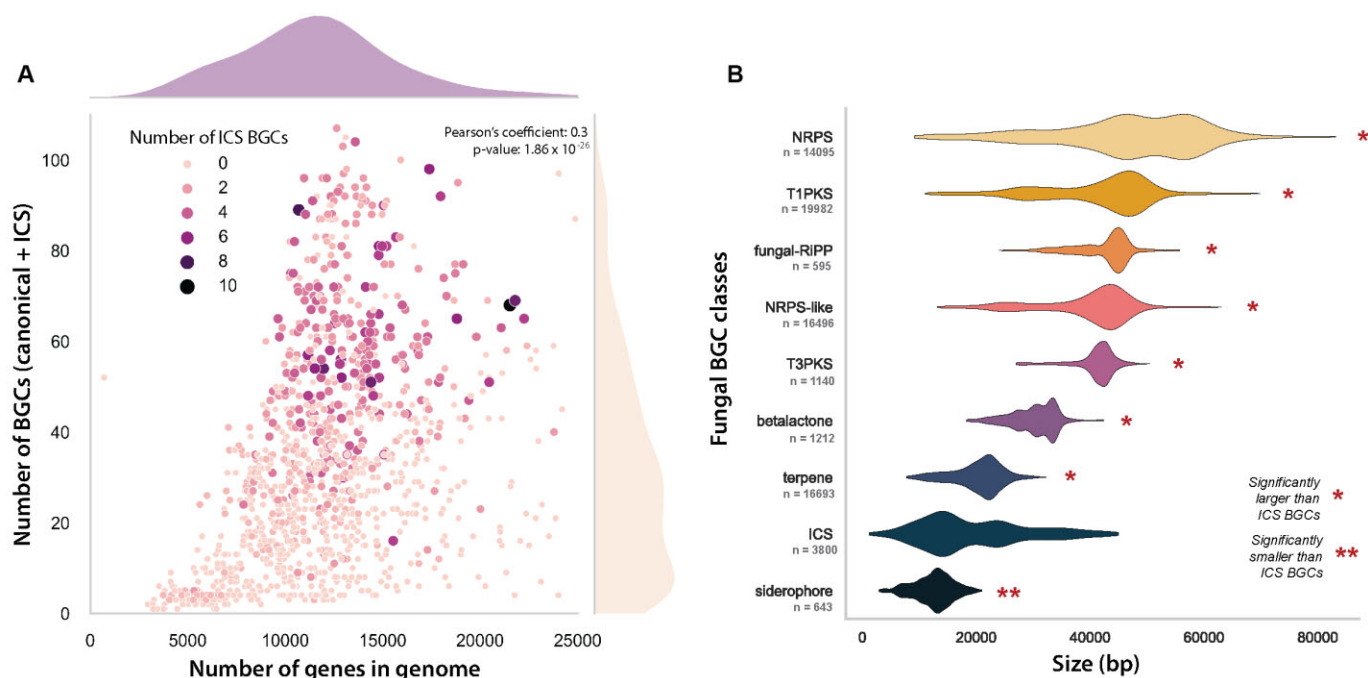
Figure 3. The relationships between isocyanide synthase (ICS) biosynthetic gene clusters (BGCs), canonical BGCs (predicted by antiSMASH), and genome sizes. (**A**) A scatterplot of the total number of BGCs (canonical + ICS) and the number of genes in a genome. Each point represents a genome and is colored and sized based on the number of ICS BGCs predicted within the genomes. Density plots are added to the margins of the scatterplot to show the distribution of the datapoints along the x and y axes. (**B**) The distribution of BGC sizes (in base pairs) for canonical BGC classes, and the ICSs elucidated in this study (n = BGCs). BGC classes that were significantly different (larger or smaller) in size (rank sums test; see Supplementary Methods) from ICS BGCs are marked on the plot.

latter GCFs are not found in every lineage (Supplementary Figure S9), similar distributional patterns of the pseurotin gene cluster have recently been explained with vertical transmission – a pattern that emphasizes the potential importance of incomplete lineage sorting and gene loss even in SMs that often have strong ecological phenotypes (82). The latter three GCFs are widespread and account for 27.5% of all fungal ICS BGCs. For example, ICS BGCs from GCF 6 were present in the genomes of Eurotiomycete, Sordariomycete, Leotiomycete, Dothideomycete, and a single Basidiomycete species. Previous studies have suggested that BGCs involved in essential biological functions, such as growth and development, tend to be highly conserved and broadly distributed (76,83). The distributional patterns we identify offer insights into the niche overlap of these lineages and raise questions about evolutionary forces maintaining these genes in specific lineages.

### Protein domains found within ICS GCFs

Canonical BGCs contain a core synthase and/or synthetase, as well as tailoring enzyme genes (4), and frequently include genes for transporters, protective enzymes, and transcription factors (70,81). Available evidence suggests that ICS BGCs share many of the same genomic signatures, as all characterized bacterial and fungal ICS BGCs contain both a backbone synthase and tailoring enzymes (19,23,27,28). However, the importance of specific protein domains in the function of ICS BGCs remains poorly understood. To identify protein domains associated with the

evolution of ICS BGCs, we analyzed protein domains of all 3800 ICS BGCs and within a subset of 10 ICS GCFs (see Materials and Methods; Figure 4).

Major facilitator superfamily (MFS) genes, which are typically associated with substrate transport (84), were the most common type of gene (after ICS backbones) in ICS BGCs, with approximately 60% of all ICS BGCs containing at least one MFS (Supplementary Table S10). The following three most common genes encoded cytochrome p450s (34.6%), zinc finger transcription factors (21.3%), and isocyanide hydratases (21.1%), the latter hypothesized to provide self-protection from isocyanide moieties (85). We further confirmed that these four domains are more frequently collocated with ICS backbone genes than can be explained by random chance (binomial distribution: $P < 0.01 \times 10^{-9}$) by comparing the observed cooccurrence of domains in a genomic window to a null distribution calculated with a binomial distribution (refer to Supplementary Methods). For a complete list of the most common protein domains found in all the ICS BGCs, refer to Supplementary Table S10.

We then looked at the biosynthetic core of 10 sampled GCFs. A GCF's biosynthetic core is highly conserved and typically represents the minimum set of genes essential to produce a given SM pathway (25,42). Therefore, we hypothesized that the ICS GCF biosynthetic cores would contain a higher proportion of genes associated with SM biosynthesis, transportation, or regulation than all the genes present within the ICS BGCs. Among the 57 core genes analysed, 48 (84.2%) were found to encode protein domains associated with SM production (Supplementary
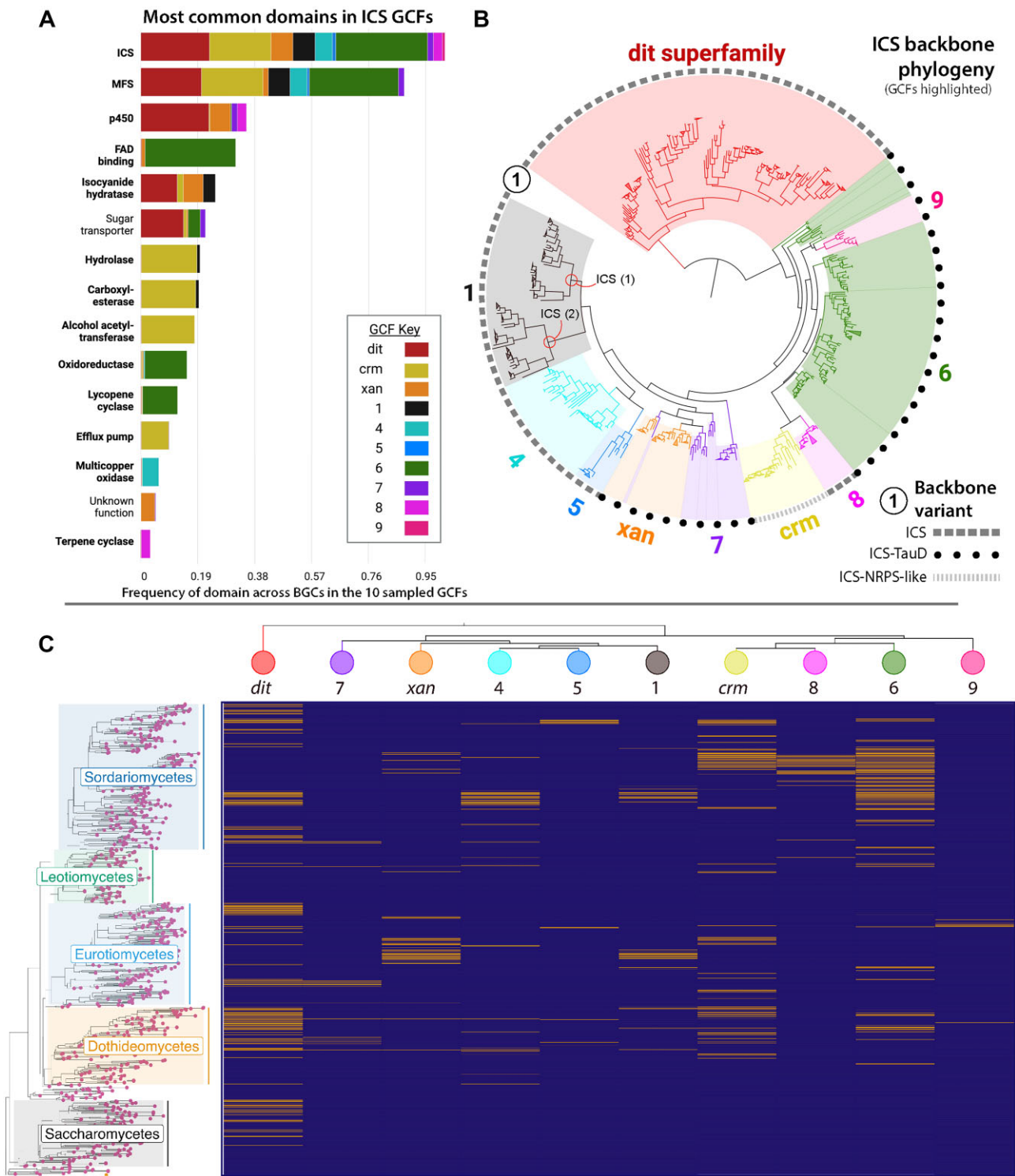
Figure 4. Architecture, evolution, and distribution of the 10 sampled isocyanide synthase (ICS) gene cluster families (GCFs) under significant linkage disequilibrium (LD). (**A**) A subset of the most common protein domains in the ICS GCFs. Domains frequently found within ICS backbones variations (e.g. TauD, NRPS-like) were filtered out, except for the ICS domain (top row) which is present in 100% of the clusters. GCFs are labeled with colors (see key), and domains implicated in specialized metabolism biosynthesis, transportation, or regulation are bolded. An un-subset version of the plot can be found in Supplementary Figure S12. (**B**) Midpoint rooted maximum likelihood phylogeny of ICS proteins from the analyzed GCFs. The tree is colored and labeled with the same key found in panel 'a'. The outer black ring conveys the variation of the ICS enzyme based on its dashed pattern. (**C**) Presence (yellow) and absence (purple) heatmap of each GCF mapped onto a species tree of Ascomycota. Columns are sorted by the relatedness of each GCF's ICS backbone using a collapsed ICS protein phylogeny.

Figure S10b; Supplementary Table S11). This is nearly 30% higher than the prevalence of these domains across all genes (including accessory) found in ICS BGCs (Supplementary Figure S10a).

The most abundant protein domains found in GCF core genes were also the most common overall domains in ICS BGCs (encoding MFSs, p450s and isocyanide hydratases), emphasizing their importance to isocyanide metabolic pathways. The functionality of other core genes was specific to individual GCFs. These genes encoded alpha/beta-hydrolases, alcohol acetyltransferases, oxidoreductases, efflux pumps, lycopene cyclases, multicopper oxidases, and terpene cyclases, among others, as shown in Figure 4A, Supplementary Figure S11, and Supplementary Figure S12. Overall, many of the commonly co-opted genes are similar to those found in canonical fungal BGCs.

Lastly, we checked for the presence of ICS BGCs that also contain a canonical synthetase and or synthase. BGCs containing two or more synthetases/synthases are not uncommon in the fungal kingdom (86) and can contribute to the chemical complexity of SMs (87). We focused our search on ICS BGCs containing NRPS, PKS, or terpene domain proteins (Table 1; Supplementary Methods). We found 12.8% ($n = 685$) of all ICS BGCs had genes encoding various NRPS domains. Most of these cases were explained by ICS-NRPS-like variants within GCF *crm* that commonly contain an ICS domain, an AMP binding domain, a phosphopantetheine (PP) loading domain, and a transferase domain (16). A total of 95 ISC BGCs contained genes encoding both NRPS-like proteins and terpene synthases; the majority (89.5%) of these cases were found in the *Fusarium*-specific GCF 8 (Supplementary Table S12). A small number of BGCs ($n = 12$) also contained a putative PKS within the ICS BGC, primarily in the genomes of select Dothideomycetes and Sordariomycetes species. Although multiple backbone genes have been shown to interact to form a single SM (88), in other cases, BGCs with multiple backbone genes have been shown to encode distinct products (89). Whether or not these genes encode activities that contribute to ICS chemistry is yet to be determined.

Selection maintains ICS GCFs

We validated BGC predictions using the principles of linkage disequilibrium (LD) to test the hypothesis that the colocalization of genes in BGCs is most likely explained by selection, not genetic drift (see Methods) (10). Between species, LD can be impacted by recombination rates, genetic drift, sampling, and selection. However, we aimed to isolate the impact of selection by comparing LD estimates within BGCs to estimates made across the full genome (we assume that drift and recombination impact the whole genome). This approach found significant LD maintaining the structure of three known ICS GCFs: *dit*, *crm* and *xan*. Based on these results, we focused our search on a small subset of 15 out of the 200 GCFs (refer to Materials and Methods section on estimating LD). This subset was necessary to capture the most taxonomically widespread ICS GCFs because high LD found within species and between closely related species offers little resolution to identify genes under selection. Two of the GCFs exhibited strong LD ($p < 0.07$)

and 12 GCFs had significant LD ($p < 0.05$) (Supplementary Figure S13). A single GCF primarily present in *Colletotrichum* species (GCF 26 in Supplementary Figure S13n) exhibited no significant LD patterns ($P = 0.15$). In this GCF, approximately 75% of BGC predictions included a core set of 7–10 SM-related genes (e.g. transcription factors, tailoring enzymes, transporters), while the remaining 25% contained only 1–3 genes. Upon manual inspection, we identified that the smaller BGC predictions were colocalized with orthologs found in the larger predictions. However, these genes did not share the same promoter motif as was found around the ICS backbone gene. These results suggest that either new genes have been recruited to the cluster in some lineages or, conversely, these genes are no longer coregulated with the ICS. Different regulatory patterns between species, like those we inferred here, have previously been associated with ICS gene clusters. The cluster-specific transcription that is part of and regulates the ICS *xan* cluster in *A. fumigatus* instead regulates the citrinin gene cluster in *Penicillium expansum* where the *xan* promoter motif has been lost (41).

The 14 validated GCFs account for nearly 50% of all fungal ICS BGCs (1865/3800). While we cannot entirely rule out other factors that influence LD, the significant co-occurrence of the same orthologous genes in these ICS GCFs relative to whole-genome levels of LD strongly suggests that selection is maintaining the structure of ICS GCFs. Similar evidence has also been used to infer the selection of BGCs by Gluck-Thaler *et al* (2020). The physical clustering of genes into BGCs is often attributed to the stepwise process of SM biosynthesis. The backbone gene synthesizes a core chemical moiety that is then adorned by tailoring enzymes. Pathway intermediates may be passaged through cellular organelles (e.g. peroxisomes) by transporters and the final SM product may then be exported from the fungus through additional transporters. The loss of any gene in this process can disrupt SM synthesis, even leading to toxic impacts on the producing fungus (90). While we were unable to validate LD acting on GCFs with narrow taxonomic distributions or differentially evolved regulatory patterns, these results validate our approach to identifying ICS gene clusters and emphasize that these clusters are maintained by strong natural selection.

The *dit* superfamily is widely distributed across ascomycota

The *dit* cluster (29) is the archetypal ICS cluster for which the associated protein domain is named and defined in protein-domain-finding algorithms used here (31,34). In yeast, this cluster is made up of two genes, *dit1* (an ICS) and *dit2* (a p450), that together synthesize dityrosine. Dityrosine is a structural component required for the normal formation of ascospore cell walls in *S. cerevisiae*. This compound has only been reported in the yeasts *S. cerevisiae* (29) and *Candida albicans* (91). While a recent study raised questions if this cluster encodes a true ICS (92), most other papers recognize this cluster as an ICS, and all proposed biosynthetic models and supporting experimental evidence suggest that Dit1 synthesizes an isocyanide intermediate (formyl tyrosine) from L-tyrosine, and thus conforms to our definition of an ICS (15,93).

Although *dit1* encodes an ICS, this backbone synthase is deeply diverged from all other ICS enzymes (31). Indeed, we confirmed Lim et al. (2018)'s finding that the *dit* ICS is more closely related to bacterial ICSs than to other fungal ICSs (Supplementary Figure S14). This divide is particularly clear when plotting this phylogeny as an unrooted tree, where two distinct clades are evident (Figure 5B). Similar deep phylogenetic divides have been used to infer convergent evolution (94). Although additional research is needed, we speculate that the *dit*-variety fungal ICSs and bacterial ICSs are descendants of a common ancestor, while other fungal ICS enzymes have originated through convergent evolution from a convergently evolved ancestor. Similar patterns of convergence in SM backbone proteins have been documented within plant and microbial terpene synthases (94,95). To explain the distribution of *dit1* in both fungi and bacteria, we suggest that either an ancient horizontal transfer event occurred or that this cluster is truly ancient and was present in a common ancestor of these two kingdoms. Investigations of bacterial ICS are needed to clarify the ancient evolutionary history of these genes. It is important to note that incomplete lineage sorting or differential selective pressures may have impacted our ability to accurately reconstruct the evolutionary history of ICS enzymes, and thus, the possibility of alternative evolutionary scenarios cannot be entirely ruled out. To shed light on the complicated evolutionary history involving bacteria, a prokaryotic ICS BGC-mining pipeline that does not rely on eukaryotic promoter motifs must be developed.

Three primary subclades were identified within the larger *dit* clade of ICS enzymes, all of which were deeply diverged from the ICSs of other GCFs (Figure 5B). The subclades within the *dit* phylogeny largely correspond to three major variations in BGC gene content (denoted as Clans 1–3). The conservation of genes flanking the *dit1*/2 genes could result from genetic hitchhiking, potentially impacting the allele frequencies of neighboring neutral genes (96). However, as flanking genes frequently shared cluster-specific promoters with the *dit1/2,* it raises questions about their potential functionality as a coregulated unit. The variation in genomic structure may reflect the divergent evolution of unique functions within specific taxa. Together, the three *dit* Clans constitute what we refer to as the *dit* superfamily (as referred to in Figures 4 and 5). Clan 1 was found primarily in Saccharomycete species, whereas Clan 2 was distributed more broadly across Eurotiomycete and Sordariomycete species, and Clan 3 was predominantly localized in the Dothideomycetes (Figure 5D). These findings shift our understanding of this cluster as it was previously studied in only yeast-like organisms (28,91,92). All three Clans are defined by the presence of both an ICS and a p450 whose genes are in the same orientation as the Saccharomycetes *dit1* (ICS) and *dit2* (p450) (Figure 5A). Therefore, we sought to verify that the p450 genes were true *dit2* orthologs as opposed to being independently co-opted p450 proteins (i.e. convergent evolution). Phylogenetic analysis of the putative Dit2 p450s, together with the many other p450s found throughout relevant genomes (including those found in other ICS BGCs), revealed a monophyletic clade of Dit2 sequences distinct from other fungal p450s (Figure 5C). A topological comparison between the Dit1 and Dit2

phylogenies revealed that the two genes closely matched, indicating that the genes have been co-evolving (Supplementary Figure S15). This data suggests that all three *dit* cluster Clans originated from the same ancestral *dit1/2* cluster. Additionally, the ubiquitous presence of *dit1* and *dit2* as core genes across all *dit* cluster Clans and taxa provides compelling evidence of their indispensable role in the biosynthesis of *dit* cluster metabolites, regardless of the specific compound(s) produced (i.e. divergent metabolites or conserved dityrosine production).

In total, 17.1% of Ascomycete species in our dataset contained a locus with co-localized *dit1/2* homologs. An additional 11.2% of Ascomycetes have a fragmented variation where the *dit1* gene is present in a different genomic locus than *dit2* (Supplementary Table S13). It is uncertain whether the observed fragmentation is entirely due to biological factors or partly due to fragmented genome assemblies. Given the ambiguity surrounding the functionality and divergence of the fragmented *dit* variants, all analyses in this study referencing the *dit* superfamily focused on species that exhibit co-localized loci with *dit1/2*.

The *dit* superfamily phylogeny is largely congruent with the fungal species tree

The *dit* superfamily is the most widely distributed ICS GCF within fungi. We sought to determine the contribution of horizontal and vertical transmission in generating its wide-ranging distribution. Phylogenetic relationships inferred from Dit proteins (rooted at most recent common ancestor of Saccharomycetes) closely matched those in the species tree among most Saccharomycete, Sordariomycete, Eurotiomycete, and Dothideomycete species (Figure 5D); this pattern is consistent with vertical transmission and lineage-specific loss. This result provides compelling evidence that the *dit* cluster's origins in Ascomycota are ancient.

Our analysis identified a specific region on the Dit tree that exhibited phylogenetic incompatibility with the species tree (Figure 5D). This clade comprises species from major Pezizomycotina lineages, including Sordariomycetes, Eurotiomycetes, Dothideomycetes, and the entire order Onygenales. Several of these incompatibilities resulted from a duplicate copy of the dit cluster, where one of the copies in a genome was congruent and the other was not. While some of the differences between trees are deep in the Dit1/2 phylogeny, the difference in Onygenales (where there were no duplications observed), are relatively shallow. In addition to incompatibilities, the branch lengths within this region were very long, suggesting rapid divergence of the Dit1/2 proteins in this group. The elongation of branch lengths in this region is primarily driven by the divergence of the Dit1 backbone ICS (Supplementary Figure S15). While we cannot fully rule out the possibility that the short sequences here do not provide full phylogenetic resolution of true evolutionary histories, the placement of these rapidly evolving Dit1/2 proteins is strongly supported by the bootstrap (Dit tree) and posterior probabilities (species tree). Onygenales are the only major lineage of Ascomycetes that are known to be dimorphic (having filamentous and yeast-like growth forms) (97,98). The observed pattern in this order could
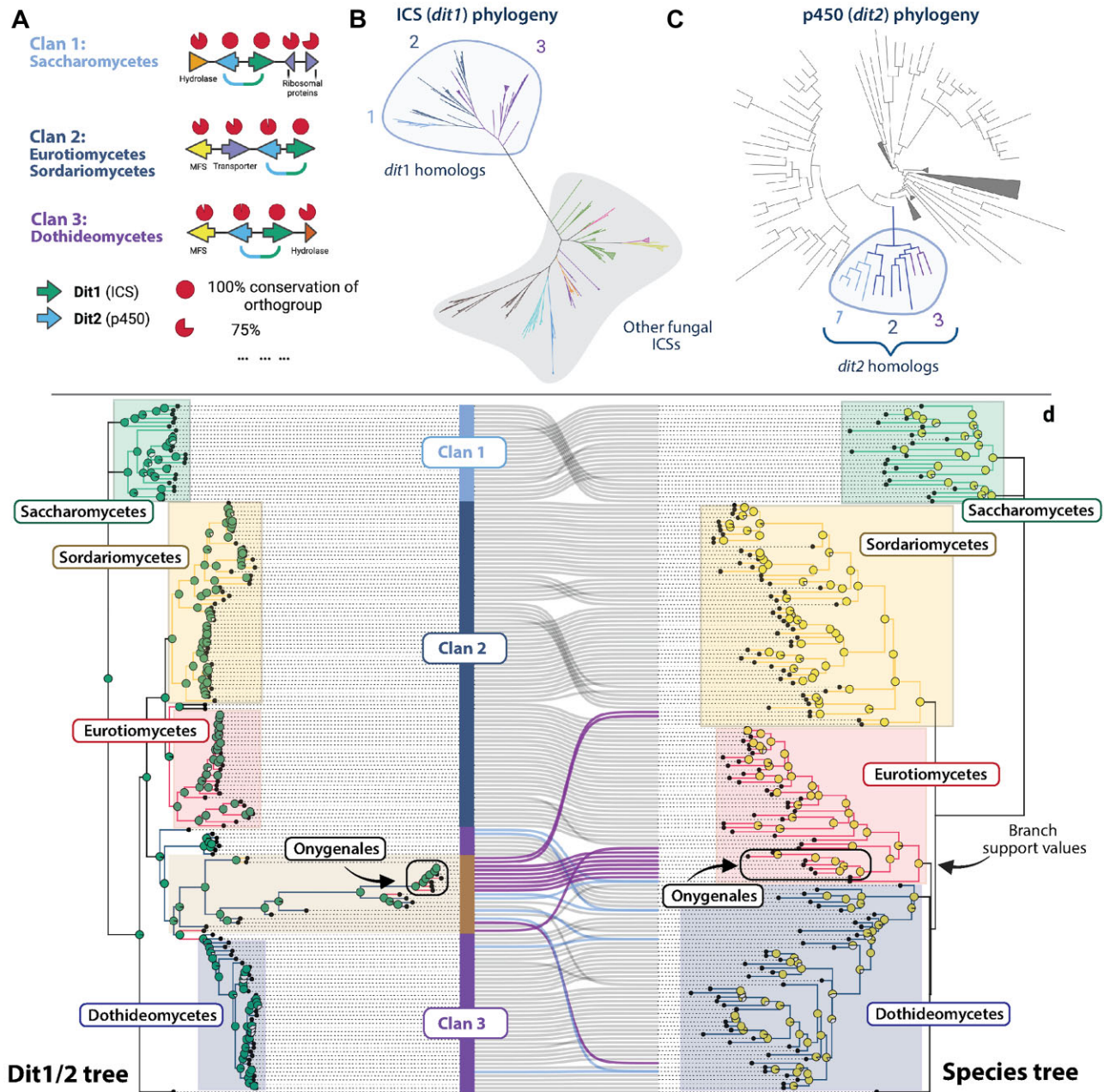
Figure 5. Reconstructed evolutionary history of the *dit* superfamily. (**A**) The biosynthetic gene cluster (BGC) architectures of the three *dit* clans that make up the *dit* superfamily. The conservation of orthologs encoding specific protein domains is indicated above and below the corresponding gene arrow, respectively. The primary taxonomic class(es) containing the most representatives of each clan is labeled. (**B**) Unrooted maximum likelihood ICS phylogeny of backbones found in the 10 sampled GCFs (the unfiltered ICS tree can be found in the Supplementary Repository). The 'lower' clade's ICS enzymes (marked in grey) are colored to show the diversity of GCFs. All ICSs within the *dit* superfamily (i.e. *dit1* homologs) are found in the 'upper' clade (marked with light-blue) and colored according to their clan. (**C**) The p450 proteins found within the *dit* superfamily BGCs were mapped onto a phylogeny of other p450s found in 10 selected genomes to establish if they are all homologs of *dit2*. As in panel 'B', the clans are colored and labeled accordingly. (**D**) Topological comparison between the concatenated maximum likelihood phylogeny of the Dit1/2 protein sequences (left) and the coalescent-based species tree (right). Branch support values are displayed as pie charts, with the left tree indicating bootstrap support values (1000 bootstrap replicates) and the right tree posterior probabilities. Phylogenetic relationships are highlighted as follows: (i) Purple: incongruencies between gene tree and species tree, (ii) Light blue: when a species has two copies of the *dit* cluster, both are highlighted regardless of incongruence. The most common *dit* clan of each taxonomic class is indicated on the Dit1/2 Tree. The highlighted clade (tan) on the Dit1/2 tree represents the region associated with most of the observed incongruencies and long branch lengths. This region also includes all the sequences corresponding to the dimorphic order, Onygenales.

be explained by the loss of the *dit1/2* cluster and the subsequent transfer of the putatively divergent *dit1/2* cluster prior to the speciation of this taxonomic order. The divergence of this Dit1/2 clade raises questions about the role of these genes in certain lineages and offers opportunity to explore the diversification of duplicated *dit* genes.

The *dit1/2* cluster in *S. cerevisiae* produces dityrosine, an SM incorporated into the protective outer-cell wall layer in its ascospores (28,29). Deleting *dit2* in *C. albicans* led to changes in hyphal morphology and increased susceptibility to antifungals, which is believed to be caused by the absence of a dityrosine cell wall layer (91). As reported by Beauvais and Latgé (2018), it remains unclear whether dityrosine is incorporated into the cell walls of filamentous fungi, as current structural models of ascospores are limited to those of *S. cerevisiae* (99). We found that up to 30% of Ascomycetes possess a co-localized or fragmented copy of *dit1* and *dit2*, with the genomes of select species (i.e. *Endocarpon pusillum*, *Botryosphaeria dothidea*) encoding multiple copies. These duplicated copies are part of a quickly evolving region of the Dit1/2 phylogeny that also includes sequences from Onygenales (discussed above), a group of fungi known to be pathogenic and have dimorphic growth patterns. Our results raise questions about the potential role of dityrosine in the formation of many Ascomycete ascospores and cell walls. Future studies should aim to characterize the *dit* superfamily SM products across a broader range of taxa, specifically in the Pezizomycotina, to better understand its chemical and ecological significance.

CONCLUSIONS

Our development of a new genome-mining pipeline has allowed us to identify ICS BGCs across the fungal kingdom, identifying evolutionary patterns that will help inform future research into these compounds. The vast majority of ICSs are part of BGCs that share regulatory motifs (3800 out of the 4341). These BGCs are maintained as contiguous clusters over evolutionary time by natural selection, similarly to canonical clusters. ICS clusters are significantly smaller than most canonical clusters, perhaps suggesting fewer chemical alterations are made to these compounds, although we cannot fully rule out the potential impact of different mining algorithms (i.e. antiSMASH vs CAS-SIS) on the observed size differences. However, the highly-reactive nature of isocyanide intermediates raises the possibility of trans-BGC interactions and yet-unexplored chemistry (16,22). Indeed, we elucidate that 15.3% of ICS BGC clusters contain synthases from other classes of SMs. While some of these clusters may represent co-localized BGCs that produce chemically distinct products as described for the fumagillin/pseurotin cluster in *A. fumigatus* (89), we speculate that some are true hybrid clusters. Our use of conserved regulatory elements to identify BGCs also allowed us to identify patterns that may help future efforts to understand the regulation and evolution of ICS BGCs.

A future goal should be to examine the relationship of isocyanide SMs with metal homeostasis. The reactive nature of isocyanide SMs has been attributed to the isocyanide functional group's ability to interact with metal compounds (21). The *A. fumigatus xan* and *crm* BGCs are regulated by copper and copper transcription factors and their SM products aid in metal-homeostasis and metal-associated competitive interactions (16,19,31). We find that the ICS BGCs are three times more numerous than siderophore BGCs, encoding the well-known iron-chelating SMs (100). The unique chemical properties of isocyanide metabolites and the abundance of ICS BGCs identified in this study underscore the potential of these pathways in shaping metal-associated microbial interactions and fungal ecology.

Finally, we find that ICS BGCs are not evenly distributed across fungi, with some taxa, specifically those within the Pezizomycotina, containing disproportionally high numbers of these clusters (Fisher's exact test: $P = 1.99^{-42}$; Supplementary Figure S2). While a similar pattern has been observed for canonical gene clusters, we identify that part of this trend can be explained by genome size. Individual GCFs show different distributional patterns, raising questions about the evolutionary histories of these clusters and their role in the niche-adaptation of specific lineages. The most widely distributed ICS GCF was the *dit* superfamily and was found in nearly 30% of all Ascomycetes. The evolutionary history of the dit gene cluster was found to closely align with the species tree, providing strong evidence for its ancient origins within Ascomycota. Furthermore, we observed a distinct group of Dit proteins that displayed incongruent evolutionary patterns and long branch lengths, highlighting its unique evolutionary trajectory. Our findings shed light on important questions for future ICS BGC research aiming to discover and utilize the diverse and valuable ecological properties of isocyanide metabolites (e.g. highly reactive metal-associated microbial natural products) for anthropocentric aims (18,20,22). With our web server (www.isocyanides.fungi.wisc.edu), researchers can easily explore, filter, and download all identified ICS BGCs and GCFs, streamlining future research and discovery.

DATA AVAILABILITY

All the genomes used in this study were obtained from the public domain (NCBI database). Versions of all software used can be found in Supplementary Table S15. The Supplementary Repository contains a reproducible script, every generated ICS BGC prediction, tree files and logs, and additional GCF tables. This can be accessed in the publication's GitHub repository https://github.com/gnick18/Fungal_ICSBGCs or https://doi.org/10.5281/zenodo.7834839.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Keller,N.P., Turner,G. and Bennett,J.W. (2005) Fungal secondary metabolism - from biochemistry to genomics. *Nat. Rev. Microbiol.*, **3**, 937–947.
2. López-Gresa,M.P., González,M.C., Ciavatta,L., Ayala,I., Moya,P. and Primo,J. (2006) Insecticidal activity of paraherquamides, including paraherquamide H and paraherquamide I, two new alkaloids isolated from *Penicillium cluniae*. *J. Agric. Food. Chem.*, **54**, 2921–2925.
3. Balba,H. (2007) Review of strobilurin fungicide chemicals. *J. Environ. Sci. Health B.*, **42**, 441–451.
4. Keller,N.P. (2019) Fungal secondary metabolism: regulation, function and drug discovery. *Nat. Rev. Microbiol.*, **17**, 167–180.
5. Khaldi,N., Seifuddin,F.T., Turner,G., Haft,D., Nierman,W.C., Wolfe,K.H. and Fedorova,N.D. (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal. Genet. Biol.*, **47**, 736–741.
6. Blin,K., Shaw,S., Steinke,K., Villebro,R., Ziemert,N., Lee,S.Y., Medema,M.H. and Weber,T. (2019) AntiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, **47**, W81–W87.
7. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., De Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
8. Robey,M.T., Caesar,L.K., Drott,M.T., Keller,N.P. and Kelleher,N.L. (2021) An interpreted atlas of biosynthetic gene clusters from 1000 fungal genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2020230118.
9. Caesar,L.K., Butun,F.A., Robey,M.T., Ayon,N.J., Gupta,R., Dainko,D., Bok,J.W., Nickles,G., Stankey,R.J., Johnson,D. *et al.* (2023) Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs. *Nat. Chem. Biol.*, **19**, 846–854.
10. Gluck-Thaler,E., Haridas,S., Binder,M., Grigoriev,I.V., Crous,P.W., Spatafora,J.W., Bushley,K. and Slot,J.C. (2020) The architecture of metabolism maximizes biosynthetic diversity in the largest class of Fungi. *Mol. Biol. Evol.*, **37**, 2838–2856.
11. Drott,M.T., Park,S.C., Wang,Y.w., Harrow,L., Keller,N.P. and Pringle,A. (2023) Pangenomics of the death cap mushroom Amanita phalloides, and of Agaricales, reveals dynamic evolution of toxin genes in an invasive range. *ISME J.*, https://doi.org/10.1038/s41396-023-01432-x.
12. Hoskisson,P.A. and Seipke,R.F. (2020) Cryptic or silent? The known unknowns, unknown knowns, and unknown unknowns of secondary metabolism. *Mbio*, **11**, e02642-20.
13. Caesar,L.K., Kelleher,N.L. and Keller,N.P. (2020) In the fungus where it happens: history and future propelling Aspergillus nidulans as the archetype of natural products research. *Fungal. Genet. Biol.*, **144**, 103477.
14. Edenborough,M.S. and Herbert,R.B. (1988) Naturally occurring isocyanides. *Nat. Prod. Rep.*, **5**, 229–245.
15. Chen,T.Y., Chen,J., Tang,Y., Zhou,J., Guo,Y. and Chang,W.c. (2021) Current Understanding toward Isonitrile Group Biosynthesis and Mechanism. *Chin. J. Chem.*, **39**, 463–472.
16. Won,T.H., Bok,J.W., Nadig,N., Venkatesh,N., Nickles,G., Greco,C., Lim,F.Y., González,J.B., Turgeon,B.G., Keller,N.P. *et al.* (2022) Copper starvation induces antimicrobial isocyanide integrated into two distinct biosynthetic pathways in fungi. *Nat. Commun.*, **13**, 4828.
17. Crawford,J.M., Portmann,C., Zhang,X., Roeffaers,M.B.J. and Clardy,J. (2012) Small molecule perimeter defense in entomopathogenic bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 10821–10826.
18. Massarotti,A., Brunelli,F., Aprile,S., Giustiniano,M. and Tron,G.C. (2021) Medicinal chemistry of isocyanides. *Chem. Rev.*, **121**, 10742–10788.
19. Raffa,N., Won,T.H., Sukowaty,A., Candor,K., Cui,C., Halder,S., Dai,M., Figueroa,J.L., Schroeder,F.C. and Keller,N. (2021) Dual-purpose isocyanides produced by Aspergillus fumigatus contribute to cellular copper sufficiency and exhibit antimicrobial activity. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2015224118.
20. Hübner,I., Shapiro,J.A., Hoßmann,J., Drechsel,J., Hacker,S.M., Rather,P.N., Pieper,D.H., Wuest,W.M. and Sieber,S.A. (2021) Broad spectrum antibiotic Xanthocillin X effectively kills *Acinetobacter baumannii* via dysregulation of heme biosynthesis. *ACS Cent. Sci.*, **7**, 488–498.
21. Pruchnik,F.P. and Duraj,S.A. (1990) Isocyanide Complexes. In: *Organometallic Chemistry of the Transition Elements*. pp. 617–645.
22. Galli,U., Tron,G.C., Purghè,B., Grosa,G. and Aprile,S. (2020) Metabolic fate of the isocyanide moiety: are isocyanides pharmacophore groups neglected by medicinal chemists?*Chem. Res. Toxicol.*, **33**, 955–966.
23. Stintzi,A., Johnson,Z., Stonehouse,M., Ochsner,U., Meyer,J.M., Vasil,M.L. and Poole,K. (1999) The pvc gene cluster of *Pseudomonas aeruginosa*: role in synthesis of the pyoverdine chromophore and regulation by PtxR and PvdS. *J. Bacteriol.*, **181**, 4118–4124.
24. Brady,S.F. and Clardy,J. (2005) Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angew. Chem. - Int. Ed.*, **44**, 7063–7065.
25. Brady,S.F., Bauer,J.D., Clarke-Pearson,M.F. and Daniels,R. (2007) Natural products from isnA-containing biosynthetic gene clusters recovered from the genomes of cultured and uncultured bacteria. *J. Am Chem. Soc.*, **129**, 12102–12103.
26. Hillwig,M.L., Fuhrman,H.A., Ittiamornkul,K., Sevco,T.J., Kwak,D.H. and Liu,X. (2014) Identification and characterization of a welwitindolinone alkaloid biosynthetic gene cluster in the stigonematalean cyanobacterium Hapalosiphon welwitschii. *Chem. Biol. Chem.*, **15**, 665–669.

27. Zhang,X., Evanno,L. and Poupon,E. (2020) Biosynthetic Routes to Natural Isocyanides. *European J Org Chem*, **2020**, 1919–1929.

28. Briza,P., Breitenbach,M., Ellinger,A. and Segall,J. (1990) Isolation of two developmentally regulated genes involved in spore wall maturation in *Saccharomyces cerevisiae*. *Genes Dev*, **4**, 1775–1789.

29. Briza,P., Eckerstorfer,M. and Breitenbach,M. (1994) The sporulation-specific enzymes encoded by the DIT1 and DIT2 genes catalyze a two-step reaction leading to a soluble LL-dityrosine-containing precursor of the yeast spore wall. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 4524–4528.

30. Friesen,H., Hepworth,S.R. and Segall,J. (1997) An Ssn6-Tup1-dependent negative regulatory element controls sporulation-specific expression of DIT1 and DIT2 in Saccharomyces cerevisiae. *Mol. Cell Biol.*, **17**, 123–134.

31. Lim,F.Y., Won,T.H., Raffa,N., Baccile,J.A., Wisecaver,J., Keller,N.P., Rokas,A. and Schroeder,F.C. (2018) Fungal isocyanide synthases and xanthocillin biosynthesis in Aspergillus fumigatus. *Mbio*, **9**, e00785-18.

32. Hang,L., Tang,M.C., Harvey,C.J.B., Page,C.G., Li,J., Hung,Y.S., Liu,N., Hillenmeyer,M.E. and Tang,Y. (2017) Reversible product release and recapture by a fungal polyketide synthase using a carnitine acyltransferase domain. *Angew. Chem. Int. Ed Engl.*, **56**, 9556–9560.

33. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. and Finn,R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.

34. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

35. Amaike,S. and Keller,N.P. (2011) Aspergillus flavus. *Annu. Rev. Phytopathol.*, **49**, 107–133.

36. Wolf,T., Shelest,V., Nath,N. and Shelest,E. (2016) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, **32**, 1138–1143.

37. Wolf,T., Shelest,V. and Shelest,E. (2013) Motif-based method for the genome-wide prediction of eukaryotic gene clusters. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.Vol. **8158**, pp. 389–398.

38. Navarro-Muñoz,J.C., Selem-Mojica,N., Mullowney,M.W., Kautsar,S.A., Tryon,J.H., Parkinson,E.I., de Los Santos,E.L.C., Yeong,M., Cruz-Morales,P., Abubucker,S. *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, **16**, 60–68.

39. Gilchrist,C.L.M., Booth,T.J., van Wersch,B., van Grieken,L., Medema,M.H. and Chooi,Y.-H. (2021) Cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinform. Adv.*, **1**, vbab016.

40. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498–2504.

41. Wang,W., Drott,M., Greco,C., Luciano-Rosario,D., Wang,P. and Keller,N.P. (2021) Transcription factor repurposing offers insights into evolution of biosynthetic gene cluster Regulation. *Mbio*, **12**, e0139921.

42. Vignolle,G.A., Schaffer,D., Zehetner,L., Mach,R.L., Mach-Aigner,A.R. and Derntl,C. (2021) FunOrder: a robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. *PLoS Compu. Biol.*, **17**, e1009372.

43. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654, .

44. Virtanen,P., Gommers,R., Oliphant,T.E., 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. *Nat Methods*, **17**, 261–272.

45. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

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

47. Nguyen,L.T., Schmidt,H.A., Von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

48. Minh,B.Q., Nguyen,M.A.T. and Von Haeseler,A. (2013) Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*, **30**, 1188–1195.

49. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., Von Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, **14**, 587–589.

50. Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

51. Yu,G., Smith,D.K., Zhu,H., Guan,Y. and Lam,T.T.Y. (2017) Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.

52. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.

53. Mirarab,S. and Warnow,T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, **31**, i44–i52.

54. Bravo,G.A., Antonelli,A., Bacon,C.D., Bartoszek,K., Blom,M.P.K., Huynh,S., Jones,G., Lacey Knowles,L., Lamichhaney,S., Marcussen,T. *et al.* (2019) Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ*, **7**, e6399.

55. Hovmöller,R., Lacey Knowles,L. and Kubatko,L.S. (2013) Effects of missing data on species tree estimation under the coalescent. *Mol. Phylogenet. Evol.*, **69**, 1057–1062.

56. Zhang,C., Rabiee,M., Sayyari,E. and Mirarab,S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 15–30.

57. Li,Y., Steenwyk,J.L., Chang,Y., Wang,Y., James,T.Y., Stajich,J.E., Spatafora,J.W., Groenewald,M., Dunn,C.W., Hittinger,C.T. *et al.* (2021) A genome-scale phylogeny of the kingdom Fungi. *Current Biology*, **31**, 1653–1665.

58. Letcher,P.M. and Powell,M.J. (2018) A taxonomic summary and revision of Rozella (Cryptomycota). *IMA Fungus*, **9**, 383–399.

59. Fitzpatrick,D.A. (2012) Horizontal gene transfer in fungi. *FEMS Microbiol Lett*, **329**, 1–8.

60. Slot,J.C. and Rokas,A. (2011) Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Current Biology*, **21**, 134–139.

61. Paradis,E., Claude,J. and Strimmer,K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

62. Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

63. Smith,R.M. (2020) TreeDist: distances between phylogenetic trees. https://doi.org/10.5281/zenodo.3528123.

64. Kück,P. and Longo,G.C. (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front Zool*, **11**, 81.

65. Rokas,A. (2022) Evolution of the human pathogenic lifestyle in fungi. *Nat. Microbiol.*, **7**, 607–619.

66. Ohm,R.A., Feau,N., Henrissat,B., Schoch,C.L., Horwitz,B.A., Barry,K.W., Condon,B.J., Copeland,A.C., Dhillon,B., Glaser,F. *et al.* (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathog*, **8**, e1003037.

67. Haas,H. (2014) Fungal siderophore metabolism with a focus on Aspergillus fumigatus. *Nat. Prod. Rep.*, **31**, 1266–1276.

68. Barry,S.M. and Challis,G.L. (2009) Recent advances in siderophore biosynthesis. *Curr. Opin. Chem. Biol.*, **13**, 205–215.

69. Visca,P., Colotti,G., Serino,L., Verzili,D., Orsi,N. and Chiancone,E. (1992) Metal regulation of siderophore synthesis in *Pseudomonas aeruginosa* and functional effects of siderophore-metal complexes. *Appl. Environ. Microbiol.*, **58**, 2886–2893.

70. Rokas,A., Mead,M.E., Steenwyk,J.L., Raja,H.A. and Oberlies,N.H. (2020) Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat. Prod. Rep.*, **37**, 868–878.

71. Stöckmann,H., Neves,A.A., Stairs,S., Brindle,K.M. and Leeper,F.J. (2011) Exploring isonitrile-based click chemistry for ligation with biomolecules. *Org. Biomol. Chem.*, **9**, 7303–7305.

72. Tang,S., Wang,J., Xiong,Z., Xie,Z., Li,D., Huang,J. and Zhu,Q. (2017) Palladium-catalyzed imidoylative cyclization of tryptophan-derived isocyanides: access to β-carbolines. *Org. Lett.*, **19**, 5577–5580.

73. Santra,S., Andreana,T., Bourgault,J.P. and Andreana,P.R. (2016) Convertible Isocyanides: application in Small Molecule Synthesis, Carbohydrate Synthesis, and Drug Discovery. In: *Domino and Intramolecular Rearrangement Reactions as Advanced Synthetic Methods in Glycoscience*. Wiley Blackwell, pp. 121–194.

74. Drott,M.T., Rush,T.A., Satterlee,T.R., Giannone,R.J., Abraham,P.E., Greco,C., Venkatesh,N., Skerker,J.M., Glass,N.L., Labbé,J.L. *et al.* (2021) Microevolution in the pansecondary metabolome of *Aspergillus flavus* and its potential macroevolutionary implications for filamentous fungi. *Proc. Natl. Acad. Sci. U. S. A.*, **118**, e2021683118.

75. Lind,A.L., Wisecaver,J.H., Lameiras,C., Wiemann,P., Palmer,J.M., Keller,N.P., Rodrigues,F., Goldman,G.H. and Rokas,A. (2017) Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol*, **15**, e2003583.

76. Bushley,K.E. and Turgeon,B.G. (2010) Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol. Biol.*, **10**, 26.

77. Boettger,D., Bergmann,H., Kuehn,B., Shelest,E. and Hertweck,C. (2012) Evolutionary imprint of catalytic domains in fungal PKS-NRPS hybrids. *Chem. Biol. Chem.*, **13**, 2363–2373.

78. Hibbins,M.S., Gibson,M.J.S. and Hahn,M.W. (2020) Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *Elife*, **9**, e63753.

79. Hoogendoorn,K., Barra,L., Waalwijk,C., Dickschat,J.S., van der Lee,T.A.J. and Medema,M.H. (2018) Evolution and diversity of biosynthetic gene clusters in Fusarium. *Front. Microbiol.*, **9**, 1158.

80. Gluck-Thaler,E. and Slot,J.C. (2018) Specialized plant biochemistry drives gene clustering in fungi. *ISME Journal*, **12**, 1694–1705.

81. Inglis,D.O., Binkley,J., Skrzypek,M.S., Arnaud,M.B., Cerqueira,G.C., Shah,P., Wymore,F., Wortman,J.R. and Sherlock,G. (2013) Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans*, *A. fumigatus*, *A. niger* and *A. oryzae*. *BMC Microbiol*, **13**, 91.

82. Tannous,J., Cosetta,C.M., Drott,M.T., Rush,T.A., Abraham,P.E., Giannone,R.J., Keller,N.P. and Wolfe,B.E. (2023) LaeA-regulated fungal traits mediate bacterial community assembly. *Mbio*, **14**, e0076923.

83. Wapinski,I., Pfeffer,A., Friedman,N. and Regev,A. (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449**, 54–61.

84. Reddy,V.S., Shlykov,M.A., Castillo,R., Sun,E.I. and Saier,M.H. (2012) The major facilitator superfamily (MFS) revisited. *FEBS Journal*, **279**, 2022–2035.

85. Goda,M., Hashimoto,Y., Shimizu,S. and Kobayashi,M. (2001) Discovery of a novel enzyme, isonitrile hydratase, involved in nitrogen-carbon triple bond cleavage. *J. Biol. Chem.*, **276**, 23480–23485.

86. Chen,L., Yue,Q., Zhang,X., Xiang,M., Wang,C., Li,S., Che,Y., Ortiz-López,F.J., Bills,G.F., Liu,X. *et al.* (2013) Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus Glarea lozoyensis. *BMC Genomics*, **14**, 339.

87. Du,L. and Shen,B. (2001) Biosynthesis of hybrid peptide-polyketide natural products. *Curr. Opin. Drug Discov. Dev.*, **4**, 215–228.

88. Park,H.B., Perez,C.E., Barber,K.W., Rinehart,J. and Crawford,J.M. (2017) Genome mining unearths a hybrid nonribosomal peptide synthetase-like-pteridine synthase biosynthetic gene cluster. *Elife*, **6**, e25229.

89. Wiemann,P., Guo,C.J., Palmer,J.M., Sekonyela,R., Wang,C.C.C. and Keller,N.P. (2013) Prototype of an intertwined secondarymetabolite supercluster. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17065–17070.

90. McGary,K.L., Slot,J.C. and Rokas,A. (2013) Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11481–11486.

91. Melo,N.R., Moran,G.P., Warrilow,A.G.S., Dudley,E., Smith,S.N., Sullivan,D.J., Lamb,D.C., Kelly,D.E., Coleman,D.C. and Kelly,S.L. (2008) CYP56 (Dit2p) in Candida albicans: characterization and investigation of its role in growth and antifungal drug susceptibility. *Antimicrob. Agents Chemother.*, **52**, 3718–3724.

92. Basiony,M., Yang,Y., Liu,G., Gao,X.D. and Nakanishi,H. (2020) Studies on the properties of the sporulation specific protein dit1 and its product formyl tyrosine. *J. Fungi*, **6**, 77.

93. Briza,P., Kalchhauser,H., Pittenauer,E., Allmaier,G. and Breitenbach,M. (1996) N,N'-bisformyl dityrosine is an in vivo precursor of the yeast ascospore wall. *Eur. J. Biochem.*, **239**, 124–131.

94. Huang,A.C., Kautsar,S.A., Hong,Y.J., Medema,M.H., Bond,A.D., Tantillo,D.J. and Osbourn,A. (2017) Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae through genome mining reveals convergent evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E6005–E6014.

95. Pichersky,E. and Lewinsohn,E. (2011) Convergent evolution in plant specialized metabolism. *Annu. Rev. Plant Biol.*, **62**, 549–566.

96. Stephan,W. (2019) Selective sweeps. *Genetics*, **211**, 5–13.

97. Lin,Z., Kakule,T.B., Reilly,C.A., Beyhan,S. and Schmidt,E.W. (2019) Secondary metabolites of onygenales fungi exemplified by *Aioliomyces pyridodomos*. *J. Nat. Prod.*, **82**, 1616–1626.

98. Van Dyke,M.C.C., Teixeira,M.M. and Barker,B.M. (2019) Fantastic yeasts and where to find them: the hidden diversity of dimorphic fungal pathogens. *Curr. Opin. Microbiol.*, **52**, 55–63.

99. Beauvais,A. and Latgé,J.-P. (2018) Special issue: fungal cell wall. *J. Fungi.*, **4**, 91.

100. Khasheii,B., Mahmoodi,P. and Mohammadzadeh,A. (2021) Siderophores: importance in bacterial pathogenesis and applications in medicine and industry. *Microbiol. Res.*, **250**, 126790.

101. Center for High Throughput Computing (2006) Center for High Throughput Computing. Center for High Throughput Computing, https://doi.org/10.21231/GNT1-HW21.