



Specialized plant biochemistry drives gene clustering in fungi

Emile Gluck-Thaler¹ · Jason C. Slot¹

Received: 14 September 2017 / Revised: 18 January 2018 / Accepted: 26 January 2018
© International Society for Microbial Ecology 2018

Abstract

The fitness and evolution of prokaryotes and eukaryotes are affected by the organization of their genomes. In particular, the physical clustering of genes can coordinate gene expression and can prevent the breakup of co-adapted alleles. Although clustering may thus result from selection for phenotype optimization and persistence, the impact of environmental selection pressures on eukaryotic genome organization has rarely been systematically explored. Here, we investigated the organization of fungal genes involved in the degradation of phenylpropanoids, a class of plant-produced secondary metabolites that mediate many ecological interactions between plants and fungi. Using a novel gene cluster detection method, we identified 1110 gene clusters and many conserved combinations of clusters in a diverse set of fungi. We demonstrate that congruence in genome organization over small spatial scales is often associated with similarities in ecological lifestyle. Additionally, we find that while clusters are often structured as independent modules with little overlap in content, certain gene families merge multiple modules into a common network, suggesting they are important components of phenylpropanoid degradation strategies. Together, our results suggest that phenylpropanoids have repeatedly selected for gene clustering in fungi, and highlight the interplay between genome organization and ecological evolution in this ancient eukaryotic lineage.

Introduction

Genome architecture is intimately linked to the trajectory of organismal evolution. The impacts of linear gene organization on prokaryotic evolution in particular have been extensively studied [1], and such organization is increasingly recognized as affecting eukaryotic fitness and evolution as well. For example, spatial clustering of functionally related genes may enable optimization of phenotypes through coordinated gene expression [2–4]. Similarly, loci composed of co-adapted alleles can facilitate inheritance of locally adapted ecotypes within populations [5, 6]. Rather than resulting from non-adaptive processes, the persistence of such organizational patterns in eukaryotic genomes suggests that they are targets of selection [3, 7]. However, the extent to which specific environmental selection

pressures drive eukaryotic genome organization, especially over macroevolutionary timescales, is unclear.

Organized genome structure is particularly apparent in fungi, a lineage of eukaryotic microorganisms whose activities impact biomass transformation and plant and animal health [8]. Fungal genomes contain many metabolic gene clusters (MGCs) composed of genes encoding enzymes, transporters, and regulators that participate in specialized metabolic processes such as nutrient acquisition, competition, and defense [9]. Although MGCs are far more rare in fungi compared with bacteria, fungal MGCs exhibit similarly sparse distributions among distantly related species with overlapping niches [10–12]. This ecological pattern of distribution suggests that conserved combinations of genes may be signatures of ecological selection in fungal genomes.

Fungal MGCs encoding specialized or secondary metabolite (SM) production have been studied extensively [13] and more recently, several reports suggest that MGCs also encode adaptations to degrade plant SMs [11, 12, 14–16]. Plant SMs mediate important biotic and abiotic interactions, including the exclusion of fungal pathogens and the rates of nutrient cycling long after the plant has died. The largest group of plant SMs are phenylpropanoids, which not only contribute to chemical defenses, but are also the main barriers to wood decay [17], and costly inhibitors of lignocellulose biofuel production [18]. As the primary

Electronic supplementary material The online version of this article (<https://doi.org/10.1038/s41396-018-0075-3>) contains supplementary material, which is available to authorized users.

✉ Jason C. Slot
slot.1@osu.edu

¹ Department of Plant Pathology, The Ohio State University, Columbus, OH, USA

colonizers of plant material, fungi are frequently in contact with phenylpropanoids, and must mitigate their inhibitory effects through sequestration, excretion, and degradation in order to grow [19]. Despite the characterization of many phenylpropanoid degradation pathways in fungi, the genomic bases of these pathways are largely unknown [19], precluding the use of currently available algorithms to investigate whether or not these metabolic processes are encoded in MGCs [9, 20].

Here, we developed a novel algorithm based on empirically derived models of fungal genome evolution to test the hypothesis that selection pressures from plant SMs impact genome organization across disparate fungal lineages. Using a database of 529 fungal genomes, we systematically detected 1110 candidate MGCs and many conserved combinations of MGCs that putatively degrade a broad array of phenylpropanoids. We tested for associations between MGCs and various fungal ecological lifestyles, and found that the presence of certain MGCs was enriched in plant pathotrophs, saprotrophs, symbiotrophs, and endophytes. Although many clusters appear to have evolved independently, we identified several gene families that are commonly associated with diverse MGCs, suggesting they play important roles in phenylpropanoid catabolism. Overall, our results suggest that phenylpropanoids are drivers of genome organization in plant-associated fungi, and that MGCs in turn determine patterns of fungal community assembly on both living and decaying plant tissues.

Materials and methods

Data acquisition, annotation, and software specifications

Publically available data from 529 assembled fungal genomes and predicted proteomes were retrieved from various sources (Supplementary Table 2). Ecological metadata were compiled from various sources (Supplementary Table 2), including the community-curated FUNGuild database (last accessed 9 September 2016) [21] and the U.S. National Fungus Collection database (last accessed 1 April 2017) [22].

Amino-acid sequence searches with cutoffs of 30% identity, 50 bitscore and where the length of the target sequence was 50–150% of the query sequence, were performed with USEARCH v8.0.1517's UBLAST algorithm (additional parameters: e-value cutoff = $1e-5$, accel = 0.8) [23] or with BLASTp v2.2.25+ (additional parameters: e-value cutoff = $1e-4$) [24]. Homology was determined using OrthoMCL v2 with an inflation value of 1.5 [25]. All amino-acid sequences from clustered homolog groups were assigned to orthologous groups from the fuNOG database

(last accessed 18 March 2016) [26] using HMMER3 [27]. Predicted functional annotations and KOG processes for each clustered homolog group were based on the most frequent fuNOG annotation assigned to proteins within that group. Clustered sequences were additionally annotated by protein family (PFAM) and gene ontology (GO) terms using InterProScan 5 v5.20–59.0 for screening purposes only [28].

All phylogenetic trees were visualized using ETE v3 [29] and all other graphs were visualized using the ggplot2 package in R [30]. All data used to generate the presented figures can be found in Supplementary Table 12.

Construction of the microsynteny tree

The evolution of microsynteny (gene content conservation over small genomic distances) does not necessarily recapitulate phylogenetic relationships determined by models of sequence evolution. In order to assess the unexpectedness of cluster distributions within a phylogenetic framework based on microsynteny conservation, we used an approach similar to Snel et al. [31] to construct a species tree based on microsyntenic distance, or pairwise comparisons of gene content conservation (Supplementary Figure 1).

To construct the microsynteny tree, we retrieved 10 genes upstream and downstream of all homologs from a randomly selected gene family (designated “gene neighborhood”). Variation in genome assembly quality occasionally resulted in gene neighborhoods smaller than the maximum neighborhood size of 21, but we required a minimum neighborhood size of 10. All genes within the retrieved neighborhoods were compared using UBLAST, then sorted into homolog groups using OrthoMCL. For each pairwise neighborhood comparison, syntenic distance was defined as $1 - (\text{the number of shared orthologs} / \text{the smallest neighborhood size})$. Pairwise syntenic distances thus range from 0 (all homolog groups are shared) to ~0.95 (only the query gene is shared). In all, 1000 neighborhoods were randomly sampled in this way. A neighbor-joining tree was constructed from the distance matrix of median pairwise syntenic distances using the ape package in R [32]. The original data set was sampled with replacement to obtain 100 trees that were then used to calculate bootstrap support on the original tree using using RAXML v8.2.0 [33], and nodes receiving <70% bootstrap support were collapsed. The final microsynteny tree was used to calculate distances covered by cluster distributions, as well as individual homolog groups.

Sampling null models of gene cluster evolution

Empirically derived null models describing the background levels of gene cluster distributions were developed for each

of the 12 largest taxonomic classes and each cluster size ranging from 4 to 24 genes (252 distributions in total). Briefly, for a given null distribution, a group of neighboring genes (i.e., query cluster) of size X was chosen at random from a randomly selected genome from taxonomic class Y . Hits to each gene in the query cluster were recovered using BLASTp. Genomes containing homologous clusters (i.e., containing hits to all genes in the query cluster, where no >6 intervening genes separated any hit from another) were then retrieved. To determine the phylogenetic distance associated with the query cluster distribution, the total non-overlapping branch length distance on the microsynteny tree connecting all genomes with homologous clusters was calculated. The above sampling process was repeated 500 times for each null distribution.

Locating clusters through unexpected synteny

All homologs of an anchor gene query sequence of interest were retrieved using BLASTp. Twenty genes upstream and downstream of all anchor gene homologs (designated “anchor gene neighborhood”) were retrieved and sorted into homolog groups using OrthoMCL after a UBLAST step. Individual homolog groups whose distribution on the microsynteny species tree had a maximum pairwise distance below 0.95 were discarded. Within the set of all anchor gene neighborhoods, the set of unique combinations of four or more homolog groups that included the anchor gene (“cluster motifs”) was determined. The genomes in which each motif occurred were identified, with the condition that genes belonging to homolog groups in the motif never be separated by >6 intervening genes. For each genome in which a given motif occurred, the probability of observing the motif in that genome, given the size of the motif and the taxonomic class of the genome, was empirically estimated by determining the proportion of samples in the appropriate null distribution that cover a total distance on the microsynteny tree greater than or equal to the distance associated with the given motif, divided by the total number of samples in the null distribution. For example, to estimate the probability of observing a motif with eight homolog groups in a Dothideomycete genome, we would compare the total distance associated with the observed motif to the null distribution of distances associated with size 8 clusters sampled from Dothideomycetes. For all such tests, the test statistic is distance on the microsynteny tree, and the null hypothesis is that the phylogenetic distribution of a given cluster motif is consistent with background rates of microsynteny evolution. The null hypothesis is rejected for motifs with an estimated probability below 0.05 in at least one genome, and all genes assigned to such motifs are designated clusters. Clusters with proteins that had fuNOG annotations or PFAM domains or GO term annotations

associated with proteins known to exclusively participate in fungal SM biosynthesis were excluded from further analysis (Supplementary Table 4). All annotations of clustered proteins were also manually inspected for evidence of exclusive participation in biosynthetic metabolism, and excluded if necessary.

Cluster models and multi-cluster model profiles (MCMPs)

Homology among clusters from the same cluster class (i.e., containing the same anchor gene) was determined by assessing similarities in homolog group content. Briefly, a matrix detailing the presence or absence of homolog groups in each cluster was used to calculate Bray–Curtis dissimilarity indices for all pairwise comparisons of clusters. Pairwise comparisons were then grouped using complete linkage clustering, and any clusters separated by under 0.6 distance units were assigned to the same cluster model. This cutoff was empirically determined after manual examination of the content of clusters assigned to the same model under various distance cutoffs. Homolog groups present in $\geq 75\%$ of clusters assigned to a given model were then used to summarize that model. The above approach was also used to group fungal species with similar combinations of clusters into MCMPs by using a matrix detailing the presence or absence of homologous clusters (as determined by cluster model) among all species with two or more clusters. MCMPs observed in >5 species were used for enrichment analyses. All above analyses were performed using the *vegan* package in R [34].

Network analyses

Amino-acid sequences from all clustered homolog groups across 13 cluster classes were combined into one set and then sorted into new homolog groups using UBLAST and OrthoMCL. Homolog groups that contained sequences from multiple cluster classes were designated as “shared”. Pairwise co-occurrences of all homolog groups in all clusters were determined, and visualized as a network with Cytoscape v.3.4.0 [35]. The network layout was determined solely by the AllegroLayout plugin with the Allegro Spring-Electric algorithm. Analyses of network modularity were performed with the spectral partitioning algorithm [36], as implemented in MODULAR [37], and the probability of network modularity was estimated by randomly sampling the network 1000 times, twice.

Enrichment tests

A one-tailed Fisher’s exact test, as implemented in the Text-NSP Perl module, was used to conduct all tests of

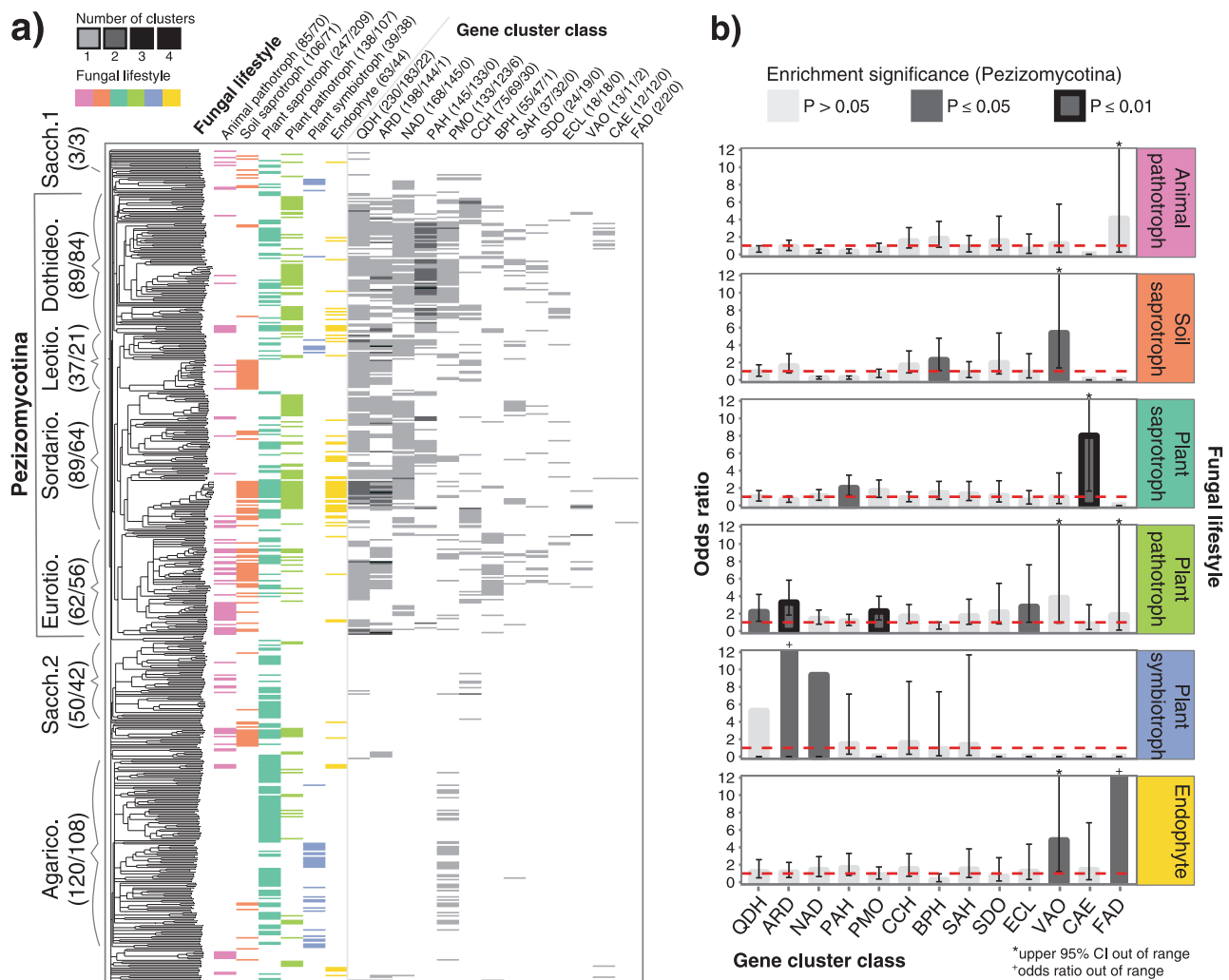


Fig. 1 Associations between gene cluster distributions and fungal lifestyle. **a** A phylogeny of 529 fungi (representing 454 species) based on pairwise microsyntenic similarity is shown to the left, annotated by taxonomic class. All taxonomic classes have been abbreviated by removing “-mycetes” suffixes. A matrix indicating the ecological lifestyle(s) associated with each fungus is shown to the immediate right of the phylogeny, followed by a heatmap indicating the number of candidate phenylpropanoid-degrading gene clusters in each genome, for each cluster class. Numbers in brackets following the taxonomic class and ecological lifestyle headers correspond to the number of genomes and species within those categories. Numbers in brackets following cluster class headers indicate the total number of clusters

assigned to that class, the number of species with at least one cluster, and the number of clusters that overlap with clusters from another cluster class. **b** Odds ratios representing the strength of the association between cluster presence and fungal ecological lifestyle are shown for each of 13 gene cluster classes and 6 lifestyles, using data at the species level from the Pezizomycotina. Dotted red lines indicate an odds ratio of 1. Dark gray bars indicate enrichment below a significance level of 0.05, whereas black outlines indicate enrichment below a significance level of 0.01. Error bars indicate the 95% confidence interval (CI) for each odds ratio measurement. CIs of 0 are not shown. The color-coding of ecological lifestyle is consistent across the entire figure.

enrichment using data from the Pezizomycotina (Supplementary Figure 1). Unless noted otherwise, cluster features (either cluster presence (Fig. 1b), cluster model presence (Supplementary Figure 4) or MCMP assignment (Fig. 2c)) were recorded at the species level, and counts of species were used to fill in contingency tables. Species assigned to multiple ecological lifestyles were considered to belong to each of those lifestyles separately for the purposes of conducting enrichment tests. Contingency tables with a zero in at least one cell had all cells incremented by 0.5 to avoid division by zero when calculating the odds ratio [38]. In

general, the odds ratio can be interpreted as the odds that fungi have a particular lifestyle, given they have a particular cluster-based feature. The precision of each odds ratio, except for those associated with contingency tables with a zero in at least one cell, was estimated by calculating the 95% confidence interval [39]. We rejected the null hypothesis that particular features are not enriched in fungi with a particular ecological lifestyle at $\alpha = 0.05$. Given the exploratory nature of this study, we did not correct for multiple testing in order to identify trends in cluster distributions to follow-up in later confirmatory studies.

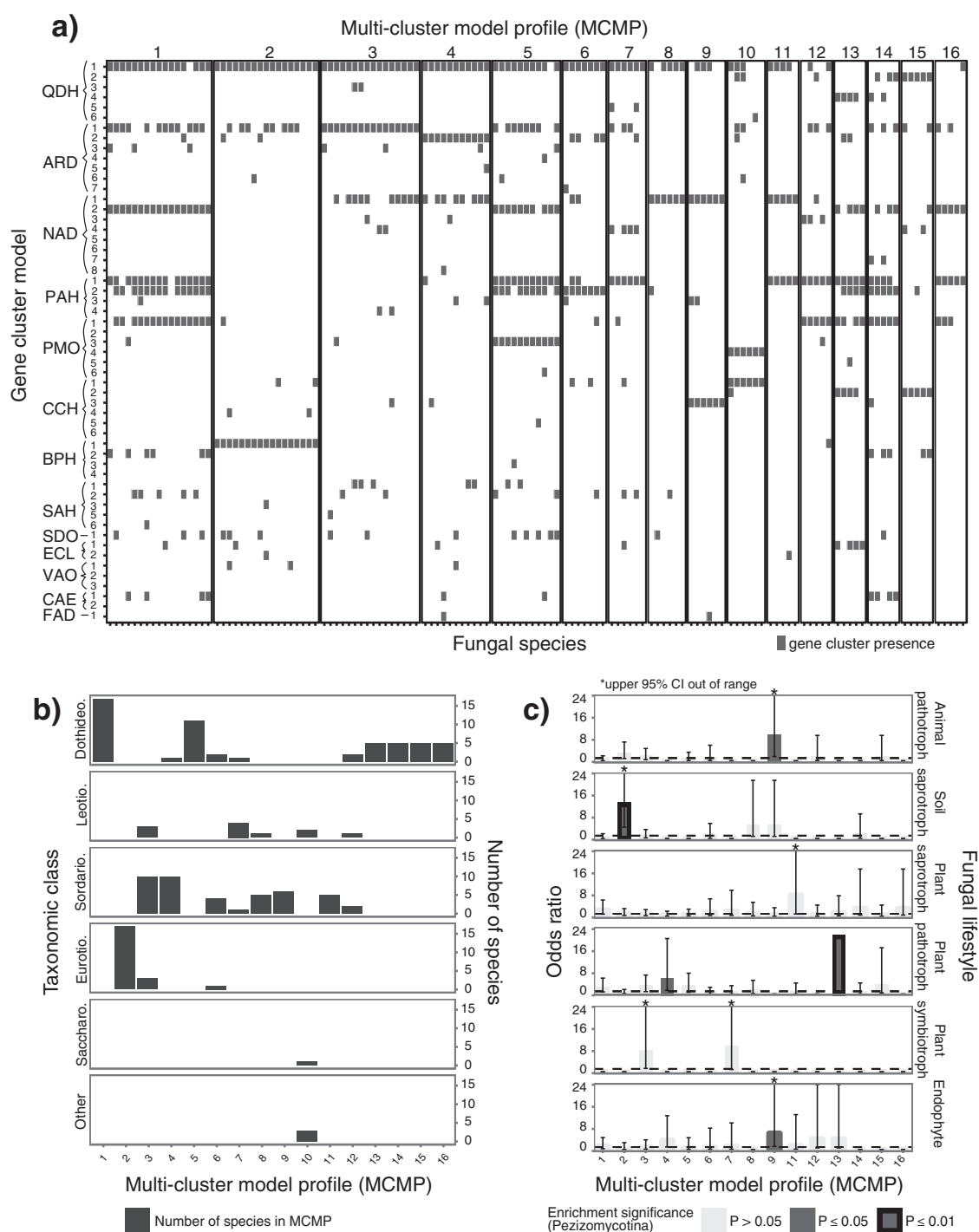


Fig. 2 Combinations of candidate phenylpropanoid-degrading gene clusters in fungal genomes. **a** A matrix describes the presence/absence of various homologous clusters (as determined by cluster model; rows) in the genomes of 133 fungal species (columns). Species are grouped into 16 multi-cluster model profiles (MCMPs) based on similarities in the combinations of clusters found in their genomes. **b** A bar chart depicts the number of fungal species from each MCMP per taxonomic class. **c** Odds ratios representing the strength of the association

between MCMP and fungal ecological lifestyle are shown, using data at the species level from the Pezizomycotina. Dotted black lines indicate an odds ratio of 1. Dark gray bars indicate enrichment below a significance level of 0.05, whereas black outlines indicate enrichment below a significance level of 0.01. Error bars indicate the 95% confidence interval (CI) for each odds ratio measurement. CIs of 0 are not shown. Enrichment data are not shown for MCMP 10, as fewer than five fungi from the Pezizomycotina are assigned to this MCMP.

Results

Diverse candidate gene clusters are associated with phenylpropanoid degradation

Using 27 different “anchor” gene families involved in phenylpropanoid degradation as separate queries (Methods; Supplementary Table 1), we searched 529 genomes from 454 fungal species for clusters using a novel gene cluster detection algorithm (Methods; Supplementary Figure 1, Supplementary Table 2). We found evidence of unexpected clustering in 13 anchor gene families, which we defined as separate cluster classes: aromatic ring-opening dioxygenase (ARD), benzoate 4-monooxygenase (BPH), catechol dioxygenase (CCH), epicatechin laccase (ECL), ferulic acid decarboxylase (FAD), ferulic acid esterase 7 (CAE), naringenin 3-dioxygenase (NAD), phenol 2-monooxygenase (PMO), pterocarpin hydroxylase (PAH), quinate 5-dehydrogenase (QDH), salicylate hydroxylase (SAH), stilbene dioxygenase (SDO), and vanillyl alcohol oxidase (VAO). We identified a total of 1110 clusters distributed across 341 fungal genomes, or 938 clusters distributed across 287 fungal species (Fig. 1, Supplementary Figure 2, Supplementary Table 5). Only 31 clusters contained multiple different anchor genes (Supplementary Table 3).

By grouping clusters together based on similarities in their gene content, we identified 56 distinct types of clusters, or cluster models (Supplementary Figure 5). The number of cluster models per cluster class varies from 1 to 8, and clusters assigned to different models generally do not overlap in terms of content (Supplementary Table 6). Additionally, the gene windows over which clusters spanned typically contained very few intervening genes not part of the cluster itself (Supplementary Figure 3). Clusters are disproportionately distributed across fungal lineages: the Pezizomycotina comprise 49.6% of species, yet contain 90.3% of all clusters, whereas Agaricomycetes represent 26.4% of species, but only 4.3% of clusters (Supplementary Table 7). However, homologous clusters assigned to the same model are typically found distributed across different taxonomic classes (Supplementary Figure 5, Supplementary Table 7).

In order to benchmark our approach, we searched for previously characterized clusters containing QDH and SDO homologs. All 8 previously predicted clusters from the SDO cluster class [12] and all 31 described clusters from the QDH class [40, 41] were recovered. Additionally, while this study was being prepared, a single cluster from the PAH class was identified by Pigné et al. [42]. To our knowledge, the remaining 1070 clusters are described here for the first time. We expect some false-negative error in our detection but not significant false-positive error to arise due to

variation in the quality of genome assemblies and annotations. Some false-positive error in our detection may, however, arise from the fact that beyond the anchor gene, no genes were required to be demonstrably involved in phenylpropanoid degradation because the loci involved in such pathways are for the most part unknown [19]. Although we removed clusters with genes known to exclusively participate in biosynthetic secondary metabolism (Supplementary Table 4), it is still possible that some remaining clusters participate in biosynthetic reactions.

Clustered homolog groups, including those distributed across cluster classes, are enriched for primary and secondary metabolic functions

Despite the lack of functional constraints on cluster composition, clustered homolog groups were enriched for seven KOG processes primarily related to primary or secondary metabolism, including “energy production and conversion” and “secondary metabolite biosynthesis, transport and catabolism”, the two largest and most significantly enriched categories (Supplementary Table 8). The KOG process “transcription” was the only non-metabolic process enriched in clustered homolog groups. Homolog groups present in more than one cluster class are also enriched for multiple functional categories related to primary and secondary metabolism (Supplementary Table 9). Conserved Pfam domains present among shared homolog groups include transport and transcription-related domains, cytochrome P450 domains, and domains related to functional group modification (e.g., dehydrogenase, transferase, and decarboxylase; Supplementary Table 10).

Phenylpropanoid degradation gene clusters are enriched in Pezizomycotina species with plant-associated lifestyles

We limited our exploratory enrichment analyses to Pezizomycotina species, as they contained the vast majority of identified clusters. Four cluster classes are enriched in plant pathotrophs, including those containing ECL, which is known to participate in flavonoid degradation (Fig. 1). Six cluster classes are enriched in species with other plant-associated lifestyles, such as plant saprotrophs and endophytes. The BPH and VAO cluster classes are enriched in soil saprotrophs. No cluster classes are enriched in animal pathotrophs. Similar patterns of lifestyle-dependent enrichment are observed at the cluster model level (Supplementary Figure 4). Notably, at least one model in each of 12 cluster classes is enriched in fungal species with a plant-associated lifestyle, and different models from the same cluster class rarely are enriched in fungi with the

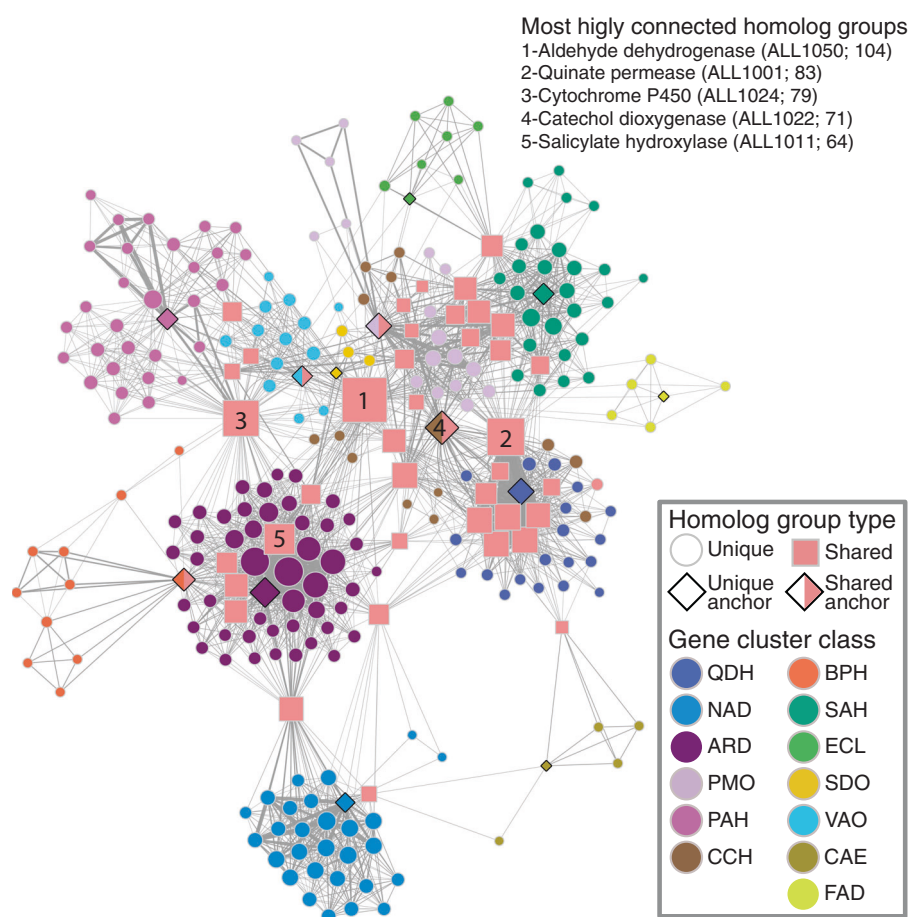


Fig. 3 Co-occurrence network of homolog groups in candidate phenylpropanoid-degrading gene clusters. Each node represents a homolog group found in a candidate phenylpropanoid-degrading gene cluster. All nodes are color-coded by the cluster class in which they are found, except for those homolog groups found associated with multiple cluster classes (i.e., “shared”), colored pink and shaped as squares. Nodes representing anchor gene families used to retrieve the clusters are shaped as diamonds, and those anchor gene families that are shared across multiple cluster classes are additionally colored pink.

Edges symbolize the co-occurrence of two homolog groups in the same gene cluster, whereas edge width is proportional to the frequency of that occurrence. Node size is proportional to the number of connections emanating from that node. The proximity of nodes to one another is proportional to the number of shared connections. The annotations, followed by the code names and number of unique connections in parentheses, of nodes with the greatest number of connections (i.e., associated with the greatest diversity of homolog groups) are indicated in the top right-hand corner of the network.

same ecology. Although ecological enrichments highlight predominant trends, clusters across all anchor genes are typically associated with diverse lifestyles, in part because any given fungal species can be associated with multiple ecological lifestyles.

A total of 191 fungal species possessed multiple clusters of different models. We assigned these fungi to MCMPs based on the combinations of cluster models found in their genomes (Methods). In all, 133 fungal species were distributed across 16 distinct MCMPs containing 5 or more species (Fig. 2). Fungi from the same MCMP tend to be closely related; however, 13 MCMPs contain fungi from different taxonomic orders, and 7 of these contain fungi from different taxonomic classes. When limiting an exploratory enrichment analysis to *Pezizomycotina* species,

we found that two MCMPs are enriched in species with plant-associated lifestyles, one is enriched in animal pathotrophs and endophytes, and one is enriched in soil saprotrophs (Fig. 2).

Homolog groups are found in modules of co-clustered genes

We visualized structural associations among all clustered homolog groups as a network where nodes represent homolog groups, and edges represent co-occurrence of homolog groups in a cluster (Fig. 3). In total, 39 homolog groups are found in clusters belonging to different classes (Supplementary Table 10), which resulted in a common network for all cluster classes. These “shared” homolog

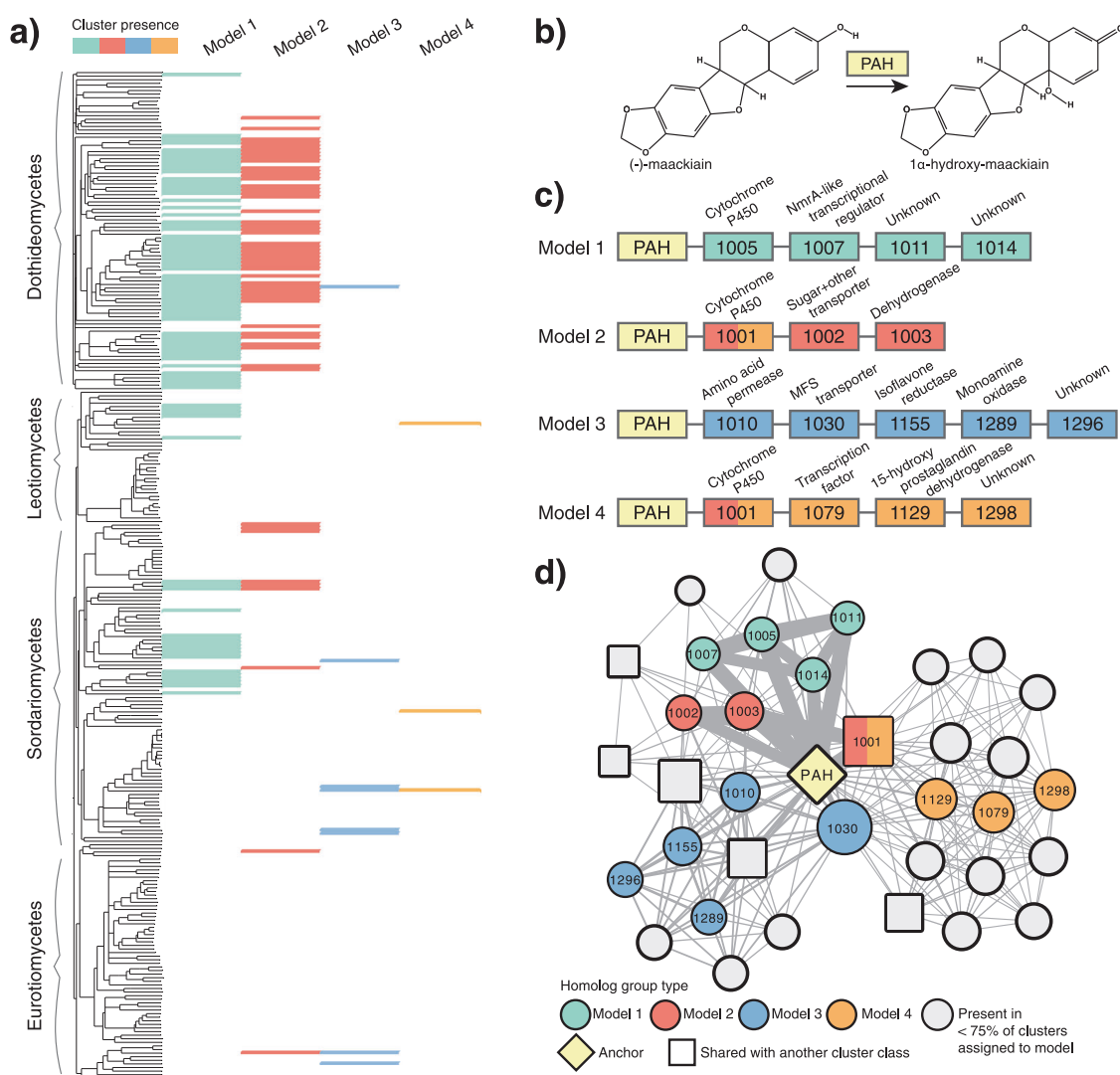


Fig. 4 The distribution of pterocarpan hydroxylase (PAH) clusters in fungi. **a** A phylogeny of the Pezizomycotina based on pairwise microsyntentic distance is shown to the left, annotated by taxonomic class. The presence/absence of clusters assigned to the four PAH cluster models is indicated in the matrix to the right of the phylogeny, color-coded by cluster model. **b** A simplified schematic of one of several reactions catalyzed by PAH. **c** Homolog groups present in ≥75% of clusters assigned to a given cluster model are depicted as

boxes color-coded by cluster model, whereas PAH homologs are indicated in yellow. The four-digit code and predicted annotation are indicated for each homolog group. **d** The depicted network follows the conventions specified in Fig. 2. Homolog groups present in ≥75% of clusters assigned to a given cluster model are color-coded by cluster model and inscribed with their code, whereas others are colored gray. Nodes representing homolog groups present in clusters from other cluster classes are drawn as squares.

groups are typically highly connected (i.e., they co-occur with a diverse set of homolog groups; Supplementary Table 10). The modularity of the network is significantly greater than that observed in networks of similar size with randomly distributed edges (Supplementary Table 11). This high degree of modularity reflects a large number of homolog groups that are unique to each cluster class (236 in total). Some homolog groups, although not shared between cluster classes, are highly connected because they are present in different cluster models (e.g., homolog group 1001 in Fig. 4).

Results spotlight: candidate PAH clusters may encode uncharacterized strategies for flavonoid degradation

The production of toxic SMs by plants is an ancient and effective strategy for deterring fungal growth. Isoflavonoids, for example, are a critical component of legume (Fabaceae) defenses against fungi. In response, some legume pathogens have evolved degradative metabolic pathways to detoxify isoflavonoids, thus facilitating host colonization [43]. PAH is involved in the degradation of

pterocarpan and medicarpin isoflavonoids, and can contribute to virulence on susceptible plant hosts [43]. We found 133 clusters in the PAH class, distributed across 94 fungal species with diverse ecological lifestyles (Fig. 4). Homology among these clusters is best described by four models containing many of the canonical functions found in fungal MGCs. Models 1, 2, and 4 all contain cytochrome P450s, a large family of enzymes involved in detoxification processes [44], whereas models 2 and 3 contain at least one transporter and model 4 contains a transcription factor. Model 3 contains a homolog of isoflavone reductase, an enzyme participating in isoflavonoid biosynthesis in plants and possibly isoflavonoid detoxification in fungi [45]. The widespread distribution of PAH clusters suggests they are involved in the degradation of flavonoids or flavonoid-like molecules found outside the Fabaceae. However, the genetics underlying isoflavonoid biosynthesis and degradation, and flavonoid metabolism in general, are not well understood for fungi.

A PAH homolog in *Alternaria brassicicola* found in a model 1 cluster was recently shown to contribute to 1,8-dihydroxynaphthalene (DHN) melanin accumulation in fungal cell walls, possibly through polymerization or cross-linking [42]. Melanins are a class of phenolic polymers that quench oxidizing radicals, and contribute to pathogenicity and survival in harsh environments [46]. Intriguingly, plant flavonoids can be polymerized in planta and in vitro into melanin by fungal laccases [47] and possibly by monooxygenases similar to PAH [48]. As the repurposing of host metabolites for the production of SMs is not unprecedented in fungi [49], we suggest that this PAH cluster is an excellent candidate for exploring hypotheses of metabolic cross-talk.

Discussion

Candidate phenylpropanoid degradation MGCs are found at over 1000 loci with unexpectedly conserved synteny

Fungi are the most ecologically significant decomposers of plant biomass, and represent some of the world's most devastating plant pathogens [8]. The presence of genes targeting plant SMs in fungal genomes has often been linked to species ecology and the ability to colonize different types of plant tissues [17]. Much less is known about how these genes are organized in fungal genomes, and how their organization impacts or may be affected by fungal ecology and evolution. Many fungal genomes tend to undergo extensive and frequent rearrangements [50, 51]. By contrast, MGCs we detected are often simultaneously present in lineages that diverged hundreds of millions of years

ago [17]. As combinations of genes that persist over long periods of time are likely to be maintained by natural selection [1, 3, 52] and often encode proteins that are functionally related [53, 54], we hypothesize that the conserved combinations of phenylpropanoid-degrading genes detected here encode adaptive metabolic phenotypes, and are signatures of selection on genome organization.

The most well-studied phenotype conferred by clustering is improved coordination of gene transcription [3]. Clustering facilitates transcription by allowing co-regulation through local chromatin modifications [55], promoter sharing [56], and avoidance of topological constraints on DNA encountered during transcription [57]. In addition to synchronizing cellular responses, coordinated gene transcription may also decrease the accumulation of toxic intermediate metabolites, such as those produced during phenylpropanoid degradation [12, 19], by optimizing enzyme stoichiometry and improving metabolic flux [4]. Gene clustering may additionally be selected because it increases the capacity for recombining populations to evolve. When genes contributing to the same trait are clustered, selection for that trait is more efficient because of decreased selective interference between genes at that locus [58]. Clustering can also prevent the breakup of co-adapted alleles in the face of gene flow [6], possibly enabling local adaptation within cryptic niches.

Similarly, clustering increases the probability of propagating adaptive combinations of genes through both vertical [58] and horizontal transfer [1, 59]. Many of the clusters observed in this study have discontinuous phylogenetic distributions consistent with evolutionary scenarios such as vertical inheritance coupled with extensive loss, convergence, duplication, or horizontal gene transfer (HGT). HGT of MGCs is predicted to be associated with rapid adaptive changes to phenotypes, including increased capacity for host colonization [10, 11] and nutrient acquisition [60]. For example, MGCs in bacteria are frequently dispersed by HGT among species inhabiting the same environment or host [1]. HGT is positively influenced by relatedness in bacteria [61] and fungi [9, 62], as well as by shared ecological niche [63]. Given that the distributions we observed here may be at least partly HGT-driven [12], the role of HGT in the dispersal of putative phenylpropanoid-degrading MGCs must be explicitly tested with phylogenetic methods in follow-up studies [62].

Gene cluster distributions suggest ecological adaptation

We detected the majority of clusters in the Pezizomycotina, which are known to possess highly clustered genomes compared with the Basidiomycota [9]. Such differences

may be due to modes of chromosomal evolution unique to the Pezizomycotina involving extensive rearrangements that could facilitate cluster formation [50, 64]. The large population sizes attributed to many Pezizomycotina lineages could then serve to increase selection efficiency for clusters (Lynch 2007) such that they would be maintained. Nevertheless, although certain clusters have distributions suggestive of lineage-specific bias (Supplementary Table 7), these conserved combinations of genes were ultimately detected because their distributions conflicted with lineage-specific phylogenetic signal, indicating that phylogeny is not the sole distribution determinant. Indeed, genome organization is an important component of fitness and may be differentially selected across ecological niches [6, 65]. Fungal MGCs, especially those enriched in plant pathogens, may contribute to pathogen fitness, as recent reports suggest that degradation of host defense compounds contributes quantitatively to plant pathogen virulence and reproduction [14, 66]. MGCs may also confer fitness benefits to saprotrophs by enabling the degradation of phenylpropanoids that inhibit fungal colonization of organic matter [17]. Our exploratory analysis indeed suggests that the presence of certain candidate MGCs is associated with ecological lifestyle (Fig. 1), and that different cluster models may be specialized for phenylpropanoids encountered in different environments, given that models from the same cluster class are rarely enriched in fungi with the same ecology (Supplementary Figure 4).

Cluster model co-occurrences may reflect simultaneous selection by multiple plant metabolites

The modularity of the structural network of clustered homolog groups suggests that different cluster classes carry out semi-independent functions [67], and may reflect differences in the phytochemicals they target (Fig. 3). Minimally, the structuring of clusters as independent modules with little overlap in content indicates that cluster classes have largely evolved independently of one another, and that selection for genome organization is likely to have occurred multiple times.

Recurrent patterns of cluster combinations among fungi from different taxonomic classes (Fig. 2) raise the intriguing possibility that conserved cluster repertoires may be important for determining ecological phenotypes, similar to how combinations of pathogenicity islands in bacteria determine host range [68]. Plants typically do not produce a single chemical in isolation; rather, defense metabolites are released as complex mixtures [69]. Some MCMPs may thus reflect the compositions of SM mixtures encountered during the colonization of specific hosts, or additionally, temporal differences in defense compound pressure [66]. A detailed

investigation of degradative MGCs from fungi colonizing the same host would be ideal for testing this hypothesis.

Do gene clusters bridge plant and fungal metabolisms?

The observation that some homolog groups are clustered with diverse sets of genes, often from different cluster classes (Fig. 3), implies they have been repeatedly recruited to different catabolic processes and may encode key strategies for phenylpropanoid degradation in fungi. Notably, some of the shared homolog groups, like those with members encoding cytochrome P450 enzymes (see groups ALL1024, ALL1032, ALL1115 in Supplementary Table 10) and multi-drug transporters (ALL1122, ALL1127), are known to be associated with evolutionary adaptability and underlie many fungal detoxification strategies [19, 44]. Intriguingly, cellular processes associated with other shared homolog groups suggest selection to integrate phenylpropanoid degradation products into central fungal metabolism, for example, carbohydrate transport (MFS monocarboxylate transporter: ALL1065; Sugar transporter: ALL1071, ALL1103), and aromatic amino-acid biosynthesis (3-dehydroshikimate dehydratase: ALL1002) and degradation (fumarylacetoacetate hydrolase: ALL1055). Combinations of both evolvable and core genes in MGCs may ultimately serve to connect diversifying chemical environments to stable metabolic networks, acting as bridges between specialized plant metabolism and central fungal metabolism.

The prevalence of candidate phenylpropanoid-degrading MGCs in plant-associated fungi suggests that specialized plant metabolism is a strong source of selection on fungal genome organization. Based on their distribution, we hypothesize that the MGCs detected here encode selectable phenotypes promoting the colonization of plant substrates by fungi. Further functional characterization of these loci will serve to accelerate discovery of metabolic processes of great interest not only for lignocellulosic biofuel production, but also for our understanding of the evolutionary dynamics between genome organization and ecological adaptation.

Data and code availability

Supplementary information is available at the ISME journal's website. Unless noted otherwise, all analyses were performed using a series of custom Perl scripts (Supplementary Figure 1). All scripts used to generate the presented results are available upon request, as are the results themselves. Fasta files of all recovered cluster model summaries, as well as a script to recover cluster models identified in this analysis from a user-supplied genome of interest, are

available for download at https://github.com/egluckthaler/cluster_retrieve.

Acknowledgements This work was supported by funds from the Ohio Agricultural Research and Development Center at the Ohio State University (EG-T, JCS), the National Science Foundation (DEB-1638999, JCS), and the Fonds de recherche du Québec-Nature et technologies (EG-T). All computational work was conducted on the Ohio State Supercomputer. We are grateful to Adrian Tsang, Alexey Grum-Grzhimaylo, Amy Jo Powell, Andrii Gryganskyi, Angel T. Martinez, Burt H. Bluhm, Christian Kubicek, Colleen Hansel, Daniele Armaleo, Daniel L. Linder, David Ezra, David S. Hibbett, Don Natvig, Francis Michel Martin, Francisco Javier Ruiz Dueñas, Francois Lutzoni, Gabor M. Kovacs, Gerald Bills, Gillian Turgeon, Gregory Bonito, Igor Grigoriev, Irina S. Druzhinina, John W. Taylor, Jon Karl Magnuson, José María Barrasa, Joseph Spatafora, Kathryn Bushley, Kerry O'Donnell, Laurie Connell, Marie-Noëlle Rosso, M. Catherine Aime, Patrik Inderbitzin, Paul Dyer, Pedro Crous, Robin Ohm, Scott E. Baker, Stephen B. Goodwin, Tom Bruns, and Trey Sato for providing access to the unpublished data produced by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and the 1000 Fungal Genomes Project.

Author contributions EG-T and JCS conceptualized the work and wrote the manuscript. EG-T developed and performed all analyses.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Baquero F. From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nat Rev Microbiol.* 2004;2:510–8.
2. Al-Shahrour F, Minguez P, Marqués-Bonet T, Gazave E, Navarro A, Dopazo J. Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS Comput Biol.* 2010;6:e1000953.
3. Hurst LD, Williams EJ, Pal C. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* 2002;18:604–6.
4. McGary KL, Slot JC, Rokas A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proc Natl Acad Sci USA.* 2013;110:11481–6.
5. Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW. Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *Populus trichocarpa*. *New Phytol.* 2016;209:1240–51.
6. Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci USA.* 2013;110: E1743–E1751.
7. Lynch M. The origins of genome architecture. Sunderland, MA: Sinauer Associates; 2007.
8. Peay KG, Kennedy PG, Talbot JM. Dimensions of biodiversity in the Earth mycobiome. *Nat Rev Microbiol.* 2016;14:434–47.
9. Wisecaver JH, Slot JC, Rokas A. The evolution of fungal metabolic pathways. *PLoS Genet.* 2014;10:e1004816.
10. Dhillon B, Feau N, Aerts AL, Beauseigle S, Bernier L, Copeland A, et al. Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. *Proc Natl Acad Sci USA.* 2015;112:3451–6.
11. Glenn AE, Davis CB, Gao M, Gold SE, Mitchell TR, Proctor RH, et al. Two horizontally transferred xenobiotic resistance gene clusters associated with detoxification of benzoxazolinones by fusarium species. *PLoS ONE.* 2016;11:e0147486.
12. Greene GH, McGary KL, Rokas A, Slot JC. Ecology drives the distribution of specialized tyrosine metabolism modules in fungi. *Genome Biol Evol.* 2014;6:121–32.
13. Hoffmeister D, Keller NP. Natural products of filamentous fungi: enzymes, genes, and their regulation. *Nat Prod Rep.* 2007;24:393–416.
14. Kettle AJ, Batley J, Benfield AH, Manners JM, Kazan K, Gardiner DM. Degradation of the benzoxazolinone class of phytoalexins is important for virulence of *Fusarium pseudograminearum* towards wheat. *Mol Plant Pathol.* 2015;16:946–62.
15. Shanmugam V, Ronen M, Shalaby S, Larkov O, Rachamim Y, Hadar R, et al. The fungal pathogen *Cochliobolus heterostrophus* responds to maize phenolics: novel small molecule signals in a plant-fungal interaction. *Cell Microbiol.* 2010;12:1421–34.
16. Wang Y, Lim L, Madilao L, Lah L, Bohlmann J, Breuil C. Gene discovery for enzymes involved in limonene modification or utilization by the mountain pine beetle-associated pathogen *Grossmannia clavigera*. *Appl Environ Microbiol.* 2014;80:4566–76.
17. Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, et al. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science.* 2012;336:1715–9.
18. Jönsson LJ, Martín C. Pretreatment of lignocellulose: formation of inhibitory by-products and strategies for minimizing their effects. *Bioresour Technol.* 2016;199:103–12.
19. Mäkelä MR, Marinović M, Nousiainen P, Liwanag AJ, Benoit I, Sipilä J, et al. Aromatic metabolism of filamentous fungi in relation to the presence of aromatic compounds in plant biomass. *Adv Appl Microbiol.* 2015;91:63–137.
20. Weber T, Blin K, Duddela S, Krug D, Kim HU, Brucoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 2015;43: W237–W243.
21. Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, et al. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecol.* 2016;20:241–8.
22. Farr DF, Rossman AY. Fungal databases. U.S. National Fungus Collections, ARS, USDA: <https://nt.ars-grin.gov/fungalatabases/>; 2017.
23. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010;26:2460–1.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
25. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of orthology groups for eukaryotic genomes. *Genome Res.* 2003;13:2178–89.
26. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 2015;44: D286–93.
27. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7:e1002195.
28. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
29. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 2016;33:1635–8.

30. Wickham H. ggplot2: elegant graphics for data analysis. New York City, NY: Springer; 2016.
31. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. *Nat Genet*. 1999;21:108–10.
32. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 2004;20:289–90.
33. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–3.
34. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara B, et al. vegan: community ecology package. *R Package Version*. 2016;2:3–5.
35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–504.
36. Newman ME. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E*. 2006;74:036104.
37. Marquitti FMD, Guimarães PR, Pires MM, Bittencourt LF. MODULAR: software for the autonomous computation of modularity in large network sets. *Ecography*. 2014;37:221–4.
38. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26:53–77.
39. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;19:227.
40. Hane JK, Lowe RG, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, et al. Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell*. 2007;19:3347–68.
41. Shalaby S, Horwitz BA, Larkov O. Structure–activity relationships delineate how the maize pathogen *Cochliobolus heterostrophus* uses aromatic compounds as signals and metabolites. *Mol Plant Microbe Interact*. 2012;25:931–40.
42. Pigné S, Zykwska A, Janod E, Cuenot S, Kerkoud M, Raulo R, et al. A flavoprotein supports cell wall properties in the necrotrophic fungus *Alternaria brassicicola*. *Fungal Biol Biotechnol*. 2017;4:1.
43. Enkerli J, Bhatt G, Covert SF. Maackiain detoxification contributes to the virulence of *Nectria haematococca* MP VI on chickpea. *Mol Plant Microbe Interact*. 1998;11:317–26.
44. Lah L, Podobnik B, Novak M, Korošec B, Berne S, Vogelsang M, et al. The versatility of the fungal cytochrome P450 monooxygenase system is instrumental in xenobiotic detoxification. *Mol Microbiol*. 2011;81:1374–89.
45. Höhl B, Arnemann M, Schwenen L, Stöckl D, Bringmann G, Jansen J, et al. Degradation of the Pterocarpan Phytoalexin (–)-Maackiain by *Ascochyta rabiei*. *Z Naturforsch C*. 1989;44:771–6.
46. Bayry J, Beaussart A, Dufrêne YF, Sharma M, Bansal K, Kniemeyer O, et al. Surface structure characterization of *Aspergillus fumigatus* conidia mutated in the melanin synthesis pathway and their human cellular immune response. *Infect Immun*. 2014;82:3141–53.
47. Fowler ZL, Baron CM, Panepinto JC, Koffas MA. Melanization of flavonoids by fungal and bacterial laccases. *Yeast*. 2011;28:181–8.
48. Desantis-Mendoza RM, Hernández-Sánchez H, Moreno A, Rojas del CE, Chel-Guerrero L, Tamariz J, et al. Enzymatic polymerization of phenolic compounds using laccase and tyrosinase from *Ustilago maydis*. *Biomacromolecules*. 2006;7:1845–54.
49. Schmalzer-Ripcke J, Sugareva V, Gebhardt P, Winkler R, Kniemeyer O, Heinekamp T, et al. Production of pyomelanin, a second type of melanin, via the tyrosine degradation pathway in *Aspergillus fumigatus*. *Appl Environ Microbiol*. 2009;75:493–503.
50. Hane JK, Rouxel T, Howlett BJ, Kema GH, Goodwin SB, Oliver RP. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol*. 2011;12:R45.
51. Plissonneau C, Stürchler A, Croll D. The evolution of orphan regions in genomes of a fungal pathogen of wheat. *mBio*. 2016;7:e01231–01216.
52. Muto A, Kotera M, Tokimatsu T, Nakagawa Z, Goto S, Kanehisa M. Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *J Chem Inf Model*. 2013;53:613–22.
53. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRINGv10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2014;43:D447–52.
54. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *Elife*. 2014;3:e03275.
55. Shwab EK, Bok JW, Tribus M, Galehr J, Graessle S, Keller NP. Histone deacetylase activity regulates chemical diversity in *Aspergillus*. *Eukaryot Cell*. 2007;6:1656–64.
56. Davila Lopez M, Martinez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE*. 2010;5:e10654.
57. Tsochatzidou M, Malliarou M, Papanikolaou N, Roca J, Nikolaou C. Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 2017;45:5818–28.
58. Pepper JW. The evolution of evolvability in genetic linkage patterns. *Biosystems*. 2003;69:115–26.
59. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*. 1996;143:1843–60.
60. Slot JC, Hibbett DS. Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. *PLoS ONE*. 2007;2:e1097.
61. Andam CP, Gogarten JP. Biased gene transfer in microbial evolution. *Nat Rev Micro*. 2011;9:543–55.
62. Gluck-Thaler E, Slot JC. Dimensions of horizontal gene transfer in eukaryotic microbial pathogens. *PLoS Pathog*. 2015;11:e1005156.
63. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*. 2011;480:241–4.
64. Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *ISME J* 2017; 11:1189–1204.
65. Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. *Genetics*. 2006;173:419–34.
66. Hammerbacher A, Schmidt A, Wadke N, Wright LP, Schneider B, Bohlmann J, et al. A common fungal associate of the spruce bark beetle metabolizes the stilbene defenses of Norway spruce. *Plant Physiol*. 2013;162:1324–36.
67. Wagner GP. Homologues, natural kinds and the evolution of modularity. *Am Zool*. 1996;36:36–43.
68. Bouyioukos C, Reverchon S, Képès F. From multiple pathogenicity islands to a unique organized pathogenicity archipelago. *Sci Rep*. 2016;6:27978.
69. Gershenzon J, Fontana A, Burow M, Wittstock U, Degenhardt J. Mixtures of plant secondary metabolites: metabolic origins and ecological benefits. In: Iason GR, Dicke M, Hartley SE, editors. *The ecology of plant secondary metabolites: from genes to global processes*. Cambridge: Cambridge University Press; 2012. p. 56–77.