RESEARCH

e-Xtra*

Genomics- and Machine Learning-Accelerated Discovery of Biocontrol Bacteria

Matthew B. Biggs,[†] Kelly Craig, Esther Gachango, David Ingham, and Mathias Twizeyimana[†]

AgBiome, Inc., P.O. Box 14069, Research Triangle Park, NC 27709

Accepted for publication 2 May 2021.

ABSTRACT

Microorganisms with antimicrobial activity have been used to successfully control various plant pathogens. The discovery of organisms with protective activity depends on empirical screenings to assess microbial activity against pathogens of interest. Machine learning can accelerate the discovery process by making screening and, thus, discovery more efficient. We developed a novel machine-learning workflow to identify genomic features associated with fungicidal activity of bacteria, and leveraged those genomic features to discover additional bacteria with the desired activity. We applied our workflow to discover solutions to two problematic fungal diseases: sorghum anthracnose and black sigatoka of banana. These diseases are problematic worldwide, with a particularly devastating impact on small-holder farmers in Sub-Saharan Africa. We screened a total of 1,227 bacterial isolates for antifungal activity against these pathogens using detached-leaf methods and identified 72 taxonomically

diverse isolates with robust activity against one or both of these pathogens. We identified biosynthetic gene clusters associated with activity against each pathogen. Machine learning improved the discovery rate of our screen by threefold, and led to the discovery of a taxonomic group in which fungicidal activity has never been reported. This work highlights the wealth of biocontrol mechanisms available in the microbial world for management of fungal pathogens, generates opportunities for future characterization of novel fungicidal mechanisms, and provides a set of genomic features and models for discovering additional bacterial isolates with activity against these two pathogens. Finally, our workflow generalizes to any discovery effort where genomic information is available to guide candidate selection.

Keywords: bacterial-fungal antagonism, biological control, black sigatoka, genomics, machine learning, sorghum anthracnose

Microorganisms such as fungi, bacteria, viruses, protozoa, and nematodes play a pivotal role in the phytobiome by influencing host physiology and development (Mendes et al. 2013). In agriculture, while many members of microbial communities protect crops by competing with pests, or increase crop yields by enhancing nutrient availability (nitrogen-fixing bacteria, mycorrhizal fungi, and plant-growth-promoting bacteria) (Afzal et al. 2019; Parnell et al. 2016), others such as pathogenic fungi, oomycetes, bacteria, and nematodes

are involved in harmful associations (Bell et al. 2019; Leach et al. 2017; Mendes et al. 2013).

Fungi are the most important plant pathogens (Doehlemann et al. 2017). Strategies available to manage fungal pathogens have primarily focused on the use of fungicides (Lamichhane et al. 2016; Strange and Scott 2005) and adopting cultural practices such as the use of disease-resistant plant cultivars and crop rotation (Barzman et al. 2015; Collinge et al. 2019). Despite the efficacy of fungicides against many plant diseases, their intensive use, especially single-mode-of-action fungicides, can result in pathogen-resistant populations, hindering management in consecutive crop seasons (Cook 2001; Hollomon 2015, 2016). We need to find alternatives to existing control strategies. Identifying biocontrol microbes—which generally act through multiple modes of action (Legein et al. 2020)—is an important part of the future management of these diseases.

Bacteria provide a rich resource for the discovery of novel tools to manage plant diseases. Bacteria have coinhabited many environments with fungal competitors for millions of years and many lineages have evolved mechanisms to antagonize and exclude fungal pathogens (Bahram et al. 2018; van Overbeek and Saikonen 2016). However, it remains difficult and expensive to search among thousands of bacterial isolates in order to identify specific ones which—by virtue of their interactions with the plant or other microbes in the

[†]Corresponding authors: M. B. Biggs; mattbiggs@gmail.com, and M. Twizeyimana; mtwizeyimana@agbiome.com

Funding: This work was supported, in whole or in part, by the Bill and Melinda Gates Foundation (OPP1151824). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

*The e-Xtra logo stands for “electronic extra” and indicates that supplementary materials are published online.

The author(s) declare no conflict of interest.

Copyright © 2021 The Author(s). This is an open access article distributed under the CC BY 4.0 International license.

phytobiome—are able to prevent a disease caused by a specific fungal pathogen. In this work, we developed an approach to more efficiently discover fungicidal bacteria, by means of genomics and machine learning paired with a large collection of environmental bacterial isolates.

Machine learning is used heavily in drug discovery screening, often with the goal of increasing screening efficiency by predicting a molecule's activity based on its chemical structure (Ekins et al. 2019; Xie et al. 2021; Zhu et al. 2020; Zorn et al. 2021). Machine learning is applied in microbiology to predict bacterial phenotypes, where the prediction of antimicrobial resistance is the most developed application area (Anahtar et al. 2021). In many cases, the primary objective of machine learning on microbial genomes is to associate genomic features with an observed phenotype such as the ability to colonize plants (Levy et al. 2018; Martínez-García et al. 2016) or to infect a host (Wheeler et al. 2018). In this work, we pioneer the use of biosynthetic gene clusters (BGCs) as the predictive genomic features, and we apply feature selection and machine learning into a novel workflow for the general-purpose discovery of microbial products.

Specifically, in this work, we applied our workflow in the context of discovering bacterial isolates and genomic features that antagonize the fungal pathogens responsible for sorghum anthracnose (SA) and black sigatoka (BS) disease. Sorghum (*Sorghum bicolor* (L.) Moench) is one of the top five cereal crops in the world agricultural economy, and second only to maize as a staple for food-insecure people in Sub-Saharan Africa (Proietti et al. 2015). The major biotic constraint to sorghum cultivation is the fungal disease SA, caused by *Colletotrichum sublineolum* P. Henn. Kabat et Bub. Yield losses ranging from 41 to 67% have been reported in Africa (Erpelding and Prom 2006). Banana and plantain crops (*Musa* spp.) are also of critical importance to food security and income generation for over 100 million people in Sub-Saharan Africa (Nkuba et al. 2015). BS, also known as black leaf streak, caused by the fungal species *Mycosphaerella fijiensis*, is regarded as the most economically important leaf disease that affects banana and plantain worldwide (African Agricultural Technology Foundation (AATF) 2003; Twizeyimana et al. 2007).

In this work, we sought to answer two questions. First, are there microbial genomic features which are statistically associated with an observable microbial interaction (in this case, fungal antagonism)? And, if so, can such genomic features be used to identify new microbial isolates with the same interaction at a significantly higher rate than a random set of microbial isolates? We discovered that our novel workflow successfully identifies BGCs associated with pathogen control, and that machine learning improves the discovery rate of our screening effort by several fold. The positive answers to these questions naturally lead to generalization in other application areas. The genomic underpinnings of interactions among any community of microbes can be identified and leveraged for further discovery.

MATERIALS AND METHODS

Microbial isolation and genome sequencing. The AgBiome isolate collection served as a resource for this work and, at the time, consisted of approximately 70,000 isolates. The AgBiome isolate collection was assembled by collecting samples in the United States and Africa from 2013 to 2019 (Fig. 1A). Samples were collected from various soils, leaves, roots, insects, plants, amphibians, tubers, grain bin dust, and excrement. After isolation, isolates were frozen in glycerol stocks at -80°C (see Supplementary Materials and Methods for details).

Isolate growth and handling. Bacterial isolates were retrieved from -80°C and cultured for 2 days in modified nutrient sporulation medium at 28°C (225 rpm). Modified nutrient sporulation medium consisted of NaCl at 5 g/liter (Thermo Fisher Scientific), tryptone at

10 g/liter (Thermo Fisher Scientific), nutrient broth at 8 g/liter (BD Biosciences), 0.14 mM CaCl_2 (Sigma-Aldrich), 0.2 mM $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich), and 0.01 mM $\text{MnCl}_2 \cdot 4\text{H}_2\text{O}$ (Sigma-Aldrich). Bacterial cells were collected and resuspended in 1 mM MgCl_2 solution before application to leaf pieces.

Isolate selection for screening. Isolates were selected from 27 states in the United States and five districts in Uganda, and were derived from 11 sample types (see Supplementary Materials and Methods). We balanced taxonomic breadth and depth by screening more deeply among taxa with well-known fungicidal activity (e.g., bacilli and pseudomonads) while also ensuring that we sampled from the broader diversity of the AgBiome collection. We refer to this first pool of semirandomly selected bacterial isolates as the “diversity screen”. In addition, we included isolates that showed activity in other screens against different fungal pathogens and refer to this pool as “spectrum” screening. Finally, in several cases when an isolate displayed activity against a pathogen, we followed up by screening additional isolates from the collection with high degrees of genomic similarity to that first active isolate. We refer to this pool of isolates as “genome similarity” screening.

SA and BS screens. Sorghum cultivar 12-GS9016-KS585 seed (Chromatin Inc.) were grown in the greenhouse for a steady supply of disease-free leaf tissue. Fully expanded sorghum leaves from 4- to 6-week-old plants were excised and cut into equal pieces, 2.5 cm wide. We obtained *C. sublineolum* (isolated from sorghum in Texas) from Dr. Thomas Isakeit's laboratory at Texas A&M University.

We used the susceptible *Musa* cultivar Grand Nain (Green Earth, Melbourne, FL, U.S.A.). Plants were maintained in the greenhouse for a constant supply of disease-free leaves. An *M. fijiensis* culture (ITC0489) obtained from the International Institute of Tropical Agriculture, Ibadan, Nigeria was used for inoculation.

The bacterial isolates were applied to the leaf pieces by spraying 120 μl of the bacterial culture suspended in buffer (1×10^8 CFU/ml). The treated leaf pieces were then plated on agar and incubated at room temperature in the dark (Fig. 1B). At 24 h posttreatment, the leaf pieces were inoculated with a droplet of pathogen spore suspension. The plates were then incubated in a growth chamber (Percival Scientific, Inc.).

In the case of SA, 7 days postinoculation, the leaf pieces were assessed for anthracnose severity on a scale of 0 to 4 according to Prom et al. (2016), with few modifications (Fig. 1B and C). In the case of BS, leaf pieces were assessed 3 to 4 weeks after inoculation (Stover 1986). To identify isolates with robust, reproducible activity, we performed a “confirmation screen” on all active isolates from the primary screen. Isolates went into confirmation screens a few weeks after being identified in the primary screen, using a freshly grown bacterial culture. The confirmation assay protocol was the same as the primary.

We attribute variation between the primary and confirmation assays to subtle changes in growth or assay conditions that affect bacterial phenotypic state (Fedele et al. 2020; Guetsky et al. 2001; Lee et al. 2020; Tasaki et al. 2017) and to inherent variability in disease progression. Therefore, isolates that did not “confirm” may still have biocontrol potential but we did not control the unknown environmental parameters to ensure reliable biocontrol activity. Isolates that controlled disease progression in both primary and confirmation assays are considered “reproducibly” active. For additional details, see Supplementary Materials and Methods.

Bacterial genomics. Bacterial isolates were grown in liquid culture and spun down, and DNA was isolated from the cell pellets. Libraries were prepared from the isolate DNA and sequenced on an Illumina HiSeq sequencing platform. Trimmed, paired-end reads were assembled using CLC Assembly Cell (for greater detail, see Supplementary Materials and Methods).

Bacterial genomes were assigned taxonomic identifiers using the Genome Taxonomy Database (GTDB) Toolkit, version 1.0.2 (Chaumeil et al. 2019; Parks et al. 2018, 2020). By way of explanation, GTDB's taxonomic system reconciles historical classifications with modern whole-genome classifications. GTDB names often indicate how the new GTDB taxonomic identifier relates to the historical one. For example, the GTDB genus *Bacillus_A* resulted from partitioning the historically overly broad *Bacillus* genus.

BGCs were annotated using antiSMASH, version 5 (Blin et al. 2019), with the following options enabled: taxon “bacteria”, Known-ClusterBlast against MIBiG (Medema et al. 2015) and General ClusterBlast, Active Site Finder, “pfam2go” mapping, and Prodigal (Hyatt et al. 2010) as the gene finding tool. Homology to well-characterized BGCs was determined based on Known Cluster Blast results against the MIBiG reference database. AntiSMASH produces multiple BGCs and BGC fragments per genome assembly. We used BiG-SCAPE with default settings (Navarro-Muñoz et al. 2020) to cluster BGCs into families (the default alignment is “glocal” and the default cutoff for clustering into a family is a distance metric of 0.3).

Enrichment analysis and machine learning. We calculated BGC family enrichment using a hypergeometric distribution. Isolates were considered “active” if they controlled disease progression by 70% or greater. Specifically, we estimated the cumulative probability

that given M isolates screened, with n of the total having a BGC family (e.g., are derived from soil), and N of the total displaying activity, that we would observe o or more active isolates with that BGC family by chance alone. Smaller probabilities indicate “enriched” BGC families that are unlikely to be so highly represented given chance alone, and that potentially have a meaningful biological role. We leveraged the hypergeometric module in SciPy to calculate enrichment P values (Virtanen et al. 2020).

We used SciKit-Learn to split training and test sets, train random forest (RF) models and generate predictions, perform cross validation, estimate permutation feature importance, and evaluate model performance (Pedregosa et al. 2011). Permutation feature importance (PFI) is a measure of how model performance is affected by randomly shuffling feature columns. PFI is a more robust measure of feature importance than RF feature importance (Altmann et al. 2010). Before implementing the RF models, we tested several configurations of RF hyperparameters, and selected ensembles of five trees with a maximum depth of 32 nodes to maximize recall in a 50-fold cross-validation analysis. Oversampling was performed by replicating all of the active observations in the training data, thereby increasing the ratio of active to inactive observations. Precision is the fraction of predicted positives that are true, and is defined as $tp/(tp + fp)$, where tp = true positive count and fp = false positive count. Recall is the

Fig. 1. Isolation and assay workflows. **A**, High-level overview of the bacterial isolation, storage, and genome sequencing pipeline. **B**, High-level overview of the leaf disk screens, comparable in principle between sorghum anthracnose (SA) and black sigatoka (BS) disease screens. Leaf disks were embedded in agar, and treated with resuspended bacterial cells. After 24 h, the leaf disks were treated with fungal spores and incubated for multiple weeks. Scores were assigned by comparing leaf disk disease progression with inoculated and noninoculated controls. Disease control >70% was considered “active”. **C**, Examples of healthy (left) and diseased (right) sorghum leaf pieces. **D**, Examples of healthy (left) and diseased (right) banana leaf pieces.

fraction of all positives that are predicted to be so, and is defined as $tp/(tp + fn)$, where fn = false negative count. F1 score is a composite of precision and recall, and is defined as $2 \times [(precision \times recall)/(precision + recall)]$. Precision, recall, and F1 scores were calculated using the SciKit-Learn functions.

We used PyTorch (Paszke et al. 2019) paired with Skorch (<https://github.com/skorch-dev/skorch>) to train a deep neural network (NN) and generate predictions. We tested several configurations and selected a three-layer design with 10 fully connected nodes and reLu activation in the first layer, 30% dropout between the first and second layers during training, 10 fully connected nodes and reLu activation in the second layer, and a single sigmoid output node to predict the probability of activity. We used a learning rate of 0.02, and trained each model for 10 epochs using the Adam optimizer and the binary cross entropy loss function.

To select additional isolates for screening, we selected representative “query” BGCs for each enriched or important BGC family and clustered them with batches of 200 BGCs from the remainder of our collection using BiG-SCAPE with the same parameters. The BGCs that clustered with a query BGC were assigned to the same family as the query.

We developed four computational approaches to prioritize bacterial isolates based on genomic BGC families: an enrichment-based (ER) prioritization method, two RF models (standard RF and trained on oversampled data [RFOS]), and a deep NN. We used only the diversity screening data to train the models because this represents the null scenario. For the ER method, we selected isolates that contained at least one member from the 12 enriched BGC families (excluding the BGC families identified by PFI). The RF and RFOS models were trained using all of the significant BGC families. For the RFOS model, we selected an oversampling factor of 13 based on cross-validation results (Supplementary Fig. S5). During 100-fold cross-validation with a 75/25% training/test split of the data, we repeated the process of feature selection described above (i.e., enrichment and PFI) using only the concurrent subset of training data. After feature selection with the RF, the same features were used to train the NN. As a null model, random isolates were predicted to be active at the same discovery rate as the diversity screen. All models were trained to predict activity in the primary screen.

The data and code developed for this work are publicly available on GitHub (https://github.com/mbi2gs/fungicidal_bact_genomics).

In vitro validation. Using the ER and RFOS methods, we prioritized as-yet-unscreened bacterial isolates from the culture collection for validation in SA assay. We required that half of the predicted active isolates come from non*Bacillus_A* genera. The prioritized isolates were evaluated using the same detached-leaf protocols as described above. As a post hoc analysis (not to select the validation set), we applied the NN model to predict activity. As a null model benchmark, for both the SA primary and confirmation screens we simulated 1,000 random activity predictions using the baseline discovery rates (all code available in the project github repository).

RESULTS

We initially screened 1,051 “diversity” isolates against SA, and 559 of those same isolates against BS. There were no statistically significant associations between isolate activity and isolation sample type or location (see Supplementary Results and Supplementary Table S1). Viewing the results from a taxonomic perspective, we discovered 156 isolates that displayed primary activity against SA, BS, or both, representing five phyla, 27 genera, and 37 species (based on GTDB labels) (Parks et al. 2020). We found 72 reproducibly (active again in the confirmation screen) active isolates against SA or BS in five phyla, 14 genera, and 22 species (Fig. 2). Isolates that

reproducibly controlled both SA and BS came from two phyla; specifically, *Proteobacteria* and *Firmicutes*. The active isolate from the phylum *Firmicutes_I* only controlled SA, the active isolate from the phylum *Bacteroidota* only controlled BS, and the active isolates from *Actinobacteriota* only controlled one disease or the other (Fig. 2B). Of the 95 genera screened against SA, 19 were active in the primary screen and 9 showed reproducible, robust control of SA in the confirmation. Of the 44 genera screened against BS, 12 were active in the primary screen and 8 showed reproducible control of BS. We define “discovery rate” as the percentage of isolates that control disease during an assay. Viewing the results from a discovery rate perspective, of the 546 isolates from the diversity screen against SA, 28 controlled SA in the primary screen (a discovery rate of 5.1%) and 8 repeated the control in the confirmation screen (a discovery rate of 1.5%). Of the 145 isolates from the diversity screen against BS, 13 controlled the pathogen in the primary screen (a discovery rate of 9.0%) and 2 repeated the control in the confirmation screen (a discovery rate of 1.4%).

To identify biosynthetic pathways associated with control of SA and BS, we clustered the annotated BGCs and fragments into 2,770 families using BiG-SCAPE (see Materials and Methods). We performed an enrichment analysis to associate isolate activity in the SA primary screen with the presence or absence of each BGC family (Table 1). Additionally, we built an RF classifier to predict isolate activity in each primary screen and calculated the PFI of each BGC family (see Materials and Methods).

Among the BGC families associated with SA control, four were selected by both the PFI and enrichment metrics. In all, 15 of the 28 selected by PFI were homologous to well-characterized BGCs, many of which are known to be fungicidal or generally involved in microbial competition (Sass et al. 2018). Many enriched BGC families did not share significant homology to well-characterized BGCs, and are described as bacteriocins, nonribosomal peptide synthetases (NRPSs), lanthipeptides, and polyketide synthases (PKSs). We found that many of the BGC families associated with activity against SA were cooccurring throughout the isolates screened (Supplementary Fig. S3). For example, we observed clear clusters of BGC families with strong associations with *Bacillus_A*. Among the BGC families associated with controlling BS (Supplementary Table S2), 11 of the 19 were homologous to well-characterized BGCs, including bacilysin. In total, 8 of the 19 were not significantly homologous to well-characterized BGCs, and are categorized as β -lactones, linear azole-containing peptides, bacteriocins, and NRPSs. We found that many of these BGC families associated with activity against BS were cooccurring throughout the isolates screened (Supplementary Fig. S4). For example, we observed clusters of BGC families strongly affiliated with the species *Pseudomonas_E protegens*. The BGC families without significant homology to well-described BGCs represent potentially novel biocontrol pathways and molecules with relevant activity against SA and BS.

To demonstrate the predictive value of the activity-associated BGC families, we developed methods to prioritize bacterial isolates for screening based on genomic BGC family content. We first compared the ER, RF, RFOS, and NN methods in silico using 100-fold cross validation (Fig. 3A). We found that all four prioritization tools performed better than the null model in terms of the F1 statistic (P value of 2×10^{-7} by Kruskal-Wallis test, followed by one-sided Mann-Whitney U tests with P values $< 3 \times 10^{-4}$) (Fig. 3B). The ER method was the poorest performer of the four methods tested, while the RFOS and the deep NN were the best performers. The RFOS average F1 score was significantly better than ER (P value < 0.05 by one-sided Mann-Whitney U test). The NN method displayed the best recall, increasing the average recall by 6.5-fold compared with the null.

Fig. 2. Scope of screening effort, and diverse isolates with confirmed fungicidal activity. **A**, Alluvial diagram illustrates the source country, sample type (the less abundant types were combined into “Other”), phylum membership, and fungicidal activity of the 1,227 isolates screened in this work. Band colors indicate phylum membership (see legend in panel B). **B**, Cladogram in the center highlights the diversity of the biocontrol isolates discovered in this work (five phyla represented by the center branches, 14 genera, and 22 species). The cladogram was constructed using whole-genome-based taxonomic IDs (see Materials and Methods). Multiple isolates displayed activity against both diseases (black leaf nodes). The outer ring shows the isolate prioritization strategies that led to each discovery, including predictive models. The “false-negative” isolate in purple was predicted to be inactive by all modeling approaches and yet was found to be reproducibly active against sorghum anthracnose.

To determine the extent to which the best-performing cross-validation iterations were due to taxonomic similarity between the training set and the test set (assuming that more related isolates between the two would improve observed model performance), we compared the Jaccard similarity of the set of genera in the training and test data to the model performance and found no significant correlation

(correlation coefficient of 0.097 and a *P* value of 0.34 by Spearman's Rank Correlation; see Supplementary Figure S6). Although this result would suggest that taxonomic overlap between training and test set does not affect model performance (which is unlikely), we also repeated the cross-validation evaluation. This time, instead of random training/test splits, we kept out a single genus at each iteration as the

TABLE 1
Biosynthetic gene cluster (BGC) families associated with sorghum anthracnose (SA) control

BGC family ID ^a	99th Percentile importance	Top enrichment scores	Predicted product class ^b	Homology to known BGC ^b	Taxonomic distribution ^c
ripps2539	True	True	Bacteriocin	None	Genus: <i>Bacillus_A</i>
nrps2550	True	True	NRPS	Elsinochrome A	Genus: <i>Bacillus_A</i>
others2545	True	True	β-Lactone	Fengycin	Genus: <i>Bacillus_A</i>
pksnrp3708	True	True	NRPS; T1PKS	Zwittermycin A	Genus: <i>Bacillus_A</i>
terpene2537	True	False	Terpene	Molybdenum cofactor	Genus: <i>Bacillus_A</i>
nrps2808	True	False	NRPS	Thiazostatin	Genus: <i>Pseudomonas_E</i>
terpene4046	True	False	Terpene	Molybdenum cofactor	Genus: <i>Bacillus_A</i>
ripps3967	True	False	Bacteriocin	None	Genus: <i>Bacillus_A</i>
nrps9204	True	False	NRPS	None	Genus: <i>Bacillus_A</i>
others2542	True	False	Siderophore	Petrobactin	Genus: <i>Bacillus_A</i>
ripps2548	True	False	Lanthipeptide	None	Genus: <i>Bacillus_A</i>
terpene6941	True	False	Terpene	None	Species: <i>gtdb_novel_strain</i>
nrps4047	True	False	NRPS;T1PKS	Zwittermycin A	Genus: <i>Bacillus_A</i>
ripps9214	True	False	Sactipeptide	Thurincin H	Species: <i>Bacillus_A thuringiensis_J</i>
nrps2540	True	False	NRPS	Zwittermycin A	Species: <i>Bacillus_A thuringiensis_J</i>
ripps2935	True	False	Bacteriocin	None	Genus: <i>Bacillus_A</i>
nrps8318	True	False	NRPS	Polyoxypeptin	Genus: <i>Bacillus_A</i>
others3835	True	False	Lanthipeptide	Zwittermycin A	Species: <i>Bacillus_A thuringiensis_J</i>
nrps2971	True	False	NRPS	None	Species: <i>Pseudomonas_E sp001547895</i>
nrps10661	True	False	NRPS	None	Species: <i>Bacillus subtilis</i>
pksother8188	True	False	T3PKS	None	Genus: <i>Bacillus</i>
nrps11511	True	False	NRPS	Surfactin	Genus: <i>Bacillus</i>
ripps2937	True	False	Bacteriocin	None	Species: <i>Bacillus_A thuringiensis_J</i>
nrps1086	True	False	NRPS	Pyoverdine	Species: <i>Pseudomonas_E chlororaphis</i>
terpene11261	True	False	Terpene	None	Genus: <i>Bacillus</i>
others11626	True	False	Lanthipeptide	Zwittermycin A	Genus: <i>Bacillus_A</i>
ripps5084	True	False	Lanthipeptide	None	Order: <i>Bacillales</i>
ripps7030	True	False	Bacteriocin	None	Species: <i>Brevibacillus reuszeri_A</i>
nrps4051	False	True	NRPS-like	None	Genus: <i>Bacillus_A</i>
ripps3390	False	True	Bacteriocin	None	Species: <i>Bacillus_A toyonensis</i>
nrps4795	False	True	NRPS	Pyoverdine	Species: <i>Pseudomonas_E protegens</i>
ripps4045	False	True	Bacteriocin; LAP	None	Genus: <i>Bacillus_A</i>
nrps2943	False	True	NRPS	Lipopeptide 8D1-1/2	Species: <i>Pseudomonas_E protegens</i>
others2810	False	True	Other	Pyrrrolnitrin	Class: <i>Gammaproteobacteria</i>
pksother4889	False	True	PKS-like	Anikasin	Species: <i>Pseudomonas_E protegens</i>
nrps4050	False	True	NRPS	Bacillibactin	Genus: <i>Bacillus_A</i>

^a Name of the BGC clusters that were most highly associated with SA control. Each BGC family in the table was associated with pathogen control by one or both of the computational methods indicated in the columns 99th Percentile importance and Top enrichment scores.

^b Derived from the consensus antiSMASH and BiG-SCAPE annotations for members of the BGC family. NRPS = nonribosomal peptide synthetase, PKS = polyketide synthase, and LAP = linear azole-containing peptide.

^c Highest taxonomic rank encompassing all members of the BGC family within our data set.

test set. We found that, under these conditions, the models failed to outperform the null model (P value = 0.13 by Kruskal-Wallis test), suggesting that the BGC families used to train the models tend to be predictive at the genus level or lower (Fig. 3C).

In parallel to the *in silico* validation of the predictive value of the enriched BGC families, we also completed an *in vivo* validation. In contrast to the *in silico* validation, where we could simulate hundreds of experiments, we performed a single *in vivo* validation. We applied two computational models to predict control of SA from among the larger AgBiome bacterial culture collection. We selected BGC families using ER and RF feature importance, based on an earlier version of the SA screen data set (the BGC features used in this *in vivo* validation exercise are listed in Supplementary Table S3 and representative sequences are provided in the project github repository). We identified 17,163 isolates in the AgBiome culture collection that contained at least one representative of the enriched BGC families. From this large set, we selected 176 isolates using the ER and RFOS methods and applied them in the SA assay. In addition, we applied the NN method post hoc to compare its performance with the ER and RFOS methods. We observed that the ER method was far more likely to predict that an isolate would be active, and uniquely prioritized 36 isolates, sharing 66 with the RFOS model and 81 with the NN model. The RFOS model prioritized 20 unique isolates, and shared 47 with the NN model. The NN model did not prioritize any unique isolates because it was applied post hoc. In all, 20 of the 176 isolates were

predicted to be inactive by all three models, one of which was found to be active (this false-negative isolate was classified as a *Bacillus altitudinis*) (Fig. 2).

We found that the RFOS primary discovery rate (precision) was threefold greater than the historical discovery rate of the diversity screen (P value of 3.7×10^{-6} by one-sided two-proportions Z test) (Fig. 4). In other words, using the RFOS model, we discovered threefold more active isolates for the effort than we would have by screening a random, diverse set of isolates from the collection. The confirmation discovery rate was even higher: 5.4-fold greater than the diversity screen (P value of 6.2×10^{-6} by one-sided two-proportions Z test) (Fig. 4). Additionally, the RFOS model achieved 72% recall, correctly predicting 23 of the 32 total active isolates (P value of 0.002, compared with 1,000 draws using a random model based on the baseline diversity screen discovery rate of 5.0%) (Supplementary Fig. S7). Notably, the RFOS method did not discover any novel taxa that were not already represented in the training set. Most of the true positives were classified as members of the genus *Pseudomonas*_E, while most of the true negatives, false negatives, and false positives were classified as members of *Bacillus*_A. Taking into account the SA confirmation screen results (which were not used to train the model), the model precision was not statistically different from the null model but the recall was exceptionally high, where the RFOS model correctly predicted 11 of the 12 isolates (92%) that reproducibly controlled SA (P value < 0.001, compared with 1,000 random draws).

Fig. 3. Machine learning workflow and *in silico* results. **A**, Machine learning workflow, starting from genome assemblies, annotating biosynthetic gene clusters (BGCs) with antiSMASH, clustering BGCs into families, aggregating the presence or absence of each BGC family in each genome, and, finally, selecting features and training models. **B**, Comparing prioritization methods by 100-fold cross validation on the primary screening data. Each iteration of the cross validation split the data randomly into training and test sets, selected features and trained models on the training data, then predicted activity in the test data. ER = enrichment-based prioritization, where any isolate with one or more enriched BGCs was predicted to be active; RF = random forest model; RFOS = RF with oversampling; NN = deep neural network; and Rand = a random, null model based on the discovery rate from the training data. All prioritization methods outperformed the null model. Both RFOS and NN outperformed the ER and RF models. The mean (μ) and median (μ) values for each set of simulations are indicated above each distribution. **C**, Each iteration of this cross validation held out all of the members of a particular genus as the test set. In this case, all methods were indistinguishable from the null (Kruskal-Wallis test), indicating that, generally speaking, the BGC families defined here are only predictive within a genus. The mean (μ) and median (μ) values for each set of simulations are indicated above each distribution.

The NN model did not perform as well as the RFOS model in the context of this post hoc analysis. We found that the NN primary discovery rate was 1.9-fold greater than the diversity screen discovery rate (P value of 0.01 by one-sided two-proportions Z test), and the confirmation discovery rate was 2.5-fold greater than the diversity screen (P value of 0.03 by one-sided two-proportions Z test) (Fig. 4). The precision was not statistically different from the null model but the recall was better than the null (P value of 0.098 compared with 1,000 random draws) (Supplementary Fig. S7). Recall was not significantly better than the null model when predicted from the results of the SA confirmation screen—the NN model correctly predicted 2 of 12 (17%) isolates that reproducibly controlled SA (P value of 0.12, compared with 1,000 random draws). In contrast to the RFOS model, the true and false positives consisted primarily of members of the genera *Bacillus_A*, while most true negatives were classified as *Bacillus* and most false negatives were classified as *Pseudomonas_E*.

Finally, we found that the ER primary discovery rate was 2.4-fold greater than the diversity screen discovery rate (P value of 7.4×10^{-5} by one-sided two-proportions Z test), and the confirmation discovery rate was 2.9-fold greater than the diversity screen (P value of 6.1×10^{-3} by one-sided two-proportions Z test) (Fig. 4). The ER method outperformed the null model in terms of recall (Supplementary Fig. S7). It correctly predicted 19 of the 32 (59.4%) SA-controlling isolates in the primary screen (P value of 7.0×10^{-3} compared with 1,000 random draws) and 3 of the 12 actives against SA in the confirmation screen (P value of 3.1×10^{-3} compared with 1,000 random draws). In taxonomic terms, most of the true and false positives were classified as genus *Bacillus_A*, most of the true negatives as *Bacillus*, and most of the false negatives as *Pseudomonas_E* and *Lysobacter*.

The ER method (uniquely among the three computational methods used) discovered active isolates from a novel genus which was not present in the training data—*Bacillus_I endophyticus*. These *Bacillus_I* isolates were prioritized by the ER method based on the presence of the BGC family “nrps2550”.

DISCUSSION

Previous work has shown that bacterial genomic features can be used to predict phenotypes (Anahtar et al. 2021; Levy et al. 2018;

Van Camp et al. 2020). To our knowledge, we are the first to apply machine learning to increase the efficiency of microbe discovery, specifically within an empirical screening framework. Our workflow is comparable with the application of machine learning to drug discovery screening (Ekins et al. 2019). The steps of our workflow could be easily replaced with multiple other methods. For example, where we used BGCs in previous work, many genomic feature types have been used to predict antimicrobial resistance, including presence or absence of nucleotide k-mers (Nguyen et al. 2019), presence or absence of homologs to preselected genes (Van Camp et al. 2020), and single nucleotide mutations (Moradigaravand et al. 2018). We expect future work to improve on this by the use of new genomic feature types such as promoters (Cassiano and Silva-Rocha 2020), non-BGC-associated gene families (Huerta-Cepas et al. 2019), and RNA aptamers (Tapsin et al. 2018). Where we applied enrichment and PFI, other feature selection methods can be substituted (Cai et al. 2018). Where we applied RF and NN models, many other predictive models exist which may perform better for a specific application (Wang et al. 2020), be more interpretable (Azodi et al. 2020), or achieve better results with fewer training examples (Barbedo 2018; Kurczab et al. 2014; Vabalas et al. 2019). Variations of this workflow can be applied to discover microbes with specific environmental roles, and to generate hypotheses as to the genomic features contributing to those roles. An obvious example of a potential application area would be probiotics discovery (McCoubrey et al. 2021).

Broader and deeper data sets are needed if we wish to predict SA or BS control with more confidence. Although every computational approach we tested performed significantly better than the null model (Fig. 3B), none of the modeling approaches achieved average F1 scores >20% or maximum F1 scores >40%. The relatively low model confidence highlights the complex nature of antifungal interactions. Small changes in BGC enzymes may result in the production of different metabolites, which may or may not be antifungal. Small changes to promoters, signaling molecules, or growth conditions can all have a large impact on the observed antifungal activity against SA or BS (Firm and Jones 2003). Furthermore, it is possible that many active isolates employ unique modes of action which are not shared with any other isolates (Jeukens et al. 2017). All of this simply points to the need for more screening data, and the use of input features capable of capturing more subtle differences between isolates. Regardless, we found that RFOS and NN models performed the best in silico (Fig. 3), and that predictive models improved the efficiency of our screening efforts (Fig. 4).

In addition to the novelty of our workflow, our application area (solutions for fungal diseases of plants, specifically SA and BS) is understudied. There are many fungicides with extensive and vital use in agriculture (Brauer et al. 2019) but the majority of research for solutions to SA and BS has been focused on identifying resistant plant varieties (Cuevas et al. 2018). Our results are a valuable step toward alternative solutions to these problematic diseases.

Many of the newly discovered isolates with confirmed activity against SA and BS are from taxonomic groups that have long been known to exhibit fungicidal properties; specifically, *Lysobacter* (Jochum et al. 2006), *Paenibacillus_B sp001894745* (sometimes labeled in NCBI as *Paenibacillus alvei* or lacking species classification) (Fatouros et al. 2018; Gkikas et al. 2021), and a variety of pseudomonads and bacilli (Supplementary Table S1) (Santoyo et al. 2012). Some taxa discovered in this work had not been reported to control fungal disease before, including MA-N2 sp002009585, a member of the family *Micrococcaceae* (labeled in NCBI as a member of the genus *Arthrobacter* and lacking a species identifier), a novel member of the genus *Glutamicibacter*, and a novel member of the family *Planococcaceae*. The broad diversity of bacterial taxa

Fig. 4. Predictive models improved discovery rate. ER = enrichment-based prioritization, RFOS = random forest with oversampling, and NN = deep neural network. The hit rate in the sorghum anthracnose primary and confirmation screens is shown by the bars, and the fold improvement compared with the diversity screen (a random set of microbes from the collection) is displayed above each bar. RFOS improved the discovery rate by threefold in the primary screen, and by 5.4-fold in the confirmation screen.

capable of controlling SA and BS suggests the widespread impact of antagonistic bacterial–fungal interactions over the course of evolutionary time (Deveau et al. 2018; Lastovetsky et al. 2020). The diversity also suggests that bacteria yet harbor many untapped opportunities for fungal antagonism in both agricultural and medical applications. Indeed, new fungicidal modes of action are reported frequently (Chen et al. 2018; Khalid et al. 2018; Li et al. 2020; Lim et al. 2017; Omoboye et al. 2019).

The BGC families we observed to be enriched and predictive of bacterial anti-SA and anti-BS activity were a mix of homologs to well-studied BGCs and novel, uncharacterized BGCs. Many of the enriched BGC families are highly homologous to the characterized BGCs for known fungicidal compounds such as fengycin (Vanittanakom et al. 1986; Wise et al. 2014), pyrrolnitrin (Nishida et al. 1965; Pawar et al. 2019), zwittermycin A (He et al. 1994; Kevany et al. 2009), bacilysin (Wang et al. 2018), and the siderophore pyoverdine (Sass et al. 2018). This is a strong validation of the feature selection approaches (enrichment and PFI) we used to rank BGC families. This validation is encouraging for future efforts to characterize the novel BGCs we observed. Many of the enriched BGC families did not share significant homology with any well-characterized BGCs and represent potentially novel biocontrol pathways. These novel BGC families are described as bacteriocins, NRPS-like lanthipeptides, and PKSs (Table 1; Supplementary Table S2). It is possible that some of the BGCs we found to be associated with antifungal activity are simply correlated with the taxonomic group that exhibited activity and are unrelated to the mode of action (MOA) (Bradley et al. 2018). A similar challenge is that singleton BGCs representing novel antifungal pathways cannot be detected by the approach used here. The nature of the training data makes it impossible to determine which BGC families actually cause anti-SA or -BS activity (Meinshausen et al. 2016). Therefore, it is necessary to do more in-depth studies to demonstrate causation, and to elucidate potential MOAs associated with BGC families discovered here.

Regardless of whether the BGC families discovered here represent the actual source of antifungal activity or are merely correlated with it, they are predictive of antifungal activity against SA and BS. Interestingly, we observed in silico that it is rare for BGC families to be predictive of activity outside the genus level (Fig. 3C). In some cases, we observed that the predictive models could correctly predict activity among isolates from a genus not represented in the training data but, in silico, these cases were rare and were not common enough to differentiate from random chance. It is unclear whether the genus-level limit to BGC predictive value is reflective of biology (i.e., antifungal MOAs tend to be limited to within a bacterial genus) or whether it is an artifact of the way we defined and ranked BGC families. For example, despite the in silico results, in the validation experiment, the enrichment method did in fact lead us to discover a novel genus. It is also likely that some antifungal BGC families span broader taxonomic groups, knowing that horizontal gene transfer does occur between microbes; even, to some extent, across phylum boundaries (Redondo-Salvo et al. 2020). An example in our training data of a more broadly distributed BGC family is “others2810” (Table 1), which can be found (in our data set) across the class *Gammaproteobacteria*. In the validation experiment, the BGC family “nrps2550”—which was limited to the genus *Bacillus* A in the training set—was the clue that led to the discovery of activity against SA by a member of the genus *Bacillus* I. A future area of research is the relationship between BGC homology (how BGC families are defined) and taxonomic dispersion. Regardless, by predicting activity based on genomic features rather than whole-genome homology or taxonomic relatedness, it becomes possible to anchor predictions on features that have been broadly inherited or horizontally dispersed (Cimermanovic et al. 2014; Kautsar et al. 2021). In this way, genomic features can

act as stepping stones to new taxa possessing similar ecological roles, hopefully paired with improved crop protection properties.

Interestingly, the BGC family “nrps2550” shares homology with the BGC known to produce elsinochrome A. Elsinochrome A is a perylenequinone, a family of molecules which are primarily known as fungal toxins involved in diseases such as citrus scab (Chung 2011), and which produce toxic effects by generating reactive oxygen species (Hu et al. 2019). Related perylenequinones such as hypocrellins display antifungal activities (Song et al. 2021; Xing et al. 2003). Family “nrps2550” is an appealing candidate for further biochemical characterization.

There is a lot left to learn and discover about bacterial interactions with their fungal neighbors in the phytobiome, and computational models will continue to accelerate the discovery effort.

ACKNOWLEDGMENTS

We thank T. Isakeit at Texas A&M University for providing isolates of *C. sublineolum*, A. E. Alakonya (a banana pathologist at the International Institute of Tropical Agriculture, Nigeria) for providing the *M. fijiensis* isolate used in this study, G. Medlock at the University of Virginia for his thoughtful comments and suggestions during the preparation of this manuscript, and the phenomenal team at AgBiome who made this work possible.

LITERATURE CITED

- Afzal, I., Shinwari, Z. K., Sikandar, S., and Shahzad, S. 2019. Plant beneficial endophytic bacteria: Mechanisms, diversity, host range and genetic determinants. *Microbiol. Res.* 221:36-49.
- African Agricultural Technology Foundation (AATF). 2003. Report of Small Group Meeting (SGM) on Improved Production of Bananas and Plantains in Sub-Saharan Africa. Theme: Agricultural Technology Interventions for Increasing the Production of Bananas and Plantains in Sub-Saharan Africa. Hotel Equatoria, Kampala, Uganda. 22 August 2003.
- Altmann, A., Tološi, L., Sander, O., and Lengauer, T. 2010. Permutation importance: A corrected feature importance measure. *Bioinformatics* 26: 1340-1347.
- Anahtar, M. N., Yang, J. H., and Kanjilal, S. 2021. Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research. *J. Clin. Microbiol.* 59:01260-20.
- Azodi, C. B., Tang, J., and Shiu, S.-H. 2020. Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.* 36:442-455.
- Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Pölme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., and Bork, P. 2018. Structure and function of the global topsoil microbiome. *Nature* 560:233-237.
- Barbedo, J. G. A. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agric.* 153:46-53.
- Barzman, M., Bärberi, P., Birch, A. N. E., Boonekamp, P., Dachbrodt-Saaydeh, S., Graf, B., Hommel, B., Jensen, J. E., Kiss, J., Kudsk, P., Lamichhane, J. R., Messéan, A., Moonen, A.-C., Ratnadass, A., Ricci, P., Sarah, J.-L., and Sattin, M. 2015. Eight principles of integrated pest management. *Agron. Sustain. Dev.* 35:1199-1215.
- Bell, T. H., Hockett, K. L., Alcalá-Briseño, R. I., Barbercheck, M., Beattie, G. A., Bruns, M. A., Carlson, J. E., Chung, T., Collins, A., Emmett, B., Esker, P., Garrett, K. A., Glenna, L., Gugino, B. K., Jiménez-Gasco, M. d. M., Kinkel, L., Kovac, J., Kowalski, K. P., Kuldau, G., Leveau, J. H. J., Michalska-Smith, M. J., Myrick, J., Peter, K., Salazar, M. F. V., Shade, A., Stopnisek, N., Tan, X., Welty, A. T., Wickings, K., and Yergeau, E. 2019. Manipulating wild and tamed phytobiomes: Challenges and opportunities. *Phytobiomes J.* 3:3-21.
- Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. 2019. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* 47:W81-W87.

- Bradley, P. H., Nayfach, S., and Pollard, K. S. 2018. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLOS Comput. Biol.* 14:e1006242.
- Brauer, V. S., Rezende, C. P., Pessoni, A. M., De Paula, R. G., Rangappa, K. S., Nayaka, S. C., Gupta, V. K., and Almeida, F. 2019. Antifungal agents in agriculture: Friends and foes of public health. *Biomolecules* 9:521.
- Cai, J., Luo, J., Wang, S., and Yang, S. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing* 300:70-79.
- Cassiano, M. H. A., and Silva-Rocha, R. 2020. Benchmarking bacterial promoter prediction tools: Potentialities and limitations. *mSystems* 5: e00439-20.
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. 2019. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36:1925-1927.
- Chen, Y., Wang, J., Yang, N., Wen, Z., Sun, X., Chai, Y., and Ma, Z. 2018. Wheat microbiome bacteria can reduce virulence of a plant pathogenic fungus by altering histone acetylation. *Nat. Commun.* 9:3429.
- Chung, K.-R. 2011. *Elsinoë fawcettii* and *Elsinoë australis*: The fungal pathogens causing citrus scab. *Mol. Plant Pathol.* 12:123-135.
- Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Lington, R. G., and Fischbach, M. A. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158:412-421.
- Collinge, D. B., Jørgensen, H. J. L., Latz, M. A. C., Manzotti, A., Ntana, F., Rojas, E. C., and Jensen, B. 2019. Searching for novel fungal biological control agents for plant disease control among endophytes. Pages 25-51 in: *Endophytes for a Growing World*. T. R. Hodkinson, F. M. Doohan, M. J. Saunders, and B. R. Murphy, eds. Cambridge University Press, Cambridge, England.
- Cook, R. T. A. 2001. First report in England of changes in the susceptibility of *Puccinia horiana*, the cause of chrysanthemum white rust, to triazole and strobilurin fungicides. *Plant Pathol.* 50:792.
- Cuevas, H. E., Prom, L. K., Cooper, E. A., Knoll, J. E., and Ni, X. 2018. Genome-wide association mapping of anthracnose (*Colletotrichum sublineolum*) resistance in the U.S. sorghum association panel. *Plant Genome* 11:170099.
- Deveau, A., Bonito, G., Uehling, J., Paoletti, M., Becker, M., Bindschedler, S., Haquard, S., Hervé, V., Labbé, J., Lastovetsky, O. A., Mieszkis, S., Millet, L. J., Vajna, B., Junier, P., Bonfante, P., Krom, B. P., Olsson, S., van Elsas, J. D., and Wick, L. Y. 2018. Bacterial-fungal interactions: Ecology, mechanisms and challenges. *FEMS Microbiol. Rev.* 42:335-352.
- Doehlemann, G., Ökmen, B., Zhu, W., and Sharon, A. 2017. Plant pathogenic fungi. Pages 701-726 in: *The Fungal Kingdom*. J. Heitman, B. J. Howlett, P. W. Crous, E. H. Stukenbrock, T. Y. James, and N. A. R. Gow, eds. ASM Press, Washington, DC.
- Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., and Clark, A. M. 2019. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* 18:435-441.
- Erpelding, J. E., and Prom, L. K. 2006. Variation for anthracnose resistance within the sorghum germplasm collection from Mozambique, Africa. *Plant Pathol. J.* 5:28-34.
- Fatouros, G., Gkizi, D., Fragkogeorgi, G. A., Paplomatas, E. J., and Tjamos, S. E. 2018. Biological control of *Pythium*, *Rhizoctonia* and *Sclerotinia* in lettuce: Association of the plant protective activity of the bacterium *Paenibacillus alvei* K165 with the induction of systemic resistance. *Plant Pathol.* 67:418-425.
- Fedele, G., Brischetto, C., González-Domínguez, E., and Rossi, V. 2020. The colonization of grape bunch trash by microorganisms for the biocontrol of *Botrytis cinerea* as influenced by temperature and humidity. *Agronomy* (Basel) 10:1829.
- Firn, R. D., and Jones, C. G. 2003. Natural products—A simple model to explain chemical diversity. *Nat. Prod. Rep.* 20:382-391.
- Gkikas, F.-I., Tako, A., Gkizi, D., Lagogianni, C., Markakis, E. A., and Tjamos, S. E. 2021. *Paenibacillus alvei* K165 and *Fusarium oxysporum* F2: Potential biocontrol agents against *Phaeoemoniella chlamydospora* in grapevines. *Plants* 10:207.
- Guetsky, R., Shtienberg, D., Elad, Y., and Dinoor, A. 2001. Combining biocontrol agents to reduce the variability of biological control. *Phytopathology* 91:621-627.
- He, H., Silo-Suh, L. A., Handelsman, J., and Clardy, J. 1994. Zwittermicin A, an antifungal and plant protection agent from *Bacillus cereus*. *Tetrahedron Lett.* 35:2499-2502.
- Hollomon, D. W. 2015. Fungicide resistance: 40 Years on and still a major problem. Pages 3-11 in: *Fungicide Resistance in Plant Pathogens: Principles and a Guide to Practical Management*. H. Ishii and D. W. Hollomon, eds. Springer Japan, Tokyo, Japan.
- Hollomon, D. W. 2016. Fungicide resistance: Facing the challenge—A review. *Plant Prot. Sci.* 51:170-176.
- Hu, J., Sarraimi, F., Li, H., Zhang, G., Stubbs, K. A., Lacey, E., Stewart, S. G., Karton, A., Piggott, A. M., and Chooi, Y.-H. 2019. Heterologous biosynthesis of elsinochrome sheds light on the formation of the photosensitive perylenequinone system. *Chem. Sci. (Camb.)* 10:1457-1465.
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. 2019. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47:D309-D314.
- Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* 11:119.
- Jeukens, J., Kukavica-Ibrulj, I., Emond-Rheault, J. G., Freschi, L., and Levesque, R. C. 2017. Comparative genomics of a drug-resistant *Pseudomonas aeruginosa* Panel and the challenges of antimicrobial resistance prediction from genomes. *FEMS Microbiol. Lett.* 364:fnx161.
- Jochum, C. C., Osborne, L. E., and Yuen, G. Y. 2006. Fusarium head blight biological control with *Lysobacter enzymogenes* strain C3. *Biol. Control* 39: 336-344.
- Kautsar, S. A., Blin, K., Shaw, S., Weber, T., and Medema, M. H. 2021. BiG-FAM: The biosynthetic gene cluster families database. *Nucleic Acids Res.* 49:D490-D497.
- Kevany, B. M., Rasko, D. A., and Thomas, M. G. 2009. Characterization of the complete zwittermicin A biosynthesis gene cluster from *Bacillus cereus*. *Appl. Environ. Microbiol.* 75:1144-1155.
- Khalid, S., Baccile, J. A., Spraker, J. E., Tannous, J., Imran, M., Schroeder, F. C., and Keller, N. P. 2018. NRPS-derived isoquinolines and lipopeptides mediate antagonism between plant pathogenic fungi and bacteria. *ACS Chem. Biol.* 13:171-179.
- Kurczab, R., Smusz, S., and Bojarski, A. J. 2014. The influence of negative training set size on machine learning-based virtual screening. *J. Cheminf.* 6: 32.
- Lamichhane, J. R., Dachbrodt-Saaydeh, S., Kudsk, P., and Messéan, A. 2016. Toward a reduced reliance on conventional pesticides in European agriculture. *Plant Dis.* 100:10-24.
- Lastovetsky, O. A., Krasnovsky, L. D., Qin, X., Gaspar, M. L., Gryganskyi, A. P., Huntemann, M., Clum, A., Pillay, M., Palanjappan, K., Varghese, N., Mikhailova, N., Stamatis, D., Reddy, T. B. K., Daum, C., Shapiro, N., Ivanova, N., Kyrpides, N., Woyke, T., and Pawlowska, T. E. 2020. Molecular dialogues between early divergent fungi and bacteria in an antagonism versus a mutualism. *MBio* 11:e02088-20.
- Leach, J. E., Triplett, L. R., Argueso, C. T., and Trivedi, P. 2017. Communication in the phytobiome. *Cell* 169:587-596.
- Lee, Y., Kim, Y. S., Balaraju, K., Seo, Y.-S., Park, J., Ryu, C.-M., Park, S.-H., Kim, J. F., Kang, S., and Jeon, Y. 2020. Molecular changes associated with spontaneous phenotypic variation of *Paenibacillus polymyxa*, a commonly used biocontrol agent, and temperature-dependent control of variation. *Sci. Rep.* 10:16586.
- Legein, M., Smets, W., Vandenheuvel, D., Eilers, T., Muyshondt, B., Prinsen, E., Samson, R., and Lebeer, S. 2020. Modes of action of microbial biocontrol in the phyllosphere. *Front. Microbiol.* 11:1619.
- Levy, A., Gonzalez, I. S., Mittelviehhaus, M., Clingenpeel, S., Paredes, S. H., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., Alvarez, B. R., Lundberg, D. S., Lu, T.-Y., Lebeis, S., Jin, Z., McDonald, M., Klein, A. P., Felcher, M. E., Rio, T. G., Grant, S. R., Doty, S. L., Ley, R. E., Zhao, B., Venturi, V., Pelletier, D. A., Vorholt, J. A., Tringe, S. G., Woyke, T., and Dangel, J. L. 2018. Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50:138-150.
- Li, X., Garbeva, P., Liu, X., Klein Gunnewiek, P. J. A., Clocchiatti, A., Hundscheid, M. P. J., Wang, X., and De Boer, W. 2020. Volatile-mediated antagonism of soil bacterial communities against fungi. *Environ. Microbiol.* 22:1025-1035.
- Lim, S. M., Yoon, M.-Y., Choi, G. J., Choi, Y. H., Jang, K. S., Shin, T. S., Park, H. W., Yu, N. H., Kim, Y. H., and Kim, J.-C. 2017. Diffusible and volatile antifungal compounds produced by an antagonistic *Bacillus velezensis* g341 against various phytopathogenic fungi. *Plant Pathol. J.* 33:488-498.

- Martínez-García, P. M., López-Solanilla, E., Ramos, C., and Rodríguez-Palenzuela, P. 2016. Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environ. Microbiol.* 18:4847-4861.
- McCoubrey, L. E., Elbadawi, M., Orlu, M., Gaisford, S., and Basit, A. W. 2021. Harnessing machine learning for development of microbiome therapeutics. *Gut Microbes* 13:1-20.
- Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., et al. 2015. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* 11:625-631.
- Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. 2016. Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. U.S.A.* 113:7361-7368.
- Mendes, R., Garbeva, P., and Raaijmakers, J. M. 2013. The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol. Rev.* 37:634-663.
- Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. 2018. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLOS Comput. Biol.* 14:e1006258.
- Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappellini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema, M. H. 2020. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* 16:60-68.
- Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., Tyson, G. H., Zhao, S., and Davis, J. J. 2019. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* 57:e01260-18.
- Nishida, M., Matsubara, T., and Watanabe, N. 1965. Pyrrolnitrin, a new antifungal antibiotic. microbiological and toxicological observations. *J. Antibiot. (Tokyo)* 18:211-219.
- Nkuba, J., Tinzaara, W., Night, G., Niko, N., Jogo, W., Ndyetabula, I., Mukandala, L., Ndayihazamaso, P., Niyongere, C., Gaidashova, S., Rwomushana, I., Opio, F., and Karamura, E. 2015. Adverse impact of banana Xanthomonas wilt on farmers' livelihoods in Eastern and Central Africa. Bishop Stuart University. <http://196.43.161.4/handle/123456789/353>
- Omboye, O. O., Oni, F. E., Batoool, H., Yimer, H. Z., De Mot, R., and Höfte, M. 2019. *Pseudomonas* cyclic lipopeptides suppress the rice blast fungus *Magnaporthe oryzae* by induced resistance and direct antagonism. *Front. Plant Sci.* 10:901.
- Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38:1079-1086.
- Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36:996-1004.
- Parnell, J. J., Berka, R., Young, H. A., Sturino, J. M., Kang, Y., Barnhart, D. M., and DiLeo, M. V. 2016. From the lab to the farm: An industrial perspective of plant beneficial microorganisms. *Front. Plant Sci.* 7:1110.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: 33rd Conf. Neural Inf. Processing Syst., Vancouver, BC Canada. <https://papers.nips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf>
- Pawar, S., Chaudhari, A., Prabha, R., Shukla, R., and Singh, D. P. 2019. Microbial pyrrolnitrin: Natural metabolite with immense practical utility. *Biomolecules* 9:443.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. 2011. Scikit-Learn: Machine learning in Python. *J. Mach. Learn. Res.* 12:2825-2830.
- Proietti, I., Frazzoli, C., and Mantovani, A. 2015. Exploiting nutritional value of staple foods in the world's semi-arid areas: Risks, benefits, challenges and opportunities of sorghum. *Healthcare* 3:172-193.
- Prom, L. K., Cuevas, H., Isakeit, T., and Droleskey, R. 2016. Excised leaf method for high volume evaluation of sorghum germplasm for resistance against *Colletotrichum sublineolum*. *Plant Pathol. J.* 15:11-16.
- Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P., and de la Cruz, F. 2020. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* 11:3602.
- Santoyo, G., Orozco-Mosqueda, M. d. C., and Govindappa, M. 2012. Mechanisms of biocontrol and plant growth-promoting activity in soil bacterial species of *Bacillus* and *Pseudomonas*: A review. *Biocontrol Sci. Technol.* 22:855-872.
- Sass, G., Nazik, H., Penner, J., Shah, H., Ansari, S. R., Clemons, K. V., Groleau, M.-C., Dieti, A.-M., Visca, P., Haas, H., Déziel, E., and Stevens, D. A. 2018. Studies of *Pseudomonas aeruginosa* mutants indicate pyoverdine as the central factor in inhibition of *Aspergillus fumigatus* biofilm. *J. Bacteriol.* 200:e00345-17.
- Song, S., Sun, X., Meng, L., Wu, Q., Wang, K., and Deng, Y. 2021. Antifungal activity of hypocrellin compounds and their synergistic effects with antimicrobial agents against *Candida albicans*. *Microb. Biotechnol.* 14:430-443.
- Stover, R. H. 1986. Measuring response of Musa cultivars to sigatoka pathogens and proposed screening procedures. Pages 114-118 in: Banana and Plantain Breeding Strategies: Proc. Int. Workshop, Cairns, QLD Australia.
- Strange, R. N., and Scott, P. R. 2005. Plant disease: A threat to global food security. *Annu. Rev. Phytopathol.* 43:83-116.
- Tapsin, S., Sun, M., Shen, Y., Zhang, H., Lim, X. N., Susanto, T. T., Yang, S. L., Zeng, G. S., Lee, J., Lezhava, A., Ang, E. L., Zhang, L. H., Wang, Y., Zhao, H., Nagarajan, N., and Wan, Y. 2018. Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes. *Nat. Commun.* 9: 1289.
- Tasaki, S., Nakayama, M., and Shoji, W. 2017. Morphologies of *Bacillus subtilis* communities responding to environmental variation. *Dev. Growth Differ.* 59:369-378.
- Twizeyimana, M., Ojiambo, P. S., Tenkouano, A., Ikotun, T., and Bandyopadhyay, R. 2007. Rapid screening of *Musa* species for resistance to black leaf streak using in vitro plantlets in tubes and detached leaves. *Plant Dis.* 91:308-314.
- Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. 2019. Machine learning algorithm validation with a limited sample size. *PLoS One* 14: e0224365.
- Van Camp, P.-J., Haslam, D. B., and Porollo, A. 2020. Prediction of antimicrobial resistance in Gram-negative bacteria from whole-genome sequencing data. *Front. Microbiol.* 11:1013.
- Vanittanakom, N., Loeffler, W., Koch, U., and Jung, G. 1986. Fengycin—A novel antifungal lipopeptide antibiotic produced by *Bacillus subtilis* F-29-3. *J. Antibiot. (Tokyo)* 39:888-901.
- van Overbeek, L. S., and Saikkonen, K. 2016. Impact of bacterial-fungal interactions on the colonization of the endosphere. *Trends Plant Sci.* 21:230-242.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archiblad, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 17:261-272.
- Wang, T., Liu, X.-H., Wu, M.-B., and Ge, S. 2018. Molecular insights into the antifungal mechanism of bacilysin. *J. Mol. Model.* 24:118.
- Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., Wang, Y., and Douiri, A., Wolfe, C. D., and Bray, B. 2020. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One* 15:e0234722.
- Wheeler, N. E., Gardner, P. P., and Barquist, L. 2018. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genet.* 14:e1007333.
- Wise, C., Falardeau, J., Hagberg, I., and Avis, T. J. 2014. Cellular lipid composition affects sensitivity of plant pathogens to fengycin, an antifungal compound produced by *Bacillus subtilis* strain CU12. *Phytopathology* 104: 1036-1041.
- Xie, Y., Meng, W.-Y., Li, R.-Z., Wang, Y.-W., Qian, X., Chan, C., Yu, Z.-F., Fan, X.-X., Pan, H.-D., Xie, C., Wu, Q.-B., Yan, P.-U., Liu, L., Tang, Y.-J., Yao, X.-J., Wang, M.-F., and Leung, E. L.-H. 2021. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Transl. Oncol.* 14:100907.

- Xing, M.-Z., Zhang, X.-Z., Sun, Z.-L., and Zhang, H.-Y. 2003. Perylenequinones act as broad-spectrum fungicides by generating reactive oxygen species both in the dark and in the light. *J. Agric. Food Chem.* 51: 7722-7724.
- Zhu, J., Wu, Y., Man Wang, K. L., Xu, L., Chen, Y., Cai, Y., and Jin, J. 2020. Integrating machine learning-based virtual screening with multiple protein structures and bio-assay evaluation for discovery of novel GSK3 β inhibitors. *Front. Pharmacol.* 11:566058.
- Zorn, K. M., Sun, S., McConnon, C. L., Ma, K., Chen, E. K., Foil, D. H., Lane, T. R., Liu, L. J., El-Sakkary, N., Skinner, D. E., Ekins, S., and Caffrey, C. R. 2021. A machine learning strategy for drug discovery identifies anti-schistosomal small molecules. *ACS Infect. Dis.* 7:406-420.