The evolution of the gliotoxin biosynthetic gene cluster in *Penicillium* fungi

Charu Balamurugan[1,2], Jacob L. Steenwyk[1,2,3,*], Gustavo H. Goldman[4], & Antonis Rokas[1,2,*]

[1] Vanderbilt University, Department of Biological Sciences, VU Station B #35-1634, Nashville, TN 37235, United States of America

[2] Vanderbilt Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN, United States

[3] Howards Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA

[4] Faculdade de Ciencias Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil

*Correspondence should be addressed to: jlsteenwyk@berkeley.edu and antonis.rokas@vanderbilt.edu

Running title: Gliotoxin Evolution in *Penicillium* Fungi

Keywords: comparative genomics; evolutionary biology; secondary metabolic gene clusters; duplication and loss; plant pathogen; secondary metabolism; specialized metabolism

Abstract

Fungi biosynthesize a diversity of secondary metabolites, small organic bioactive molecules that play diverse roles in fungal ecology. Fungal secondary metabolites are often encoded by physically clustered sets of genes known as biosynthetic gene clusters (BGCs). Fungi in the genus *Penicillium* produce diverse secondary metabolites that have been both useful (e.g., the antibiotic penicillin and the cholesterol-lowering drug mevastatin) and harmful (e.g., the mycotoxin patulin and the immunosuppressant gliotoxin) to human affairs. BGCs often also encode resistance genes that confer self-protection to the secondary metabolite-producing fungus. Some *Penicillium* species, such as *Penicillium lilacinoechinulatum* and *Penicillium decumbens*, are known to produce gliotoxin, a secondary metabolite with known immunosuppressant activity; however, an evolutionary characterization of the BGC responsible for gliotoxin biosynthesis among *Penicillium* species is lacking. Here, we examine the conservation of genes involved in gliotoxin biosynthesis and resistance in 35 *Penicillium* genomes from 23 species. We found homologous, less fragmented gliotoxin BGCs in 12 genomes, mostly fragmented remnants of the gliotoxin BGC in 21 genomes, whereas the remaining two *Penicillium* genomes lacked the gliotoxin BGC altogether. In contrast, we observed broad conservation of homologs of resistance genes that reside outside the BGC across *Penicillium* genomes. Evolutionary rate analysis revealed that BGCs with higher numbers of genes evolve slower than BGCs with few genes. Even though the gliotoxin BGC is fragmented to varying degrees in nearly all genomes examined, ancestral state reconstruction suggests that the ancestor of *Penicillium* species possessed the gliotoxin BGC. Our analyses suggest that genes that are part of BGCs can be retained in genomes long after the loss of secondary metabolite biosynthesis.

Introduction

Gliotoxin is a secondary metabolite produced by certain fungi, including the major opportunistic human pathogen *Aspergillus fumigatus* (Raffa and Keller 2019). Secondary metabolites are bioactive molecules of low molecular weight that are not required for the organism's growth but aid survival in harsh environments (Raffa and Keller 2019). Genes that participate in the biosynthesis of secondary metabolites, including gliotoxin, typically reside next to each other in fungal genomes and form biosynthetic gene clusters (BGCs) (Rokas *et al.* 2020). The gliotoxin BGC is implicated in human pathogenicity because gliotoxin suppresses the immune response of the mammalian host through diverse mechanisms, including by inhibiting protein complexes necessary for the generation of antimicrobial reactive oxygen species, decreasing cytotoxic activities of T lymphocytes, and preventing integrin activation (Dolan *et al.* 2015; Raffa and Keller 2019). Gliotoxin's role in modulating host biology suggests that it is a virulence factor (Raffa and Keller 2019). For example, virulence is attenuated in certain animal models of disease when *gliP*, the non-ribosomal peptide synthetase gene involved in gliotoxin biosynthesis, is deleted (Sugui *et al.* 2007).

Fungi that produce gliotoxin need to be resistant to the toxin. Several genes contribute to resistance, such as the thioredoxin reductase gene *gliT*, located within the gliotoxin BGC (Schrettl *et al.* 2010). *gliT* deletion strains of *A. fumigatus* exhibit resistance to gliotoxin oxidation and unchecked methylation (Owens *et al.* 2015). As a result, *gliT*-deficient *A. fumigatus* are hypersensitive to gliotoxin (Owens *et al.* 2015). Other resistance genes encoding transcription factors, transporters, and oxidoreductases, reside outside the BGC and – like *gliT* –

are found in both gliotoxin-producing and non-producing species (Castro *et al.* 2022). For example, the transcription factor RglT is the primary regulator of *gliT* (Ries *et al.* 2020). Seven other genes are known to be regulated by *rglT* and contribute to gliotoxin resistance: *gtmA* (encodes a *bis*-thiomethyltransferase, AFUA_2G11120), *kojR* (transcription factor, AFUA_5G06800), *abcC1* (ABC-transporter, AN7879/AFUA_1G10390), *mtrA* (methyltransferase, AN3717/AFUA_6G12780), AN9051 (oxidoreductase, AFUA_7G00700), AN1472 (MFS transporter, AFUA_8G04630), and AN9531 (NmrA-transcription factor, AFUA_7G06920) (Castro *et al.* 2022).

Though progress has been made in understanding the mechanisms and functions of the gliotoxin biosynthetic pathway, several questions remain, especially concerning the evolutionary and ecological significance of this BGC in lineages that contain a mix of biotechnologically and medically relevant fungi, such as *Penicillium* (Steenwyk *et al.* 2019). For example, *Penicillium camemberti* and *Penicillium roqueforti* contribute to cheese production (Nelson 1970; Lessard *et al.* 2012), whereas *Penicillium expansum*, *Penicillium digitatum*, and *Penicillium italicum* are postharvest pathogens of citrus fruits, stored grains, and other cereal crops (Marcet-Houben *et al.* 2012; Ballester *et al.* 2015; Li *et al.* 2015). Examination of the gliotoxin BGC in the genomes of *Penicillium* species will shed light on the evolution of the gliotoxin BGC within Aspergillaceae, the family encompassing both *Aspergillus* and *Penicillium* species.

Considering the close relatedness of *Penicillium* and *Aspergillus*, it is interesting that evidence of gliotoxin production is scant within the former. To fill this gap, we employed a genome-scale approach to infer the evolutionary history of the gliotoxin BGC among 35 strains of 23

Penicillium species. We found that most *Penicillium* genomes examined contained fragmented gliotoxin BGCs and two lacked a BGC. However, some *P. expansum* strains had two homologous gliotoxin BGCs. Codon optimization analysis reveals that genes *in Penicillium* BGCs are lowly optimized, whereas genes in *Aspergillus* gliotoxin BGCs are highly optimized.

In contrast, gliotoxin resistance genes in *Penicillium* and *Aspergillus* fungi have similar degrees of codon optimization, suggesting that *Penicillium* species encounter exogenous gliotoxin in their environments. Examination of evolutionary rates revealed that genes from highly fragmented gliotoxin BGCs evolved at significantly higher rates than genes from less fragmented BGCs, suggesting that less fragmented BGCs have been experiencing relaxation of selective constraints for longer. Ancestral state reconstructions indicate that the *Penicillium* ancestor possessed a less fragmented gliotoxin BGC, followed by distinct trajectories of duplication and loss, highlighting the diverse evolutionary pathways of the gliotoxin BGC in *Penicillium* species.

Materials and Methods

I. Data collection and quality assessment

We retrieved the genomes and gene annotations of 35 *Penicillium* strains from 23 species as well as of two outgroups (*Aspergillus fumigatus* and *Aspergillus fischeri*) from NCBI (https://www.ncbi.nlm.nih.gov/) (Table S1).

Genome assembly and annotation quality were examined to evaluate whether the dataset is sufficient for comparative genomics. The quality and characteristics of the genomes (N50, L50, assembly size, number of scaffolds, and gene count) were evaluated using BioKIT (v0.1.0)

(Steenwyk *et al.* 2022) (Figure S1). The average N50 value was 1,850,972.1 bases, where 46% of proteomes consisted of N50 values greater than 1 Mb, and the lowest N50 value was 31,119 bases for *P. expansum* CMP 1. Gene annotation completeness was assessed using BUSCO (v5.0.0) (Waterhouse *et al.* 2018) (Figure S2). BUSCO uses a predetermined set of near-universally conserved single-copy genes (or BUSCO genes) to identify their presence in a query proteome (characterized as single-copy, duplicated, or fragmented) or absence. We used the 4,181 BUSCO genes from the Eurotiales OrthoDB dataset (Manni *et al.* 2021; Zdobnov *et al.* 2021). Nearly all the genomes have high BUSCO gene coverage (average: 95.9% ± 3.1%), with the lowest percentages being for *P. coprophilum* (87.9%) and *P. decumbens* (85.3%).

II. Identification and characterization of gliotoxin BGC and resistance genes

a. Identification of gliotoxin GBC and resistance genes

The representative gliotoxin BGC (BGC0000361, Download date: April 2022) from the *Aspergillus fumigatus* Af293 reference strain was downloaded from the Minimum Information about a Biosynthetic Gene Cluster (MiBIG) database (Kautsar *et al.* 2019). Command-line NCBI BLASTP (Camacho *et al.* 2009) searches for the Af293 gliotoxin BGC against the proteome of each species were executed. Highly similar sequences were identified using an expectation value threshold of 1e-4 and a query coverage of 50%. The resulting BLAST outputs were then cross-referenced with the NCBI feature table file, which contains genome location information for each gene, and parsed to identify clusters of homologs. Less fragmented BGCs are defined as having at least 7 / 13 genes from the query gliotoxin BGC present, including *gliP*, encoding the core nonribosomal peptide synthetase (Castro *et al.* 2022); mostly fragmented clusters are defined as having at least 3 /13 genes from the gliotoxin BGC without a requirement for this

cluster to include *gliP*. When identifying BGCs, we allowed up to four genes between each pair of adjacent homologs using the *A. fumigatus* Af293 BGC from the MiBIG database (Kautsar *et al.* 2019) as reference (Castro *et al.* 2022).

To rule out gene annotation errors in cases where we infer genes to be absent, we conducted command-line NCBI tBLASTn searches for the Af293 gliotoxin BGC against the genome sequences. Highly similar sequences were identified using an expectation value threshold of 1e-10. The resulting outputs were analyzed, and no new presence/absence information was found.

Sequence similarity searches were also conducted for eight gliotoxin resistance genes (*abcC1*/AN7879, *mtrA*/AN3717, AN9051, AN1472, AN9531, *rglT*, *gtmA*/AFU2G11120, *kojR*/AFUA_5G06800), three of which were transcription factors (AN9531, *rglT*, *kojR*). We used an expectation value threshold of 1e-3 and a query coverage threshold of 50%; we used a lower query coverage threshold of 40% for the three transcription factors.

b. **Codon bias**

To estimate the potential functional significance of the partial gliotoxin BGCs present in *Penicillium* genomes, mean gene-wise relative synonymous codon usage (gRSCU) was determined for each clustered *gli* gene across all proteomes using BioKIT (Steenwyk *et al.* 2022). This provides insight into how codon usage bias influences the expression level of a particular gene. The percentile rankings of each of the present and clustered *gli* genes were calculated using the R package *dplyr* (v1.0.9) (Wickham *et al.* 2022), and these values, for each species, were then plotted using the R package *ggplot2* (Wickham 2016).

c. Synteny Analysis

Alignments of representative *Penicillium* genomes with less and more fragmented gliotoxin BGCs were generated using a GenomeDiagram in Biopython (Cock *et al.* 2009). Five genomes (*A. fumigatus* Af293, *P. flavigenum* IBT 14082, *P. roqueforti* FM164, *P. nordicum* DAOMC 185683, and *P. expansum* CMP1) with the largest number of different, homologous *gli* cluster genes above seven, and including *gliP*, were chosen to visualize the conservation of synteny of less fragmented gliotoxin BGCs across the phylogeny. Similarly, the five genomes (*P. steckii* IBT 24891, *P. vulpinum* IBT 29486, *P. rubens* 43M1, *P. camemberti* FM 013, and *P. italicum* PHI 1) with the greatest number of different, homologous *gli* cluster genes above three and below seven, and not needing to include *gliP*, were chosen to visualize synteny of mostly fragmented BGCs across the phylogeny.

III. Phylogenetic Analysis

a. Species Tree Inference

The evolutionary relationships of *Penicillium* species were obtained from a previous study (Steenwyk *et al.* 2019) using treehouse (Steenwyk and Rokas 2019). For three species with population-level data, within-species relationships were inferred using phylogenomics. To do so, protein sequences of BUSCO genes were first aligned using MAFFT (v7.490) with the *--auto* parameter (Katoh and Standley 2013). Codon-based alignments were generated by threading the corresponding DNA sequences onto the protein alignment with the *thread_dna* function in PhyKIT (v1.11.2) (Steenwyk *et al.* 2021). The resulting nucleotide alignments were trimmed using ClipKIT (v1.3.0) (Steenwyk *et al.* 2020) with default parameters. The resulting aligned and

trimmed sequences were concatenated into a supermatrix with 8,124,861 sites using the *create_concat* function in PhyKIT. We then inputted the concatenated matrix into IQ-TREE 2 (v2.0.6), a software that implements a maximum likelihood framework for inferring phylogenies. All other evolutionary relationships between species were constrained following the relationships inferred in a previously published study (Steenwyk and Rokas 2019). The best-fitting substitution model (GTR+F+I+G4) was determined using ModelFinder (Kalyaanamoorthy *et al.* 2017).

b. **Single-gene tree inference**

To infer the evolutionary history of genes in the gliotoxin BGCs, individual *gli* genes were compiled and aligned with MAFFT (v7.490) using the --*auto* parameter (Katoh and Standley 2013). The corresponding nucleotide sequences for each file were obtained from the CDS files for each species, using the *faidx* function of BioKIT (v0.1.0) (Steenwyk *et al.* 2022). These nucleotide sequences were then threaded onto the protein alignments using the *thread_dna* function of PhyKIT (Steenwyk *et al.* 2021), resulting in a codon-based alignment. All individual codon-based gene alignments were trimmed with ClipKIT (Steenwyk *et al.* 2020) with default parameters. The trimmed alignments were used to construct a phylogeny using IQ-TREE 2 (Minh *et al.* 2020). The best-fitting substitution model was chosen for each *gli* gene using Bayesian information criteria (BIC) implemented in ModelFinder (Kalyaanamoorthy *et al.* 2017) from IQ-TREE 2. Branch support in each phylogenetic tree was assessed by 1000 bootstraps using ultrafast bootstrapping approximation (Hoang *et al.* 2018). Tree visualization was carried out using the R packages *ape* (v5.6.2) (Paradis and Schliep 2019) and *phytools* (v1.0.3) (Revell 2012).

205

206 To characterize variation in the evolution of individual genes of the gliotoxin BGC, the trimmed
207 alignments and maximum-likelihood trees from IQ-TREE 2 were used as input into the
208 *evolutionary_rate*, *total_tree_length*, and *pairwise_identity* functions of PhyKIT to estimate two
209 tree-based measures of evolutionary rate and one sequence-based measure. Evolutionary rate is
210 defined as the total tree length divided by the number of terminals (Telford *et al.* 2014; Steenwyk
211 *et al.* 2021). The total tree length is the sum of all branches (Steenwyk *et al.* 2021).

212

213 c. **Ancestral state reconstructions**

214 Ancestral state reconstruction for each gene of the gliotoxin BGC across three discrete characters
215 ("Presence clustered," "Presence unclustered," and "Absence") was estimated using *phytools*
216 (v1.0.3) (Revell 2012). Presence generally indicates that a homolog of the particular gene was
217 identified. "Presence clustered" identifies an existing homolog of the specific gene within a
218 maximum distance of four genes from other homologs of the gliotoxin BGC. "Presence
219 unclustered" identifies an existing homolog of the particular gene without clustering. "Absence"
220 indicates that no homolog of the specific gene was identified. Estimation of ancestral character
221 states was done using the Dollo parsimony method. This method assumes that a complex
222 character lost during the evolution of a particular lineage cannot be regained (Rogozin *et al.*
223 2006). Count, a software package for the evolutionary analysis of homolog family sizes, was
224 used to generate these ancestral state reconstructions (Csűös 2010).

225

226 d. **Tree Topology Testing**

Tree topology testing was used to determine whether the duplication event resulting in the two less fragmented, homologous gliotoxin BGCs in *P. expansum* strains MD 8 and d1 occurred solely in the lineage of *P. expansum* or deeper in the tree. IQTREE 2 (Minh *et al.* 2020) was used to compute log-likelihoods of a constrained tree (monophyly of *P. expansum gliP* homologs) and the observed tree in which a polyphyly of *gliP* in both clusters is seen (inconsistent with the known species tree). 1000 RELL replicates (Kishino *et al.* 1990) were performed. The AU test results (Shimodaira 2002) was used for comparison.

Results and Discussion

I. The gliotoxin BGC is fragmented in *Penicillium* species

Presence / absence data of the 13 genes in the gliotoxin BGC among the 23 *Penicillium* species analyzed reveals that the cluster is largely fragmented in the genus *Penicillium* (Figure 1). The proteomes of 12 strains from 5 *Penicillium* species (*P. arizonense*, *P. flavigenum*, *P. roqueforti*, *P. nordicum*, *P. expansum*), possessed less fragmented BGCs, and the proteomes of 23 strains from 18 *Penicillium* species had mostly fragmented BGCs (Figure S3-S15). Two less fragmented BGCs, which contained 10 / 13 genes and 7 / 13 genes, were identified in *P. expansum* strains d1 and MD 8, respectively. Regardless of the number of less fragmented BGCs found, to our knowledge, none of the *Penicillium* species in question are known to produce gliotoxin, except *P. decumbens* (Feng *et al.* 2018), suggesting that the absence of clustering in this species may be due to strain heterogeneity and requires further exploration.

II. A complete gliotoxin BGC was present in the ancestor of *Penicillium* species

Ancestral state reconstruction revealed the presence of all 13 genes in the gliotoxin BGC in the ancestor of the *Penicillium* species used in our study (Figure 1). We infer that the first gene lost was *gliH*, which is absent from 25 of the 35 *Penicillium* strains examined. The *gliH* gene encodes an acetyltransferase that, when deleted, results in a loss of gliotoxin production in *A. fumigatus* (Schrettl *et al.* 2010; Castro *et al.* 2022). Thus, the early loss of *gliH* in the genus *Penicillium* may have been the key determinant of a lack of gliotoxin production. Further, the synteny of genes in the BGC is mostly conserved and similar to the arrangement of the *A. fumigatus* Af293 gliotoxin BGC across representative, less fragmented BGCs, such as *P. flavigenum* IBT 14082 *and P. expansum* CMP 1 (Figure 2). In contrast, there is extensive divergence in synteny conservation among mostly fragmented BGCs (Figure 2). To our knowledge, none of the *Penicillium* species examined are known to produce gliotoxin, except *P. decumbens* (Feng *et al.* 2018), suggesting that the absence of clustering in this species may be due to strain heterogeneity and requires further exploration.

III. Resistance genes are broadly conserved

The presence/absence results of the eight resistance genes, portrayed in Figure 1, suggest that their origins predate the *Aspergillus* and *Penicillium* genera (Figure S16-S23). All species possessed *abcC1*, AN1472, AN9051, AN9531, and *kojR* homologs. In addition, only *Penicillium* species with mostly fragmented gliotoxin BGCs lacked at least one resistance gene, such as *gtmA*, *mtrA*, and *rglT*. *Penicillium chrysogenum* lacked both *rglT* and *gliT*, an observation consistent with the transcriptional dependency of *gliT* to *rglT* (Ries *et al.* 2020).

IV. *Penicillium* species have experienced changes in gliotoxin BGC synteny over time

All genes of the gliotoxin BGC were broadly found within the genus *Penicillium*, except for *gliH*, yet most were sparsely clustered (Figure 1). More specifically, 12 out of 35 *Penicillium* species/strains were found to have a less fragmented, homologous BGC. Two strains of *Penicillium expansum* (d1 and MD 8) were found to have two BGCs. Evidence of variation in gene presence / absence is also evident within species. For example, *Penicillium roqueforti* shows population variation in the presence of *gliZ*, a major transcriptional regulator of gliotoxin biosynthesis (Bok *et al.* 2006); five strains of *P. roqueforti* lack *gliZ* whereas one strain has the gene. As a result, we can conclude that the ancestor of *P. roqueforti* had a *gliZ* homolog, but the gene was lost over time in most of the strains, highlighting the importance of population-level sampling. Overall, we can see that the gliotoxin BGC has experienced relocations and duplications of its genes, specifically in *Penicillium expansum* strains d1 and MD 8, as is expected in the formation of most secondary metabolite-producing BGCs (Rokas *et al.* 2018).

V. Few *Penicillium* species contain codon-optimized gliotoxin BGCs

Compared to the two outgroup *Aspergillus* species, *A. fumigatus* and *A. fischeri*, *Penicillium* species have much lower gRSCU value rankings (Figure 3). Specifically, the mean gRSCU percentile rank of gliotoxin BGC genes among the *Aspergillus* outgroups is 0.81, while that among the *Penicillium* species is 0.35; these scores suggest that *gli g*enes from *Aspergillus* are more codon-optimized than *gli* genes from *Penicillium*. Regardless of mean gRSCU values, *gliT* and *gliA* homologs, when present, are ranked consistently in the top three to four clustered genes. However, when considering resistance genes, the spread and range of their gRSCU values are similar across all species. The mean gRSCU percentile rank of gliotoxin resistance genes among the *Aspergillus* outgroups is 0.58, while that among the *Penicillium* species is 0.53. This allows

us to infer that these *Penicillium* species may ecologically encounter exogenous gliotoxin, making *gliT*, encoding a gliotoxin-neutralizing enzyme, *gliA*, encoding a transporter that exports gliotoxin, and non-TF resistance genes such as *abcC1*, encoding an ABC-transporter, rank in the top percentiles among each of the species' gene sets.

VI. *gli* genes in less fragmented clusters are evolving at a slower rate than mostly fragmented clusters

In the comparison of tree-based and sequence-based measures of evolutionary rate, *gli* genes from less fragmented clusters are evolving at a significantly slower pace ($p<0.0001$) than those from mostly fragmented clusters across all three metrics, as seen by a two-way ANOVA with an additive model (Figure 4, Figure S3-S15). This difference highlights a notable feature of many BGCs, the fact that they are rapidly evolving, hinted at by their high variability and narrow taxonomic range (Rokas *et al.* 2020).

VII. A duplication of the gliotoxin BGC may have occurred before the divergence between *P. flavigenum* and *P. roqueforti*

We conducted a tree topology test to infer when the gliotoxin BGCs found in *P. flavigenum* occurred. The maximum likelihood phylogeny suggests that this duplication occurred before the divergence between *P. flavigenum* and *P. roqueforti*. An alternative hypothesis is that duplication occurred within *P. expansum*. This alternative hypothesis would be supported by monophyly of *P. expansum* homologs of BGC genes. After conducting a tree topology test comparing log likelihood values between the maximum likelihood phylogeny and an alternative tree wherein *P. expansum* gliP homologs were constrained to be monophyletic, we found that the

constrained topology was significantly rejected (Approximately Unbiased test, p = 7.34e-110) (Figure S24). In other words, it is unlikely duplication occurred within *P. expansum* lineage; instead, duplication likely occurred more anciently, prior to the diversification of *P. expansum*.

Conclusions

The ancestor of *Penicillium* species likely possessed a complete gliotoxin BGC. A duplication event of the BGC occurred in one lineage, likely prior to the divergence of *P. flavigenum* and *P. roqueforti*. Also, the presence/absence results of the eight resistance genes suggest that their origins predate the *Aspergillus* and *Penicillium* genera suggesting that resistance has long been important among these species. The genes in *Penicillium* gliotoxin BGCs are less codon optimized (gRSCU percentile rank mean: 0.35) compared to their *Aspergillus* counterparts (gRSCU percentile rank mean: 0.81) suggesting that *gli* genes are much more often expressed in *Aspergillus* species than in *Penicillium*. However, less fragmented BGCs within *Penicillium* species are evolving at a slower rate than mostly fragmented clusters, suggestive of potential functionality.

Although informative, this work only utilizes publicly available protein annotations of biotechnologically and medically relevant *Penicillium* fungi, making it important to expand upon the species/strains studied. Moreover, this same targeted gliotoxin analysis within a larger phylogeny of *Aspergillus* species, for which there is greater evidence of the production of this secondary metabolite, may be helpful. An analysis of gliotoxin BGCs encoded in all fungi would also provide us with more insight into the evolutionary mechanisms that give rise to BGC diversity. In addition, expanding on the causes of conservation of less fragmented gliotoxin BGCs within a variety of *Penicillium* strains may be important, especially because evidence of

production is lacking. As a result, this exciting reality encourages further understanding of the motivating hypothesis that individual secondary metabolites are "cards" of virulence in a larger "hand" that fungi possess.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, and supplemental material.

Acknowledgements

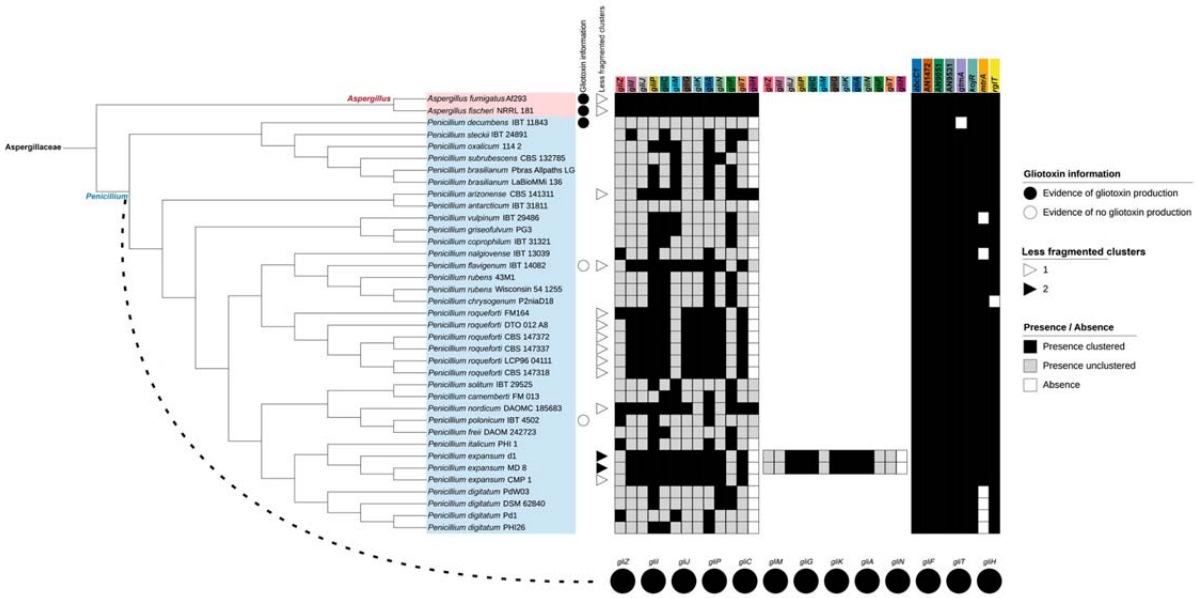
We thank members of the Rokas Laboratory at Vanderbilt University for support and feedback on this work. We also thank the Vanderbilt Data Science Institute for their undergraduate enrichment opportunities. This work was performed in part using resources contained within the Advanced Computing Center for research and Education at Vanderbilt University in Nashville, TN.

Funding

C.B. was supported by the Vanderbilt University Data Science Institute—Summer Research Program. J.L.S. and A.R. were funded by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program. Research in A.R.'s lab is supported by grants from the National Science Foundation (DEB-2110404), the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund.

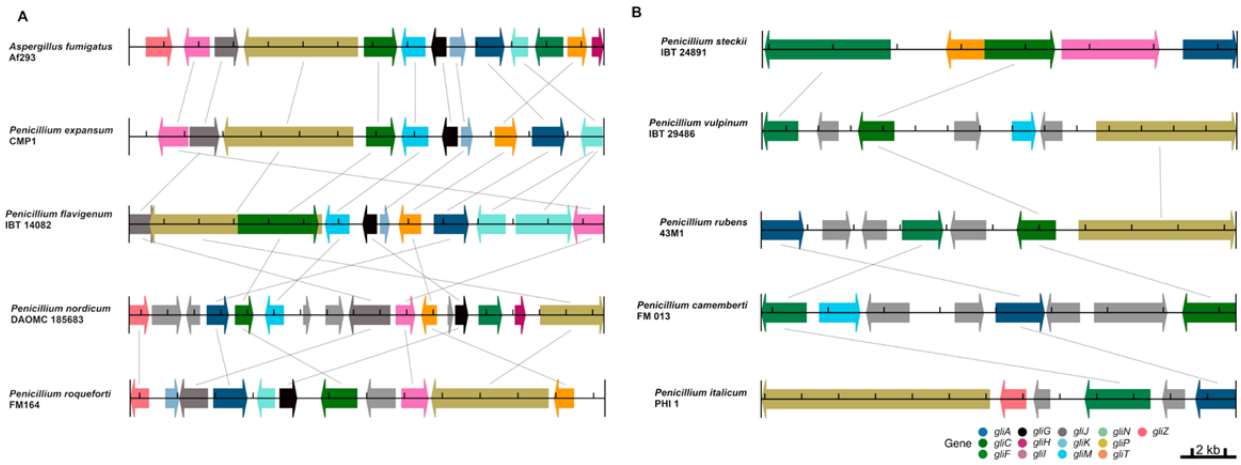
Conflicts of interest

J.L.S. is a scientific consultant for Latch AI Inc. J.L.S. is a scientific advisor for WittGen Biotechnologies. A.R. is a scientific consultant for LifeMine Therapeutics, Inc.



367 **Fig. 1 Phylogeny of *Penicillium* genomes.** Different genera are depicted using different-colored
368 boxes. *Aspergillus* is shown in red and *Penicillium* in blue. Shaded circles next to species / strain
369 names indicate gliotoxin production information from the literature, or lack thereof (Fischer *et al.*
370 2000; Spikes *et al.* 2008; Knowles *et al.* 2020; Redrado *et al.* 2022). Shaded squares in the
371 second column depict number of clusters identified. Remaining color strips depict gene presence
372 clustered (black), presence unclustered (gray), and absence (white) according to the requirements
373 outlined in the *Methods* section. Ancestral state reconstructions of each gene of the gliotoxin
374 BGC (for the ancestor of *Penicillium* species) are presented in pie charts below the phylogeny.

375



376

377 **Fig. 2 Conservation of gliotoxin BGC synteny for representative *Penicillium* species.**

378 Synteny analysis of representative genomes with less fragmented **(A)** and mostly fragmented **(B)**

379 gliotoxin BGCs. Each interval along the track represents 2 kb.

380

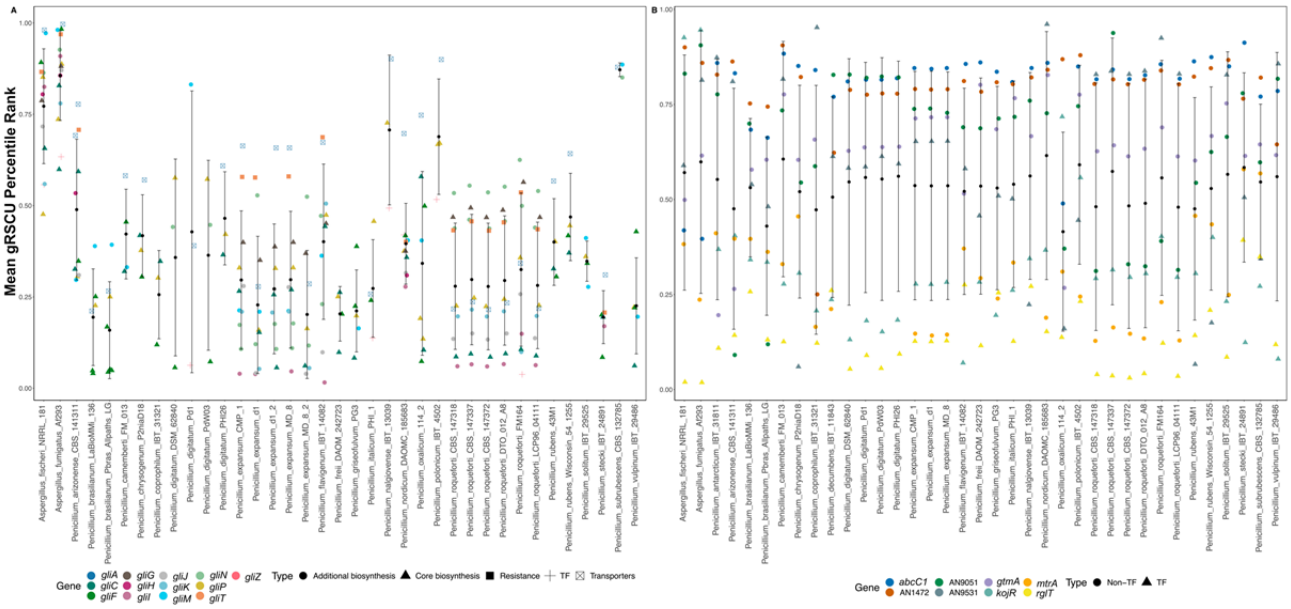


Fig. 3 Gene-wise relative synonymous codon usage (gRSCU) for gliotoxin BGC and resistance genes. (A) Percentile rankings of gene-wise relative synonymous codon usage (gRSCU) among gliotoxin BGC genes, in comparison to all other genes. Types / functionality of each gene of the gliotoxin BGC is depicted by shape in the categories of “Core biosynthesis”, “Additional biosynthesis”, “Resistance”, “Transcription Factor”, “Transporter” **(B)** Percentile ranking of gene-wise relative synonymous codon usage (gRSCU) among gliotoxin resistance genes, in comparison to all other genes. Types / functionality of each resistance gene is depicted by shape in the categories of “Non-Transcription Factor and Transcription Factor”.

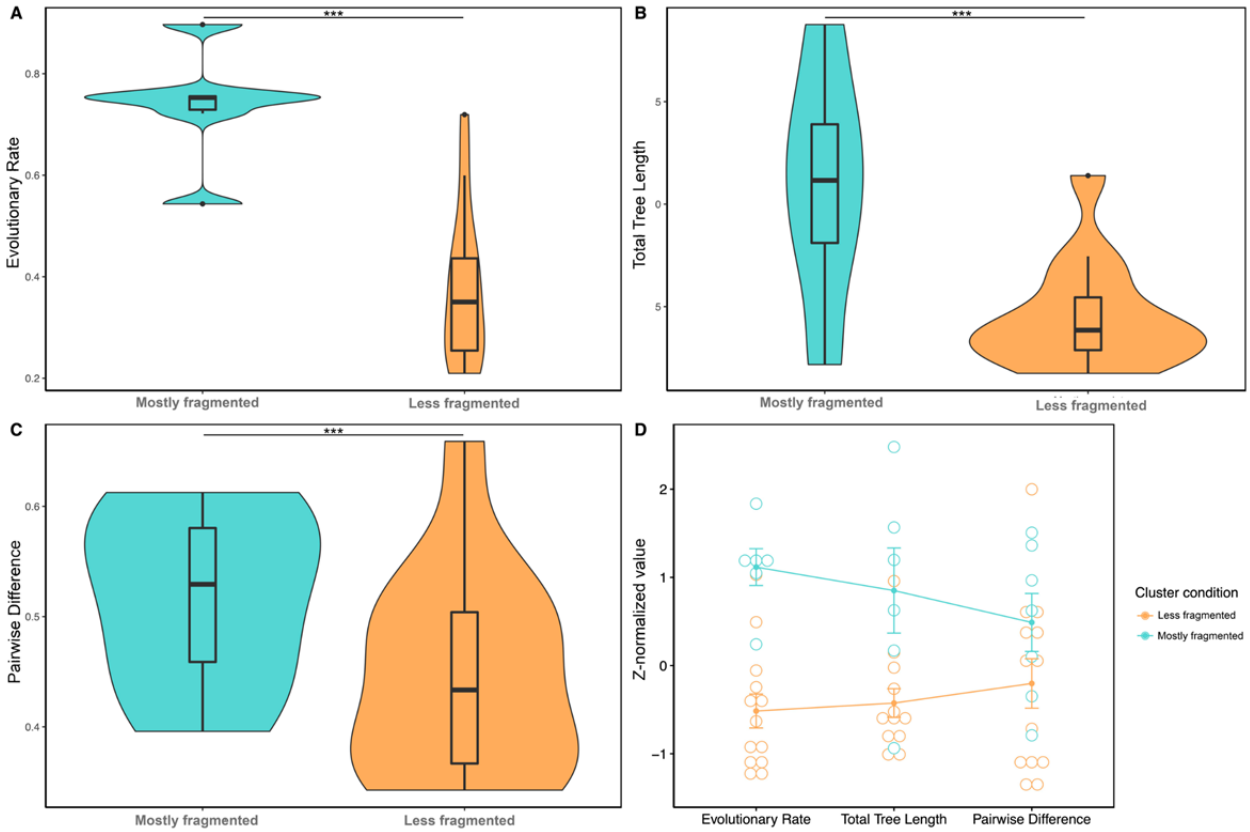


Fig. 4 Evolutionary rate comparison across gliotoxin BGCs. Multi-method comparison of evolutionary rates between less fragmented and mostly fragmented gliotoxin BGCs. Less fragmented clusters were required to contain a *gliP* ortholog and at least 7 different genes of the cluster. Mostly fragmented clusters had no requirement to contain a *gliP* ortholog and only needed to contain at least 3 different genes of the cluster. **(A)** Comparison of evolutionary rates, as a function of total tree length divided by the number of taxa, between less fragmented and mostly fragmented gliotoxin BGCs. **(B)** Comparison of total tree length between less fragmented and mostly fragmented gliotoxin BGCs. **(C)** Comparison of pairwise identity between less fragmented and mostly fragmented gliotoxin BGCs.

403 **Works Cited**

404 Ballester, A.-R., M. Marcet-Houben, E. Levin, N. Sela, C. Selma-Lázaro *et al.*, 2015 Genome,
405 Transcriptome, and Functional Analyses of Penicillium expansum Provide New Insights
406 Into Secondary Metabolism and Pathogenicity. MPMI 28: 232–248.

407 Bok, J. W., D. Chung, S. A. Balajee, K. A. Marr, D. Andes *et al.*, 2006 GliZ, a transcriptional
408 regulator of gliotoxin biosynthesis, contributes to Aspergillus fumigatus virulence. Infect
409 Immun 74: 6761–6768.

410 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+:
411 architecture and applications. BMC Bioinformatics 10: 421.

412 Castro, P. A. de, A. C. Colabardini, M. Moraes, M. A. C. Horta, S. L. Knowles *et al.*, 2022
413 Regulation of gliotoxin biosynthesis and protection in Aspergillus species. PLOS
414 Genetics 18: e1009965.

415 Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox *et al.*, 2009 Biopython: freely
416 available Python tools for computational molecular biology and bioinformatics.
417 Bioinformatics 25: 1422–1423.

418 Csűös, M., 2010 Count: evolutionary analysis of phylogenetic profiles with parsimony and
419 likelihood. Bioinformatics 26: 1910–1912.

420 Dolan, S. K., G. O'Keeffe, G. W. Jones, and S. Doyle, 2015 Resistance is not futile: gliotoxin
421 biosynthesis, functionality and utility. Trends in Microbiology 23: 419–428.

422 Feng, H., S. Liu, M. Su, E. L. Kim, J. Hong *et al.*, 2018 Gliotoxin is Antibacterial to Drug-
423 resistant Piscine Pathogens. Nat Prod Sci 24: 225–228.

424 Fischer, G., T. Müller, R. Schwalbe, R. Ostrowski, and W. Dott, 2000 Species-specific profiles
425 of mycotoxins produced in cultures and associated with conidia of airborne fungi derived

426 from biowaste. International Journal of Hygiene and Environmental Health 203: 105–
427 116.
428 Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh, 2018 UFBoot2:
429 Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35: 518–522.
430 Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, 2017
431 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:
432 587–589.
433 Katoh, K., and D. M. Standley, 2013 MAFFT Multiple Sequence Alignment Software Version 7:
434 Improvements in Performance and Usability. Mol Biol Evol 30: 772–780.
435 Kautsar, S. A., K. Blin, S. Shaw, J. C. Navarro-Muñoz, B. R. Terlouw *et al.*, 2019 MIBiG 2.0: a
436 repository for biosynthetic gene clusters of known function. Nucleic Acids Research
437 gkz882.
438 Kishino, H., T. Miyata, and M. Hasegawa, 1990 Maximum likelihood inference of protein
439 phylogeny and the origin of chloroplasts. J Mol Evol 31: 151–160.
440 Knowles, S. L., M. E. Mead, L. P. Silva, H. A. Raja, J. L. Steenwyk *et al.*, 2020 Gliotoxin, a
441 Known Virulence Factor in the Major Human Pathogen Aspergillus fumigatus, Is Also
442 Biosynthesized by Its Nonpathogenic Relative Aspergillus fischeri. mBio 11: e03361-19.
443 Lessard, M.-H., G. Bélanger, D. St-Gelais, and S. Labrie, 2012 The Composition of Camembert
444 Cheese-Ripening Cultures Modulates both Mycelial Growth and Appearance. Applied
445 and Environmental Microbiology 78: 1813–1819.
446 Li, B., Y. Zong, Z. Du, Y. Chen, Z. Zhang *et al.*, 2015 Genomic Characterization Reveals
447 Insights Into Patulin Biosynthesis and Pathogenicity in Penicillium Species. MPMI 28:
448 635–647.

449 Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov, 2021 BUSCO: Assessing Genomic
450 Data Quality and Beyond. Current Protocols 1: e323.
451 Marcet-Houben, M., A.-R. Ballester, B. de la Fuente, E. Harries, J. F. Marcos *et al.*, 2012
452 Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest
453 pathogen of citrus. BMC Genomics 13: 646.
454 Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams *et al.*, 2020 IQ-
455 TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic
456 Era. Mol Biol Evol 37: 1530–1534.
457 Nelson, J. Howard., 1970 Production of Blue cheese flavor via submerged fermentation by
458 Penicillium roqueforti. J. Agric. Food Chem. 18: 567–569.
459 Owens, R. A., G. O'Keeffe, E. B. Smith, S. K. Dolan, S. Hammel *et al.*, 2015 Interplay between
460 Gliotoxin Resistance, Secretion, and the Methyl/Methionine Cycle in Aspergillus
461 fumigatus. Eukaryotic Cell 14: 941–957.
462 Paradis, E., and K. Schliep, 2019 ape 5.0: an environment for modern phylogenetics and
463 evolutionary analyses in R. Bioinformatics 35: 526–528.
464 Raffa, N., and N. P. Keller, 2019 A call to arms: Mustering secondary metabolites for success
465 and survival of an opportunistic pathogen. PLOS Pathogens 15: e1007606.
466 Redrado, S., P. Esteban, M. P. Domingo, C. Lopez, A. Rezusta *et al.*, 2022 Integration of In
467 Silico and In Vitro Analysis of Gliotoxin Production Reveals a Narrow Range of
468 Producing Fungal Species. Journal of Fungi 8: 361.
469 Revell, L. J., 2012 phytools: an R package for phylogenetic comparative biology (and other
470 things). Methods in Ecology and Evolution 3: 217–223.

Ries, L. N. A., L. Pardeshi, Z. Dong, K. Tan, J. L. Steenwyk *et al.*, 2020 The Aspergillus fumigatus transcription factor RglT is important for gliotoxin biosynthesis and self-protection, and virulence. PLoS Pathog 16: e1008645.

Rogozin, I. B., Y. I. Wolf, V. N. Babenko, and E. V. Koonin, 2006 Dollo parsimony and the reconstruction of genome evolution.

Rokas, A., M. E. Mead, J. L. Steenwyk, H. A. Raja, and N. H. Oberlies, 2020 Biosynthetic gene clusters and the evolution of fungal chemodiversity. Nat Prod Rep 37: 868–878.

Rokas, A., J. H. Wisecaver, and A. L. Lind, 2018 The birth, evolution and death of metabolic gene clusters in fungi. Nat Rev Microbiol 16: 731–744.

Schrettl, M., S. Carberry, K. Kavanagh, H. Haas, G. W. Jones *et al.*, 2010 Self-Protection against Gliotoxin—A Component of the Gliotoxin Biosynthetic Cluster, GliT, Completely Protects Aspergillus fumigatus Against Exogenous Gliotoxin. PLOS Pathogens 6: e1000952.

Shimodaira, H., 2002 An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology 51: 492–508.

Spikes, S., R. Xu, C. K. Nguyen, G. Chamilos, D. P. Kontoyiannis *et al.*, 2008 Gliotoxin Production in Aspergillus fumigatus Contributes to Host-Specific Differences in Virulence. The Journal of Infectious Diseases 197: 479–486.

Steenwyk, J. L., T. J. Buida III, C. Gonçalves, D. C. Goltz, G. Morales *et al.*, 2022 BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. Genetics 221: iyac079.

492 Steenwyk, J. L., T. J. Buida III, A. L. Labella, Y. Li, X.-X. Shen *et al.*, 2021 PhyKIT: a broadly
493 applicable UNIX shell toolkit for processing and analyzing phylogenomic data.
494 Bioinformatics 37: 2325–2331.
495 Steenwyk, J. L., T. J. B. Iii, Y. Li, X.-X. Shen, and A. Rokas, 2020 ClipKIT: A multiple
496 sequence alignment trimming software for accurate phylogenomic inference. PLOS
497 Biology 18: e3001007.
498 Steenwyk, J. L., and A. Rokas, 2019 Treehouse: a user-friendly application to obtain subtrees
499 from large phylogenies. BMC Research Notes 12: 541.
500 Steenwyk, J. L., X.-X. Shen, A. L. Lind, G. H. Goldman, and A. Rokas, 2019 A Robust
501 Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the
502 Genera Aspergillus and Penicillium. mBio.
503 Sugui, J. A., J. Pardo, Y. C. Chang, K. A. Zarember, G. Nardone *et al.*, 2007 Gliotoxin is a
504 virulence factor of Aspergillus fumigatus: gliP deletion attenuates virulence in mice
505 immunosuppressed with hydrocortisone. Eukaryot Cell 6: 1562–1569.
506 Telford, M. J., C. J. Lowe, C. B. Cameron, O. Ortega-Martinez, J. Aronowicz *et al.*, 2014
507 Phylogenomic analysis of echinoderm class relationships supports Asterozoa.
508 Proceedings of the Royal Society B: Biological Sciences 281: 20140479.
509 Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO
510 Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol
511 Evol 35: 543–548.
512 Wickham, H., 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
513 Wickham, H., R. François, L. Henry, K. Müller, and RStudio, 2022 dplyr: A Grammar of Data
514 Manipulation.

515 Zdobnov, E. M., D. Kuznetsov, F. Tegenfeldt, M. Manni, M. Berkeley *et al.*, 2021 OrthoDB in
516 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Research 49:
517 D389–D393.

518