An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes

Matthew T. Robey[a], **Lindsay K. Caesar**[b], **Milton T. Drott**[c], **Nancy P. Keller**[c,d], **and Neil L. Kelleher**[a,b,e,1]

[a]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208; [b]Department of Chemistry, Northwestern University, Evanston IL 60208; [c]Department of Medical Microbiology and Immunology, University of Wisconsin–Madison, Madison, WI 53706; [d]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI 53706; and [e]Proteomics Center of Excellence, Northwestern University, Evanston, IL 60208

Edited by Jens Nielsen, BioInnovation Institute, Copenhagen, Denmark, and approved April 5, 2021 (received for review September 26, 2020)

Fungi are prolific producers of natural products, compounds which have had a large societal impact as pharmaceuticals, mycotoxins, and agrochemicals. Despite the availability of over 1,000 fungal genomes and several decades of compound discovery efforts from fungi, the biosynthetic gene clusters (BGCs) encoded by these genomes and the associated chemical space have yet to be analyzed systematically. Here, we provide detailed annotation and analyses of fungal biosynthetic and chemical space to enable genome mining and discovery of fungal natural products. Using 1,037 genomes from species across the fungal kingdom (e.g., Ascomycota, Basidiomycota, and non-Dikarya taxa), 36,399 predicted BGCs were organized into a network of 12,067 gene cluster families (GCFs). Anchoring these GCFs with reference BGCs enabled automated annotation of 2,026 BGCs with predicted metabolite scaffolds. We performed parallel analyses of the chemical repertoire of fungi, organizing 15,213 fungal compounds into 2,945 molecular families (MFs). The taxonomic landscape of fungal GCFs is largely species specific, though select families such as the equisetin GCF are present across vast phylogenetic distances with parallel diversifications in the GCF and MF. We compare these fungal datasets with a set of 5,453 bacterial genomes and their BGCs and 9,382 bacterial compounds, revealing dramatic differences between bacterial and fungal biosynthetic logic and chemical space. These genomics and cheminformatics analyses reveal the large extent to which fungal and bacterial sources represent distinct compound reservoirs. With a >10-fold increase in the number of interpreted strains and annotated BGCs, this work better regularizes the biosynthetic potential of fungi for rational compound discovery.

natural products | fungi | biosynthesis | genome mining | secondary metabolism

Fungi have been an invaluable source of bioactive compounds with a wide variety of societal impacts. Mycotoxins such as aflatoxin, ochratoxin, and patulin, pharmaceuticals including penicillin, cyclosporine, and lovastatin, and agrochemicals like paraherquamide and strobilurin are all derived from fungi (1, 2). Recent genome sequencing efforts have revealed that <3% of the biosynthetic space represented by fungal genomes has been linked to metabolite products (3). In both bacteria and fungi, secondary metabolic pathways are typically encoded by biosynthetic gene clusters (BGCs). BGCs encode for backbone enzymes responsible for creating the core metabolite and tailoring enzymes that modify this scaffold along with regulatory transcription factors and transporters that transport metabolites and necessary precursors (4). In fungi, the most common backbone enzymes include nonribosomal peptide synthetases (NRPSs), polyketide synthases (PKSs), dimethylallyltransferases (DMATs), and terpene synthases.

Over the last decade, genome mining has emerged as an approach that utilizes genome sequencing and bioinformatics for targeted compound discovery based on genes of interest or biosynthetic novelty. Natural product discovery is poised to expand from using a single or few genomes to using many genomes interpreted together using increasingly sophisticated tools (5–9). The interpretation step can infuse knowledge of BGC phylogenetic distribution, inferences about the molecules encoded (e.g., prevalence and structural variance), and avoidance of known compounds (dereplication). To date, the application of such large-scale genome mining approaches to fungi has been largely limited to individual biosynthetic enzymes (10) or datasets of <100 genomes from well-studied taxonomic groups (11–15).

The concept of a gene cluster family (GCF) has emerged as an approach for large-scale analysis of BGCs (5–8). The GCF approach involves comparing BGCs using a series of pairwise distance metrics, then creating families of BGCs by setting an appropriate similarity threshold. This results in a network structure that dramatically reduces the complexity of BGC datasets and enables automated annotation based on experimentally characterized reference BGCs. Depending on the similarity threshold, BGCs within a family are expected to encode identical or similar metabolites and therefore serve as an indicator of new chemical scaffolds. The use of GCFs represents a logical shift from a focus on single genomes of interest to large genomics datasets, providing a means of regularizing collections of BGCs and their encoded chemical space (Fig. 1*A*). The use of GCF networks has been utilized for global analyses of bacterial biosynthetic space (6), bacterial genome mining at the >10,000 genome scale (9, 16), and integrated with metabolomics datasets for large-scale compound and BGC discovery (5, 7). Together with advances in large-scale metabolomics data analysis such as molecular networking (17), the GCF paradigm has helped in the modernization of natural products discovery.

Application of GCFs to fungal genomes has been largely limited to datasets of <100 genomes from well-studied genera such as *Aspergillus*, *Fusarium*, and *Penicillium* (13–15). Despite the availability of thousands of genomes representing a broad sampling of the fungal kingdom, global analyses of the BGC content of these genomes are lacking. As such, our knowledge of

Significance

Fungi represent an underexploited resource for new compounds with applications in the pharmaceutical and agriscience industries. Despite the availability of >1,000 fungal genomes, our knowledge of the biosynthetic space encoded by these genomes is limited and ad hoc. We present results from systematically organizing the biosynthetic content of 1,037 fungal genomes, providing a resource for data-driven genome mining and large-scale comparison of the genetic and molecular repertoires produced in fungi, and compare to those present in bacteria.

Author contributions: M.T.R., L.K.C., and N.L.K. designed research; M.T.R. and L.K.C. performed research; M.T.R. and L.K.C. analyzed data; and M.T.R., L.K.C., M.T.D., N.P.K., and N.L.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: n-kelleher@northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2020230118/-/DCSupplemental.

Published May 3, 2021.

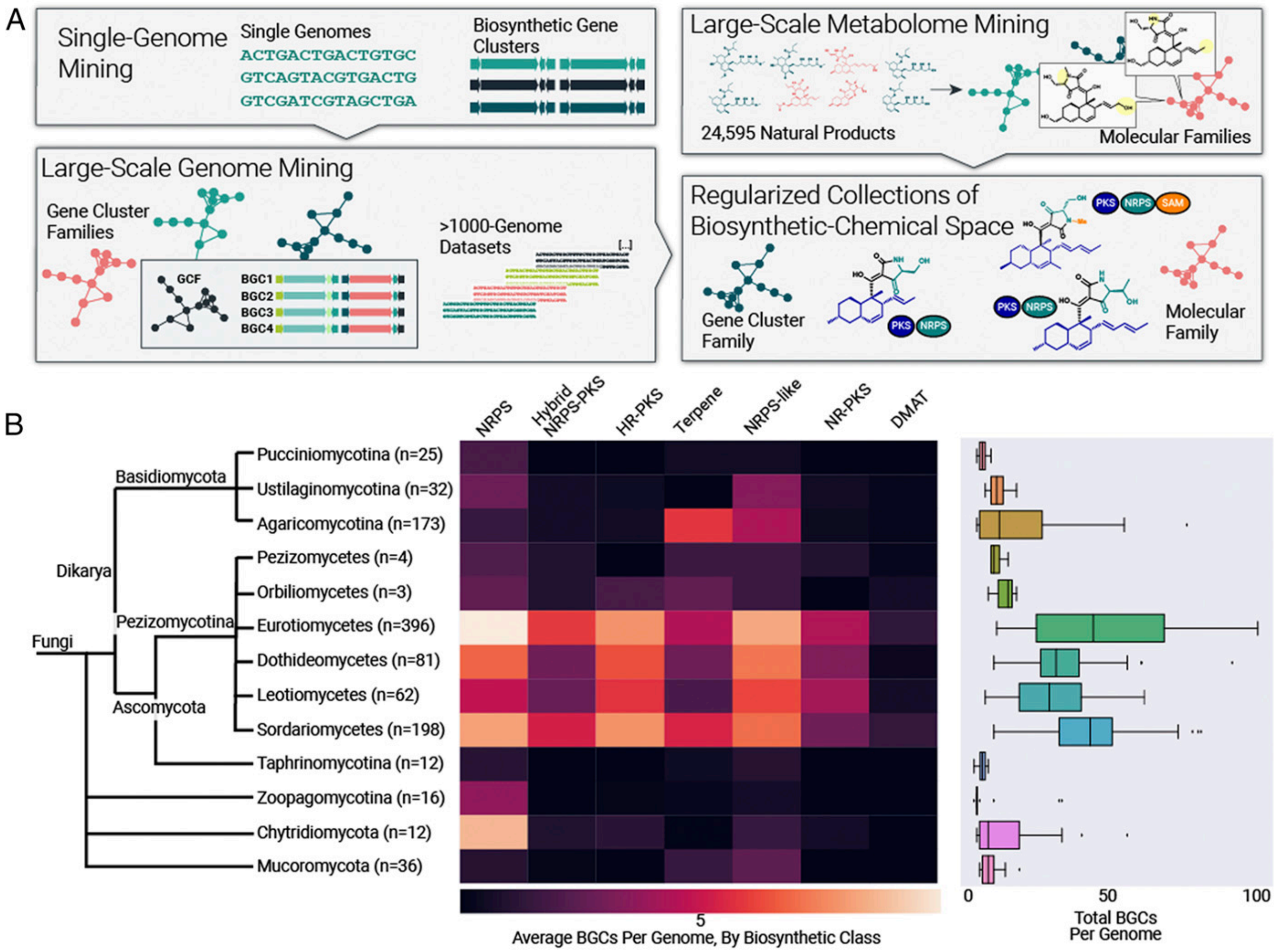

Fig. 1. Organizing BGCs from 1,037 fungal genomes. (*A*) Exploring fungal diversity using networks of GCFs and MFs. A GCF is a collection of similar BGCs aggregated into a network and predicted to use a similar chemical scaffold and create a family of related metabolites. An MF is a collection of metabolites that likewise represent chemical variations around a chemical scaffold. This networking approach enables hierarchical analysis of BGCs and their encoded metabolite scaffolds from large numbers of interpreted genomes. (*B*) Distribution of BGCs across the fungal kingdom. The BGC content of fungal genomes varies dramatically with phylogeny. Organisms within Pezizomycotina have more BGCs per genome and a greater diversity of biosynthetic types than organisms in Basidiomycota and non-Dikarya phyla.

the overall phylogenetic distribution of GCFs in fungi is limited, and many taxonomic groups have no experimentally characterized BGCs. Therefore, we performed a global analysis of BGCs and their families from a dataset of 1,037 genomes from across the fungal kingdom. Across fungi, the vast majority of GCFs are species specific, indicating that species-level sampling for genome sequencing and metabolomics will yield significant returns for natural products discovery.

To relate this now-available set of fungal GCF-encoded metabolites to known fungal scaffolds, we performed network analysis of 15,213 fungal compounds, organizing these into 2,945 molecular families (MFs) (Fig. 1*A*). Analysis of this joint genomic–chemical space revealed dramatic differences between both major fungal taxonomic groups, as well as between bacteria and fungi. This lays the groundwork for systematic discovery of new compounds and their BGCs from the fungal kingdom.

Results

A Reference Set of Fungal Biosynthetic Gene Clusters. Despite the availability of thousands of fungal genomes, the biosynthetic space represented within them has yet to be surveyed systematically. To address this gap, we curated a dataset of 1,037 fungal genomes, covering a broad phylogenetic swath (*SI Appendix*, Table S1 and Dataset S1). This selection includes well-studied taxonomic groups such as Eurotiomycetes (*Aspergillus* and *Penicillium* genera), Sordariomycetes (*Fusarium*, *Cordyceps*, and *Beauveria* genera), and taxa in which little is known about their BGCs, such as Basidiomycota or Mucoromycota. This genomic sampling likewise covers a large swath of ecological niches, from forest-dwelling mushrooms to plant endophytes and extremophiles (18).

Each of the 1,037 genomes was analyzed using antiSMASH (19), yielding an output of 36,399 BGCs ranging from 5 to 220 kb in length. As has been previously observed (20), the number of BGCs per genome varies dramatically across fungi (Fig. 1*B* and *SI Appendix*, Table S1). Eurotiomycetes average 48 BGCs per genome, with 25% of organisms within this class possessing >60 BGCs. Organisms outside of Pezizomycotina possess significantly fewer BGCs, with organisms from the non-Dikarya phyla averaging <15 BGCs per genome. The distribution of biosynthetic classes across the fungal kingdom also varies dramatically and unexpectedly. Organisms within the Pezizomycotina classes Eurotiomycetes, Dothideomycetes, Leotiomycetes, and Sordariomycetes average approximately five each of NRPS, hybrid NRPS–PKS, NRPS, HR–PKS, terpene, NRPS-like, and NR–PKS and two

DMAT BGCs per genome (Fig. 1*B*). Basidiomycota have far fewer BGCs that together encode a relatively limited chemical repertoire, with terpene BGCs being the most abundant in Agaricomycotina as previously suggested (10).

Organizing Gene Clusters into Families to Map Fungal Biosynthetic Potential. To further assess the ability of fungi to produce new chemical scaffolds, we grouped BGCs into families using the pairwise distance between BGCs and a clustering algorithm to yield GCFs. BGCs from antiSMASH were converted to arrays of protein domains then compared based on the fraction of shared domains and backbone protein domain sequence identity (7, 8). Density-based spatial clustering of applications with noise (DBSCAN) clustering was performed on the resulting distance matrix, resulting in a set of 12,067 GCFs (Fig. 2*A*) organized into a network (Fig. 3*A*). This total number allows hybrid BGCs to be present in more than one family, an approach used in previous GCF analyses (7, 21). Restricting hybrid BGCs to only a single GCF resulted in 3,556 nonredundant families, indicating a large amount of mixing of biosynthetic logic in fungi. Across the fungal kingdom, the distribution of GCFs shows a clear relationship with phylogeny (yellow streaks in Fig. 2*A* and *SI Appendix*, Figs. S1–S5). Evidence from studies of well-characterized strain sets of *Aspergillus* and *Penicillium* has suggested that GCFs are largely genus or species specific (13, 22, 23); however, here we show that several GCFs span entire subphyla or classes (Fig. 2*A*). The fraction of GCFs that two organisms share is likewise correlated with phylogenetic distance, evidenced by sets of shared GCFs between closely related taxonomic groups (*SI Appendix*, Fig. S6). In order to facilitate visualization of these phylogenetic patterns, we created Prospect, a web-based application for hierarchical browsing of fungal GCFs, BGCs, and proteins as well as MFs and compounds (prospect-fungi.com). Additional details of the site are available in *SI Appendix*, *Methods*.

We then sought to quantify the relationship between phylogeny and shared GCF content. To accomplish this, we used the protein sequence identity of 290 shared single-copy orthologous genes from the fungal BUSCO dataset (24) as a proxy for whole-genome distance, then we counted the fraction of GCFs shared within each genome in pairwise comparisons (Fig. 2*B*). A result was a clear relationship between genomic distance and shared GCF content, with an average of 75% shared GCFs at the species level but less than 5% shared GCFs at taxonomic ranks higher than family (Fig. 2*C*). A similar trend exists for individual phyla and taxonomic classes (*SI Appendix*, Fig. S7). Across the fungal kingdom, 76% of GCFs are species specific, and only 16% are genus specific (*SI Appendix*, Fig. S8), supporting the hypothesis that most BGCs enable fungi related at the species level to secure their respective ecological niches with highly specialized compounds (4).

GCF-Enabled Annotation of Fungal Biosynthetic Repertoire Anchored by Known BGCs. Identifying BGCs that have known metabolite products is an important component of genome mining, enabling researchers to prioritize known versus unknown biosynthetic pathways for discovery (25, 26). These “genomic dereplication” efforts have been bolstered by the development of the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository (27), which contains 213 fungal BGCs with known metabolites (at the time of these analyses, June 2019). When anchored with known BGCs, the GCF approach enables large-scale annotation of unstudied BGCs based on similarity to reference BGCs, identifying clusters likely to produce known metabolites or derivatives of knowns.

Within our dataset, 154 GCFs contained known BGCs from MIBiG, ∼1% of the 12,067 total GCFs reported here (*SI Appendix*, Fig. S9). These families collectively include a total of 2,026 BGCs (*SI Appendix*, Fig. S9) whose approximate metabolite products can now be inferred, a ∼10-fold increase in the number of annotated BGCs over the experimentally characterized clusters available in MIBiG (27). To make this expanded set of annotated BGCs and their families available for routine genome mining, we created a section within the Prospect website that highlights these newly annotated BGCs.

Large-Scale Comparison of GCFs and Fungal Compounds. To assess the relationship between GCFs and their chemical repertoire, we next compared GCF-encoded scaffolds to a dataset of known fungal scaffolds. Analogous to our GCF analysis, we utilized network analysis of fungal metabolites, organizing these compounds into molecular families (MFs) based on Tanimoto similarity, a commonly used metric for determining chemical relatedness (28, 29). To directly relate GCF- and MF-encoded metabolite scaffolds, we determined the relationship between chemical similarity and BGC similarity for a set of 154 fungal GCFs with known metabolite products (*SI Appendix*, Fig. S10). We chose a MF similarity threshold that resulted in similar levels of chemical similarity represented by GCF and MF metabolite scaffolds.

Using this compound network analysis strategy, we organized a dataset of 15,213 fungal metabolites from the Natural Products Atlas (30) into 2,945 MFs (Fig. 3*A*). We annotated each compound within this network with chemical ontology information using ClassyFire, a tool for classifying compounds into a hierarchy of terms associated with structural groups, chemical moieties, and functional groups (refer to *SI Appendix*, Fig. S11 for a breakdown of this chemical ontology analysis for fungal metabolites) (31). The number of MF scaffolds (2,945) is only 25% the number of GCF-encoded scaffolds (12,067) in our 1,000-genome dataset. This suggests that even this small genomic sampling of the entire fungal kingdom, estimated to have >1 million species (32), possesses biosynthetic potential that significantly dwarfs known fungal chemical space—not only in terms of individual metabolites but also in terms of metabolite scaffolds. In this joint GCF–MF dataset, MFs and GCFs represent complementary approaches for representing the same metabolite scaffold, such as the tenellin/desmethylbassianin structural class, whose GCF and MF contains both BGCs and compounds, respectively (Fig. 3 *A*, *Middle*) (33, 34). Exploring such pairings of GCF and MFs is a proven strategy for large-scale assignment of BGCs to their metabolite products (35), an activity that will provide a basis for improved compound discovery and for identifying the biosynthetic mechanisms fungi use for diversifying their bioactive scaffolds. An example of the latter is described below.

Diversification of the Equisetin Scaffold Inferred from Gene Cluster Families. To further explore the link between metabolite scaffolds as represented by MFs and GCFs, we looked to the decalin–tetramic acids, a structural class well represented in our BGC and metabolite datasets. Containing a tetramic acid moiety commonly found in both bacteria and fungi (36), decalin–tetramic acids such as equisetin, altersetin, phomasetin, and trichosetin (*SI Appendix*, Fig. S12) (37–39) have a wide range of reported biological activities, including antibiotic, anti-cancer, phytotoxic, and HIV integrase inhibitory activity (40). We reasoned that further exploration of the decalin–tetramic acid structural class would yield insights into the biosynthetic mechanisms for variation of this bioactive scaffold.

Using Prospect, we examined a GCF comprised of 81 decalin–tetramic acid BGCs which encode NRPS–PKS hybrids. This family contains known BGCs responsible for biosynthesis of equisetin (41), trichosetin (42), and phomasetin (43) as well as BGCs from *Alternaria spp.* that are likely responsible for the biosynthesis of altersetin (38, 44). While most fungal GCFs are confined to single species or genera (Fig. 2), the equisetin GCF has an exceptionally broad phylogenetic distribution, with clusters found in the four Pezizomycotina classes: Eurotiomycetes,

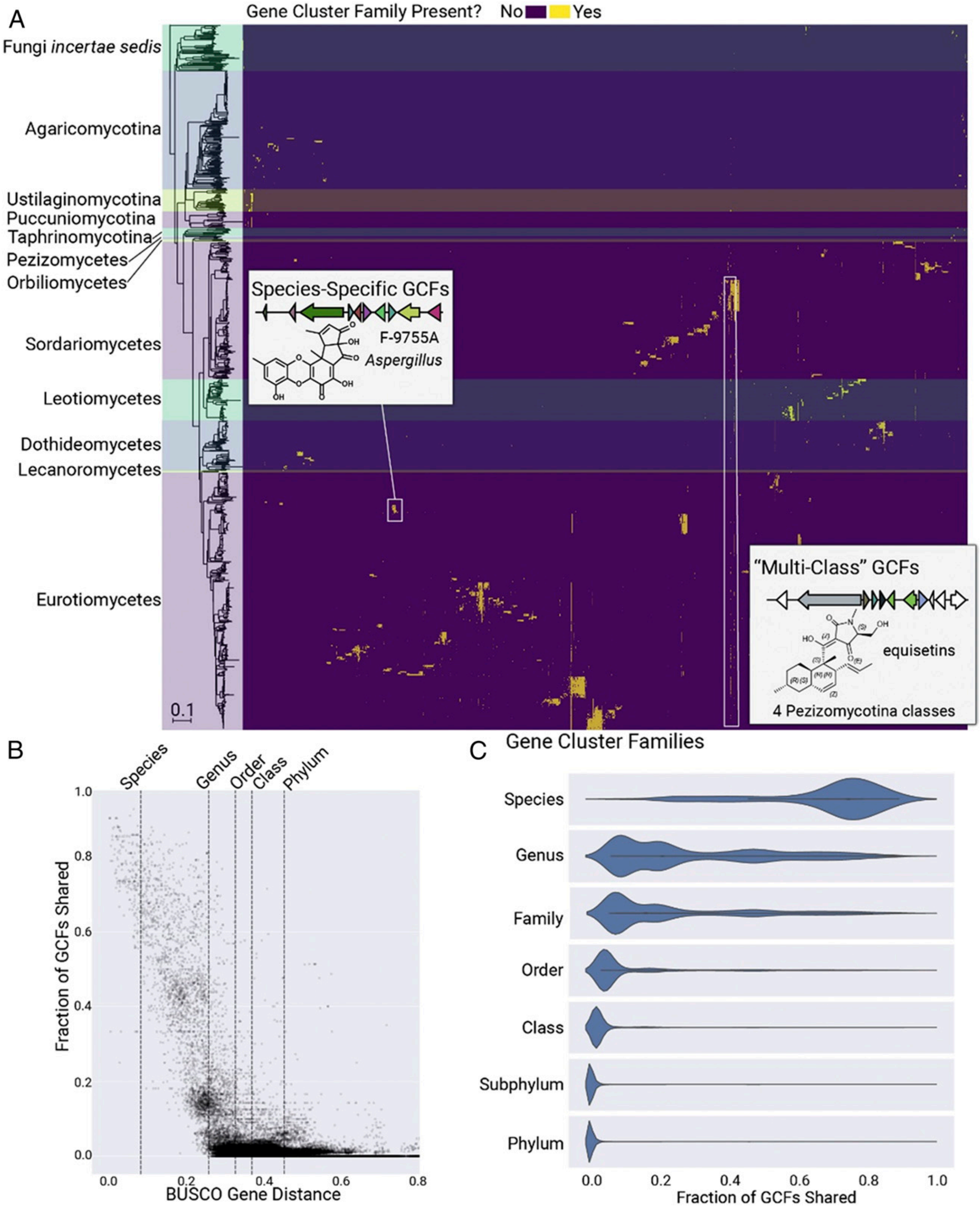

Fig. 2. The distribution of 12,067 GCFs across the fungal kingdom. (*A*) Heatmap of GCFs across Fungi. The phylogram to the left shows a Neighbor Joining species tree based on 290 shared orthologous genes across 1,037 genomes; horizontal shaded regions across the heatmap correspond to each labeled taxonomic group. The order of GCF columns is the result of hierarchical clustering based on the GCF presence/absence matrix. Across Fungi, the distribution of GCFs largely follows phylogenetic trends, with most GCFs confined to a specific genus or species. (*B*) Relationship between genetic distance and GCF content. The dotted lines indicate median genetic distance values for organisms within the same species, genus, order, class, or phylum. Each point in the scatterplot represents a pair of genomes and the fraction of the pair's GCFs that are shared. (*C*) Relationship between taxonomic rank and shared GCF content across the fungal kingdom. Violin plots show the fraction of GCFs shared between all pairs of organisms within our 1,000-genome dataset, with each pair classified based on the lowest taxonomic rank shared between the two organisms.

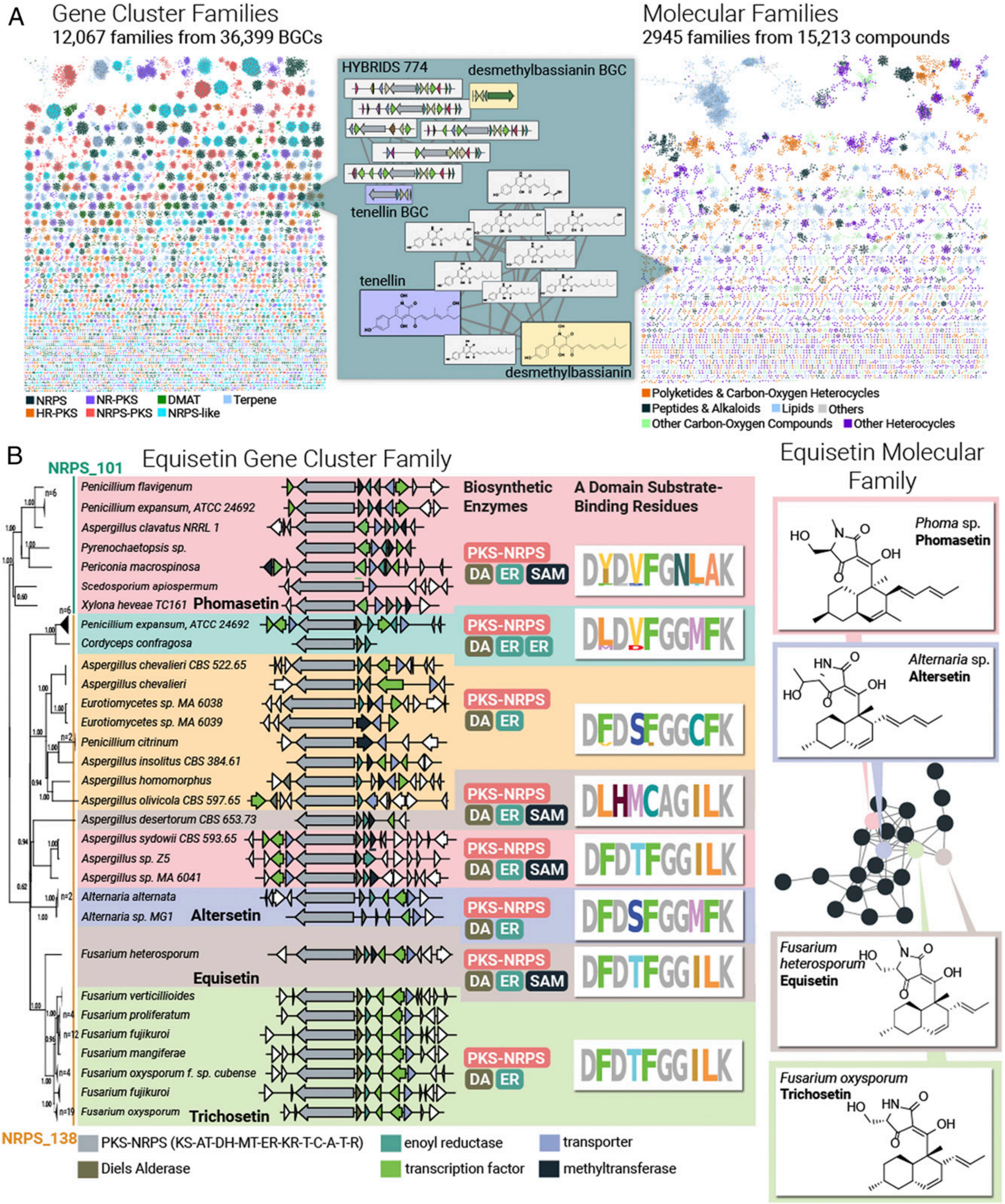

Fig. 3. Large-scale analysis of fungal genome-encoded and known metabolite scaffolds. (*A*) Colliding large-scale collections of fungal genetic content (*Left*) and fungal natural products (*Right*) using a network of GCFs interpreted from 1,037 genomes (*Left*) and 15,213 metabolites arranged into 2,945 molecular families based on their Tanimoto similarity score (*Right*). Note that 92% of these 12,067 GCFs remain unassigned to their metabolite products. (*B*) Variations in adenylation domain substrate-binding residues and tailoring enzyme composition facilitate modifications to the equisetin GCF (*Left*) and MF (*Right*). The phylogram to the left represents a maximum likelihood tree based on the hybrid NRPS–PKS backbone enzyme. All branches in this tree have >50% bootstrap support.

Dothideomycetes, Xylonomycetes, and Sordariomycetes (Fig. 3 *B*, *Left*). The associated equisetin MF is likewise found in a variety of Dothideomycete and Sordariomycete species (Fig. 3 *B*, *Right*). Filtering for biosynthetic class (NRPS versus PKS), this set of 81 BGCs grouped through automated analysis into 18 PKS families but just two NRPS families (NRPS_138 and NRPS_101), indicating a higher degree of similarity in their NRPS domains.

The equisetin biosynthetic pathway involves three major steps: assembly of a decalin core via the action of PKS enzyme domains and a Diels Alderase, formation of an amino acid–derived tetramic acid moiety catalyzed by NRPS domains, and N-methylation of the tetramic acid moiety (*SI Appendix*, Fig. S13) (43, 45). While the domain structure of the PKS contained in the equisetin GCF remains consistent across fungi, differences in backbone enzyme amino acid sequence and the presence/absence of tailoring enzymes mediate structural variations to the scaffold. The PKS enzymes from *Fusarium oxysporum* and *Pyrenochaetopsis* sp. RK10-F058 share 50% sequence identity, which likely result in the additional ketide unit and C-methylation observed in equisetin versus phomasetin (Fig. 3*B*). In the NRPS module of the hybrid NRPS–PKS, changes to adenylation domain substrate-binding residues are predicted to mediate incorporation of serine (trichosetin, equisetin, and phomasetin) and threonine (altersetin). The *Aspergillus desertorum* BGC contains adenylation domain substrate-binding residues that are highly variant from those found in other clusters within the GCF, indicating its tetramic acid moiety is likely diversified with a different amino acid. The equisetin GCF contains additional variations in the number of enoyl reductase enzymes (one additional in the uncharacterized *Penicillium expansum* clade), indicating possible differences in the degree of saturation, and a methyltransferase that is expected to mediate changes in tetramic acid N-methylation.

This pattern of biosynthetic variation within a GCF resulting in metabolite diversification suggests that exploring such pairs of GCFs and MFs with knowledge of their taxonomic distribution will be valuable to guide genome mining in the identification of new analogs of compounds with proven therapeutic or agrochemical value. The equisetin GCF is one of only 90 GCFs (representing 0.75% of total GCFs) within our dataset that spanned multiple taxonomic classes (*SI Appendix*, Table S2). This includes bioactive scaffolds such PR-toxin, swainsonine, chaetoglobosin, and cytochalasin (*SI Appendix*, Fig. S14) which contain variations in tailoring enzyme composition expected to diversify these scaffolds. Given the observed biosynthetic diversity within such "multi-class" GCFs, exploring such pairs of GCFs and MFs represents an attractive approach for discovering new analogs of bioactive metabolites.

A "Bird's Eye" View of Fungal versus Bacterial Biosynthetic Space. Having surveyed GCFs across the fungal kingdom, we sought to compare and contrast this genomic and chemical repertoire to the well-established bacterial canon. We gathered 5,453 bacterial genomes whose BGCs were publicly available in the antiSMASH bacterial BGCs database (46), resulting in a dataset of 24,024 bacterial BGCs to compare to our dataset of 36,399 fungal BGCs. To visualize the biosynthetic space encompassed by these BGCs, we determined the frequency of protein domains within BGCs for each major taxonomic group (*SI Appendix*, *Methods*). Principle component analysis (PCA) of these encoded BGCs showed a phylogenetic bias in this biosynthetic space, with bacteria and fungi occupying distinct regions (Fig. 4*A*).

We quantified the dramatic differences in bacterial versus fungal NRPS and PKS assembly line logic. Bacterial and fungal PKS enzymes are known to differ in aromatic polyketide assembly logic (47, 48), and the vast majority of characterized fungal PKS enzymes are iterative (49). This iterative PKS pattern observed in characterized fungal PKS enzymes holds true across this dataset, with fungal PKSs most often encoding a single-backbone PKS enzyme, compared to bacterial PKS BGCs which contain a median of 1.7 PKS backbone enzymes per cluster (Fig. 4 *B*, *Right*). Fungal NRPS BGCs also usually encode single-backbone proteins, compared to the multiple-backbone enzymes more typically observed in bacterial systems (Fig. 4 *B*, *Left*). This observation is consistent with those fungal NRPS proteins that have been characterized (3), the majority of which have single-backbone enzymes. Fungal NRPS and PKS enzymes also average ∼150% the size of bacterial backbones (*SI Appendix*, Fig. S15). In addition to these contrasting backbone enzyme compositions, we observed systematic differences in the most common NRPS domain organizations (*SI Appendix*, Fig. S16), particularly in NRPS termination domains (Fig. 4*C*). The most common fungal NRPS termination domains are C-terminal condensation domains, recently found to catalyze release of peptide intermediates via intramolecular cyclization (50–52). The next most common are terminal thioester reductase domains that perform either reductive release to aldehydes or alcohols or release via cyclization (53). This is in stark contrast to bacterial NRPS BGCs, which most commonly terminate with type I thioesterase domains that release intermediates as linear or cyclic peptides (Fig. 4*C*).

These collective differences between fungal and bacterial BGCs show systematic differences in NRPS biosynthetic logic between these two kingdoms. In bacterial NRPS canon, a pathway is comprised of multiple NRPS genes whose chromosomal order (and the order of catalytic domain "modules" within the encoded polypeptide) corresponds to the order of amino acid monomers in the metabolite product (Fig. 4 *D*, *Right*) (54). In the field of bacterial natural products, the use of this "collinearity rule" to predict metabolite scaffolds is commonplace (19, 55, 56); however, the large number of exceptions to this rule reduces the accuracy of these predictions. The prototypical fungal NRPS (Fig. 4*D*, *Left*) primarily involves the action of biosynthetic domains within the same backbone enzyme rather than multiple NRPS backbones acting in concert. This suggests that future efforts to predict fungal NRPS scaffolds will be able to largely bypass the need to account for permutations of multiple NRPS genes, raising the possibility of increased predictive performance compared to bacteria.

Uncovering Distinct Natural Product Reservoirs. Having shown that fungi and bacteria are distinct biosynthetically, we sought to compare these genomics-based insights to the chemical space of known metabolites. We added 9,382 bacterial compounds to our dataset of 15,213 fungal metabolites, analyzing these bacterial compounds using the same network analysis and chemical ontology workflow described above. We performed PCA to visualize the chemical space of major fungal and bacterial taxonomic groups within this compound dataset (further described in *SI Appendix*, *Methods*).

PCA of bacterial and fungal compounds (Fig. 5*A*) revealed a trend that parallels our analysis of fungal and bacterial biosynthetic space (Fig. 4*A*). Bacteria and fungi occupy separate regions of chemical space, differing dramatically in terms of chemical ontology superclass, a high-level descriptor of general structural type (Fig. 5*B*). Fungi have twice the frequency of lipids and nearly twice the frequency of heterocyclic compounds, a structural group that includes aromatic polyketide-related moieties such as furans and pyrans. Many of the chemical moieties and structural classes that are highly enriched in bacteria or fungi are vital in bioactive scaffolds. This includes moieties such as the bacterial aminoglycoside antibiotics (57), thiazoles present in the bacterial anti-cancer bleomycin family (58), and the steroid ring that forms the core scaffold of steroid drugs such as the fungal metabolite fusidic acid (59) (Fig. 5*B*). PCA loading plots similarly reveal differences between bacterial and fungal chemical space, including a high prevalence of peptide-associated chemical ontology terms in bacteria and lipid and aromatic polyketide terms in

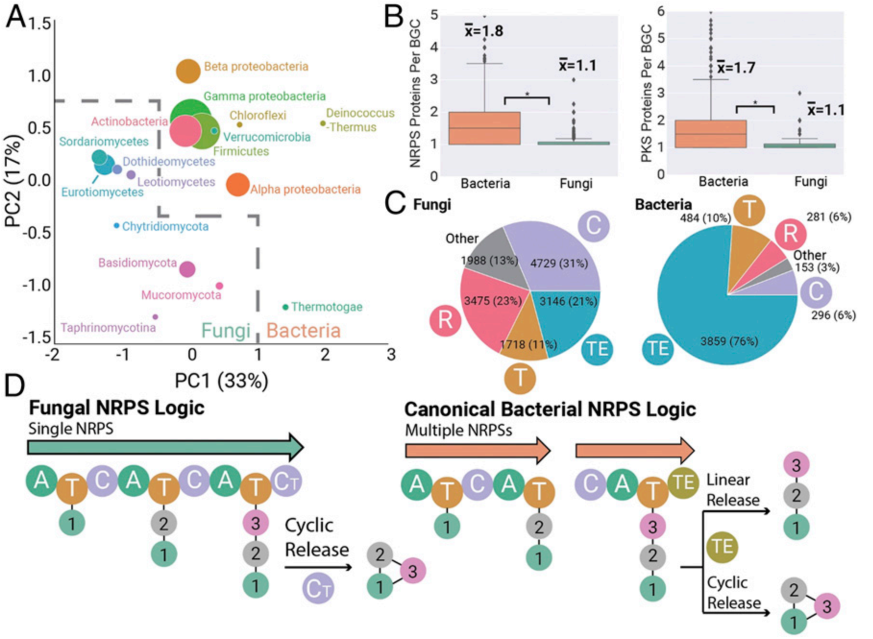

Fig. 4. Fungal BGCs are distinct from their canonical bacterial counterparts. (*A*) PCA of 36,399 fungal and 24,024 bacteria BGCs, with points sized according to the number of BGCs analyzed. Fungal and bacterial taxonomic groups occupy distinct regions of this biosynthetic space. (*B*) Fungal and bacterial BGCs differ in backbone enzyme composition, with fungal NRPS and PKS clusters typically encoding only a single backbone, compared to multiple-backbone enzymes found in bacterial BGCs. (*C*) Fungal and bacterial NRPS BGCs differ dramatically in their use of termination domains for release of peptide intermediates. (*D*) Fungal NRPS logic is distinct from bacterial canon. Most fungal NRPS pathways involve a single NRPS enzyme that utilizes a terminal condensation domain to produce a cyclic peptide. In contrast, bacterial NRPS enzymes contain multiple NRPS enzymes that operate in a colinear fashion and typically utilize thioesterase domains to produce linear or cyclic peptides.

fungi (*SI Appendix*, Fig. S17). While there are known cases of shared compounds between bacteria and fungi (several highlighted in *SI Appendix*, Fig. S18), similar compounds (i.e., Tanimoto similarity > 0.6) between bacteria and fungi were completely absent from our dataset, reinforcing the rarity of such shared chemical space.

Within the fungal kingdom, PCA revealed differences in the chemical repertoire of major taxonomic groups (*SI Appendix*, Fig. S19). Pezizomycotina classes grouped together in chemical space, largely due to a higher proportion of polyketide- and peptide-related chemical moieties (*SI Appendix*, Fig. S20). Basidiomycota are distinct chemically, possessing a much higher proportion of chemical moieties and descriptors associated with terpenes and other lipids. These observations based on chemical space are consistent with the higher proportion of NRPS and PKS BGCs within Pezizomycotina and the prevalence of terpene BGCs within Basidiomycota groups such as Agaricomycotina (Fig. 2*B*) and further supported by PCA of fungal BGCs, in which fungal phyla represent distinct groups (*SI Appendix*, Figs. S21 and S22).

Discussion

A Framework for Exploring Fungal Scaffolds Using Gene Cluster Families. The GCF approach enables the systematic mapping of the biosynthetic repertoire encoded by large groups of fungal genomes. The fungal kingdom is a wealth of untapped biosynthetic potential, with the 1,000 genomes analyzed here representing a reservoir of >12,000 GCF-encoded scaffolds. This genome dataset is only a small subset of the >1 million predicted fungal species (32), indicating that the total biosynthetic potential of the fungal kingdom far surpasses that assembled here.

By organizing biosynthetically related BGCs into families, the GCF approach provides a means of cataloguing and dereplicating genome-encoded MFs. In the field of bacterial natural products discovery, this GCF paradigm has been expanded for automated linking of GCFs to MFs detected by metabolomics and molecular networking analysis, enabling high-throughput genome mining from industrial-scale strain collections (5, 7, 32, 35). Establishing the GCF approach for fungal genomes lays the groundwork for similar GCF-driven large-scale compound discovery efforts from fungi.

Data-Driven Prospecting for Fungal Natural Products. Large-scale genome sequencing projects such as the 1000 Fungal Genomes project, whose stated goal is sampling every taxonomic family within fungi (60), will uncover a large amount of biosynthetic and chemical novelty. However, as 76% of fungal GCFs are species- and 16% are genus specific, such genome sequencing efforts focused on taxonomic families will miss the majority of GCFs. Additional large-scale efforts to sample this biosynthetic space based on "depth" rather than "breadth" is suggested to more efficiently access these genomes; indeed, a recent report on metabolic diversity of just two clinical isolates of the model fungus *Aspergillus nidulans* revealed six novel clusters (61). Future "1,000-genome" projects, now feasible for academic research groups due to ever-decreasing genome sequencing costs, should focus on expanding this dataset with species-level sequencing of taxonomic groups.

The GCF approach provides a means of selecting fungi for compound and BGC discovery via approaches such as heterologous expression (21) based not on taxonomic or phylogenetic markers but with a strategy that focuses on efficient sampling of biosynthetic pathways. The distribution of GCFs shows groups of organisms with shared GCFs (*SI Appendix*, Fig. S6), and sampling based on these organism "groups" reduces the number of genomes required to capture the majority of fungal biosynthetic space. Our simulated sampling based on shared GCFs indicated that 80% of GCFs from the 386 Eurotiomycete genomes are represented in a sample of only 145 genomes. By contrast, to represent the same number of GCFs, species-level sampling required 189 genomes and random sampling required 263 genomes

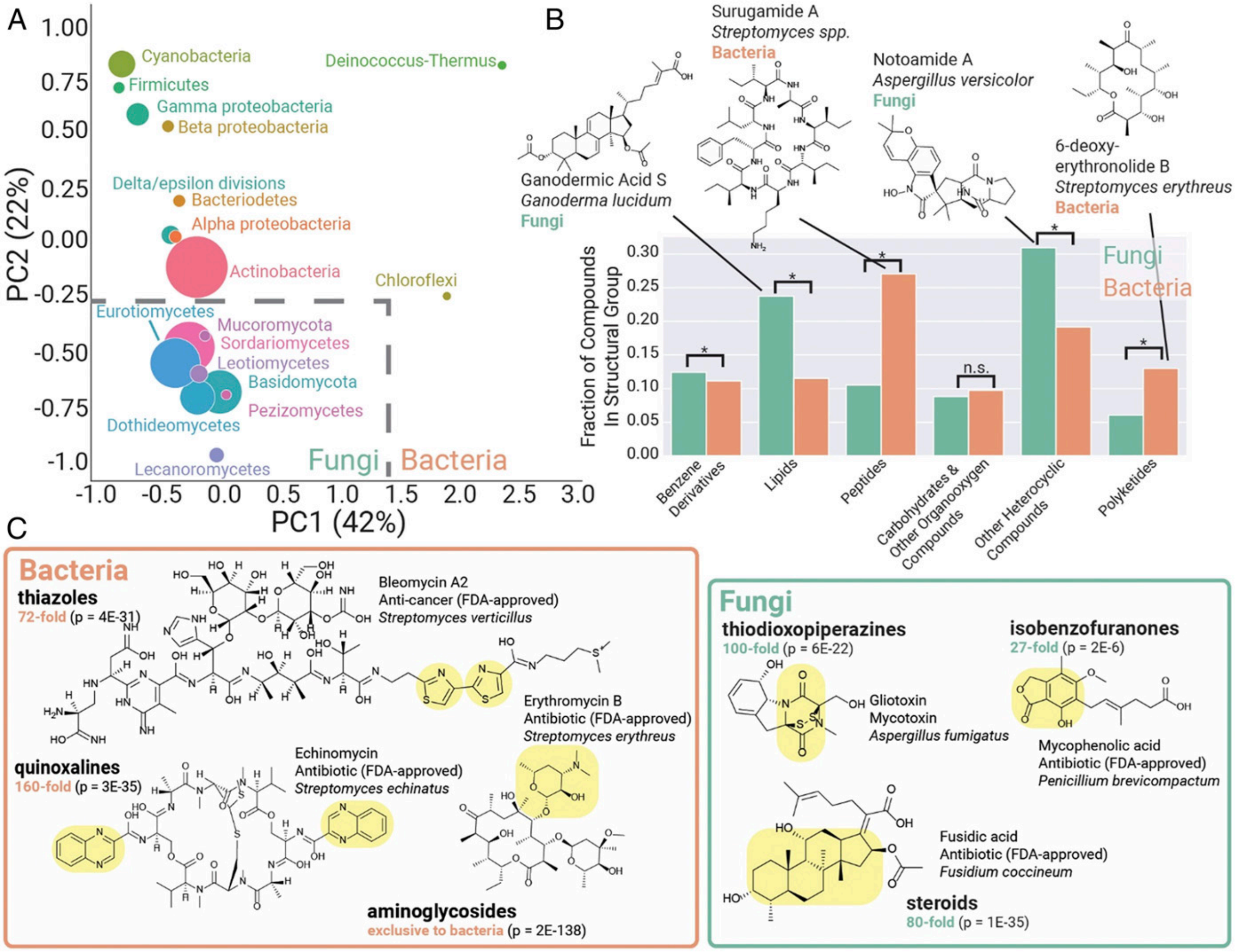

Fig. 5. Bacteria and fungi are distinct sources for natural product scaffolds. (*A*) PCA of 24,595 known bacterial and fungal compounds, with points sized according to the number of compounds. Fungal and bacterial taxonomic groups occupy distinct regions in this representation of chemical space for natural products. (*B*) Quantitative comparison of structural classifications in bacterial versus fungal compounds. (*C*) Bacteria and fungi represent distinct pools for bioactive compounds and scaffolds. Selected chemical moieties enriched and characteristic of each taxonomic group are highlighted in yellow. The fold enrichment of the chemical moiety is indicated in green, with *P* values from a chi-squared test indicated.

(*SI Appendix*, Fig. S23). This indicates that the GCF approach can be used as a roadmap for systematic characterization of new fungal biosynthetic pathways and their compounds.

Unearthing New Medicines. These analyses of both chemical and biosynthetic space show that bacteria and fungi represent chemically distinct sources for natural products discovery. Interestingly, fungal compounds are closer to US Food and Drug Administration–approved compounds than bacterial compounds in terms of several chemical properties, including three out of four "Lipinsky Rule of Five" properties often used as guidelines for predicting oral bioavailability (*SI Appendix*, Fig. S24) (62). While many of the most successful natural products violate these rules of thumb, these data suggest that fungal metabolites may be more "druglike" than those occupying bacterial chemical space.

Major compound discovery campaigns can be initiated with the understanding that different biological sources will enrich for different types of metabolite scaffolds. The fungal kingdom is rich in aromatic polyketides, while bacteria harbor a higher proportion of peptidic scaffolds. Within the fungal kingdom, Basidiomycota is a rich reservoir of terpenes, while BGC-rich Pezizomycotina classes bias toward polyketides and peptides. In sum, an atlas like that assembled here allows quantification of previously anecdotal trends about the phylogenetic distribution of metabolite subtypes and taxonomy-informed mining of specific scaffolds with therapeutic potential (*SI Appendix*, Fig. S25).

Conclusion

We have mapped the landscape of 12,067 GCFs across 1,037 fungal genomes, revealing the phylogenetic distribution of these families and establishing a framework for high-throughput genome mining from fungi. This framework introduces a fundamental biosynthetic unit—the gene cluster family—for cataloguing and annotating the rapidly increasing number of fungal genomes available. Network analysis at the GCF level advances the field of fungal genome mining with an approach scalable to industrial-scale strain collections, providing an approach for systematically mapping known and unknown fungal biosynthetic space and associated metabolite scaffolds. The GCF paradigm further provides an atlas for exploring metabolite scaffolds and their derivatives, enabling targeted genome mining focused on scaffolds with proven value. These collective analyses reveal that genomes across the fungal kingdom represent a rich resource for discovery of natural products. In both under-explored and well-studied fungal taxa, a wide variety of metabolite scaffolds awaits discovery, and the ever-decreasing cost of genome sequencing will help usher in a wave of large-scale fungal genome mining efforts that rival those currently underway in bacteria.

Materials and Methods

Methods and additional necessary information are available in *SI Appendix, Methods*. This includes descriptions of genome dataset curation, gene cluster family network analyses, web portal creation, phylogenetic trees, cheminformatics analyses, and principal component analysis.

Data Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by the National Center for Complementary and Integrative Health (R01 AT009143). L.K.C. was supported by the National Institute of General Medical Sciences (F32 GM134679). M.T.D. was supported by US Department of Agriculture, National Institute of Food and Agriculture postdoctoral fellowship award 2019-67012-29662.

1. L. B. Bullerman, Significance of mycotoxins to food safety and human health. *J. Food Prot.* **42**, 65–86 (1979).
2. G. F. Bills, J. B. Gloer, Biologically active secondary metabolites from the fungi. *Microbiol. Spectr.* **4**, 1087–1119 (2016).
3. Y. F. Li *et al*., Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genet. Biol.* **89**, 18–28 (2016).
4. N. P. Keller, Fungal secondary metabolism: regulation, function and drug discovery. *Nat. Rev. Microbiol.* **17**, 167–180 (2019).
5. D. D. Nguyen *et al*., MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2611–E2620 (2013).
6. P. Cimermancic *et al*., Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
7. J. R. Doroghazi *et al*., A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
8. J. C. Navarro-Muñoz *et al*., A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
9. S. A. Kautsar, J. J. Van Der Hooft, D. De Ridder, M. H. Medema, BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *bioRxiv* [Preprint] (2020) 10.1093/gigascience/giaa154 (Accessed 19 August 2020).
10. X.-L. Li *et al*., Rapid discovery and functional characterization of diterpene synthases from basidiomycete fungi by genome mining. *Fungal Genet. Biol.* **128**, 36–42 (2019).
11. S. Gao *et al*., Genome-wide analysis of Fusarium verticillioides reveals inter-kingdom contribution of horizontal gene transfer to the expansion of metabolism. *Fungal Genet. Biol.* **128**, 60–73 (2019).
12. I. Kjærbølling, U. H. Mortensen, T. Vesth, M. R. Andersen, Strategies to establish the link between biosynthetic gene clusters and secondary metabolites. *Fungal Genet. Biol.* **130**, 107–121 (2019).
13. J. C. Nielsen *et al*., Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. *Nat. Microbiol.* **2**, 17044 (2017).
14. K. Hoogendoorn *et al*., Evolution and diversity of biosynthetic gene clusters in Fusarium. *Front. Microbiol.* **9**, 1158 (2018).
15. S. Theobald *et al*., Uncovering secondary metabolite evolution and biosynthesis using gene cluster networks and genetic dereplication. *Sci. Rep.* **8**, 17957 (2018).
16. K.-S. Ju *et al*., Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12175–12180 (2015).
17. J. Y. Yang *et al*., Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).
18. S. A. Cantrell, J. C. Dianese, J. Fell, N. Gunde-Cimerman, P. Zalar, Unusual fungal niches. *Mycologia* **103**, 1161–1174 (2011).
19. K. Blin *et al*., antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
20. N. Khaldi *et al*., SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
21. K. D. Clevenger *et al*., A scalable platform to identify fungal secondary metabolites and their gene clusters. *Nat. Chem. Biol.* **13**, 895–901 (2017).
22. I. Kjærbølling *et al*., Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E753–E761 (2018).
23. T. C. Vesth *et al*., Investigation of inter- and intraspecies variation through genome sequencing of Aspergillus section Nigri. *Nat. Genet.* **50**, 1688–1695 (2018).
24. 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).
25. C. L. M. Gilchrist, H. Li, Y.-H. Chooi, Panning for gold in mould: Can we increase the odds for fungal genome mining? *Org. Biomol. Chem.* **16**, 1620–1626 (2018).
26. N. Ziemert, M. Alanjary, T. Weber, The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
27. M. H. Medema *et al*., Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
28. D. Butina, Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
29. C. R. Pye, M. J. Bertin, R. S. Lokey, W. H. Gerwick, R. G. Linington, Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5601–5606 (2017).
30. J. A. van Santen *et al*., The natural products atlas: An open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
31. Y. Djoumbou Feunang *et al*., ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
32. M. Blackwell, The fungi: 1, 2, 3 ... 5.1 million species? *Am. J. Bot.* **98**, 426–438 (2011).
33. L. M. Halo *et al*., Authentic heterologous expression of the tenellin iterative polyketide synthase nonribosomal peptide synthetase requires coexpression with an enoyl reductase. *ChemBioChem* **9**, 585–594 (2008).
34. K. M. Fisch *et al*., Rational domain swaps decipher programming in fungal highly reducing polyketide synthases and resurrect an extinct metabolite. *J. Am. Chem. Soc.* **133**, 16635–16641 (2011).
35. A. W. Goering *et al*., Metabologenomics: Correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent. Sci.* **2**, 99–108 (2016).
36. M. Jiang, S. Chen, J. Li, L. Liu, The biological and chemical diversity of tetramic acid compounds from marine-derived microorganisms. *Mar. Drugs* **18**, 114 (2020).
37. R. F. Vesonder, L. W. Tjarks, W. K. Rohwedder, H. R. Burmeister, J. A. Laugal, Equisetin, an antibiotic from Fusarium equiseti NRRL 5537, identified as a derivative of N-methyl-2,4-pyrollidone. *J. Antibiot. (Tokyo)* **32**, 759–761 (1979).
38. V. Hellwig *et al*., Altersetin, a new antibiotic from cultures of endophytic Alternaria spp. Taxonomy, fermentation, isolation, structure elucidation and biological activities. *J. Antibiot. (Tokyo)* **55**, 881–892 (2002).
39. E. C. Marfori, S. Kajiyama, E. Fukusaki, A. Kobayashi, Trichosetin, a novel tetramic acid antibiotic produced in dual culture of Trichoderma harzianum and Catharanthus roseus Callus. *Z. Natforsch. C J. Biosci.* **57**, 465–470 (2002).
40. R. Schobert, A. Schlenk, Tetramic and tetronic acids: An update on new derivatives and biological aspects. *Bioorg. Med. Chem.* **16**, 4203–4221 (2008).
41. J. W. Sims, J. P. Fillmore, D. D. Warner, E. W. Schmidt, Equisetin biosynthesis in Fusarium heterosporum. *Chem. Commun. (Camb.)* 186–188 (2005).
42. S. Janevska *et al*., Establishment of the inducible Tet-on system for the activation of the silent trichosetin gene cluster in Fusarium fujikuroi. *Toxins (Basel)* **9**, 126 (2017).
43. N. Kato *et al*., Control of the stereochemical course of [4+ 2] cycloaddition during trans-decalin formation by Fsa2-family enzymes. *Angew. Chem. Int. Ed. Engl.* **57**, 9754–9758 (2018).
44. J. J. Kellogg *et al*., Biochemometrics for natural products research: Comparison of data analysis approaches and application to identification of bioactive compounds. *J. Nat. Prod.* **79**, 376–386 (2016).
45. X. Li, Q. Zheng, J. Yin, W. Liu, S. Gao, Chemo-enzymatic synthesis of equisetin. *Chem. Commun. (Camb.)* **53**, 4695–4697 (2017).
46. K. Blin *et al*., The antiSMASH database version 2: A comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **47**, D625–D630 (2019).
47. R. Thomas, A biosynthetic classification of fungal and streptomycete fused-ring aromatic polyketides. *ChemBioChem* **2**, 612–627 (2001).
48. R. Thomas, Examination of potential exceptions to the F and S biosynthetic classification of fused-ring aromatic polyketides. *ChemBioChem* **17**, 2208–2215 (2016).
49. C. D. Campbell, J. C. Vederas, Biosynthesis of lovastatin and related metabolites formed by fungal iterative PKS enzymes. *Biopolymers* **93**, 755–763 (2010).
50. X. Gao *et al*., Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. *Nat. Chem. Biol.* **8**, 823–830 (2012).
51. J. A. Baccile *et al*., Diketopiperazine formation in fungi requires dedicated cyclization and thiolation domains. *Angew. Chem. Int. Ed. Engl.* **58**, 14589–14593 (2019).
52. L. K. Caesar *et al*., Heterologous expression of the unusual terreazepine biosynthetic gene cluster reveals a promising approach for identifying new chemical scaffolds. *MBio* **11**, e01691-20 (2020).
53. M. W. Mullowney, R. A. McClure, M. T. Robey, N. L. Kelleher, R. J. Thomson, Natural products from thioester reductase containing biosynthetic pathways. *Nat. Prod. Rep.* **35**, 847–878 (2018).
54. G. L. Challis, J. H. Naismith, Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Struct. Biol.* **14**, 748–756 (2004).
55. M. A. Skinnider, N. J. Merwin, C. W. Johnston, N. A. Magarvey, PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
56. M. A. Skinnider *et al*., Genomes to natural products prediction informatics for secondary metabolomes (PRISM). *Nucleic Acids Res.* **43**, 9645–9662 (2015).
57. K. M. Krause, A. W. Serio, T. R. Kane, L. E. Connolly, Aminoglycosides: An overview. *Cold Spring Harb. Perspect. Med.* **6**, a027029 (2016).
58. U. Galm *et al*., Antitumor antibiotics: Bleomycin, enediynes, and mitomycin. *Chem. Rev.* **105**, 739–758 (2005).
59. L. Verbist, The antimicrobial activity of fusidic acid. *J. Antimicrob. Chemother.* **25** (suppl. B), 1–5 (1990).
60. I. V. Grigoriev *et al*., MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
61. M. T. Drott *et al*., Diversity of secondary metabolism in Aspergillus nidulans clinical isolates. *MSphere* **5**, e00156-20 (2020).
62. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).