HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Nat Chem Biol. 2023 July ; 19(7): 846–854. doi:10.1038/s41589-023-01276-8.

Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs

Lindsay K. Caesar[1], **Fatma A. Butun**[1], **Matthew T. Robey**[2], **Navid J. Ayon**[1], **Raveena Gupta**[1], **David Dainko**[1], **Jin Woo Bok**[3], **Grant Nickles**[3], **Robert J. Stankey**[4], **Don Johnson**[4], **David Mead**[4], **Kristof B. Cank**[5], **Cody E. Earp**[5], **Huzefa A. Raja**[5], **Nicholas H. Oberlies**[5], **Nancy P. Keller**[3,6], **Neil L. Kelleher**[1,2,7,✉]

[1]Department of Chemistry, Northwestern University, Evanston, IL, USA.

[2]Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA.

[3]Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI, USA.

[4]Terra Bioforge, Madison, WI, USA.

[5]Department of Chemistry and Biochemistry, University of North Carolina at Greensboro, Greensboro, NC, USA.

[6]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA.

[7]Proteomics Center of Excellence, Northwestern University, Evanston, IL, USA.

Abstract

✉**Correspondence and requests for materials** should be addressed to Neil L. Kelleher. n-kelleher@northwestern.edu.

Author contributions

L.K.C. led the project, organized data collection and analyzed data for both large scale correlations and targeted biosynthetic studies. F.A.B. was responsible for fungal culture and DNA extraction and helped prepare and run MS samples. M.T.R. assembled genomes and conducted bioinformatic analysis for both GCF networking and metabologenomics correlations. N.J.A. grew fungi, extracted metabolomes and ran MS samples. R.G. and D.D. prepared and ran MS samples and assisted with metabolomics analysis. J.W.B. completed fungal transformations for heterologous expression and knockout studies. G.N. compiled correlations plots and assisted with GCF optimization. R.J.S., D.J. and D.M. designed, cloned and validated plasmids for heterologous expression. K.B.C., C.E.E. and N.H.O. assisted with metabolite dereplication and NMR analysis. H.A.R. provided expertise for fungal growth and extraction and taxonomic identification of fungal strains. N.P.K. and N.L.K. supervised the project after its initiation by N.L.K. The manuscript was written by L.K.C. and N.L.K., with all authors providing substantial edits and commentary throughout.

Competing interests

The authors declare financial conflicts of interest with MicroMGx (N.L.K.), Varigen Biosciences (D.M.) and Terra Bioforge (N.L.K., D.M., M.T.R. and N.P.K.). Further, N.L.K. is a consultant for Thermo Fisher Scientific focusing on the use of Fourier-transform Mass Spectrometry in multi-Omics research. Finally, N.H.O. and H.A.R. are on the Scientific Advisory Board of Clue Genetics, and N.H.O. is on the Scientific Advisory Board of Mycosynthetix. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41589-023-01276-8.

Peer review information *Nature Chemical Biology* thanks Hosein Mohimani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-023-01276-8.

Natural products research increasingly applies -omics technologies to guide molecular discovery. While the combined analysis of genomic and metabolomic datasets has proved valuable for identifying natural products and their biosynthetic gene clusters (BGCs) in bacteria, this integrated approach lacks application to fungi. Because fungi are hyper-diverse and underexplored for new chemistry and bioactivities, we created a linked genomics–metabolomics dataset for 110 Ascomycetes, and optimized both gene cluster family (GCF) networking parameters and correlation-based scoring for pairing fungal natural products with their BGCs. Using a network of 3,007 GCFs (organized from 7,020 BGCs), we examined 25 known natural products originating from 16 known BGCs and observed statistically significant associations between 21 of these compounds and their validated BGCs. Furthermore, the scalable platform identified the BGC for the pestalamides, demystifying its biogenesis, and revealed more than 200 high-scoring natural product–GCF linkages to direct future discovery.

In nature, fungi live in dynamic communities and adapt to changing environmental conditions and competition. Eons of evolutionary selection have turned fungi into expert chemists capable of biosynthesizing intricate natural products with specific bioactivities fine-tuned to suit their ecological needs[1]. Compounds from fungi have had major societal effects, providing the famous therapeutics penicillin and lovastatin[2], the deadly poisons aflatoxin and gliotoxin[3], and agrochemicals including azoxystrobin and derquantel[4]. This remarkable functional diversity derives from the structural complexity of fungal natural products. While there is ample evidence showing that the enzymatic machinery responsible for forming most of these compounds is encoded by biosynthetic gene clusters (BGCs), most fungal natural products have not yet been linked to their BGCs[5]. Understanding the pathways by which natural products are biosynthesized can enable their effective production and manipulation, facilitate fungal disease management and increase understanding of fungal ecology and evolution[6].

Natural products research has evolved by embracing genomics-guided strategies for charting unexplored biosynthetic space[7]. Originally focused on mining single genomes, scientists now analyze tens, hundreds and even thousands of genomes simultaneously[8,9]. After BGCs have been identified using the chosen BGC detection algorithm[10], BGCs from different genomes can be grouped into gene cluster families (GCFs) based on similarities in their overall gene content and sequence identities[9,11,12]. The choice of similarity threshold for grouping BGCs is an important one that can affect interpretation of inferred GCF biosynthetic products. GCFs formed using strict similarity thresholds will be smaller families composed of BGCs that produce identical metabolites, while those produced with permissive thresholds will be larger and contain BGCs that instead encode structurally related natural product families[11]. Such analyses enable researchers to uncover patterns of BGC prevalence and phylogenetic distribution, and also anchor GCF networks to experimentally characterized BGCs[8,9,11,13,14].

While genome sequences can discern the biosynthetic potential of a group of organisms, metabolomics profiles represent chemical phenotypes revealing structural information of expressed downstream natural products. When combined, genomics and metabolomics have the power to not only identify BGCs or natural products of interest, but to link natural

products to their biosynthetic machinery[15–29]. Integrated analyses of -omics datasets are growing steadily, and initiatives such as the Paired Omics Data Platform have recently become available[17]. Integrated ‘metabologenomics’ analyses have been applied to hundreds of bacterial strains and used to deorphanize dozens of BGCs, discovering new metabolite scaffolds and their biosynthetic machinery in the process[7,18–20]. Whereas these approaches have advanced in bacteria, they have not yet been applied to fungi.

Several computational strategies have been developed to link GCFs to their metabolites that can be categorized as either feature- or correlation-based[5,7,21–29]. Feature-based approaches use BGC sequences to predict specific core structures and functional groups and compare these building blocks to predicted molecular scaffolds from chemical datasets. Alternatively, correlation-based approaches compare BGC and metabolite profiles across a set of strains to assert associations between BGCs and metabolite products[21–24]. Current feature-based approaches have been developed largely for bacteria and depend on product type or heavily rely on homology to known BGCs to predict molecular structures[25–29]. As such, we chose to compare three correlation-based approaches for extension of metabologenomics to fungi: pattern matching, correlation scoring and intensity ratio analysis (Table 1, Fig. 1 and Supplementary Table 1).

Recently, our group conducted a global analysis of 1,037 publicly available fungal genomes and reported 12,000 GCFs (containing roughly 37,000 BGCs) that have not been linked to their encoded metabolites[9]. As the last decade of -omics research in bacteria has taught us, successful integration of metabolomics or genomics datasets requires considerable process optimization and overall results improve with increased strain number[15,16,20]. Given the stark differences in BGC content between fungi and bacteria[9], extension of ‘metabologenomics’ to the fungal kingdom has yet to be reported. Here, we surveyed extracts of 110 fungal strains using MS-based metabolomics and correlated metabolite signals to their cooccurrence with BGCs in genomics datasets (Fig. 1). Using published natural product–BGC pairs[30], we assess the impact of GCF networking thresholds on three scoring models for metabologenomics in fungi. We validate the methods using 25 known linkages and identify more than 200 new natural product–GCF pairs. As a specific case, the correct biosynthetic pathway of pestalamides was identified unambiguously using this platform, clarifying its himeic acid-like biosynthesis.

Results

Creation and analysis of GCF networks

Each genome in our dataset, derived from 64 publicly available genomes and 46 assembled genomes new to this study (Supplementary Table 2), was analyzed using antiSMASH v.6.0 (ref. [31]), yielding a dataset of 7,020 BGCs of ten biosynthetic types. The total number of BGCs detected per genome ranged from 20 to 114 (Supplementary Table 3). When producing GCF networks for correlation to metabolites, one must first apply a cutoff to their distance matrix to determine when a group of BGCs is ‘similar enough’ to be clustered into a single GCF. As the similarity of BGCs in a GCF can directly affect the interpretation of analysis (that is, the relatedness of encoded metabolites), we compared results from nine GCF networks (each containing GCFs from BGCs of ten biosynthetic types) produced using

different similarity thresholds (ranging from a permissive 50 to strict 90%). To do this, the 7,020 BGCs were converted into protein domain arrays and their similarity to other BGCs in the dataset was calculated based on the fraction of shared domains and sequence identity[9]. To identify GCFs with known metabolite products, we dereplicated this dataset using known fungal BGCs published in the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository[30].

The choice of similarity threshold had a major impact on the final numbers of GCFs in the nine resulting networks, with the strictest cutoff of 90% yielding a network of 5,837 GCFs (93 containing known BGCs from MIBiG) and the most permissive cutoff of 50% yielding a predictably smaller network of just 1,353 GCFs (109 anchored to known MIBiG clusters). While this analysis made clear that GCF networking parameters drastically affect the number and size of GCFs, it was unclear which parameters were optimal for correlation-based scoring of GCF-natural product pairs. To answer this, each of the nine GCF networks was subjected to correlation-based scoring for optimization of parameters for our final metabologenomics model (below).

Acquisition and dereplication of metabolomics datasets

To measure the presence-absence patterns of detectable fungal metabolites, we used liquid chromatography with tandem mass spectrometry (LC–MS/MS) to measure secondary metabolites of 110 strains grown on each of three conditions[32–34]. The MS profiles of 330 fungal extracts were combined and the total number of unique MS signals for each strain determined. After background subtraction and data processing, 30–627 unique ions were detected per strain (Supplementary Table 3) for a total of 9,301 ions in the MS^1 dataset. Using in-house dereplication libraries[35,36], we confidently identified more than 350 known metabolites. We combined these data with publicly available databases[37,38] and electrospray ionization–MS/MS spectral prediction tools[39], identifying a subset of 25 known metabolites (represented by 42 MS ion signals) that were detected in at least two strains and had known BGCs (Supplementary Figs. 1 and 2 and Supplementary Data 1).

Strengths and biases of correlation-based scoring methods

In the first approach, pattern matching, *P* values are calculated by comparing presence or absence patterns of GCFs and detected ions across strains using a chi-square test[7]. Although it is easy to interpret the significance of resulting correlations and compare results from multiple datasets, the pattern-matching approach may miss linkages in which metabolite expression and/or ion detection is low. The second approach, correlation scoring, overcomes this by weighing particular presence or absence patterns more heavily than others to rank the quality of GCF–metabolite pairs[15]. By using correlation scoring, the presence of an ion without its GCF is penalized more heavily than the GCF without the ion, whereas the presence of both a GCF and an ion is rewarded more strongly than the absence of both the GCF and ion across strains. We also assessed the use of a third approach developed specifically for this project, intensity ratio analysis, in which GCF–metabolite pairs are ranked based on the ratio of the averaged abundance of an ion across all replicates of GCF-containing strains divided by the averaged abundance of the same ion across all strains without the GCF. Although biased toward abundant metabolites, this quantitative metric

can overcome instances of column bleed during mass spectrometry acquisition (when a metabolite is retained on the column during a run and elutes during the analysis of the next sample in the queue) as well as the occasional tendency for peak picking algorithms to assign nonzero values to noise.

Optimization of GCF networking parameters

To identify optimal BGC similarity thresholds for creating GCFs for metabologenomics in our dataset, we correlated the nine GCF networks produced using different similarity cutoffs from the metabolomics data and evaluated the ability of these models to identify validated metabolite–GCF pairs used as ground truth in this study. The resulting correlations were plotted to visualize the overall distribution of scores in our dataset and the strength of correlations of validated matches (Fig. 2 and Supplementary Figs. 3–12).

GCF similarity thresholds had a major impact on the metabologenomics models regardless of scoring approach, either masking or enhancing the ability to detect true metabolite–GCF pairs. There was no ‘one-size-fits-all’ threshold; instead, BGCs of different biosynthetic types required different similarity thresholds for optimal clustering. An example of this is seen when comparing correlation distribution plots of nonribosomal peptide synthetase (NRPS) and nonreducing polyketide synthase (NRPKS) containing GCFs calculated at different thresholds (Fig. 2). For NRPS BGCs, a similarity threshold of 60% maximizes the number and strength of validated matches. For example, notoamide B, acetylaszonalenin and chrysogine all have significant correlations with their validated GCFs, visualized by purple points in Fig. 2a. When this threshold is increased to 70%, all these correlations get removed from analysis (because no experimentally detected BGCs were similar enough to the MIBiG cluster to be grouped into the same GCF) (Fig. 2b). Notably, as thresholds increased beyond 60%, the total number of MIBiG-anchored GCFs decreased because no BGCs in our dataset were considered related to the MIBiG cluster when using these stricter parameters (Fig. 2c). Although the optimal similarity threshold for maximizing the number and significance of validated correlations for NRPS-containing BGCs was 60% (Fig. 2c), this did not translate to the NRPKS BGCs (Fig. 2d–f). Instead, metabologenomics analysis using GCFs produced at the 60% threshold assigned nonsignificant correlations to many validated matches (purple points, Fig. 2d). With a 70% threshold, the number of significant correlations for validated matches nearly doubled (Fig. 2e). The significance of the association between mycophenolic acid and its GCF shifted from a nonsignificant $-\log_{10}(P)$ of 4.4 ($P > 0.05$) to a highly significant $-\log_{10}(P)$ of 25 ($P = 3.1 \times 10^{-21}$ after multiple-hypothesis correction) when using GCFs formed with 60 and 70% thresholds, respectively.

Instead of choosing one cutoff for all GCFs in our final network, we optimized similarity thresholds by BGC type to maximize the number of validated links with higher confidence levels in our score distributions for the entire dataset (see Methods for the final parameters implemented, Supplementary Figs. 3–12). The optimized set of GCFs for each biosynthetic type were combined into a single GCF network to use for metabologenomics. This final network contained 3,007 GCFs of ten biosynthetic types (Supplementary Figs. 13 and 14). Owing to the hybrid nature of many fungal BGCs, 1,303 BGCs in this dataset were assigned to multiple GCFs of different biosynthetic types and the total number of

nonredundant GCFs was 2,405. Only around 5% of GCFs (131 total) could be anchored to published BGCs from MIBiG. Notably, when comparing correlations calculated using different GCF networking parameters, our optimized dataset had the highest number of significant correlations for validated matches when compared to datasets using one cutoff for all biosynthetic types (Supplementary Fig. 15). These results emphasize the importance of GCF threshold selection for maximizing results from metabologenomics.

Correlation of GCF networks with metabolite data

With an optimized network of 3,007 GCFs, we evaluated the performance of three scoring methods to identify known metabolite–BGC pairs: pattern matching (*P* values), correlation scoring and the intensity ratio-based approach developed specifically for this project (Table 1). These results are shown in the integrated display format in Fig. 3a, with the cooccurrence data underpinning the method shown for three knowns in Fig. 3b–d. Beginning with the pattern-matching approach, we observed statistically significant correlations for 21 out of 25 of the known MIBiG compounds (84%) to at least one of their associated GCFs after applying the conservative Bonferroni correction for multiple-hypothesis testing (Fig. 3a, *y* axis, Supplementary Table 4). Validated matches were among the top five best correlations for 20 out of 25 (80%) when using a pattern-matching approach. Correlation scoring resulted in scores ranging from −50 to roughly 500, with an average score of 100 for the entire dataset and 125 for validated linkages (Fig. 3a, *x* axis); validated matches were among the top five correlations in 19 out of 25 (76%) metabolites (Supplementary Table 4). The intensity ratio was less effective, with validated matches ranking among the top five in only 9 out of 25 (36%) cases. However, because of the sheer number of potential correlations (3,007 GCFs × 9,301 ions or >27 million individual tests), we used it to filter for ion-GCF pairs with an intensity ratio ≥5. This reduced the total number of hypotheses from roughly 27 million down to roughly 420,000 to focus on more promising metabolite–GCF associations and improving the statistical power of the resulting correlations.

To assess the capacity of the three scoring approaches to assign natural product–GCF pairs, we plotted the true positive rates (sensitivity) against the false positive rates (specificity) using receiver operating characteristic curves. The performance of the three models for the 25 known pairs ranged from fair (intensity ratio model, area under curve (AUC) = 0.78) to good (pattern matching and correlation scoring models, AUC = 0.83 and 0.89, respectively) (Supplementary Fig. 16), quantifying the performance and stage of development for these three models. For the pattern-matching approach (AUC = 0.83), the total number of false positives for each metabolite (linkages to GCFs not known to be involved in biosynthesis) ranged from 0 to 59 correlations (with an average of 17, or roughly 0.6% of the 3,007 possible GCF linkages, Supplementary Table 5). While this level of false positive rate is tenable, efforts to disentangle true positives from several false ones requires considerable manual interpretation of MS and BGC data underlying putative correlations. Therefore, an integrated scoring approach was sought using the two top-performing scores.

Because the pattern-matching and correlation scoring approaches have different biases and can be calculated using the same data input (Supplementary Table 1), we used them in parallel to minimize false positives in our dataset. We first filtered out all nonsignificant

correlations using the pattern-matching approach, after which we ranked all the remaining correlations for each metabolite by correlation score and evaluated whether the true match was ranked first. In 16 out of 25 cases, the true GCF–metabolite linkage was ranked first, with zero confounding false positives. In one additional case, the true GCF–metabolite linkage was tied for first with a single false positive, and in a second case, it was ranked third (Supplementary Table 5). These results highlight the complementary nature of these approaches, and both models and metrics for correlation-based associations will continue to be developed[21–24]. Because of the labor required to confirm predictions, we selected a correlation scoring threshold that minimized false positives at the expense of increasing false negatives (Supplementary Fig. 17), choosing to focus initial discovery efforts on linkages to metabolites that had a correlation score of at least 140 and were among the top three correlation scores after filtering for significance. At the 110-strain level, we are able to assert a final set of high-confidence BGC associations for 238 ions in our dataset, with an estimated false positive rate of roughly 1% using knowns (Supplementary Fig. 17).

Untargeted metabologenomics reveals biosynthetic machinery for the pestalamides

Having characterized the specificity and sensitivity of fungal metabologenomics at the 110-strain level, we used it to uncover a new GCF–metabolite pair. We targeted an ion with an m/z of 343.129 that was detected above baseline levels in six of the 110 strains. The metabolite dereplicated as pestalamide B, a bioactive natural product belonging to a small but growing class of metabolites (roughly 24 analogs) containing 4-oxo-1,4-dihydropyridine-3-carboxamide and 4-oxo-4H-pyran-3-carboxamide core structures (Supplementary Fig. 18)[40–44]. Following a series of gene knockout and complementation experiments in *Aspergillus niger*, it was recently proposed that the NRPKS BGC *epa* is responsible for pestalamide biosynthesis[45]. Notably, the top producer of pestalamide B in our dataset (*A. brasiliensis* Westerdijk Fungal Biodiversity Institute (CBS) 101.740) lacked the *epa* BGC purportedly responsible for pestalamide biosynthesis (GCF NRPKS_59, Fig. 3e and Supplementary Table 6). Instead, the top GCF linkage to pestalamide B was TERPENE_288 ($-\log_{10}(P\text{ value})$ of 21 and a correlation score of 163) (Fig. 3e). Given the lack of terpene elements in the pestalamide B structure, we turned to our second-ranked linkage between pestalamide B and the GCF ‘HYBRID_85’ (a NRPS-PKS family of BGCs) that had a $-\log_{10}(P\text{ value})$ of 17 ($P = 4.6 \times 10^{-13}$ after Bonferroni correction) and a correlation score of 161. All six pestalamide B-producing strains possessed the GCF HYBRID_85, and three additional strains possessed this cluster but did not produce pestalamide B (Fig. 3e and Supplementary Table 6).

MS^2 analysis and molecular networking revealed a family of highly similar molecules that are also produced by a subset of the six pestalamide B-producing strains (Supplementary Fig. 19). This molecular family includes pestalamides A, B, D and additional analogs. Four molecules (pestalamide A, B and two putative new derivatives, pestalamide E and F) had significant correlations to the same GCF, HYBRID_85. Pestalamide A was found in seven out of nine strains with this hybrid GCF, and pestalamides E and F were detected in three and four of these nine strains, respectively (Supplementary Table 6). Pestalamide B was purified from *A. brasiliensis* CBS 101.740 and subjected to MS and nuclear magnetic resonance (NMR) analysis. Fragmentation patterns (Supplementary Fig.

20a) were consistent with those predicted using CFM-ID peak assignment algorithms, and NMR analyses matched previous publications (Supplementary Table 7 and Supplementary Figs. 21–25)[46]. Pestalamide A was identified through dereplication of its MS^2 data (Supplementary Figs. 20b and 26). Pestalamides E and F were inferred as hydrated and carboxylated forms of pestalamide B (Supplementary Fig. 20c,d); mass defect filtering was supportive of these assignments[36].

Based on the observation that HYBRID_85 was associated with pestalamide detection, we rigorously tested the assertion that BGCs in this GCF produced the pestalamides. Given that a previous publication implicated a different BGC[45], we heterologously expressed the HYBRID_85 BGC from *A. brasiliensis* CBS 101.740 (Supplementary Table 8) in *A. nidulans* after BGC cloning using a new CRISPR–Cas9 mega-endonuclease-based in vitro technology[47]. Notably, the transformed *A. nidulans* strain (*A. nidulans-pst*) showed production of pestalamide B and several analogs (Supplementary Fig. 19), while the extract from the *A. nidulans* host did not have detectable pestalamides (Fig. 4a). Production of pestalamide B in the heterologous expression strain was similar to several native producers (Supplementary Table 6), but was lower than the parent strain. Knockout of the backbone synthase in the heterologous host (*A. nidulans*-Δ*pstD*) resulted in a complete loss of pestalamide production, confirming the necessity of this BGC for pestalamide biosynthesis (Fig. 4a).

Next, we grew the heterologous expression strain *A. nidulans-pst* with stable isotopically labeled precursors. Labeling with [$^{13}C_6$]-leucine and with phenylacetic acid (phenyl-d_5) resulted in m/z shifts of +6 and +5 Da, respectively, in pestalamide B (Fig. 4b). This is consistent with adenylation domain specificity predictions that suggested that the A domain in the hybrid NRPS-PKS backbone recognizes leucine and indicates that the free-standing AMP-binding enzyme may act as a phenylacetyl-CoA ligase. Inspection of MS^2 spectra following leucine feeding studies suggested that leucine undergoes substantial rearrangement during pestalamide biosynthesis, with fragment ions showing +1, +4, +5 and +6 Da shifts (Fig. 4c). Similar shifts were evident on inspection of MS^2 spectra for pestalamides A, E and F (Supplementary Fig. 27). The biosynthetic pathway for himeic acid A, a structurally related metabolite to the pestalamides, was recently described[41]. Through a series of NMR-based stable isotope feeding and gene disruption experiments, researchers reported that the *him* BGC, containing a hybrid NRPS-PKS backbone, is responsible for himeic acid biosynthesis. The HimA synthetase incorporates leucine into the NRPS-PKS product, which then is rearranged to form the pyrone intermediate[41]. This rearrangement shows excellent agreement with our biosynthetic feeding studies (Supplementary Fig. 28a,b), suggesting that the pestalamides are formed through a similar mechanism. While himeic acid and the pestalamides share a 4-oxo-4H-pyran/dihydropyridine-3-carboxamide core structure (Supplementary Fig. 28a), their deviations in structure (that is, a long fatty acyl side chain in place of a phenylacetic acid moiety) are mirrored by deviations in BGC content (Supplementary Fig. 28c). For example, the *him* BGC and the *A. brasiliensis* BGC belonging to HYBRID_85 (named *pst*) are only partially similar, with three of their encoded genes sharing moderate sequence identity (Supplementary Fig. 28d). Although the overall sequence identity between the HimA and PstD NRPS-PKS synthetases and their domains is moderate (Supplementary Fig. 28e), both the domain architecture of these

genes and the A-domain active site residues are nearly identical (Supplementary Fig. 28f). These findings suggest that the enzymes involved in leucine incorporation and subsequent 4-oxo-4H-pyran/dihydropyridine-3-carboxamide formation are functionally equivalent in these two biosynthetic pathways.

Taken together, these results indicate that pestalamide biosynthesis occurs through a different pathway than described by Wang et al.[45]. Based on our data, we propose a biosynthetic scheme for the pestalamides in Fig. 5. Unlike himeic acids, pestalamides contain a phenylacetic moiety instead of a highly reduced alkyl chain. As such, we speculated that the AMP-binding protein PstH may function as a phenylacetyl-CoA ligase. This gene has moderate sequence identity (29%) to the phenylacetyl-CoA ligase PhlB involved in penicillin G biosynthesis[48]. We suggest that first, the PKS portion of the hybrid synthetase PstD, together with the phenylacetyl-CoA ligase PstH, catalyze the formation of the polyketide part of the molecule, which is then condensed with leucine by the NRPS portion of PstD. The resulting intermediate, formed from one phenylacetyl-CoA unit, two malonyl-CoA units and one leucine, is then cyclized and released from PstD (perhaps facilitated by the reductase PstG). PstM or PstN (an amidohydrolase and amidotransferase, respectively) may then initiate a ketone to imine conversion, after which the monooxygenase PstC catalyzes an α-oxidation of the tetramic acid ring, prompting the himeic acid-like rearrangement of the NRPS portion of the molecule to yield the pyrone or pyridone intermediate (for pestalamide A or B, respectively) (Fig. 5 and Supplementary Fig. 28). The subsequent dehydration reaction is catalyzed by the dehydrogenase PstA, after which the molecule is carboxylated by the P450 PstL. Because the putative derivative pestalamide F possesses an additional CO_2, it is possible that PstL catalyzes two carboxylation steps.

Discussion

Genomics and metabolomics datasets illuminate the biosynthetic capacity of living organisms, providing ample opportunity for scientific discovery. In the last decade, a growing number of laboratories have integrated -omics datasets to systematically mine the untapped chemical potential of the domain Bacteria. Such studies have illustrated that paired -omics analysis can help improve predictions, prioritize biosynthetic pathways and target discovery of new natural products[21–24].

Illustrated by this first application of paired -omics in fungi, optimization of scoring was required to suppress false positives from metabologenomics datasets. Our comparative analysis demonstrated the importance of optimizing GCF networking parameters before integrated analysis to ensure the robustness of high-scoring matches for targeted study. We found that GCF networking parameters can be optimized by BGC subtype, with certain classes requiring higher degrees of BGC overlap than others. We speculate that the biosynthetic classes with the high flexibility for domain structure and those with iterative activity require higher similarity thresholds for GCF grouping than those with a more rigid and/or modular structure. Notably, our choice of antiSMASH as a BGC detection algorithm limits us to the detection of BGCs containing core enzymes that are incorporated into this program[7]. Likely, there are hundreds of BGCs in our dataset that were not detected due to their non-canonical nature. Given the relatively small size of the 110-strain dataset and the

fact that we validated our process using established BGCs from MIBiG, it is possible that our manually optimized parameters led to model overfitting. Future studies will be required to see whether the thresholds chosen for this project will expand to larger datasets and to fungal genera not included in this study. Future efforts to automate the optimization of GCF networking parameters would be worthwhile to minimize likelihood of overfitting and reduce manual interpretation.

It is worth acknowledging that 16% of validated metabolite–GCF pairs did not have significant correlations after multiple-hypothesis correction even after dataset optimization. Because correlation-based approaches rely on overlapping ion/GCF detection profiles, BGCs for metabolites with low detection levels cannot be identified confidently. For example, aspernidine A had a $-\log_{10}(P)$ value of 4.7 ($P > 0.05$) (Fig. 3a and Supplementary Table 4) due to its low level of detection across strains. Out of the 11 strains with the GCF responsible for aspernidine biosynthesis, only two of them had detectable levels of aspernidine A. Future studies aimed at maximizing silent BGC expression (for example, with the inclusion of epigenetic modifiers to fungal culture[49] or by fungal–fungal coculture[50]), may help to improve correlations for metabolite with low detection levels.

Notably, all correlation-based scoring approaches share an inherent limitation in that they cannot differentiate associations between ions and GCFs that share the same presence or absence patterns across strains. While useful for ranking potential matches, manual interpretation of the chemical and BGC data for top-scoring hits is still required to determine how well the structural features of the target ion align with the correlated BGC (as illustrated with pestalamide B). One can imagine that future efforts to develop feature-based metrics for fungi will provide a complementary methodology to assign BGC-metabolite linkages and minimize manual interpretation of datasets for targeted analysis.

As is the case with metabologenomics in bacteria, correlation scores will improve substantially with even two- to threefold larger datasets[7,15]. To illustrate this, we recalculated correlations using subsets of fungi from our 110-strain dataset, and plotted the resulting AUC receiver operating characteristic values against strain number in a power curve (Supplementary Fig. 29). From Supplementary Fig. 29 we can clearly see that the AUC values increase sharply with strain number for known metabolite–GCF pairs. The variability in performance between datasets of the same size containing different subsets of strains is high when ≤20 strains are included but stabilizes at the >100-strain level (Supplementary Fig. 29). We found that some smaller datasets had higher AUCs than our final dataset, illustrating that strain selection is particularly important and can be guided by phylogenetic information. Specifically, one seeks a balance between the extent of BGC overlap within a strain collection and the coverage of new BGC types.

In a recent global analysis of roughly 1,000 fungal genomes, nearly 12,000 fungal GCFs (from roughly 37,000 BGCs) without known metabolite products were reported[9], underscoring that even well-characterized fungi have the potential to biosynthesize compounds greatly exceeding known fungal chemical space. With this work, we illustrate that correlative metabologenomics is a promising tool to measure the diversity of fungal secondary metabolism and link natural products to their BGCs. Even in our modest 110-

strain dataset, correlative metabologenomics analysis enabled us to revise the pestalamide biosynthetic pathway and revealed more than 200 high-confidence metabolite–GCF pairs to inspire future discovery.

Methods

Growth and extraction

Strains were acquired from private and public strain libraries, including the Agricultural Research Service (NRRL) collection, the American Type Culture Collection and the CBS collections. All fungi were grown in three growth conditions to increase the likelihood of secondary metabolite production. Glycerol stocks or agar plugs of each fungus were used to inoculate potato dextrose agar plates that were incubated at 21 °C until the mycelial mat was fully grown at 5–7 days. Agar plugs were then cut from plates and transferred to 10 ml of YESD broth and cultivated at 21 °C at 150 rpm for 3 days. Three seed cultures were prepared for each strain and used to inoculate three 250 ml Erlenmeyer flasks that contained autoclaved rice, oats or Cheerios, which were grown at 21 °C for 2–5 weeks (refs. [32–34]).

To extract secondary metabolites, cultures were processed using methods established in the literature[32–34]. Briefly, 60 ml of MeOH–$CHCl_3$ (1:1) were added to each flask, followed by chopping with a spatula, sonicating briefly and leaving overnight at room temperature. Cultures were then filtered in vacuo after brief sonication, and 90 ml of $CHCl_3$ and 150 ml of H_2O were added to the filtrate. The mixture was transferred to separatory funnel, where the organic layer (bottom) was collected and evaporated under N_2. The dried organic layer was reconstituted in 100 ml of MeOH–CH_3CN (1:1) and 100 ml of hexanes transferred to a separatory funnel and shaken vigorously. The defatted organic layer (MeOH–CH_3CN) was dried under N_2.

DNA extraction and sequencing

Fungal genomic DNA was extracted from lyophilized hyphal mycelial mats using the Basic Protocol 4 phenol-chloroform extraction described in a previous publication[51]. The quality of the DNA was checked by running extracts on a 1% agarose gel before sending for sequencing at the University of Illinois at Urbana-Champaign Roy J. Carver Biotechnology Center. These genomes were sequenced in two batches using Illumina NovaSeq 6000 sequencing technology. The shotgun gDNA libraries were constructed from 300 ng of DNA after sonication with a Covaris ME220 (Covaris) to an average fragment size of 400 bp with the Hyper Library Preparation Kit from Kapa Biosystems (Roche). To prevent index switching, the libraries were constructed using unique dual indexed adapters from Illumina. The individually barcoded libraries were amplified with three cycles of PCR and run on a Fragment Analyzer (Agilent) to confirm the absence of free primers and primer dimers, and to confirm the presence of DNA of the expected size range. Libraries were pooled in equimolar concentration and the pool was quantified by quantitative PCR on a BioRad CFX Connect Real-Time System (BioRad Laboratories, Inc.).

The pooled barcoded shotgun libraries were loaded on a NovaSeq SP lane for cluster formation and were sequenced for 250 cycles from each side of the DNA fragments. The

FASTQ read files were generated and demultiplexed with the bvl2fastq v.2.20 Conversion Software (Illumina). Paired reads were assembled using SPAdes[52], and gene prediction was performed using AUGUSTUS with default parameters[53]. The remaining 64 genomes were downloaded from the National Center for Biotechnology Information (NCBI) or Joint Genome Institute[54,55]. A table of strains in this study and genome accession numbers, including 46 newly sequenced genomes for this project, can be found in Supplementary Table 2. Genomes sequenced for this project are available under BioProject accession no. PRJNA852164.

Gene cluster prediction and network analysis

Gene clusters were predicted using antiSMASH six via command line with the argument '—taxon fungi.' For GenBank genomes that contained gene predictions, we downloaded .gbff files to use as input to antiSMASH along with the '–genefinding-tool none' argument to prevent gene calling on contigs that lacked predicted genes. For genomes sequenced for this project and those downloaded from the Joint Genome Institute, we instead used .fasta sequence files and .gff gene prediction files as the input to antiSMASH, with the '—genefinding-gff3' argument.

Following BGC prediction using antiSMASH, we performed protein domain prediction and alignment as previously described[9]. Protein domains were detected using hmmscan[56] via the command line, using the arguments '—cut-tc', for trusted cutoffs, '—no-ali' to skip alignments and '—domtblout' to output a tabular format. Following detection of protein domains, we used domain-profile alignments rather than all-versus-all sequence alignments to decrease compute time. Each predicted protein domain was aligned to its Pfam model using hmmalign with default parameters. For each genome, we stored the resultant aligned protein domains in JSON format for later use. These protein prediction and alignment steps were implemented as scripts and command line wrappers written in C#10 running on .NET 6.

GCF network creation first required classification of BGCs into biosynthetic types. Each gene within a BGC was classified according to the presence or absence of biosynthetic domains according to Supplementary Table 9, and the BGC was given a biosynthetic designation based on the combined classifications of its genes. BGCs within each biosynthetic type were compared based on sequence identity and domain composition. We first removed from consideration all gene cluster pairs without shared adjacent protein domains pairs. For each pair of gene clusters within a biosynthetic type, we computed the sequence identity of each pair of protein domains with the same Pfam model. The average of each domain sequence identity value was used as a final similarity metric for each gene cluster pair. These similarity thresholds were optimized by biosynthetic type to maximize the number of validated linkages with significant scores. We chose a 60% cutoff for NRPS GCFs; a 65% cutoff for dimethylallyl tryptophan synthase (DMAT) and polyketide synthase (PKS)-like GCFs; a 70% cutoff for HRPKS, NRPKS, PRPKS and terpene GCFs, and a 75% cutoff for NRPS-like GCFs. For GCF types without validated matches (RiPP and hybrid NRPS-PKS in this dataset), we chose a threshold of 65%. Edges that passed the

similarity threshold were removed, resulting in GCF networks for each biosynthetic type. These similarity comparisons were implemented in C#10 running on .NET 6.

LC–MS analysis

Dried samples were resuspended in MeOH at 1 mg ml^{-1} and transferred into filter vials. Each strain was represented by three distinct extracts (for cultures grown on rice, oats and Cheerios), with most extracts being injected once on the mass spectrometer. A subset of seven extracts were injected in triplicate. LC–MS was used to analyze samples on a Thermo Q Exactive mass spectrometer equipped with an inline Agilent 1290 Infinity II ultrahigh performance liquid chromatograph separated using a Kinetix 1.3 μm C18 100 Å particle-size column with dimensions 50 × 2.1 mm^2. Column temperature was maintained at 40 °C with a flow rate of 0.3 ml min^{-1} and an injection volume of 5 μl. Mobile phases of 5% CH_3CN in H_2O with 0.1% formic acid as buffer A and 100% acetonitrile with 0.1% formic acid as buffer B were used with the following gradient: 5–100% B from 0 to 8.00 min; 100% B from 8.00–9.00 min and 100–5% B from 9.00–9.20 min. Analyte detection by electrospray ionization–MS was completed in positive mode using the following settings: capillary temperature 320 °C, sheath gas 10 (arbitrary units) and spray voltage 3.6 kV. Full scan MS spectra were acquired at a resolving power of 17,500 for the mass range of 150–2,000 m/z. MS^2 analysis was conducted in a data-dependent mode, selecting for the top five ions of each scan for fragmentation. A normalized collision energy of 25 was used for higher-energy collisional dissociation.

To assess how well our approach identified known natural product–gene cluster pairs, we sought to identify fungal metabolites whose gene clusters are published in the MIBiG database. Metabolite identification was accomplished by dereplicating the MS data using in-house libraries from Northwestern University and data from over 625 fungal secondary metabolites from the University of North Carolina at Greensboro[35,36]. To identify additional metabolites, we downloaded the fungal database from Natural Product Atlas[38] and targeted compounds whose accurate masses matched those of fungal metabolites from the MIBiG database[30]. Fragmentation spectra from these metabolites were then compared to experimental spectra in the Global Natural Products Social Molecular Networking (GNPS) database[37]. For ions that did not have a spectral match to in-house or public databases, we compared in silico MS^2 spectra using CFM-ID v.4.0 algorithms to experimental spectra and those with at least three matched fragments were assigned tentative identifications[39].

ProteoWizard[57] was used to convert ThermoRAW MS data to .mzXML and uploaded onto MZmine v.2.53 (ref. [58]). The ADAP workflow (chromatogram deconvoluted wavelengths) was used for peak detection followed by deisotoping and alignment of the peaks[59]. Finally, the peaks with MS/MS scans were retained and filtered. The MS^1 level that contained the m/z values, retention time and peak height associated with each feature was exported into a .csv file for final correlations (Supplementary Data 2). The same .mzXML files used for creating the MZmine dataset were uploaded to GNPS for molecular networking analysis as described below. All .mzXML files are available through the MassIVE repository under accession no. MSV000089848. Detailed parameters for MZmine processing are outlined in Supplementary Table 10.

A molecular network was created using the online workflow on the GNPS website (http://gnps.ucsd.edu). The data were filtered by removing all MS^2 fragment ions within ±17 Da of the precursor *m*/*z*, and only the top six fragment peaks were compared for analysis. The precursor ion mass and the MS^2 fragment ion tolerance were each set to 0.02 Da. Consensus spectra containing fewer than ten spectra were discarded. Because it is possible for related metabolites to be formed by separate gene clusters, we chose a strict similarity threshold for molecular networking analysis to avoid grouping metabolites that were not biosynthetically related. Ions in the resulting molecular network were grouped into the same molecular family if they had a cosine score (similarity score) above 0.95 and more than three matched fragment peaks. The maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from the families until the family size was below this threshold. To dereplicate against spectral libraries in the GNPS database, library spectra were filtered in the same way as the input data and matches kept between network spectra and library spectra were required to have at least three matched peaks and a cosine score ≥0.7.

Correlation of metabolomics and genomics datasets

The input to the correlations process was a presence or absence matrix of GCFs across the 110 strains, as well as the metabolite feature table. csv exported from MZmine. To remove ubiquitous GCFs such as the fungal lysine biosynthetic pathway, we excluded all GCFs present in >80% of the strains in our dataset. As intensity values of zero would result in division by zero for the intensity ratio, a zero-filling strategy was required. We used a simple routine that assumed that undetected metabolites were below the limit of detection. For each zero in the metabolite feature table, we replaced it with a value near the limit of detection by randomly generating a value between the minimum intensity and 0.01% of the maximum intensity for the file. For each strain, we took the highest observed metabolite intensity across the three growth conditions, resulting in a metabolite intensity table based on strain rather than LC–MS file. Metabolites detected in fewer than two strains with at least 1×10^7 intensity were removed from further analysis.

For each biosynthetic type, we computed correlation metrics between every pair of appropriate GCFs and filtered metabolites. For each metabolite–GCF pair, we computed the number of strains with both the GCF and the metabolite, with the metabolite but not the GCF, without the metabolite but with the GCF and without the GCF and without the metabolite. These data were scored using the previously described correlation score[15] using a 1×10^7 intensity threshold. As previously described[20], we also used a chi-squared test (implemented in the C# framework Accord.NET) to test the significance of each metabolite–GCF correlation. To compute the intensity ratio, we determined the average intensity of the metabolite in strains that contained the GCF, as well as the average intensity in strains without the GCF. The ratio of intensities in strains with versus without the GCF was used as the intensity ratio score, only considering metabolite–GCF pairs with intensity ratios ≥5. All analyses were implemented in C#10 running on .NET 6.

Metabolite purification and structure elucidation

A scale up culture of *Aspergillus brasiliensis* CBS 101.740 was grown in 25 flasks of Cheerios and extracted, yielding 1.9 g of dried extract. This extract was subjected to normal-

stage flash chromatography using a Teledyne ISCO CombiFlash NextGen 300 system and evaluated using ultraviolet absorbance at 254 and 280 nm using a 60-min hexane/$CHCl_3$/MeOH gradient on a silica 40-g gold column (Teledyne ISCO) at a flow rate of 40 ml min^{-1}; the resulting fractions were combined into 22 pools. Fraction 19 (112 mg) had the highest concentration of the target metabolite and was purified further using reversed-phase flash chromatography on a 15.5 g HP C18 gold RediSep Rf column at a flow rate of 30 ml min^{-1}. A 50-min gradient of CH_3CN/H_2O was used, starting at 5% CH_3CN and increasing to 25% over 11 min. It then increased to 50% from 11 to 38 min, after which it was increased to 100% CH_3CN over 2 min and held for the remainder of the run. Of the 14 resulting fractions, fraction 3 (15 mg) had the most target metabolite and was subjected to a final round of reversed-phase chromatography on an Agilent 1200 HPLC and analyzed with OpenLAB CDS Chemstation software (v.1.8.1, Agilent Technologies). Fraction 19–3 was chromatographically separated on a Kinetix C18 semipreparatory column (5 μm; 100 Å; 250 × 10.0 mm^2) with a 2.5 ml min^{-1} flow rate. Approximately 200 μl of a 10 mg ml^{-1} solution were injected per run and subsequent fractions pooled for final analysis. The 45-min run began at 15:85 CH_3CN-H_2O and was held isocratically for 38 min. The gradient was increased to 100% CH_3CN over the next 2 min and held at 100% for the remainder of the run. Fourteen fractions were collected, and fraction 3 (15 mg) was subjected to a final round of reversed-phase high-performance liquid chromatography. Compound **1** (pestalamide B) eluted at 17.5 min (1.5 mg, 95% purity, 0.08% yield). NMR spectra for the target compound were collected on a Bruker Avance III 500-MHz system equipped with a DCH CryoProbe at 298.2 K.

Pestalamide B (**1**): yellow, amorphous solid; HRESIMS *m*/*z* 343.1290 $[M+H]^+$ (calculated for $C_{18}H_{19}N_2O_5^+$, 343.1294). Fragmentation patterns are provided in Supplementary Fig. 20. ^1H, ^{13}C, correlation spectroscopy, heteronuclear single quantum correlation and heteronuclear multiple bond correlation data (dimethylsulfoxide-d_6) are provided in Supplementary Table 7 and Supplementary Figs. 21–25 and match favorably to previous publications[46].

Plasmid fungal artificial chromosome (pFAC) vector design

The large-insert cloning vector pSMART-BAC (NCBI accession no. EU101022) was modified into the fungal artificial chromosome vector pFAC by linearizing adjacent to the oriV site and inserting a cassette composed of kanR (*Escherichia coli* selection), AMA1 (fungal plasmid replicator), gpdA-HygR-trpC terminator (fungal selection) and pyrG (fungal selection). See Supplementary Table 11 for a complete list of plasmids used for this study, and Supplementary Fig. 30 for graphical maps of plasmids pFAC-*pst* and pFAC-Δ*pstD*.

pst Cas9 cleavage and assembly to pFAC

A. brasiliensis gDNA (50–150 ng μl^{-1}) was restricted with Cas9 (33–133 ng μl^{-1}) combined with equimolar guide RNAs directing cleavage upstream and downstream of the *pst* BGC locus. Digestion was performed in a 20 μl reaction containing TA buffer (33 mM tris acetate, 66 mM potassium acetate and 10 mM magnesium acetate, pH 7.5) at 37 °C for 90 min. The cleaved sample was mixed with linearized pFAC containing overlap sequences specific to the released *pst* BGC fragment in an isothermal DNA assembly reaction. The reaction

product was transformed to *E. coli* BACOpt2.0 (Lucigen) cells and plated to Luria-Bertani (LB) agar containing 12.5 μg ml^{-1} chloramphenicol. Colonies were screened by PCR using primer pairs specific to the cloning junctions (Supplementary Fig. 31a). Sanger sequencing of the colony PCR amplicons further confirmed successful cloning. Positive clones were cultured in 100 ml of LB medium containing 12.5 μg ml^{-1} chloramphenicol and midiprep of the BAC DNA was performed using Zymo-PURE II Plasmid Midiprep Kit (Zymo Research). One μg of BAC DNA was restricted by either EcoRI or BamHI and compared to simulated digests, confirming correct cloning of the *pst* BGC into pFAC (Supplementary Fig. 32a). See Supplementary Table 12 for a list of primers and oligonucleotides purchased from IDT used in these studies, and Supplementary Table 13 for target sequences for digestion. Guide RNAs were obtained from IDT by providing the required target sequences and preparing them according to recommended procedures.

pFAC-*pst*-Δ*pst* construct design

Purified pFAC-*pst* DNA (50 ng μl^{-1}) was restricted with Cas9 (30 ng μl^{-1}) combined with equimolar guide RNAs (IDT) directing cleavage upstream and downstream of the *pstD* gene to create a knockout construct. Digestion was performed in 60 μl of reaction containing TA buffer at 37 °C for 30 min. The cleaved sample was added to an isothermal DNA assembly reaction with a PCR product of the apramycin resistance gene ApramR amplified from pSMART-BAC-S and containing overlaps specific to the cleaved FAC. The reaction product was transformed to *E. coli* BACOpt2.0 (Lucigen) cells and plated to LB agar containing 50 μg ml^{-1} apramycin. Colonies were screened by PCR using primer pairs specific to the cloning junctions (Supplementary Fig. 31b). Sanger sequencing of the colony PCR amplicons further confirmed successful cloning. Positive clones were cultured in 100 ml of LB medium containing 50 μg ml^{-1} apramycin and midiprep of the BAC DNA was performed using Zymo-PURE II Plasmid Midiprep Kit (Zymo Research). One μg of BAC DNA was restricted by EcoRI or BamHI and compared to simulated digests, confirming correct deletion of the *pstD* gene (Supplementary Fig. 32b).

Heterologous expression of pFAC plasmids in host *A. nidulans*

The pFACs were transformed in the *A. nidulans* host using a modified PEG-calcium based transformation method reported to improve transformation yield[60]. Transformants with the *pst* cluster were confirmed by PCR with a forward primer AbrasF and a reverse primer AbrasR located in the key gene *pstD* ORF (Supplementary Fig. 33), and transformants with *pstD* deleted from the *pst* cluster were confirmed with a forward primer ΔpstF located outside the KO cassette and a reverse primer aprR marker (apramycin resistance) gene for *pstD* gene replacement (Supplementary Fig. 34). Two confirmed transformants per group were chosen for subsequent metabolite analysis. All primer sequences used for the confirmation are listed in Supplementary Table 14 and fungal strains used for heterologous expression are listed in Supplementary Table 15. For the identification of the expressed target metabolite, triplicated plates of each *A. nidulans* transformant with and without target pFACs were inoculated and incubated for 7 d at 37 °C on glucose minimal medium. Subsequently, plates were collected and lyophilized for 48 h and extracted following established procedures[60].

Stable isotope feeding studies

For targeted biosynthetic studies, pyroxidine supplemented potato dextrose agar plates with and without 0.1 mM [$^{13}C_6$]-l-leucine or 0.5 mM phenylacetic acid (phenyl-d_5) were prepared. *Aspergillus nidulans* control and transformant strains (containing the pestalamide BGC) were grown on plates for 1 week at room temperature in darkness. Agar plugs from these plates were then used to inoculate pyridoxine supplemented LMM media with and without 1 mM l-leucine or 5 mM phenylacetic acid. Cultures were grown for 1 week at room temperature without any rotation. Mycelia were collected in a falcon tube and placed in −80 °C for roughly 4 h after which it was lyophilized, crushed into a fine powder and extracted using 15 ml of 10% methanol in ethyl acetate. Tubes were incubated for 3.5 h at room temperature and sonicated every 30 min. The resulting extracts were then filtered and evaporated to under N_2.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Genome sequencing for this project was conducted at the Roy J. Carver Biotechnology Center at the University of Illinois-Urbana-Champaign. This research was supported in part by the National Institutes of Health grant nos. F32 GM132679 to L.K.C., R01 GM112739–05A1 to N.P.K., T32 GM135066 to G.N., R44 AI140943–03 to J.W.B., P01 CA125066 to N.H.O. and 2R01 AT009143 to N.L.K. This work also made use of the IMSERC NMR facility at Northwestern University, which has received support from the Soft and Hybrid Nanotechnology Experimental Resource (grant no. NSF ECCS-2025633) grant. Figures 1, 4 and 5 were created using BioRender.com.

Data availability

All genomes that were sequenced for this work are available via NCBI under BioProject PRJNA852164. The metabolomics data (as .mzXML files) for the 110-strain dataset are available via the MassIVE repository under accession no. MSV000089848. Additionally, we have included Supplementary Data 1, which includes.html files for all MIBiG-anchored GCFs with detected metabolites, as well as the pestalamide GCF discovered in this work. The processed MZmine peak list that we used for correlations (generated using the publicly available .mzXML files) is provided as Supplementary Data 2.

References

1. Bernardini S, Tiezzi A, Laghezza Masci V & Ovidi E Natural products for human health: an historical overview of the drug discovery approaches. Nat. Prod. Res. 32, 1926–1950 (2018). [PubMed: 28748726]
2. Hyde KD et al. The amazing potential of fungi: 50 ways we can exploit fungi industrially. Fungal Divers. 97, 1–136 (2019).
3. Ráduly Z, Szabó L, Madar A, Pócsi I & Csernoch L Toxicological and medical aspects of Aspergillus-derived mycotoxins entering the feed and food chain. Front. Microbiol. 10, 2908 (2020). [PubMed: 31998250]
4. Bills GF, Gloer JB, Heitman J, Howlett BJ & Stukenbrock EH Biologically active secondary metabolites from the fungi. Microbiol. Spectr. 4, 4.6.01 (2016).
5. Li YF et al. Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. Fungal Genet. Biol. 89, 18–28 (2016). [PubMed: 26808821]

6. Keller NP Fungal secondary metabolism: regulation, function and drug discovery. Nat. Rev. Microbiol. 17, 167–180 (2019). [PubMed: 30531948]
7. Caesar LK, Montaser R, Keller NP & Kelleher NL Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities. Nat. Prod. Rep. 38, 2041–2065 (2021). [PubMed: 34787623]
8. Kautsar SA, van der Hooft JJJ, de Ridder D & Medema MH BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience 10, giaa154 (2021). [PubMed: 33438731]
9. Robey MT, Caesar LK, Drott MT, Keller NP & Kelleher NL An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes. Proc. Natl Acad. Sci. USA 118, e2020230118 (2021). [PubMed: 33941694]
10. Chavali AK & Rhee SY Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. Brief. Bioinform. 19, 1022–1034 (2017).
11. Navarro-Muñoz JC et al. A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. 16, 60–68 (2020). [PubMed: 31768033]
12. Nielsen JC et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. Nat. Microbiol. 2, 17044 (2017). [PubMed: 28368369]
13. Kautsar SA, Blin K, Shaw S, Weber T & Medema MH BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. 49, D490–D497 (2021). [PubMed: 33010170]
14. Drott MT et al. Microevolution in the pansecondary metabolome of *Aspergillus flavus* and its potential macroevolutionary implications for filamentous fungi. Proc. Natl Acad. Sci. USA 118, e2021683118 (2021). [PubMed: 34016748]
15. Doroghazi JR et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat. Chem. Biol. 10, 963–968 (2014). [PubMed: 25262415]
16. Goering AW 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). [PubMed: 27163034]
17. Schorn MA et al. A community resource for paired genomic and metabolomic data mining. Nat. Chem. Biol. 17, 363–368 (2021). [PubMed: 33589842]
18. Duncan KR et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. Chem. Biol. 22, 460–471 (2015). [PubMed: 25865308]
19. Maansson M et al. An integrated metabolomic and genomic mining workflow to uncover the biosynthetic potential of bacteria. mSystems 1, e00028–15 (2016).
20. Tryon JH et al. Genome mining and metabolomics uncover a rare d-capreomycidine containing natural product and its biosynthetic gene cluster. ACS Chem. Biol. 15, 3013–3020 (2020). [PubMed: 33151679]
21. Männle D et al. Comparative genomics and metabolomics in the genus *Nocardia*. mSystems. 5, e00125–20 (2020). [PubMed: 32487740]
22. Handayani I et al. mining indonesian microbial biodiversity for novel natural compounds by a combined genome mining and molecular networking approach. Mar. Drugs 19, 316 (2021). [PubMed: 34071728]
23. Cao L, Shcherbin E & Mohimani H A metabolome- and metagenome-wide association network reveals microbial natural products and microbial biotransformation products from the human microbiota. mSystems 4, e00387–19 (2019). [PubMed: 31455639]
24. Cao L et al. MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. Cell Syst. 9, 600–608 (2019). [PubMed: 31629686]
25. Hjörleifsson Eldjárn G et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLoS Comput. Biol. 17, e1008920 (2021). [PubMed: 33945539]
26. Johnston CW et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. Nat. Comm. 6, 8421 (2015).

27. Kersten RD & Weng J-K Gene-guided discovery and engineering of branched cyclic peptides in plants. Proc. Natl Acad. Sci. USA 115, E10961–E10969 (2018). [PubMed: 30373830]
28. Merwin NJ et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. Proc. Natl Acad. Sci. USA 117, 371–380 (2020). [PubMed: 31871149]
29. Mohimani H et al. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. J. Nat. Prod. 77, 1902–1909 (2014). [PubMed: 25116163]
30. Medema MH et al. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol. 11, 625–631 (2015). [PubMed: 26284661]
31. Blin K et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 49, W29–W35 (2021). [PubMed: 33978755]
32. Al Subeh ZY et al. Media and strain studies for the scaled production of cis-enone resorcylic acid lactones as feedstocks for semisynthesis. J. Antibiotics. 74, 496–507 (2021).
33. Flores-Bocanegra L et al. Cytotoxic naphthoquinone analogues, including heterodimers, and their structure elucidation using LR-HSQMBC NMR experiments. J. Nat. Prod. 84, 771–778 (2021). [PubMed: 33006889]
34. Knowles SL et al. Opportunities and limitations for assigning relative configurations of antibacterial bislactones using GIAO NMR shift calculations. J. Nat. Prod. 84, 1254–1260 (2021). [PubMed: 33764773]
35. El-Elimat T et al. High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. J. Nat. Prod. 76, 1709–1716 (2013). [PubMed: 23947912]
36. Paguigan ND et al. Enhanced dereplication of fungal cultures via use of mass defect filtering. J. Antibiot. 70, 553–561 (2017).
37. Wang M et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. 34, 828–837 (2016). [PubMed: 27504778]
38. Van Santen JA et al. The natural products atlas: an open access knowledge base for microbial natural products discovery. ACS Cent. Sci. 5, 1824–1833 (2019). [PubMed: 31807684]
39. Wang F et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. Anal. Chem. 93, 11692–11700 (2021). [PubMed: 34403256]
40. Ding G et al. Pestalazines and pestalamides, bioactive metabolites from the plant pathogenic fungus *Pestalotiopsis theae*. J. Nat. Prod. 71, 1861–1865 (2008). [PubMed: 18855443]
41. Hashimoto M, Kato H, Katsuki A, Tsukamoto S & Fujii I Identification of the biosynthetic gene cluster for Himeic acid A: a Ubiquitin-Activating Enzyme (E1) inhibitor in Aspergillus japonicus MF275. Chem. Bio. Chem. 19, 535–539 (2018).
42. Hiort J et al. New natural products from the sponge-derived fungus *Aspergillus niger*. J. Nat. Prod. 67, 1532–1543 (2004). [PubMed: 15387655]
43. Zhou H et al. Penipyridones a–f, pyridone alkaloids from *Penicillium funiculosum*. J. Nat. Prod. 79, 1783–1790 (2016). [PubMed: 27359163]
44. Zhou X et al. Aspernigrins with anti-HIV-1 activities from the marine-derived fungus *Aspergillus niger* SCSIO Jcsw6F30. Bioorg. Med. Chem. Lett. 26, 361–365 (2016). [PubMed: 26711143]
45. Wang B et al. Deletion of the epigenetic regulator GcnE in *Aspergillus niger* FGSC A1279 activates the production of multiple polyketide metabolites. Microbiol. Res. 217, 101–107 (2018). [PubMed: 30384904]
46. Chiang Y-M et al. Characterization of a polyketide synthase in Aspergillus niger whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-γ-pyrone. Fungal Genet. Biol. 48, 430–437 (2011). [PubMed: 21176790]
47. Montaser R & Kelleher NL Discovery of the biosynthetic machinery for stravidins, biotin antimetabolites. ACS Chem. Biol. 15, 1134–1140 (2019).
48. Wang F-Q et al. Molecular cloning and functional identification of a novel phenylacetyl-CoA ligase gene from *Penicillium chrysogenum*. Biochem. Biophys. Res. Comm. 360, 453–458 (2007). [PubMed: 17612506]

49. Albright JC et al. Large-scale metabolomics reveals a complex response of *Aspergillus nidulans* to epigenetic perturbation. ACS Chem., Biol. 10, 1535–1541 (2015). [PubMed: 25815712]
50. Knowles SL, Raja HA, Roberts CD & Oberlies NH Fungal–fungal co-culture: a primer for generating chemical diversity. Nat. Prod. Rep. 39, 1557–1573 (2022). [PubMed: 35137758]
51. Nickles G, Ludwikoski I, Bok JW & Keller NP Comprehensive guide to extracting and expressing fungal secondary metabolites with Aspergillus fumigatus as a case study. Curr. Protoc. 1, e321 (2021). [PubMed: 34958718]
52. Bankevich A et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477 (2012). [PubMed: 22506599]
53. Stanke M et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34, W435–W439 (2006). [PubMed: 16845043]
54. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW GenBank. Nucleic Acids Res. 44, D67–D72 (2016). [PubMed: 26590407]
55. Nordberg H et al. The genome portal of the department of energy joint genome institute: 2014 updates. Nucleic Acids Res. 42, D26–D31 (2014). [PubMed: 24225321]
56. Potter SC et al. HMMER web server: 2018 update. Nucleic Acids Res. 46, W200–W204 (2018). [PubMed: 29905871]
57. Chambers MC et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. 30, 918–920 (2012). [PubMed: 23051804]
58. Pluskal T, Castillo S, Villar-Briones A & Orešič M MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf. 11, 395 (2010).
59. Du X, Smirnov A, Pluskal T, Jia W & Sumner S in Computational Methods and Data Analysis for Metabolomics (ed. Li S.) 25–48 (Springer, 2020).
60. Bok JW et al. Fungal artificial chromosomes for mining of the fungal secondary metabolome. BMC Genomics 16, 343 (2015). [PubMed: 25925221]

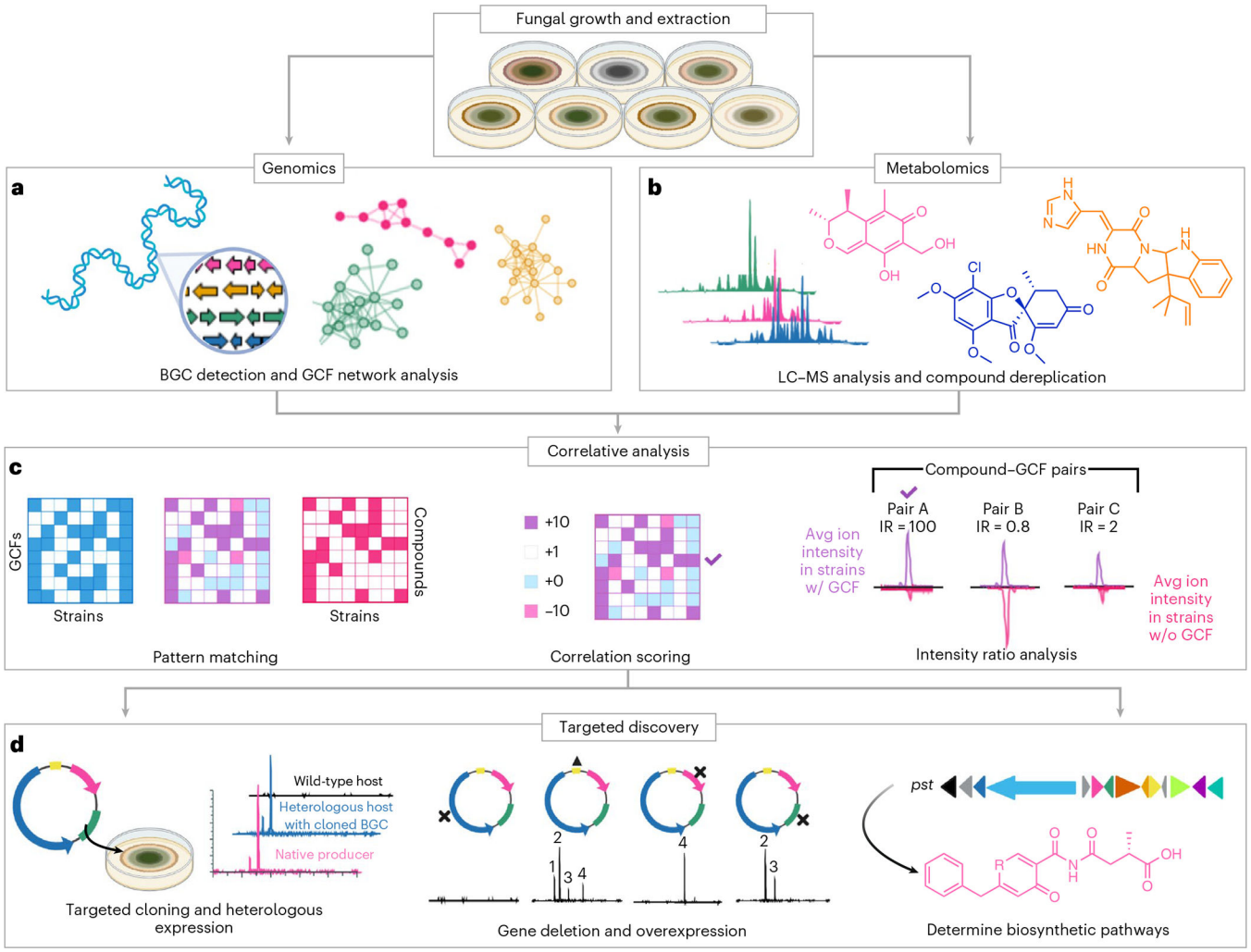

Fig. 1 |. Workflow for the metabologenomics approach for natural products discovery in fungi. **a**, Using interpreted sequences of 110 assembled fungal genomes, we grouped BGCs into GCFs. **b**, In tandem, LC–MS/MS profiles (that is, both MS^1 and MS^2 datasets) were collected from extracts from all 110 strains, each grown on three conditions (oats, rice and Cheerios). **c**, Correlative analyses were completed using three scoring methods (pattern matching, correlation scoring and intensity ratio analysis). **d**, Gene cluster-metabolite linkages identified through correlative analysis were then confirmed through targeted biosynthetic studies. All panels were created with BioRender.com. Avg, average; w/, with; w/o, without.

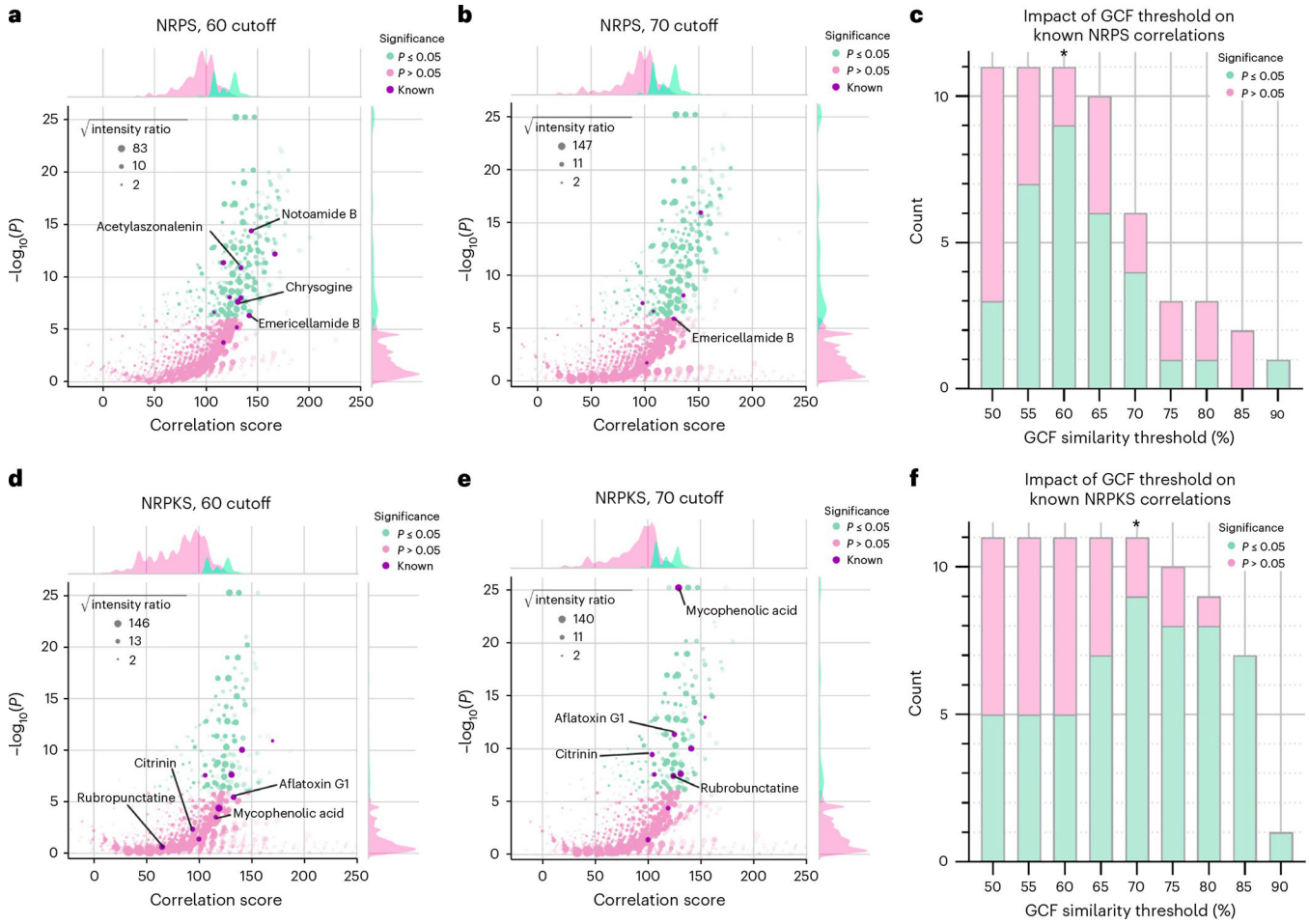

Fig. 2 |. Comparison of GCF-natural product score distributions using different GCF grouping parameters.

For **a**, **b**, **d** and **e**, each point represents a unique metabolite–GCF pair, and the location on the plot reveals the strength of the associated weighted correlation scores (*x* axis), $-\log_{10}(P)$ values (*y* axis) and intensity ratios (point size). *P* values are the result of the chi-squared test with Bonferroni correction. Significant correlations ($P \leq 0.05$ after multiple-hypothesis correction) are colored green, nonsignificant correlations are colored pink and known metabolite–GCF pairs are colored purple. **a**, Distributions at the 60% similarity threshold for NRPS-containing GCFs. **b**, Distributions at the 70% similarity threshold for NRPS-containing GCFs. **c**, Total number of significant (green) and nonsignificant (pink) correlations for knowns with validated BGCs belonging to NRPS-containing GCFs calculated with the nine different GCF similarity thresholds using the pattern-matching approach. The 60% similarity threshold maximizes the number and significance of validated matches. **d**, Distributions for NRPKS-containing GCFs at the 60% cutoff. **e**, The 70% cutoff for NRPKS-containing GCFs. **f**, Total number of correlations to validated knowns of the NRPKS biosynthetic type calculated with different GCF similarity thresholds using the pattern-matching approach, illustrating that the 70% threshold is optimal for this biosynthetic type.

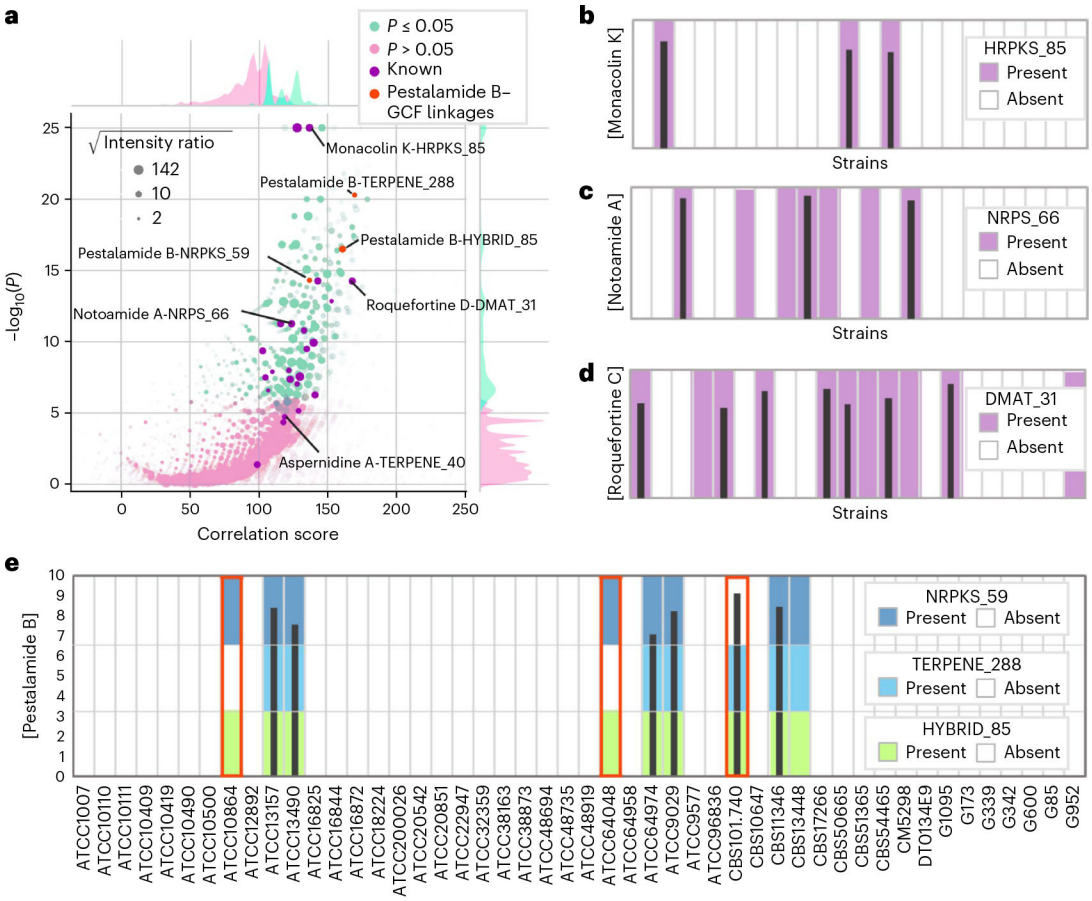

Fig. 3 |. Compiled metabolite–GCF correlations using optimized GCF network.

a, Each point represents a unique metabolite–GCF pair and its location corresponds to the strength of the association. Weighted correlation scores are on the *x* axis and $-\log_{10}(P$ values) calculated using the pattern-matching approach on the *y* axis. *P* values are the result of a chi-squared test with a Bonferroni correction. Point size corresponds to the square root of the intensity ratio. Significant correlations ($P \leq 0.05$ after multiple-hypothesis correction) and nonsignificant correlations are colored in green and pink, respectively. Correlations for validated metabolite–GCF pairs are in purple (with selected known linkages labeled with metabolite–GCF names) and correlations between the pestalamide B and three GCFs of interest are colored in orange. **b**–**e**, Cooccurrence plots for monacolin K to HRPKS_85 (3 of 3 strains with the BGC produce monacolin K) (**b**), notoamide A to NRPS_66 (3 of 7 BGC-containing strains produce notoamide A) (**c**), roquefortine C to its GCF; 7 of 11 strains with DMAT_31 BGCs produce roquefortine C (**d**) and pestalamide B-GCF linkages (**e**). Strains are on the *x* axis and log-transformed peak heights on the *y* axis. Presence/absence patterns for three candidate GCF linkages are highlighted along the gridlines. NRPKS_59 (dark blue) is missing in the top-producing strain (orange box, far right). TERPENE_288 (light blue) was the top-ranked linkage (due to smaller GCF size than other high-scoring GCFs, orange boxes, left and middle), but ruled out using MS^2 data. HYBRID_85 (neon green), the second-ranked linkage, was targeted for follow up studies (for **b**–**e**, only a subset

of strains are shown for clarity). HRPKS, highly reducing polyketide synthase; DMAT, dimethylallyl tryptophan synthase.

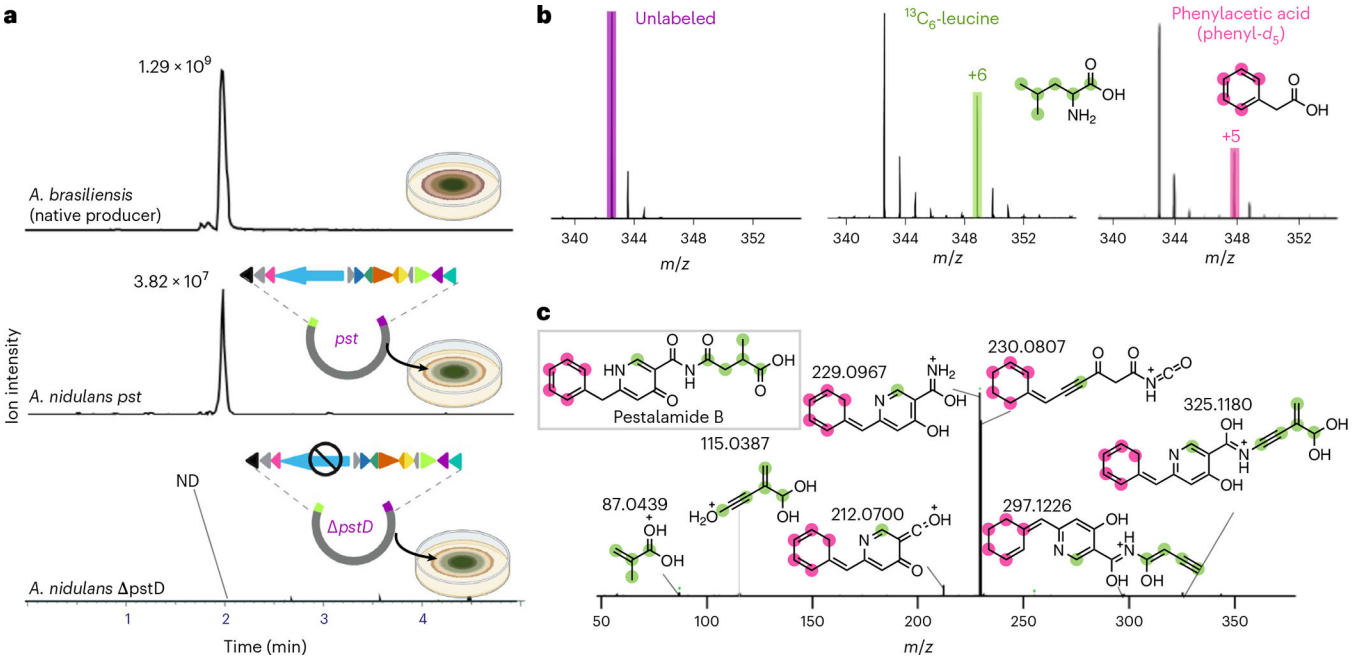

Fig. 4 |. Heterologous expression of pestalamide B in *Aspergillus nidulans*.

a, The expression strain *A. nidulans-pst* produces pestalamide B at a high MS titer, but lower than the native producer *A. brasiliensis* CBS 101.740; the expression strain lacking the backbone synthase (*A. nidulans-ΔpstD*) does not produce this metabolite. **b**, MS^1 spectral shifts of pestalamide B following feeding with [$^{13}C_6$]-leucine (green), phenylacetic acid (phenyl-d_5) (pink) or without heavy isotopes (purple). **c**, MS^2 spectral shifts of pestalamide B fragments following feeding with [$^{13}C_6$]-leucine and phenylacetic acid (phenyl-d_5). Fragment ions are annotated with their unlabeled *m*/*z* values, but putative structures have been color-coded based on the proposed incorporation of isotopically labeled precursors. Notably, while fragment ions show either +0 or +5 Da shifts on phenylacetic acid feeding (pink circles on structures), fragment ions often show +1, +4, +5 or +6 Da shifts (green circles on structures) following leucine feeding, consistent with substantial rearrangement of leucine during pestalamide biosynthesis. All panels were created with BioRender.com.

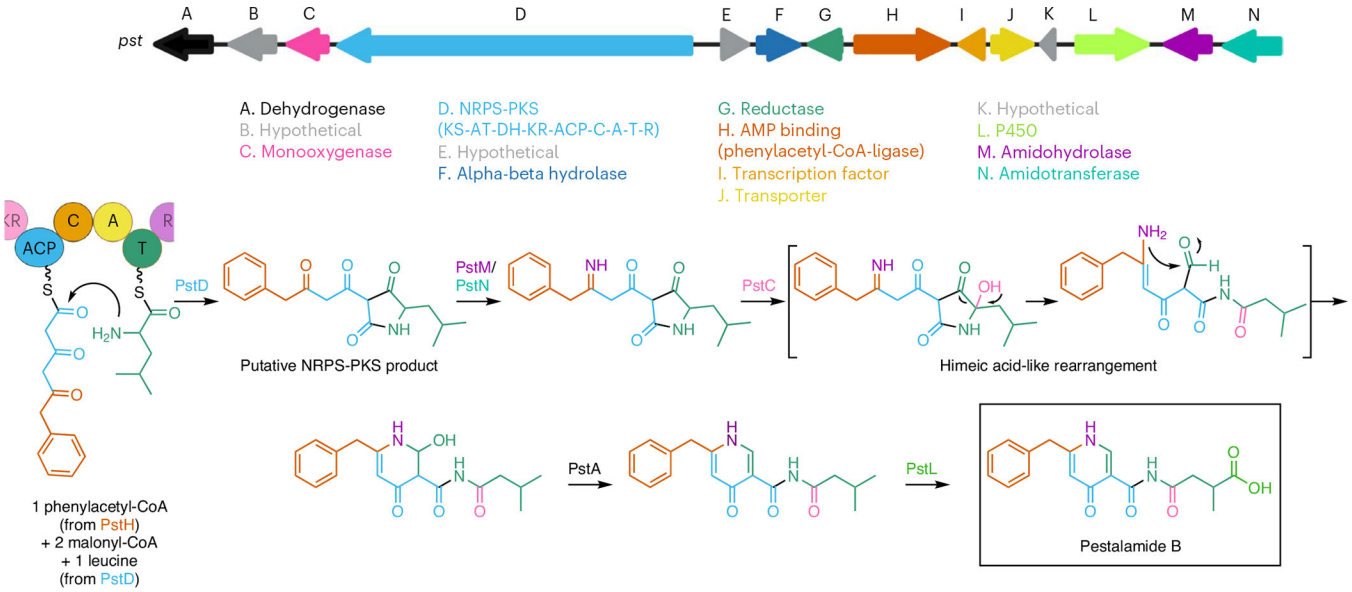

Fig. 5 |. Proposed biosynthesis of pestalamide B.
All panels were created with BioRender.com.

Table 1 |

Comparison of correlation-based scoring methods for metabologenomics

	Approach 1	**Approach 2**	**Approach 3**
Name	Pattern matching	Correlation scoring	Intensity ratio analysis
Weighted?	No	Yes	Yes
GCF data matrix input	Binary (presence or absence)	Binary (presence or absence)	Binary (presence or absence)
Metabolomics data matrix input	Binary (presence or absence)	Binary (presence or absence)	Quantitative (peak height)
Calculation	Pearson's chi-squared test	GCF present, ion present +10; GCF absent, ion present −10; GCF present, ion absent 0; GCF absent, ion absent +1	$\frac{\text{avg peak height in strains w GCF}}{\text{avg peak height in strains w/o GCF}}$
Score output	*P* value	Correlation score	Intensity ratio score