Published on Mar 05, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-9768-f6c5

A structurally divergent actin conserved in fungi has no association with specific traits

We outline a comparative approach to investigate protein function by correlating the presence or absence of a protein with species-level phenotypes. We applied this strategy to a novel actin isoform in fungi but didn't find an association with any of the phenotypes we considered.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Brae M. Bigge, Megan L. Hochstrasser, Ilya Kolb, David G. Mets, Manon Morin, Austin H. Patton, Taylor Reiter, Dennis A. Sun, Ryan York

Version 2 · Mar 31, 2025

Purpose

We were curious to see if phylogenetic trait mapping might be a reliable way to uncover the function of structural variants of actin that we identify via our ProteinCartography pipeline **[1]**. ProteinCartography leverages recent advances in protein folding prediction **[2]** to identify structurally similar proteins, independent of their sequence similarity. Actin is an ancient and highly conserved protein in

eukaryotes and is essential to multiple cellular processes. In previous work **[3]**, we identified a set of actin proteins that are present in a large number of fungi yet are structurally distinct from the primary cytoskeletal actin, suggesting these proteins may serve a different function.

We wondered if the presence or absence of these non-canonical, divergent fungal actins (DFAs) correlates, across species, with biologically relevant fungal traits. A strong correlation would suggest that this actin isoform is related to a given trait, potentially suggesting a novel structure-function relationship within this protein family. We identified six fungal traits, available in public databases, that we thought DFAs might influence. However, we found that none of these traits predicted the presence of a DFA.

While we decided not to continue this project, we believe it could spark interest in many audiences (e.g., fungal ecologists, evolutionary biologists, cell biologists). At the end of this pub, we discuss potential follow-up directions for anyone interested in studying DFAs.

- This pub is part of the **platform effort**, "Annotation: Mapping the functional landscape of protein families across biology." Visit the platform narrative for more background and context.
- **Data**, including the inputs and outputs from our ProteinCartography run, are available on Zenodo.
- All associated **code**, plus lists of divergent actins, associated species, and trait information, is available in this GitHub repository.

Background and goals

Actins are some of the most conserved proteins among eukaryotes and support essential functions including cell division, cellular trafficking, cell shape, and motility **[4]**. In fungi, primary actin is known to be essential to many cellular processes (apical growth, endocytosis, exocytosis, cellular trafficking, cytokinesis, and possibly pathogenicity in pathogenic species) **[5]**. While investigating the structural similarity of

actin, actin-like proteins, and actin-related proteins with ProteinCartography (a tool for clustering structurally similar proteins across diverse organisms **[1]**), our functional annotation team identified a well-defined and distinct cluster that contained around 290 proteins **[3]** (Figure 1). The vast majority of the proteins in this cluster are fungal, annotated as Actin-2 or actin-like proteins, and are found in species that also possess another, structurally canonical actin (Figure 1). We therefore refer to these as “divergent actins.”

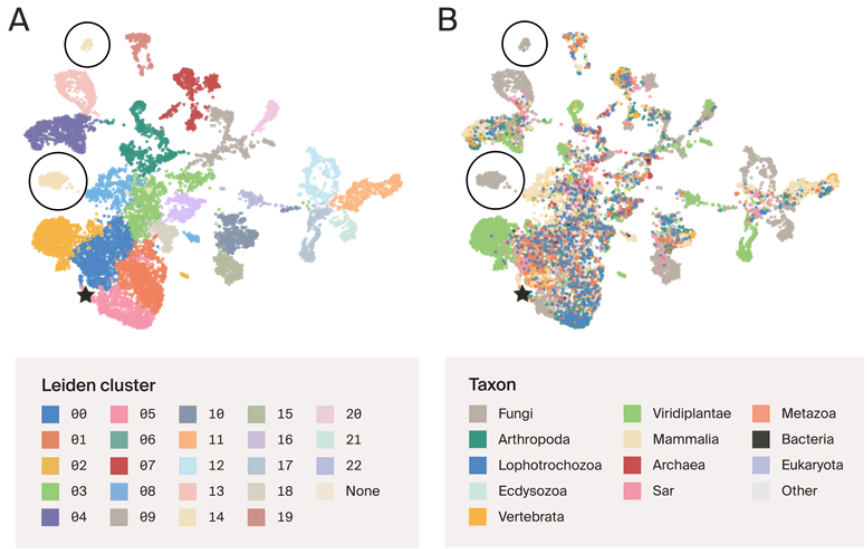

Figure 1

UMAP plot for the human cytoplasmic actin (ACTB).

(A) Cluster overlay. Leiden cluster identity (LC number) is indicated by color for each of the proteins in the study.

(B) Broad taxon overlay. Color indicates the taxon to which each protein belongs.

The black circles indicate the cluster (LC14) that contains the divergent fungal actins. The star represents the human actin structure we used to seed the ProteinCartography run.

It’s not rare for organisms to possess multiple actin isoforms (for instance, humans have six nearly identical actin isoforms **[6]** and *Arabidopsis thaliana* has at least 10 isoforms **[7]**). However, some species, like the malaria-causing parasite *Plasmodium*,

have structurally divergent isoforms known to have functions that are distinct from their canonical isoform **[8][9]**.

Identifying a class of structurally similar actin isoforms that diverge from canonical actin and are present in more than 200 fungal species raises a question — what function(s) do these divergent actins perform in fungi? The proteins in this cluster of divergent actins have conserved ATP-binding residues, but the residues required for polymerization are not well-conserved **[1]**. These residues are important for the biochemical functions of actin and contribute to the overall role that the protein plays in the cell. We wondered whether these divergent actins have an uncharacterized function or role required by some shared biological feature of the fungi that possess them. Thus, we sought to identify biologically relevant fungal traits that predicted the presence or absence of these divergent actins within species, a pattern that would hint at the function of these actins. To do so, we tested for statistical associations between the presence or absence of a divergent actin and each selected phenotype using the workflow outlined in Figure 2 (and detailed in the next section, "The approach"). Ultimately, we didn't identify any correlations between the divergent actin and these traits. Thus, the function of these actins remains mysterious (described in "The results"), but we hope our trait-mapping strategy offers a useful approach for future functional annotation efforts or that others in the community with a particular interest or expertise in this space can make additional progress.

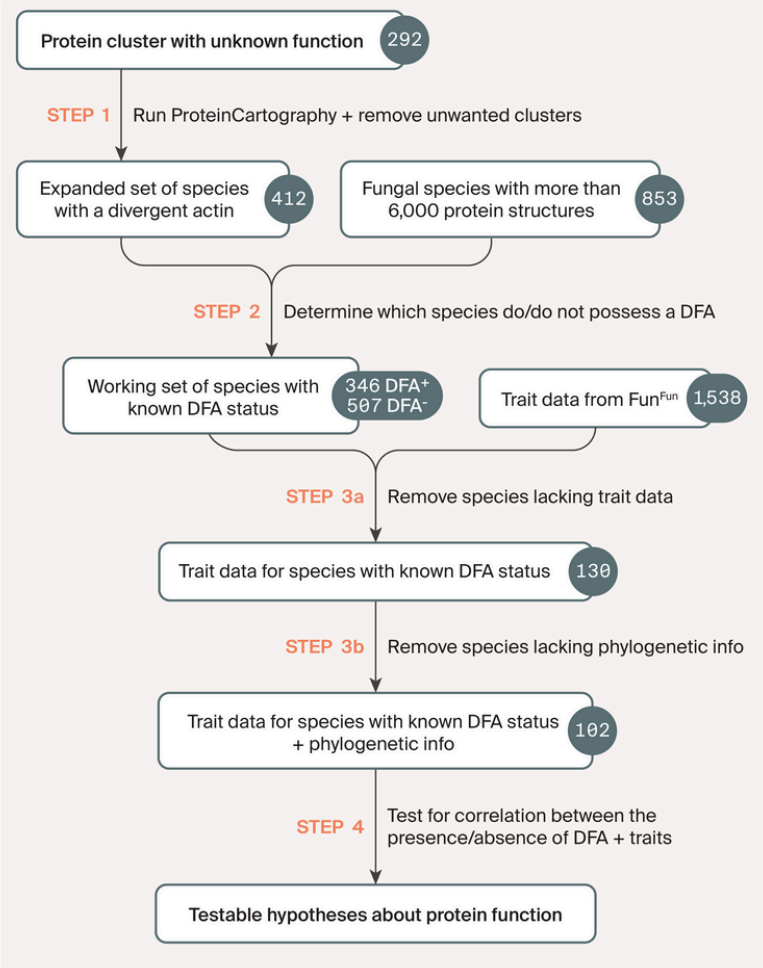

Figure 2

Workflow for generating hypotheses about protein function using ProteinCartography.

Step 1: Enrich the initial set of divergent fungal actins

Step 2: Identify a working set of fungal species with known DFA status

Step 3: Curate trait data

Step 4: Statistically model the association of DFAs and fungal traits

The approach

To investigate the functions of these divergent fungal actins (DFAs) **[3]**, we decided to test the association of a trait and the presence or absence of DFAs to generate hypotheses about their role(s). For example, if all of the fungal species that possess a DFA also possess a specific spore-bearing structure, we might guess that DFA is involved in spore storage and/or release. To be successful, we'd need both trait information and genomic information about the presence or absence of DFA across as many species as possible.

Our approach consisted of four main steps (Figure 2). First, we expanded the set of fungal species in our analysis by running a new ProteinCartography analysis focused on these divergent actins and removing non-fungal species. While this allowed us to confidently identify fungal species that possess a divergent actin, it was also necessary to be able to confidently identify fungal species that don't possess one. Therefore, in step two, we defined our working set of species: the set of fungal species for which we could determine whether or not they possess a DFA (for details on how we determined the presence or absence of a DFA, jump to the section, "Identifying a working set of fungal species"). Third, we curated public fungal databases to gather trait and phylogenetic information for as many species as possible in our working set. The last step then consisted of running statistical models to test for the correlation between the presence or absence of the DFA and six different fungal traits: growth form, trophic mode, ascus dehiscence, presence of an auxin-responsive promoter, spore length, and spore width.

We discuss each of these four steps below. Keep reading or skip straight to the results.

1) Enriching the initial set of divergent fungal actins

We identified six representative divergent actins from an initial ProteinCartography run (available on Zenodo in "actin_older_version.zip"). We then performed a single ProteinCartography analysis with these six proteins as the input to capture as many structurally similar DFAs as possible.

Clustering the original set of divergent actins and selecting representatives

We first identified divergent fungal actins when we ran human ß-actin (UniProt ID: P60709) through ProteinCartography and noticed a cluster, LC14, that was distinct within the map and mostly contained fungal proteins **[3]** (note that this original run used ProteinCartography version v0.4.0-alpha, available on Zenodo). In this work, we clustered all 292 protein sequences from cluster LC14 using MMseqs2 (version 14.7e284) and the clustering module **[10][11]**. This generated six clusters with sizes ranging from one sequence to 281 sequences. From each cluster, we extracted the longest sequence as the representative sequence (cluster 1: A0A401L4A6, cluster 2: A0A0C9N219, cluster 3: A0A2N1JBK3, cluster 4: A0A5B0SCN5, cluster 5: A0A226D8X1, cluster 6: A0A7J6TT41).

> All associated **code** and **related files** are available in our GitHub repository (DOI: 10.5281/zenodo.10779267).

Running ProteinCartography

We aimed to expand the existing LC14 cluster by running ProteinCartography (version v0.4.0-alpha) on our six representative proteins listed above. We used each of the six divergent fungal actins as inputs for “search mode” in the pipeline. Full details on the ProteinCartography pipeline can be found in the associated GitHub repository and pub.

Briefly, ProteinCartography “search mode” starts with an input protein(s) and searches for proteins with either similar sequences using BLAST **[12]**, or structures using Foldseek **[13]**. The pipeline downloads all available structures from the AlphaFold database and compares every downloaded structure to every other downloaded structure, creating an all-v-all matrix of structural similarity scores **[13][2][14]**. The pipeline then uses Leiden clustering on this similarity matrix to group these proteins **[15]**. In our ProteinCartography analysis, we used “search mode” with standard parameters on these six divergent actins **[1]**. We requested 3,000 Foldseek hits per input protein and 6,000 total proteins per input. The run generated 3,596 unique structure hits grouped into 17 clusters.

ProteinCartography compares pairs of protein structures using the TM-align algorithm **[13]** to calculate their structural similarity **[1]**. This comparison yields a TM-score (template modeling score) between zero and one. A TM-score above 0.5 suggests structural similarity, while a score below 0.17 indicates unrelated proteins. For a given protein cluster, the "cluster compactness" score reflects the average TM-score for all pairs of compared proteins within the cluster. Increasing “cluster compactness” scores (on the diagonal of the similarity matrix (Figure 3, B)) indicates increasing similarity within a cluster. The average cluster compactness (average of the diagonal) indicates how well protein structures have been sorted, and thus represents the overall quality of the results. In previous work **[1]**, 25 different runs of ProteinCartography yielded cluster compactness scores ranging from 0.35–0.86. Considering this range, we consider that the average cluster compactness of our run, 0.6, is a reasonable score, underlying an overall useful clustering of the proteins. For this study specifically, we considered any cluster whose compactness is greater than 0.6 to be “well-defined.” We identified eight well-defined clusters: LC01, LC03, LC04, LC10, LC11, LC12, LC14, and LC15.

> The **ProteinCartography inputs and outputs** are available on Zenodo (DOI: 10.5281/zenodo.10211653).

Defining the extended set of divergent actin proteins

We identified two clusters that contained the divergent actin structures used as input, LC04 and LC11, representing a total of 407 proteins. We then combined this set of proteins with cluster LC14 from the original human actin ProteinCartography analysis and obtained an extended set of structurally similar actin proteins containing 436 proteins, spanning 412 strains.

Taxonomic analysis of the extended set of divergent actins and selection of the fungal divergent actin set

For each protein that ProteinCartography identifies, it returns a set of metadata, including the organism in which the protein is found and the associated information on taxonomy or lineage.

For each protein in our extended set of divergent actins, we determined the kingdom, phylum, and order of its species. As some proteins belong to organisms that do not have a kingdom reported in UniProt, we manually curated them and added corresponding clade information instead. This includes Discoba, SAR, Amoebozoa, and Opisthokonta.

We removed all proteins associated with kingdoms other than fungi, leaving us with 406 DFA proteins.

These 406 DFAs were present in a total of 385 unique strains. Among them, 16 strains contained two or more DFA hits: one strain with six DFA hits, one strain with three DFA hits, and 14 strains with two DFA hits. We aimed to verify whether these strains really possess multiple DFAs in their genomes or if this is an artifact of inaccurate protein annotation or low genome sequencing and assembly quality. For half of the strains, a single protein sequence had been annotated by different groups and thus resulted in multiple entries into the PDB. In these cases there was clearly only one DFA in the species. For the other strains, protein-to-nucleotide BLAST (tBLASTn) alignments failed to identify discrete genomic locations. We believe this could be because of low genome sequencing coverage and low-quality genome assembly. Nevertheless, the great majority (≥ 95%) of the fungal species associated with divergent actin seem to possess only one DFA in their genome.

2) Identifying a working set of fungal species

To test for any correlation between fungal traits and DFAs, we needed to establish a "working set" of species where we confidently knew the presence or absence of DFAs. While ProteinCartography allowed us to expand the set of species in which we knew a DFA was present, we had to identify other fungal species from which DFAs were absent.

There are two possible reasons a species was not present in the output of ProteinCartography: 1) the species encodes the protein but that information was not available in UniProt or the AlphaFold database, and 2) the species truly does not have a DFA. Studies have shown that some fungi have as few as 6,000 proteins and a typical fungal genome contains 10,000 protein-coding genes **[16][17]**. We considered DFAs to be absent in any species that didn't have a DFA hit if that species also had more than 6,000 proteins in UniProt. Our selection criteria are liberal and are likely to cause false

negative errors where we determine DFA to be absent when it is actually present. This is particularly true for those fungal species that possess large numbers of proteins (i.e. ≫ 6,000 proteins). That is, we likely will have underestimated the prevalence of these DFAs across the fungal tree of life for species with typical fungal genome sizes (i.e. ~10,000 genes **[18]**, and thus > 10,000 proteins), a fact that may have limited our ability to recover DFA–trait associations.

To identify the fungal species with 6,000 or more protein structures in the UniProtKB and AlphaFold databases, we first conducted an advanced search in UniProt using the following query: “Fungi” in the “Taxonomy” field and “*” for the field “AlphaFoldDB cross-reference” (found within the “Cross reference/3D structure” field), to obtain all the fungal proteins with available structures in AlphaFold. We then counted the number of proteins per fungal species from this search. Finally, after filtering for fungal species that have more than 6,000 proteins with available structures, we obtained their taxonomic classification from NCBI. This yielded 853 total fungal species. Among them, 346 species were also present in our extended set of species that possess a DFA (41%) and the 507 remaining species don’t possess a DFA (59%).

To assess whether our 6,000-protein threshold introduced a sampling bias (independent of taxonomy), we varied the count threshold from 6,000 proteins to 25,000 proteins and compared the proportions of species with and without a DFA. We found that the ratio of species with vs. without a DFA does not drastically change across this threshold range (ratio for a threshold at 10,000 proteins: 40%:60%; ratio for a threshold at 20,000 proteins: 44%:56%; ratio for a threshold at 25,000 proteins: 42%:58%).

> The **list of all fungal proteins and structures available in UniProt** is available on Zenodo (10.5281/zenodo.10211653).

Distribution of divergent fungal actins in the fungal kingdom

We obtained the phylogenetic relationships of the fungal orders represented in our working set of species from the TimeTree database’s web interface (timetree.org; **[19]** (Figure 4). The resulting tree represented 85 fungal orders. We next investigated the

distribution of DFAs in the fungal kingdom by calculating and visualizing the distribution of DFAs at the order level.

We were able to recover the order for 783 of the 853 species. For each order, we calculated the fraction of associated species that possess a DFA and mapped this information onto the tree (Figure 4, B).

3) Curating trait data

We used the database Fun^{Fun} as the source of fungal trait information **[20]**. This database contains a large amount of species-level information compiled from different studies. In addition to FUNGuild information (classification of fungi based on their ecological function and classification of fungi based on their trophic mode) **[21]**, it includes ecological, cellular, and biochemical traits.

We decided to focus on six traits: growth form, trophic mode, ascus dehiscence, auxin-responsive promoter, spore length, and spore width. We chose these traits specifically to maximize the overlap between the species for which we could obtain trait information and for which we could determine DFA status, and to include biological features for which actin was relevant. We extracted information on these traits for the species present in the database that were also in our working set.

A total of 143 species from our working set had information for at least one of the six selected traits in Fun^{Fun}. Of these species, 36 had multiple strains in the ProteinCartography DFA dataset. However, we do not have trait information for individual strains, just species. For 23 of these species, a DFA was present in all of the strains. For the 13 species where DFA status varied across individual strains, we attempted to determine whether this variation across strains resulted from real biology or was caused by some bioinformatic error — e.g., a strain was incorrectly identified as not possessing a DFA when it actually did. For all the strains that don't possess a DFA, we conducted a protein BLAST (BLASTp) search in NCBI as well as a protein-to-nucleotide BLAST (tBLASTn) to identify whether there was evidence that a DFA was encoded in the genome of the strain. However, these attempts proved uninterpretable and the variation in DFA status across strains may have resulted from undersampling the genetic material from some of these species and noisy assembly data. We thus removed these 13 species from the study. Intersecting the remaining species with

those in our phylogeny led to the removal of an additional 28 species not present in TimeTree.

Altogether, we were able to collect DFA status, trait information, and phylogenetic relationship information for a total of 102 species.

> All associated **code** and **related files** are available in our GitHub repository.

4) Statistical modeling of the association of DFAs and fungal traits

To test whether each of our six traits predicted the presence of DFAs, we applied several statistical models, including generalized linear models for continuous traits and discrete-state Markov models for categorical/binary traits. These approaches are described in more detail below.

For discrete traits, we used a model selection approach comparing the likelihoods of two models: one where the evolutionary trajectory of DFA (i.e., its presence/absence in any given species across evolutionary time) and the similar trajectory of another trait are the same, and a second model where DFA and the trait of interest evolved independently. For continuous traits, we estimated the portion of variation in the presence or absence of a DFA that can be accounted for by variation in the trait of interest while controlling for shared evolutionary history. For a summary of the input data, see Table 1.

Trait	Data type	Number of categories with ≥ 4 species	Number of species
Growth form	Discrete	3 (agaricoid, microfungus, yeast)	24
Trophic mode	Discrete	3 (saprotroph, pathotroph, symbiotroph)	63
Ascus dehiscence	Discrete	2 (deliquescent, poricidal)	13
Auxin-responsive promoter	Discrete	2 (present/absent)	71
Spore length	Continuous	–	10
Spore width	Continuous	–	10

Table 1

Description of the data used for statistical modeling of DFA presence/absence and fungal traits.

Testing the association of DFAs with discrete or binary traits

We re-defined categorical trait data from the Fun[Fun] database to maximize the number of categories containing four or more species, as categories with fewer than four species would not have enough data to accurately model the association between DFA status and the trait:

- For “growth form,” we collapsed the categories “yeast” and “facultative yeast” into a single level: “yeast.” We removed the categories ergot, cordyceptoid, rust and xylaroid.
- For “trophic mode,” we defined three levels: “saprotroph,” “pathotroph,” and “symbiotrioph,” and parsed any species with multiple trophic modes into each individual mode (for instance, if a species was labeled as “saprotroph-pathotroph,” we counted it as “saprotroph” and “pathotroph”).
- For “ascus dehiscence,” we removed the categories fissitunicate and rostrate.
- For “auxin-responsive promoter,” we transformed the number of auxin-responsive promoters into a simple binary variable: presence or absence of promoters.

To determine whether DFA status and a discrete trait are associated, we used an evolutionary model selection procedure. As mentioned above, we fit two classes of models to the data: a “correlated” model in which we assumed the evolution of DFA presence/absence correlates with the trait of interest and an “independent” model where we assumed a DFA and the trait of interest evolved independently. We then compared the likelihood of these models using the Akaike information criterion (AIC), a measure of likelihood that penalizes for model complexity. Under this paradigm, if the correlated model was more likely, we would take this as evidence that the evolution of DFA could be explained in part by the trait of interest, and conversely, if the independent evolutionary model was more likely, it would suggest that DFA and that particular trait evolved independently.

We used this model selection procedure for two classes of models, a discrete-time Markov model (DTMM) and a hidden Markov model (HMM), both commonly used for modeling the evolution of discrete traits over time **[22]**. DTMMs assume that the evolutionary rate of change for a trait is constant independent of the state of that trait. For example, the probability that a DFA will be lost as a function of evolutionary time is the same as the probability that a DFA will be gained in that same amount of time. Alternatively, HMMs allow for multiple evolutionary rates dependent on the current trait status (e.g., DFA presence or absence). Our HMMs allowed for two different evolutionary rates for each observed trait status.

Altogether, using the R corHMM package (version 2.8) **[22]**, we fit four models for each trait: DTMM with assumed independent evolution of DFA and trait (labeled as “independent_model_fit” in the package output), DTMM with assumed correlated evolution of DFA and trait (labeled as “correlated_model_fit” in the corHMM package output), HMM with assumed independent evolution of DFA and trait (labeled as “hidden_Markov_independent_model_fit” in the package output), HMM with assumed correlated evolution of DFA and trait (labeled as “hidden_Markov_correlated_model_fit” in the package output).

Testing the association between DFA and continuously variable traits

We evaluated the correlation between DFA presence with continuously variable traits (e.g. spore size) using phylogeny-corrected generalized linear mixed models (pglmm). Specifically, the pglmm_compare function from the R package phyr (version 1.1.2) **[23]**.

These models test whether variation in the trait (i.e., the predictor variable) can account for variation in DFA status while controlling for the evolutionary non-independence among species due to their shared evolutionary history. Specifically, they implement a linear model (a logistic regression) to determine whether changes in the continuous predictor trait account for the presence or absence of a DFA. The model equation is typically structured as follows:

$$logit(P(DFA = 1)) = \beta 0 + \beta 1 * Trait + Zu + \epsilon$$

Where:

- logit(*P*(*DFA*=1)) is the logit transformation of the probability that DFA equals one (i.e., the probability that DFA is present in a species). The logit link function is used to model the relationship between the probability of the binary outcome and the continuous predictor, ensuring that the predicted probabilities lie between zero and one.
- $\beta 0$ is the intercept: the predicted log odds of the DFA outcome when the continuous trait is at zero.
- $\beta 1$ (or slope) is the unknown coefficient for the continuous trait indicating the effect size of the trait on the log odds of DFA being one.
- *Trait* is the known vector of continuous trait values (e.g., spore length or spore width).
- *Z* is the known evolutionary variance-covariance matrix capturing the average relatedness among species. It represents the random effects due to phylogenetic relatedness among observations, capturing the unobserved phylogenetic variance.
- u is the vector of unknown coefficients on the Z matrix.
- ϵ is the residual error term.

To evaluate whether a given continuous fungal trait is a predictor of DFA status, we focused on the coefficient for the continuous trait (or slope $\beta 1$) that a fitted pglmm returns. Any slope that is significantly different from zero indicates that changes in trait values change the probability of the DFA outcome, indicating that, to some degree, the continuous trait is a predictor of DFA status.

All **code** we generated and used in this pub is available in our GitHub repository, including notebooks for the analysis of the ProteinCartography run (filtering of the extended set and its phylogenetic analysis), the definition of the working set of species and their DFA status, the analysis of the DFA distribution within fungal orders, the curation of trait information, and the statistical analysis of DFA–trait correlation.

Additional methods

We used ChatGPT to help write some code.

The results

SHOW ME THE DATA: You can find the inputs and outputs from our ProteinCartography run on Zenodo and lists of divergent actins, associated species, and trait information on GitHub.

ProteinCartography identifies clusters of divergent actins

We expected the initial set of divergent actins identified in our original work to be incomplete. Thus, we first aimed to look for other proteins that are structurally similar to our proteins of interest using ProteinCartography.

We identified six representative divergent actins to seed ProteinCartography, which generated 3,596 unique hits grouped into 17 clusters (Figure 3, A), eight of which were well-defined (LC01, LC03, LC04, LC10, LC11, LC12, LC14, and LC15 — Figure 3, A and B). These clusters contain hits from three main kingdoms: Metazoa, Fungi, and Viridiplantae (Figure 3, C). Semantic analysis shows that they are mainly associated with the actin family, and they contain proteins with similar length distribution. Together, these findings indicate that the well-defined clusters contain proteins that

belong to the actin protein family but are sufficiently structurally different to cluster separately, suggesting that these are structurally distinct isoforms.

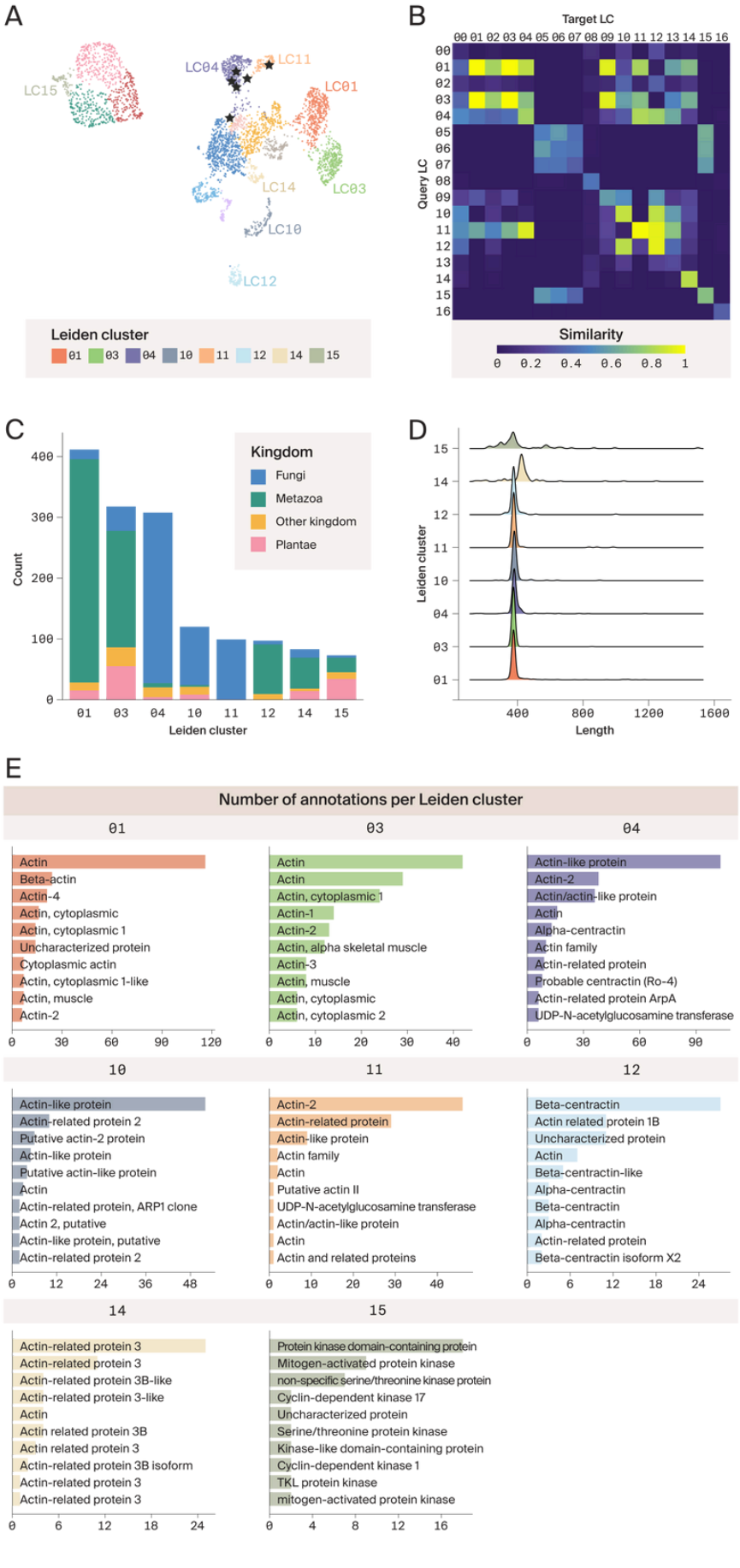
A
LC15
LC04
LC11
LC01
LC14
LC03
LC10
LC12
Leiden cluster
01 03 04 10 11 12 14 15
B
Target LC
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Query LC
Similarity
0 0.2 0.4 0.6 0.8 1
C
Kingdom
Fungi
Metazoa
Other kingdom
Plantae
Count
Leiden cluster
D
Leiden cluster
Length
E
Number of annotations per Leiden cluster
01
Actin
Beta-actin
Actin-4
Actin, cytoplasmic
Actin, cytoplasmic 1
Uncharacterized protein
Cytoplasmic actin
Actin, cytoplasmic 1-like
Actin, muscle
Actin-2
03
Actin
Actin
Actin, cytoplasmic 1
Actin-1
Actin-2
Actin, alpha skeletal muscle
Actin-3
Actin, muscle
Actin, cytoplasmic
Actin, cytoplasmic 2
04
Actin-like protein
Actin-2
Actin/actin-like protein
Actin
Alpha-centractin
Actin family
Actin-related protein
Probable centractin (Ro-4)
Actin-related protein ArpA
UDP-N-acetylglucosamine transferase
10
Actin-like protein
Actin-related protein 2
Putative actin-2 protein
Actin-like protein
Putative actin-like protein
Actin
Actin-related protein, ARP1 clone
Actin 2, putative
Actin-like protein, putative
Actin-related protein 2
11
Actin-2
Actin-related protein
Actin-like protein
Actin family
Actin
Putative actin II
UDP-N-acetylglucosamine transferase
Actin/actin-like protein
Actin
Actin and related proteins
12
Beta-centractin
Actin related protein 1B
Uncharacterized protein
Actin
Beta-centractin-like
Alpha-centractin
Beta-centractin
Alpha-centractin
Actin-related protein
Beta-centractin isoform X2
14
Actin-related protein 3
Actin-related protein 3
Actin-related protein 3B-like
Actin-related protein 3-like
Actin
Actin related protein 3B
Actin related protein 3
Actin-related protein 3B isoform
Actin-related protein 3
Actin-related protein 3
15
Protein kinase domain-containing protein
Mitogen-activated protein kinase
non-specific serine/threonine kinase protein
Cyclin-dependent kinase 17
Uncharacterized protein
Serine/threonine protein kinase
Kinase-like domain-containing protein
Cyclin-dependent kinase 1
TKL protein kinase
mitogen-activated protein kinase

Figure 3

ProteinCartography analysis of six representatives of the divergent actin.

(A) UMAP of the ProteinCartography clustering output with cluster identity indicated by color. Black stars indicate the six proteins that were the input.

(B) Similarity matrix for the clustering of the divergent actins. For each cluster pair, we calculated the mean TM-score of the structures in a cluster vs. structures of proteins in the other cluster.

(C) Kingdom distribution of the proteins within clusters.

(D) Distribution of protein lengths within clusters.

(E) Semantic analysis of keywords describing proteins in each cluster.

We next examined the proteins that co-clustered with our representative divergent actin proteins. The representative divergent actin proteins fell into two well-defined (high within-cluster compactness score in similarity matrix; Figure 3, B) clusters, LC04 and LC11. Proteins in both clusters are largely fungal and are annotated as “Actin-like protein” (Figure 3, C and E). Therefore, we considered any protein in these two clusters to be a divergent actin similar to the divergent actins used in this search, which inspired this project. Altogether, clusters LC04 and LC11 represent 407 proteins, 144 of which were not part of the original set of divergent actins, and they span 139 additional strains and species. Combining the original set and the new hits generated an extended set of 436 divergent actins spanning 412 strains.

The extended set of divergent actins still contains mainly fungal proteins

What caught our attention in the original set of divergent actins was the fact that nearly all (285/292) are fungal proteins. We analyzed the kingdom or clade distribution (as defined by NCBI Taxonomy when kingdom rank was not available) for the proteins in

the extended set of divergent actins (Figure 4, A) to see if we were still looking at mostly fungal proteins. While the percentage of non-fungal proteins is higher, more than 93% of the proteins are found in fungal species. The second-most represented kingdom is Metazoa, which represents just 2% of the proteins. This confirms that these divergent actins are mostly found in fungi. We therefore refer to them as divergent fungal actins (DFAs). Additionally, most of the fungi seem to possess only one divergent actin in their genome, suggesting that there is usually only one DFA per species (in addition to a more conserved primary actin).

The distribution of DFAs across species is highly variable

We next investigated the distribution of DFA within the fungal kingdom. We examined how consistently DFAs are present in orders or phyla and if they were gained and lost frequently across the fungal tree. The latter is a characteristic pattern of an evolutionarily labile trait (in contrast to a conserved trait). The distribution of DFA across species in the fungal kingdom will indicate whether DFA is associated with fundamental, conserved traits or if it is more evolutionarily labile and potentially important for adaptive responses to the environment.

We started by determining a working set of fungal species for which we could reliably determine whether a DFA is present or absent (see "The approach"). This working set is composed of 853 fungal species: 346 species that possess a DFA (these are from the extended set of divergent actin species) and 507 species that don't possess a DFA. These species span eight fungal phyla: Ascomycota (611 species), Basidiomycota (186 species), Mucoromycota (30 species), Blastocladiomycota (two species), Chytridiomycota (16 species), Zoopagomycota (13 species), Microsporidia (two species), and Cryptomycota (one species). We visualized the phylogeny of fungal orders and mapped the fraction of species that possess a DFA in each fungal order (Figure 4, B).

Overall, the distribution of DFAs is highly variable across fungal orders. For many orders, the fraction of species possessing one or more DFA is neither zero (i.e., no species have a DFA) nor one (i.e., all species have a DFA), indicating that DFA distribution is also variable within orders. Thus, DFA seems evolutionarily labile. This lability suggests that DFA could have an alternative function to the canonical actin,

which is extremely evolutionarily conserved. It's possible that the presence/absence of a DFA can rapidly change in response to natural/environmental pressures, and thus DFAs may be associated with specific adaptive fungal traits. Our next step was to look for any such associations. We note, however, that these findings may be impacted by our definition of DFA absence defined earlier. That is, by potentially overestimating the number of species for which DFAs are absent, we may have in turn overestimated the evolutionary lability of the trait.

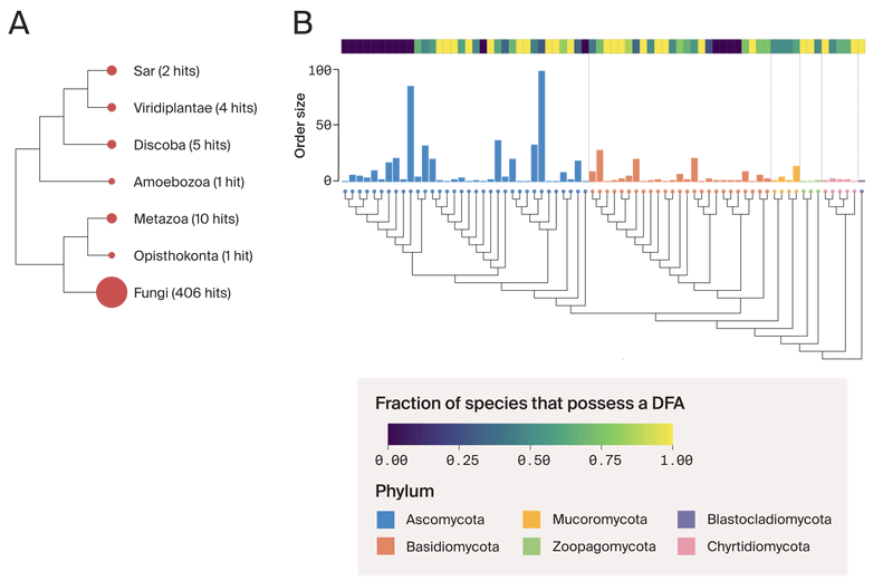

Figure 4

Taxonomic analysis of the organisms possessing the divergent actin form of interest in the extended set.

(A) Kingdom (or clade) analysis. Each branch is one representative species from a given clade/kingdom.

(B) Phylogenetic tree that highlights, for each fungal order, the fraction of species that possess a DFA (heatmap) and the number of species analyzed per order (bar plot). Bar and tree tip color indicate their phylum.

None of the six tested fungal traits correlate with DFA status

We then took an evolutionary modeling approach to identify biological processes that DFA may be involved in. We looked for evidence that DFA and specific adaptive traits are correlated. We started by curating public databases to gather trait information that we believe to be relevant to the protein we are investigating. For this project, we chose to use Fun^{Fun} **[20]**, a recently established database that aggregates trait information from multiple databases.

We chose to focus on six available traits (Figure 5). Four traits are discrete traits that take on categorical values: growth form, trophic mode (source from which a fungus derives its nutrients), ascus dehiscence (mechanism to release the ascospores), and the number of auxin-responsive promoters (the ability to respond to auxin-based signals from the environment **[24]**. The two other traits are continuous traits associated with spore morphology: spore length and spore width. We chose to look at these traits because each one is associated with either morphological structures, cell architecture, cell dynamics, or cell trafficking — all areas where actin could play a pivotal role. Furthermore, these traits are widely distributed across the fungal species in our working set. Thus, we believe that DFA could be associated with one of these traits (see below).

Altogether, we were able to collect high-confidence DFA status, phenotypic data for at least one trait, and phylogeny information for a total of 102 species, allowing us to pursue statistical modeling of the evolutionary trajectory of DFA status and traits in these species **[24]**.

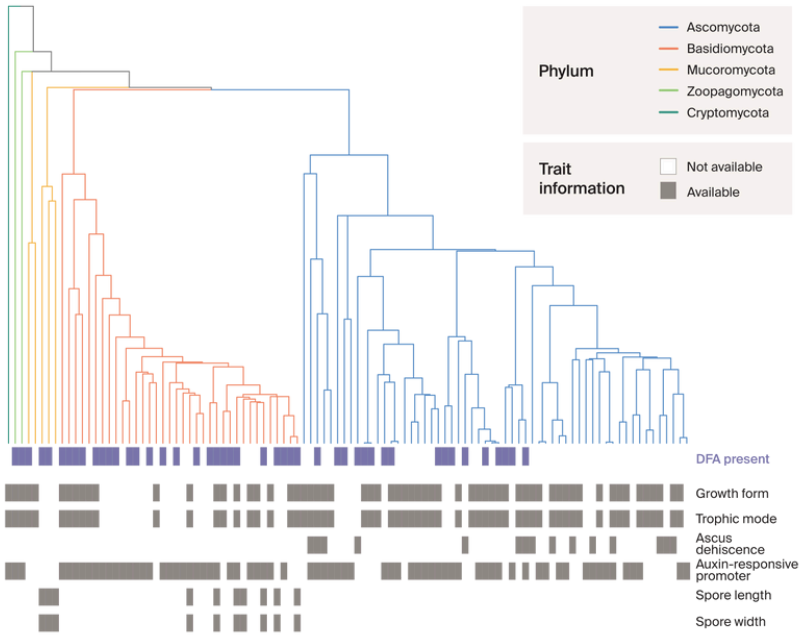

Figure 5

Tree representation of the fungal species for which we looked for a correlation between DFA status and specific traits.

Tree branches are colored based on phylum. The row "DFA present" indicates whether the species possesses a DFA (purple) or not (empty). In subsequent rows, a grey square indicates that information on a given trait is available for the species.

Next, we developed an evolutionary modeling strategy to find evidence of correlated evolution between DFA and one of these traits. For the discrete traits (Figure 6, A–D), we compared statistical models that assumed either correlated or independent evolution of the trait and DFA for two classes of model: the discrete-time Markov model (DTMM) and the hidden Markov model (HMM). We used the Akaike information criterion (AIC) to evaluate the models, where the model that describes the best association of a trait and DFA is the one with the lowest AIC (Table 2). For all discrete traits, we found the model in which DFA and a trait of interest did not have correlated evolutionary histories to be more likely.

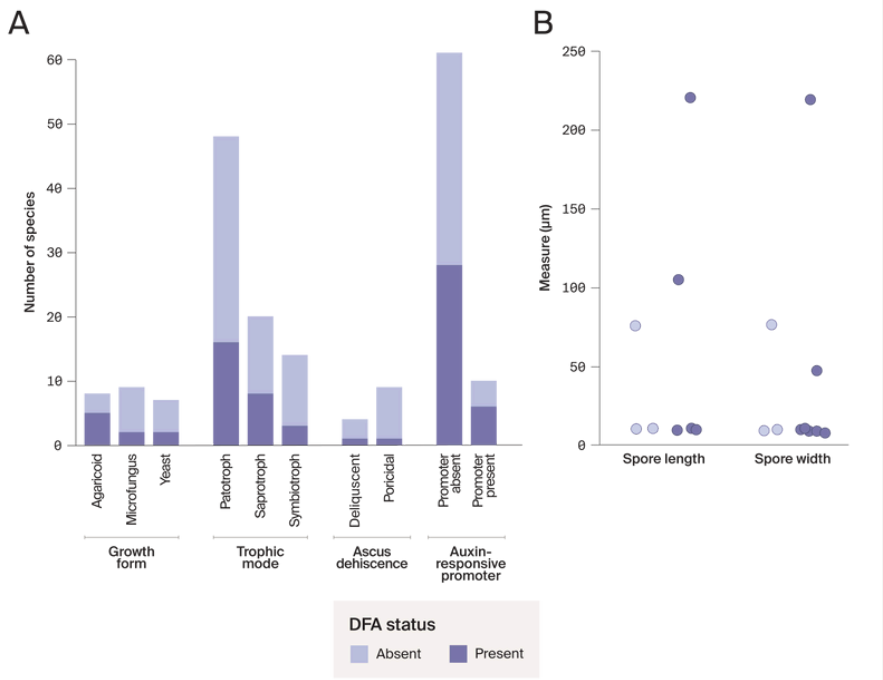

Figure 6

Distribution of DFA status within trait data.

For each trait, we're showing the number of species that have (purple) and don't have (light purple) DFAs. For discrete traits (A) we show the number of species in each trait category. For continuous traits (B), we show the trait value for each species.

For continuous traits (Figure 6, E–F), we used a generalized linear mixed effects model that accounts for the evolutionary non-independence of species and their traits, and quantifies the degree to which a continuous variable explains the presence or absence of DFA. It provides a statistical test for the influence of a trait on DFA status, and a significant p-value (≤ 0.05) indicates a correlation between the trait and DFA (Table 3). None of the continuous traits explained the presence or absence of DFA in a given species.

In conclusion, we did not detect a correlation between the presence of a DFA and the traits investigated in this study.

Trait	Model class	Evolution of DFA and trait	AIC
Growth form	DTMM	Independent	69.34
	HMM	Independent	87.13
	DTMM	Correlated	86.65
	HMM	Correlated	125.62
Trophic mode	DTMM	Independent	211.79
	HMM	Independent	222.46
	DTMM	Correlated	224.42
	HMM	Correlated	256.31
Ascus dehiscence	DTMM	Independent	30.06
	HMM	Independent	41.02
	DTMM	Correlated	37.49
	HMM	Correlated	57.15
Auxin-responsive promoter	DTMM	Independent	144.88
	HMM	Independent	146.13
	DTMM	Correlated	149.33
	HMM	Correlated	159.73

Table 2

Akaike information criterion (AIC) for the different models used to model the evolution of DFA and discrete fungal traits.

Trait	Parameters	Values	p-values
Spore length	Intercept	0.8882429	0.41
	Length	0.0060128	0.64
Spore width	Intercept	1.003899	0.35
	Width	0.0045372	0.71

Table 3

Results of the phylogeny-corrected generalized linear mixed models for continuous traits.

Limitations

We did not find a correlation between the presence of a DFA in a fungal species and any fungal traits. Thus, we failed to support any preliminary hypotheses about the function of DFAs. We've identified a handful of limitations and weaknesses in our study that may have contributed to this negative result.

Our failure in identifying a correlation between DFAs and any fungal trait most likely stems from the fact that we have only investigated six traits, and did not include traits that were biologically relevant to the DFAs in our work. The restricted scope of this work is a direct consequence of one of the main challenges in any trait mapping project: collecting a large amount of accurate data. We only explored a small number of traits because of the limited availability and quality of the data we could obtain. Furthermore, these trait data were not originally collected with the goals of the present study in mind, and thus are likely limited in relevance for DFAs.

The scarcity of reliable trait information not only limited the breadth of our investigation but also impacted the depth to which we could explore the relationships between DFA and fungal traits, as it significantly reduced our statistical power. For instance, starting from 36,253 fungi with at least one protein structure in UniProt, we were only able to gather reliable trait information (DFA status, one of the six fungal traits, and phylogeny) for an average of 34 fungi.

Finally, our ability to link a phenotype and the presence or absence of a DFA is limited by our ability to determine whether DFAs are present or absent. While we can accurately identify species that have a DFA, our determination of DFA absence is impacted by the quality and coverage of genomic sequence data. Errors in the assessment of DFA status reduce our ability to identify significant associations between DFAs and phenotypes.

Key takeaways

We hoped to use trait mapping and evolutionary modeling as a way to generate hypotheses about the potentially undiscovered, new function of the divergent fungal actin (DFA) discovered in our previous work. We found that the distribution of this DFA is variable within the fungal kingdom, suggesting DFA has a more adaptive function compared to canonical actin, which is highly conserved in the fungal kingdom. We tried an evolutionary modeling strategy to see if we could correlate the presence or absence of this actin variant with a set of fungal traits, since any correlation could provide insight into the function of DFAs.

Our results showed no correlation between any of the tested traits and DFAs, so the function of this variant remains unknown. While we didn't find anything conclusive, we're still excited by the potential to use trait mapping to generate hypotheses about unknown protein functions in the future.

Next steps

We've decided to put this project on ice. We think there may be interesting biology underlying divergent fungal actins, but the approach we took here to elucidate it was limited by the availability of relevant trait information. Nevertheless, we would greatly appreciate any feedback and comments on this work.

While we're not pursuing this topic, several investigative paths are possible for others. To keep investigating the function of DFAs, one obvious follow-up is to expand the range of traits to test for correlation with the presence of DFAs. This would require more complete datasets, including information for multiple species whose DFA status can be established. Fungal ecology groups and mycologists may have the tools and knowledge to generate such information. Another approach would be to focus on

genetic traits and rely on public genomic information. One could use available genomes of fungal species that we're confident either have or don't have a DFA and search for any correlation with the presence/absence of gene families.

Someone could also probe DFA function by using molecular biology techniques to knock out the DFA in a given species and characterize the resulting phenotype(s), though this would require genetically tractable organisms and technical knowledge.

References

1. Avasthi P, Bigge BM, Celebi FM, Cheveralls K, Gehring J, McGeever E, Mishne G, Radkov A, Sun DA. (2024). ProteinCartography: Comparing proteins with structure-based maps for interactive exploration. https://doi.org/10.57844/ARCADIA-A5A6-1068

2. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

3. Avasthi P, Bigge BM, Sun DA, York R. (2024). Exploring the actin family: A case study for ProteinCartography. https://doi.org/10.57844/ARCADIA-A7CB-9F5C

4. Pollard TD. (2016). Actin and Actin-Binding Proteins. https://doi.org/10.1101/cshperspect.a018226

5. Berepiki A, Lichius A, Read ND. (2011). Actin organization and dynamics in filamentous fungi. https://doi.org/10.1038/nrmicro2666

6. Perrin BJ, Ervasti JM. (2010). The actin gene family: Function follows isoform. https://doi.org/10.1002/cm.20475

7. McDowell JM, Huang S, McKinney EC, An Y-Q, Meagher RB. (1996). Structure and Evolution of the Actin Gene Family in *Arabidopsis thaliana*.

https://doi.org/10.1093/genetics/142.2.587

8 Yee M, Walther T, Frischknecht F, Douglas RG. (2022). Divergent Plasmodium actin residues are essential for filament localization, mosquito salivary gland invasion and malaria transmission. https://doi.org/10.1371/journal.ppat.1010779

9 Vahokoski J, Bhargav SP, Desfosses A, Andreadaki M, Kumpula E-P, Martinez SM, Ignatev A, Lepper S, Frischknecht F, Sidén-Kiamos I, Sachse C, Kursula I. (2014). Structural Differences Explain Diverse Functions of Plasmodium Actins. https://doi.org/10.1371/journal.ppat.1004091

10 Steinegger M, Söding J. (2018). Clustering huge protein sequence sets in linear time. https://doi.org/10.1038/s41467-018-04964-5

11 Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988

12 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009). BLAST+: architecture and applications. https://doi.org/10.1186/1471-2105-10-421

13 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2022). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1101/2022.02.07.479398

14 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

15 Traag VA, Waltman L, van Eck NJ. (2019). From Louvain to Leiden: guaranteeing well-connected communities. https://doi.org/10.1038/s41598-019-41695-z

16 Mohanta TK, Bae H. (2015). The diversity of fungal genome. https://doi.org/10.1186/s12575-015-0020-z

17 Mohanta TK, Mishra AK, Khan A, Hashem A, Abd-Allah EF, Al-Harrasi A. (2021). Virtual 2-D map of the fungal proteome. https://doi.org/10.1038/s41598-021-86201-6

18 Elliott TA, Gregory TR. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. https://doi.org/10.1098/rstb.2014.0331

19 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. https://doi.org/10.1093/molbev/msac174

20 Zanne AE, Abarenkov K, Afkhami ME, Aguilar-Trigueros CA, Bates S, Bhatnagar JM, Busby PE, Christian N, Cornwell WK, Crowther TW, Flores-Moreno H, Floudas D, Gazis R, Hibbett D, Kennedy P, Lindner DL, Maynard DS, Milo AM, Nilsson RH, Powell J, Schildhauer M, Schilling J, Treseder KK. (2019). Fungal functional ecology: bringing a trait-based approach to plant-associated fungi. https://doi.org/10.1111/brv.12570

21 Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, Schilling JS, Kennedy PG. (2016). FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. https://doi.org/10.1016/j.funeco.2015.06.006

22 Beaulieu JM, O'Meara BC, Donoghue MJ. (2013). Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. https://doi.org/10.1093/sysbio/syt034

23 Li D, Dinnage R, Nell LA, Helmus MR, Ives AR. (2020). phyr: An <scp>r</scp> package for phylogenetic species-distribution modelling in ecological communities. https://doi.org/10.1111/2041-210x.13471

24 Chanclud E, Morel J. (2016). Plant hormones: a fungal point of view. https://doi.org/10.1111/mpp.12393