


Population structure in a fungal human pathogen is potentially linked to pathogenicity

Received: 22 July 2024

Accepted: 30 July 2025

Published online: 15 August 2025



E. Anne Hatmaker^{1,2}, Amelia E. Barber³, Milton T. Drott⁴, Thomas J. C. Sauters^{1,2}, Adiyantara Gumilang^{1,2}, Ana Alastruey-Izquierdo^{5,6}, Dea Garcia-Hermoso⁷, Justin L. Eagan⁸, Nancy P. Keller^{8,9}, Dimitrios P. Kontoyiannis¹⁰, Oliver Kurzai^{11,12} & Antonis Rokas^{1,2} 

Aspergillus flavus is a clinically and agriculturally important saprotrophic fungus responsible for severe human infections and extensive crop losses. Here, we analyze genomic data from 300 (117 clinical and 183 environmental) *A. flavus* isolates from 13 countries, including 82 clinical isolates sequenced in this study, to examine population and pan-genome structure and their relationship to pathogenicity. We use single nucleotide polymorphisms to build a phylogeny, analyze admixture, and perform discriminant analysis of principal components. We identify five *A. flavus* populations, including a new population, D, corresponding to distinct clades in the genome-wide phylogeny. Strikingly, > 75% of clinical isolates were in population D and <5% in population B. We also use orthogroup clustering to identify core and accessory genes within the pan-genome. Accessory genes, including genes within biosynthetic gene clusters, were significantly more common in some populations but rare in others. Our functional annotations show that population D is enriched for genes associated with carbohydrate metabolism, lipid metabolism and certain types of hydrolase activity, whereas a non-clinical population is depleted in genes related to zinc ion binding. In contrast to previous results from the major human pathogen *Aspergillus fumigatus*, isolation of *A. flavus* from human specimens is associated with population structure, providing a promising system for future investigations into the contributions of population-specific genetic differences to human infection.

The fungal genus *Aspergillus* (subphylum Pezizomycotina, phylum Ascomycota) comprises some of the most important human opportunistic fungal pathogens. Invasive aspergillosis and chronic pulmonary aspergillosis¹ impact ~250,000 and 3 million patients annually on a global scale, respectively². Invasive aspergillosis mainly afflicts individuals with compromised immunity or other underlying conditions^{3–7}. Mortality due to invasive aspergillosis varies among patient populations; ICU patients, as well as those with lung cancer, generally exhibit

~50% mortality⁸. In contrast to invasive aspergillosis, keratitis caused by *Aspergillus* spp. mainly occurs in immunocompetent patients after ocular trauma or contact lens use and can result in visual impairment and even blindness⁹. Fungal keratitis is estimated to cause over one million cases of blindness annually¹⁰.

Molecular barcoding studies from the last decade suggest that the most common infectious agents are *Aspergillus fumigatus* (4–52% of cases), followed by *Aspergillus flavus* (13–40%), *Aspergillus niger*

(8–35%), and *Aspergillus terreus* (0–7%)^{11–13}, depending on the disease. Despite belonging to the same genus, these species exhibit high levels of genomic sequence divergence; for example, *A. fumigatus* and *A. flavus* are as diverged as human and fish genomes¹⁴. Pathogenicity in *Aspergillus* has evolved independently multiple times, with species following various evolutionary trajectories to enable pathogenicity¹⁵. Alongside evolutionary differences, diverse geographic regions exhibit distinct epidemiological patterns in *Aspergillus* species prevalence. For example, the percentage of invasive aspergillosis cases caused by *A. flavus* varies by region, with ~10% of cases in the USA and Canada¹⁶ but ~40% in India¹³ attributed to *A. flavus*. The incidence of keratitis also exhibits regional differences in etiological agents^{9,17}.

Despite several *Aspergillus* species infecting humans, research to date has focused primarily on *A. fumigatus*. Unlike *A. fumigatus*, *A. flavus* is of both clinical and agricultural interest¹⁸. *A. flavus* is notorious for producing highly carcinogenic mycotoxins known as aflatoxins. Aflatoxins are associated with billion-dollar crop losses annually, and consumption by humans and other animals is associated with cancer, stunted growth, liver failure, and even death¹⁹. The production of aflatoxins and an arsenal of other small, bioactive molecules termed secondary metabolites varies among populations of *A. flavus*, suggesting niche adaptation to specific microenvironments or competition²⁰. In some fungi, including *A. fumigatus*, other secondary metabolites like gliotoxin can suppress host immune systems, serving as virulence factors^{15,21}. Some strains of *A. flavus* and *A. terreus* also produce gliotoxin or its precursors²². While there is some evidence that

kojic acid produced by *A. flavus* increases the toxicity of aflatoxin in some insect species²³, to our knowledge, virulence factors that impact human infections have yet to be described for *A. flavus*. Due to the nature of opportunistic pathogens, traits enabling human infection are likely to benefit the fungus in certain environmental niches, as human infections arise from the environment rather than person-to-person transmission.

Individual isolates of *A. flavus* exhibit substantial phenotypic variation, termed strain heterogeneity. Environmental (i.e., plant- or soil-associated) isolates can cause disease in animal models of aspergillosis²⁴ and keratitis²⁵, but strains vary in clinically relevant traits such as growth under iron starvation conditions and virulence in animal models of fungal disease^{24,25}. Additionally, *A. flavus* isolates may produce large or small sclerotia (a hardened mass of compacted mycelium capable of long-term survival under stressful conditions), resulting in L and S morphotypes, respectively; sclerotial morphotypes correlate with aflatoxin production and other cellular processes, including conidiation²⁶.

Although genetic diversity within *A. flavus* has been studied using microsatellite markers in environmental^{27–29}, veterinary³⁰, and clinical³¹ contexts, these studies focused on a few loci and therefore examine only a small fraction of the total genomic divergence among the isolates. For example, a recent large study of *A. flavus* found high levels of genetic diversity and genetically isolated populations that vary in extent of recombination³². However, this study focused on agricultural isolates from the USA and did not include clinical (i.e., human-

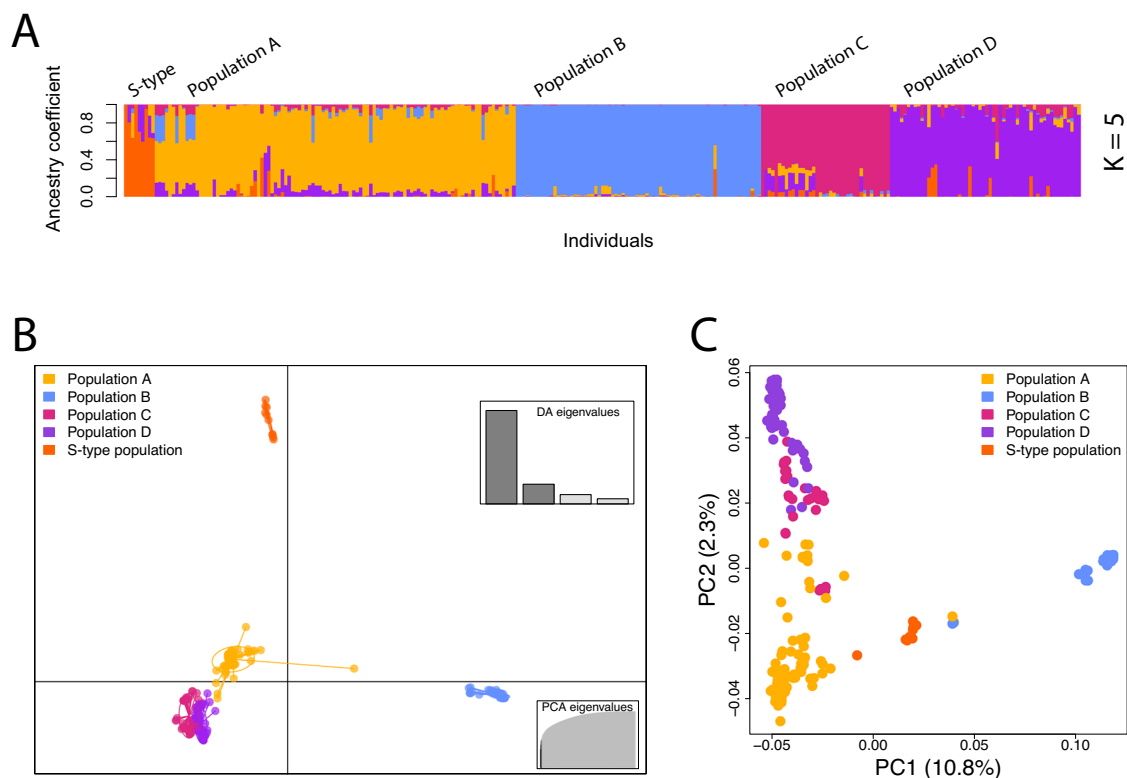


Fig. 1 | Population structure of *Aspergillus flavus* reveals genetic isolation reflecting five populations, including a new population, D. Analyses are based on 1,941,481 biallelic single nucleotide variants. **A** Estimates of individual ancestry, with $K = 5$, conducted using the software package LEA⁶⁷, which estimates individual admixture coefficients from a genotypic matrix⁹⁸. We estimated admixture for $K = 2$ through 10, with $K = 5$ providing the best fit for our data according to the cross-entropy criterion^{68,69}. **B** Discriminant analysis of principal components shows admixture among populations A, C, and D as well as clear separation of populations B and S-type. Dots represent individuals and ellipses indicate group clustering of individuals. Populations are color coded as indicated in the top right. The

discriminant analysis eigenvalues are shown on the bottom left, with the darker bars showing eigenvalues retained. **C** A principal coordinates analysis displaying relative genetic distances of individual isolates, here represented by dots, using Nei's genetic distance matrix. Axes indicate the two principal coordinates retained and the percentage of variance explained by each coordinate. Populations A, C, and D varied primarily along PC2 rather than PC1; population B showed genetic differentiation from other populations and varied primarily along PC1. Single nucleotide polymorphisms can be inferred from the alignment provided as our Source Data file.

Tree scale: 0.1

Source

- Clinical
- Environmental

Country

- Ethiopia
- Kenya
- France
- Germany
- India
- Japan
- Netherlands
- Pakistan
- Spain
- USA
- Other

Genetic Cluster

- S-type
- Population A
- Population B
- Population C
- Population D

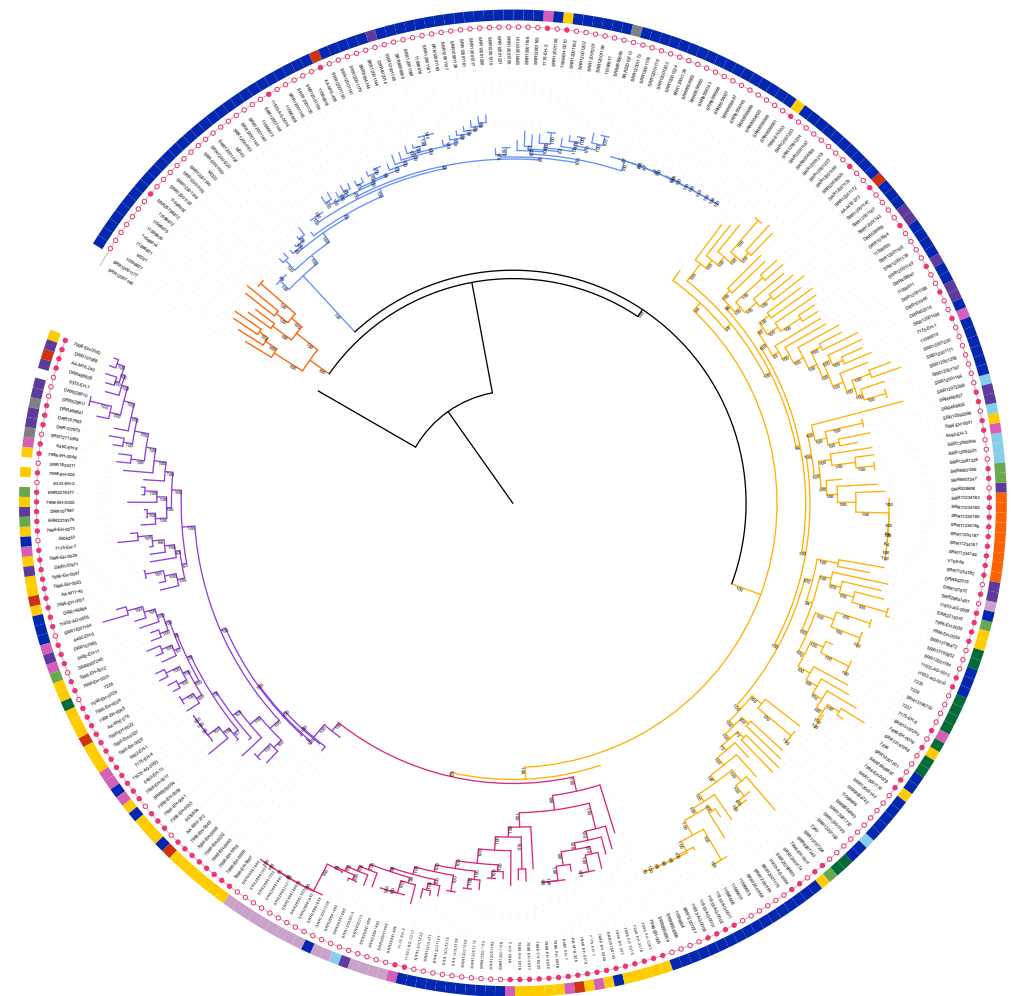


Fig. 2 | Maximum likelihood phylogeny supports the existence of five populations and non-random distribution of clinical isolates across populations. Filled in circles along the outer track indicate clinical isolates; empty circles indicate environmental isolates. Branch colors correspond with population assignment based on the discriminant analysis of principal components (DAPC; Fig. 1): the S-type population is indicated in orange; population A in yellow; population B in

blue, population C in pink, and population D in purple. The outgroup, *Aspergillus minisclerotigenes*, is represented in black. Apart from the outgroup, each tip represents an *A. flavus* isolate and branch lengths denote sequence divergence. Ultrafast bootstrap values are indicated below nodes. The phylogeny was constructed using 2,018,259 SNPs from a dataset of 300 *A. flavus* isolates and the outgroup *A. minisclerotigenes*. Source data are provided as a Source Data file.

associated) isolates, leaving the relationship between environmental and clinical isolates unexplored. Although public databases such as NCBI's GenBank and Sequence Read Archive contain many *A. flavus* genome assemblies and whole genome sequencing (WGS) datasets, only a small proportion of these are from clinical isolates^{33,34}. WGS data provide opportunities to study not only fine-scale population structure (population genomics), but also gene presence and absence across all available genomes in a species (pan-genomics). In a pan-genome, the core genome is defined as genes present in nearly all individuals while genes that are absent from some individuals are called accessory genes. The strong conservation of core genes is thought to represent the core metabolism and housekeeping functions of a species³⁵, with accessory genes encoding non-essential functions and possible local niche adaptations. Pan-genomic analyses are widespread in bacteria but have only recently been adopted in fungi³⁶. In *A. fumigatus*, pan-genomes have been used to identify genetic variants associated with human pathogenicity, recombination rates, and relatedness of clinical and environmental isolates^{37–39}.

In this study, we sequenced the genomes of 82 clinical isolates of *A. flavus* representing a diversity of human infection types. We combined our genomic data with publicly available genomic data to create

a dataset of 300 (117 clinical and 183 environmental) genomes. We analyzed the genomes of clinical and environmental isolates to (a) infer the population structure of the species, (b) investigate the relationship between population structure and isolation environment (clinical vs. environmental), (c) define the pan-genome of *A. flavus*, and (d) identify genetic elements that are associated with clinical isolates.

Results

Five populations of *A. flavus* were identified, with clinical isolates overrepresented in one population

For genomes from clinical isolates sequenced in this study, short-read sequencing resulted in over 10 million paired-end reads for each of the isolates (Supplemental Data 2). Trimming resulted in 10,326,680 to 53,093,447 paired reads per isolate (Supplemental Data 2).

To explore the population structure of *A. flavus*, we analyzed 1,941,481 biallelic SNPs from 281 isolates, after clone correction. We found evidence of five populations based on admixture and DAPC analyses (Fig. 1). In addition to the previously described A, B, C, and S-type populations, we discovered a new population, D (Fig. 1; Supplemental Data 3). We calculated admixture coefficients for each isolate (Fig. 1A), revealing higher levels of admixture in populations A, C, and

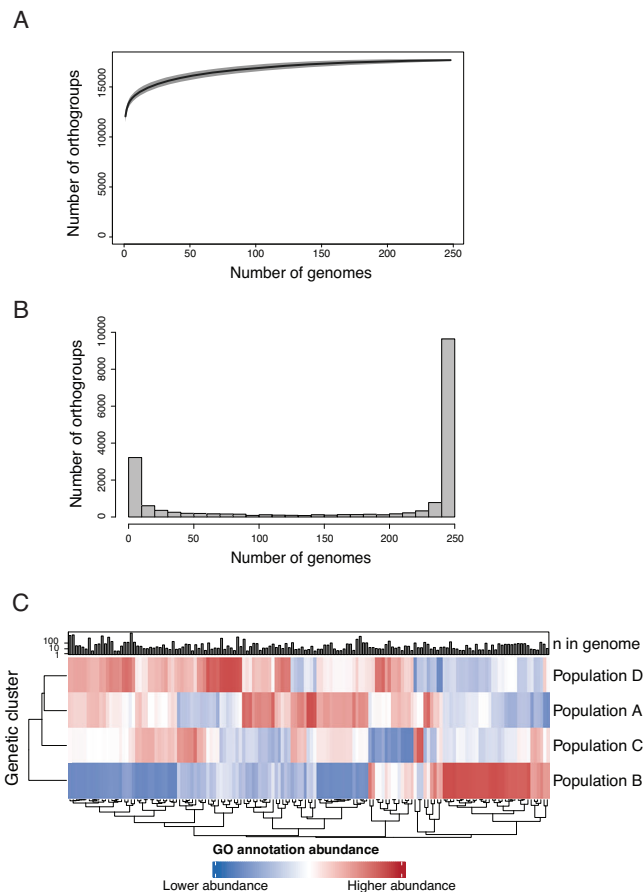


Fig. 3 | The pan-genome of *Aspergillus flavus* is closed and contains 17,676 orthogroups. **A** Rarefaction curve of number of orthogroups added with each additional genome, excluding singletons. Data are presented as mean of 100 permutations with random ordering of genomes added with standard deviation indicated. **B** Histogram of orthogroup frequency determined by number of genomes in which each orthogroups is present. The core genome contains 10,161 orthogroups. The accessory genome of *A. flavus* contains 7515 orthogroups. **C** Heatmap of normalized abundance of gene ontology (GO) annotations with significant differences in abundance among populations. Significance determined by one-way ANOVA. Bonferroni-corrected, $p < 0.05$. The bar chart shows the mean number of genes containing each GO term across all genomes. Source data in the form of predicted proteomes are provided in the associated Figshare repository and Supplemental Data 6.

D than in populations B and S-type. As with the admixture analysis, the DAPC also provided evidence of five populations in our dataset (Fig. 1B), based on BIC scores for each cluster. A principal coordinates analysis using Euclidian distance showed considerable overlap between populations A, C, and D, but populations B and S-type were distinct from others along PC1 (Fig. 1C). The S-type population was the smallest ($n = 9$) and included isolates previously confirmed to produce S-type sclerotia. Population A ($n = 114$) was the largest, followed by population B ($n = 68$), population D ($n = 54$), and population C ($n = 35$). The reference strain, *A. flavus* NRRL 3357, was placed in population A. The designated type strain for the species, *A. flavus* NRRL 1957, was placed in population D.

An overwhelming majority of isolates in both the S-type and B populations were from the USA, whereas the other three populations (A, C, and D) each contained isolates from at least five countries representing three or more continents. To examine the impact of geography on the population structure, we tested for isolation by distance⁴⁰. The Mantel test statistic, or Pearson's product-moment

correlation r , lies between -1 and 1 , ranging from a perfect negative correlation between the metrics tested and a perfect positive correlation, with 0 indicating no correlation. We saw evidence of isolation by distance in our dataset (Mantel test; $r = 0.173$, $p = 0.001$). We also saw evidence of isolation by distance within populations C and D, but not populations A or B (Fig. S1).

Population B was diverged from other populations, as Nei's genetic distance was high between population B and all other populations (D was above 0.111 for all comparisons between population B and all others). Nei's genetic distance was lowest between populations A and D ($D = 0.043$), indicating genetic similarities between the populations. Genetic differentiation across all five populations was consistent with the results of an AMOVA (Analysis of Molecular Variance) (global ϕ -statistic = 0.765 ; $p = 0$). Populations A, C, and D contained over 95% of the clinical isolates, whereas the S-type population contained exclusively environmental isolates and population B contained only three clinical isolates. Population D had the highest proportion of clinical isolates, with $>85\%$ of isolates in the population being patient-derived. Less than 40% of isolates assigned to population A, the largest population, were human-associated, in comparison. In population C, 51% of isolates were clinical. Clinical isolates originated from seven different countries: India, Japan, Netherlands, Germany, France, Spain, and the USA (Supplemental Data 1).

To examine the phylogenetic relationships among the isolates, we next constructed a maximum likelihood phylogeny using the full SNP dataset (1,941,481 SNPs), with *A. miniscerotigenes* (section *Flavi*) serving as an outgroup (Fig. 2). Clinical isolates were in all five clades, with each clade corresponding to a genetic population. All isolates with the "S" sclerotial morphotype were placed in a single monophyletic group, along with other isolates we infer to also be S-type (Fig. 2). Phylogenetic placement of isolates mostly supported the genetic populations (Fig. 2). Two clinical isolates were assigned to population A in the DAPC but exhibited considerable admixture with populations A and D (Fig. 1A). The maximum likelihood phylogeny recapitulated the DAPC results, with both isolates branching within the population A clade. However, in a neighbor-net network, one isolate was placed within population D rather than population A (Fig. S2). Unlike phylogenetic trees, neighbor-net networks capture additional nuance in relationships by including recombination. Based on the admixture analysis and neighbor-net network, these two clinical isolates, one from Germany, and one from India, possibly represent hybridizations between the A and D populations.

To test whether clinical isolates were non-randomly distributed across the *A. flavus* phylogeny, we calculated Fritz and Purvis's D statistic⁴¹; a value of 0 indicates a clumping of the observed trait (in this case, human pathogenicity) as expected under the Brownian motion model, whereas a value of 1 indicates a random distribution across the phylogeny. The model tests D against significant departure from 0 (Brownian motion model of evolution), as well as departure from 1 (random distribution). We calculated D for the dataset ($n = 300$ isolates) and observed a lack of random distribution ($D = 0.245$), which was significantly different from a value of 1 expected under random distribution ($p < 0.0001$), but not significantly different from a value of 0 expected under Brownian motion ($p = 0.067$). These results suggest that patient-derived isolates are not randomly distributed across the phylogeny; rather, they appear to predominately be found in certain populations (e.g., populations A, C, and D) and largely absent from others (e.g., population B).

Although we observed enrichment of clinical isolates in some clades, as supported by our D statistic, we recognize that our sampling of environmental and clinical isolates was uneven due to hospital culture collection location and availability of public data. We did however, include both clinical and environmental data from multiple areas of the globe. Despite the uneven sampling inherent in studies utilizing public data, populations with a large proportion of clinical isolates (A, C, and D)

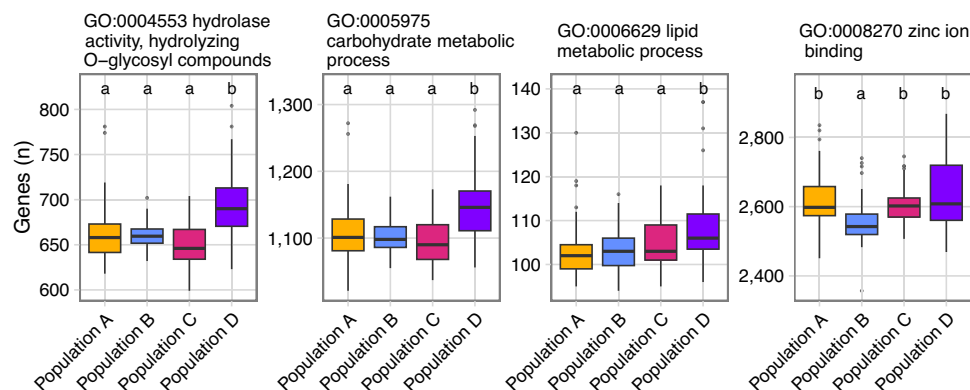


Fig. 4 | Gene ontology (GO) terms more prevalent in population D than other populations include zinc ion binding and hydrolase activity, among others.

Boxplots indicate the number of genes annotated with each GO term per isolate, by population assignment. The Y axis scale is adjusted for each GO term to better show differences among populations. In the box-and-whisker plots, the horizontal line in the box indicates the 50th percentile and the box extends from the 25th to the 75th percentile. The whiskers encompass the lowest and highest values within 1.5x the interquartile range. Individual dots represent statistical outliers. Statistical

significance was determined using a one-way analysis of variance followed by *post hoc* testing using Tukey's honest significance test (one-sided). The letters denote significances as a compact letter display where groups that are not significantly different from each other are indicated with the same alphabet letter; $P < 0.05$. After removing clones, low quality assemblies, and the S-type isolates, statistics were calculated for 232 isolates: population A ($n = 81$), population B ($n = 69$), population C ($n = 29$), population D ($n = 53$). Source data are provided in Supplemental Data 6.

include isolates from multiple countries across several continents (Fig. 3; Supplemental Data 1). Furthermore, even though the majority of clinical isolates are from Europe, we still observe a higher proportion of clinical isolates in population D (54%) than in A or C (25% and 8%, respectively) following exclusion of European isolates. Population D includes isolates from seven countries, including clinical isolates from Japan, India, and the USA. In contrast, population B isolates are almost entirely environmental from the USA, although the isolates represent diverse regions within the country, and we did not see strong evidence for isolation by distance in this population, as discussed above. We also did not see evidence of isolation by distance in population A, our largest and most geographically distributed population. Notably, no Kenyan or Ethiopian isolates, all of which were environmental, were placed in population D. With the available data, we conclude that populations enriched in clinical isolates are more globally widespread than populations with fewer clinical isolates (S-type and B populations), although this finding warrants further testing through sequencing of isolates from additional geographic regions.

A. flavus isolates exhibit heterogeneity in gene content and genome size

Strain heterogeneity in gene content can impact diverse traits, including virulence and secondary metabolite production, so we also examined our dataset in using methods independent of the reference genome. Genomes from 82 clinical isolates sequenced in this study contained 17 to 4235 scaffolds (Supplemental Data 2). We also assembled genomes for an additional 16 clinical and 12 environmental isolates from publicly available sequencing data, resulting in genomes containing 700 to 3615 scaffolds (Supplemental Data 2). Publicly available genome assemblies of environmental *A. flavus* isolates contained 8 to 1821 scaffolds^{32–34,42–46}. Genomes with low completeness based on BUSCO analysis (<95%) of the assemblies were excluded from the pan-genome analysis (Supplemental Data 4).

We examined variability in genome size among populations, which could indicate genetic expansions or streamlining. The mean genome size of population D was higher than both populations A and B (Tukey's multiple comparisons test, adjusted $p = 0.0175$), with all other pairwise comparisons of population means being non-significant (Fig. S3).

We annotated all genomes using Funannotate⁴⁷ to obtain consistent annotations for comparison. To ensure our annotation pipeline

resulted in high-quality proteomes, we compared the Funannotate predicted proteome of NRRL 3357 to the RefSeq reference annotation and an additional transcriptome-based annotation of the same strain⁴⁸; the two published proteomes contained 11 and 66 orthogroups that were not present in the Funannotate prediction, respectively, accounting for a tiny fraction of the gene content. Minor differences in gene prediction are expected due to gene fragmentation contributing to orthogroup variation. Overall, we are confident that the annotations predicted by the Funannotate pipeline are consistently high quality, enabling comparisons across isolates. The number of protein-coding genes predicted by Funannotate ranged from 11,461 to 15,501 across isolates (Supplemental Data 5).

The *A. flavus* pan-genome is closed and contains 17,676 orthogroups

To quantify the degree of gene presence-absence variation among isolates, we constructed a pan-genome of *A. flavus*. We used OrthoFinder to cluster predicted proteins into orthogroups, which were then compared across isolates and populations. Our pan-genome of *A. flavus* is closed (Heap's law, $\alpha = 1.000023$), with each genome after the 200th adding fewer orthogroups (Fig. 3A). We identified a total of 17,676 orthogroups. Of these, 10,161 (57.5%) orthogroups were in at least 95% of isolates; we consider these orthogroups to be the core genome. Within the core genome, 3375 orthogroups were single-copy and present in all isolates. The pan-genome of *A. flavus* exhibits a U-shaped distribution, as expected (Fig. 3B). The accessory pan-genome of *A. flavus* consists of 7515 orthogroups, of which 3387 (19.1% of all orthogroups) were in <5% of isolates and which we consider the "cloud" genome.

To explore which functional annotations were over or under-represented within populations, we examined presence or absence and abundance of InterPro annotations and gene ontology (GO) terms. Populations A, C, and D, which are enriched in clinical isolates, shared much of their gene content and did not show any population-specific patterning of orthogroup presence or absence in a PCoA, but population B, did (Fig. S4). We infer that gene content among population B isolates is more conserved and distinctive from other populations, likely due to low diversity within the population. In addition, we examined the abundance of GO terms and InterPro annotations and compared the mean among populations, excluding the S-type

population due to its small size. Populations had substantial differences in annotations and several GO terms were differentially abundant among populations (Fig. 3C). Given the over-representation of clinical isolates in population D, we focused on interpreting differences in functional annotations between population D and all other populations. Isolates in this population had a higher abundance of genes involved in many cellular processes, including certain types of hydrolase activity, nucleoside metabolic and carbohydrate metabolic processes, DNA-binding transcription factor activity, regulation of transcription, lipid metabolic process, NAD binding, catalytic activity, and acyltransferase activity (Supplemental Data 6; Fig. 4 and S5). Genes annotated with ferric iron binding functionalities were found in lower abundance in population D compared to other populations. Population B, the population with the fewest clinical isolates, was depleted in genes related to zinc ion binding (Fig. 4).

Genes in a putative non-ribosomal peptide synthesis (NRPS) BGC with an unknown product were absent in >90% of isolates within the S-type and B populations. The backbone gene for the NRPS cluster was present in all isolates in all populations, but additional biosynthetic or transporter genes within the BGC were absent from isolates within the S-type and B populations (G4B84_009247, G4B84_009246, G4B84_009245, and G4B84_009244 in *A. flavus* NRRL 3357, where all genes in the BGC are present). The BGC was previously identified as “BGC_44” on chromosome 6 but was not explored in depth²⁰. GO terms associated with multiple genes within the BGC (OG0011498 [GO:0000981; GO:0006355; GO:0008270]; OG0011868 [GO:0003824, GO:0006807]; OG0011918 [GO:0003824]) were also differentially abundant among populations (Supplemental Data 6).

Additionally, orthogroups related to aflatoxin biosynthesis (GO:0045122) were abundant at different levels among populations (Supplemental Data 6). Predictions from antiSMASH indicate the aflatoxin BGC was present in 68.5% of isolates ($n = 166/242$). In line with previous research²⁰, the BGC was present in almost all isolates in populations A, C, and S-type, but absent or degraded in many isolates within population B (Fig. S6). Interestingly, we found that the aflatoxin BGC was also absent or degraded in the newly defined population D, which has the most clinical isolates of any population.

Although virulence factors for human pathogenicity have not been identified in *A. flavus*, we examined 20 genes known to increase virulence in plant pathogenicity assays. All 20 genes were present in the isolates, but nucleotide percent similarity to the reference varied. For example, the stress-responsive transcription factor *skn7* showed substantial variation among isolates; *skn7* is part of the phosphorelay signal transduction system⁴⁹ associated with GO:0000160, which we saw in lower abundance in population A compared to our other populations (Supplemental Data 6).

Three orthogroups with high variability in number of gene family members were correlated with human-association

We used phylogenetic generalized least squares (PGLS) models to examine the relationships between multiple variables including isolate source (clinical or environmental), genome size, number of tRNAs, number of predicted genes, and number of predicted biosynthetic clusters. Using these phylogenetically informed linear regression models, we observed a correlation between genome size and number of predicted genes ($p = 0$; adjusted $R^2 = 0.4891$). We also examined the 10 orthogroups with the largest variability of gene family members from each isolate. We found that only three were significantly associated with isolate source (OG0000011, OG0000060, and OG0000270), all with low adjusted R-squared values (Supplemental Data 6); BLASTp results linked OG0000060 and OG0000270 with hypothetical proteins and implicated OG0000011 in natural product biosynthesis. OG0000011 was annotated with gene ontology terms “GO:0003824 catalytic activity” and “GO:0016746 transferase activity, transferring acyl groups,” which were both significantly differentially

abundant among populations (Supplemental Data 7). None of the PGLS models showed significant correlation between orthogroups and DAPC population assignment (Supplemental Data 3).

Discussion

In this study, we examined the population structure, phylogeny, and pan-genome of *A. flavus* using genomic data from 300 isolates, including 82 clinical isolates sequenced by us. By combining genomic data for both clinical and environmental isolates of this pathogen, our study provides a rich dataset for future study and revealing fine-scale differences in pathogenicity within *A. flavus*.

Previous research using genomic data of environmental isolates from the USA described three populations of *L*-type (isolates producing large sclerotia) *A. flavus* isolates: A, B, and C³². With the inclusion of additional isolates, notably clinical isolates, our study identifies another distinct population, here termed population D, which contains the majority of clinical isolates. Clinical isolates were present in all *L*-type populations, but at different abundances—populations A, C, and D were enriched in clinical isolates, with population D containing the highest proportion. Isolates with the small sclerotial morphotype (S-type) grouped together in all analyses, as seen previously³², and only a single clinical isolate was placed in this group. We did not expect many clinical isolates to be part of the S-type population, as S-type isolates produce conidia at a far lower rate than *L*-type isolates²⁶—interaction with these airborne asexual spores is how most patients are infected with *Aspergillus* species, so isolates producing fewer conidia are less likely to have spores interact with a human host.

Populations contained a combination of isolates from several different infections or microenvironments including soil, infections of different plant hosts (e.g., peanuts, corn, almonds), and different types of human infections (keratitis, aspergillosis, otomycosis, etc.), indicating a lack of specialization in populations. Previous work on environmental isolates showed no evidence of host specialization in *A. flavus*, with a single isolate able to infect both plant and animal hosts⁵⁰, which is consistent with our observation of the lack of clustering of isolates from the same microenvironment.

Interestingly, clinical isolates were concentrated in populations A, C, and D, with population D containing the majority of clinical isolates and few environmental ones. *Cryptococcus neoformans*, another opportunistic fungal pathogen of humans, shows similar enrichments of clinical isolates in some clades compared to others⁵¹, whereas the most common human pathogen in the genus *Aspergillus*, *A. fumigatus*, does not³⁷. In *A. flavus* populations A, C, and D, isolates did not cluster by country of origin, but we did observe a positive correlation between genetic and geographic distances in populations C and D, indicating that genetically divergent isolates were likely to be geographically distant. Geographic sampling within our dataset was not balanced, due to our use of public data and heavy sampling of clinical isolates within Europe. No environmental isolates from Europe had public data available, and none were available from large fungal culture collections, such as the NRRL or the Westerdijk Institute. The addition of environmental isolates from Europe in future studies will further test the validity of the inferred link between pathogenicity and population structure observed in our analyses.

We observed several important differences between *A. flavus* and the major human pathogen *A. fumigatus*. Our finding that some populations are highly enriched for clinical isolates in *A. flavus* contrasts from observations in *A. fumigatus*, wherein clinical isolates are more evenly distributed across all clades³⁷, highlighting the importance of studying *A. flavus* as a pathogen rather than assuming that pathogenicity in the two species evolved similarly. Ecological differences among species contribute to the various clinical presentations and prevalence of *Aspergillus* species causing aspergillosis, such as the ability to form biofilms, or the size of conidia⁵². Likewise, genetic differences among species may explain some of the variance and

prevalence of *A. flavus* related to *A. fumigatus*. *A. flavus* has a larger pan-genome than *A. fumigatus*, likely due to the difference in overall genome size between the species, with accessory genes composing a similar percentage of the pan-genome. Several pan-genomic studies have focused on *A. fumigatus*, with the core genome ranging from 55% to 69% of the pan-genome^{37–39}, compared to our finding of 57% in *A. flavus*. A recent review stated that *A. flavus* clinical isolates were more similar to one another and exhibited lower diversity than clinical isolates of *A. fumigatus*, which were more genetically diverse⁵³. *A. fumigatus* has many more genomes available from clinical isolates than *A. flavus* and the isolates are not associated with population structure as in *A. flavus*. However, our observation that clinical isolates are constrained to populations A, C, and D, which share genetic similarities and overlap in a principal coordinates analysis, supports a level of similarity among *A. flavus* clinical isolates despite deep phylogenetic divergences. We advocate for additional sequencing of clinical isolates, particularly from South America and Africa, as no genomes of clinical isolates are currently publicly available from these regions. Nevertheless, our analyses suggest that the core genome of *A. flavus* will remain similar even with the addition of new data; among the pan-genomic studies of *A. fumigatus* the core genome was consistent whereas the accessory genome varied based on input data^{37–39}.

Within the pan-genome of *A. flavus*, we observed several differences among populations. Several GO terms were enriched in population D, often in biological processes like carbohydrate, nucleoside, and lipid metabolism. Molecular functions enriched in population D included hydrolyzing O-glycosyl compounds, a function of exo-polygalacturonases, which are involved in the degradation of plant cell wall polysaccharides⁵⁴. None of the GO terms were directly implicated in functions typically related to pathogenicity, but several GO terms associated with metal acquisition were differentially abundant among populations. For example, zinc ion binding was depleted in population B. Zinc is an essential micronutrient for many fungal processes, and in *A. fumigatus*, deletion of a zinc acquisition factor attenuated virulence in a mouse model of aspergillosis⁵⁵; however, we do not yet know whether the lower abundance of genes annotated to involve zinc ion binding would correlate with lower virulence in population B of *A. flavus*.

In other fungal infections of humans, secondary metabolites have been implicated in virulence⁵⁶, most notably the role of gliotoxin in *A. fumigatus* infections. However, no secondary metabolites have been associated with *A. flavus* human infections. The most famous secondary metabolite produced by *A. flavus* is aflatoxin, which is not thought to be important for human infections as the optimum temperature for production of aflatoxins is below 37 °C, with transcription of the BGC dropping at higher temperatures⁵⁷. The predicted BGC for aflatoxin follows previously reported population-specific patterns²⁰, with presence or absence of the aflatoxin genes explained by clade and population.

Although both clinical and environmental isolates within populations A and C contained the aflatoxin BGC, isolates in population D often lacked genes related to aflatoxin biosynthesis. Population B, which included almost entirely environmental isolates, also lacked aflatoxin biosynthesis genes. Hospitals do not measure aflatoxin production for clinical isolates, leading to a paucity of production data for clinical isolates. However, it appears that although some clinical isolates in populations A and C may have the potential to produce aflatoxin, many clinical isolates in population D lack the necessary genes, and we expect them to be non-aflatoxigenic. The absence of the aflatoxin BGC within many clinical isolates of *A. flavus*, including almost all within population D, reinforces the apparent lack of association between aflatoxin and virulence in the context of human infections. Other predicted biosynthetic genes and gene clusters, such as BGC₄₄²⁰, which had accessory biosynthetic genes more prevalent in

population D than in other populations, have not been connected to metabolites and therefore their potential role in infection remains unknown.

In summary, we present evidence that clinical isolates of *A. flavus* share genetic similarities and are concentrated in certain populations rather than distributed across the phylogeny, particularly in an apparently non-aflatoxigenic, newly defined clade, which we have named population D. Clinical isolates from many countries and infection types are present in population D. We acknowledge that sampling was uneven and did not cover the full distribution of *A. flavus*, and advocate for additional sampling from regions under-represented in this dataset (e.g., environmental isolates from Europe and clinical isolates from South America and Africa). Additionally, accessory genes and the aflatoxin BGC differ between populations, possibly providing future opportunities for distinct agricultural and clinical treatments. Although we did not discover a single genetic element that could explain the difference between clinical and environmental isolates of *A. flavus*, we did discover a new clade of *A. flavus* that appears to be enriched in clinical isolates, with distinct genetic features. This *A. flavus* genomic dataset and pan-genome provide a valuable tool for understanding why isolates from some populations of *A. flavus* appear to be more commonly responsible for human infections than other populations.

Methods

Retrieval of publicly available data

We retrieved data for 180 *A. flavus* isolates with paired-end Illumina whole genome sequencing data available on National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) in July 2021, including data from 25 clinical isolates. In October 2024, we also retrieved paired-end whole genome sequencing data for 37 additional isolates deposited in SRA since the last search, composed of 10 clinical and 27 environmental isolates (Supplemental Data 1). *A. flavus* NRRL 3357, which has a chromosome-level genome assembly, was used as a reference. The publicly available data were from 10 countries: China (1), Ethiopia (7), India (7), Israel (1), Japan (21), Kenya (16), Netherlands (9), Pakistan (12), the USA (141), and Vietnam (1). Country of origin was unavailable for four isolates. Isolates represent diverse sources including patient-derived, soil, seed, and plant-associated micro-environments (Supplemental Data 1).

Collection of *A. flavus* clinical isolates and genome sequencing

We also sequenced 82 patient-derived isolates for this study, from 4 different countries: France (15 from the culture collection of the National Reference Center for Invasive Mycoses and Antifungals [CNRMA] at the Institut Pasteur), Germany (48 from the National Reference Center for Invasive Fungal Disease [NRZMyk]), Spain (7 from the National Centre for Microbiology [CNM] culture collection), and the USA (12 from the University of Texas M. D. Anderson Cancer Center). Isolates were from patients diagnosed with keratitis, aspergillosis, and otomycosis and were obtained through a variety of methods (Supplemental Data 1). Culture and DNA extraction methods varied and are available in the Supporting Information.

DNA libraries were prepared using the Nextera DNA Library PrepKit (Illumina, San Diego, CA, USA), according to manufacturer's guidelines. Sequencing for the isolates from Germany, France, and the USA (75 total) was performed at Vanderbilt University's sequencing facility, VANTAGE, using the Illumina NovaSeq 6000 instrument, following manufacturer's protocols. Sequencing for the seven Spanish CNM isolates was performed using the Illumina MiSeq system, following the manufacturer's protocols. All sequencing resulted in 150 bp paired-end reads. All isolates included in the study were confirmed to be *A. flavus* using whole genome sequencing data, in addition to preliminary classification based on morphology or marker genes.

Read mapping

By combining the publicly available data from 217 isolates with clinical isolates sequenced in this study, we compiled a dataset of 300 *A. flavus* isolates. Raw reads for all isolates were trimmed using Trimmomatic v0.39⁵⁸ for paired-end data. Trimmed reads were mapped to the NRRL 3357 reference (GCA_009017415.1)⁵⁹ using Bowtie2 v2.3.4.1 with default parameters⁶⁰. We used SAMtools v1.6⁶¹ to convert the resulting data files to BAM format and sort the BAM files. The AddOrReplaceReadGroups option in Picard tools v2.17.10 (<https://broadinstitute.github.io/picard/>) was used to append read group labels to BAM files. The Genome Analysis Tool Kit v3.8 (GATK) RealignerTargetCreator and IndelRealigner options were used to produce realigned BAM files⁶² and duplicates were removed using the MarkDuplicates option in Picard. We called variants for each genome using the GATK HaplotypeCaller option with -ploidy 1 for haploid organisms. GVCF files were combined using the CombineGVCFs option and the combined file genotyped using the GenotypeGVCFs option. Variants include single nucleotide polymorphisms, insertions, and deletions, so only SNPs were selected and retained. SNPs were filtered using the VariantFiltration option, with -filter-expression parameters “QD < 2.0”, “QUAL < 30”, “MQ < 40.0”, “MQRankSum < -12.5”, “SOR > 3.0”, “FS > 60.0”, and “ReadPosRankSum < -8.0”; other parameters were set as -cluster 8 and -window 10, according to the GATK best practices workflow (<https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4>).

Population genomics

Biallelic SNPs that passed hard filters were retained for further analysis for 300 isolates. Biallelic loci refer to loci which at the population level only have two alleles: the reference and an alternative. Clonal isolates were identified through genetic distance calculations in poppr v2.8.3⁶³ and were excluded from the dataset, leaving 281 isolates. We conducted a principal components analysis in R v4.3.1 using adegenet v2.1.10⁶⁴. Adegenet was also used for the discriminant analysis of principal components (DAPC), a multivariate method to determine the optimal number of genetic clusters for a given dataset⁶⁵. The Bayesian Information Criterion (BIC) score was used to evaluate a range of possible numbers of genetic clusters from 1–15. The optimal number of clusters was determined by graphing the BIC score for each possible number of clusters. Calculations of missing data per population and sample, minor allele frequency, and Nei’s genetic distance were conducted using the R package SambaR v1.09⁶⁶. The R package LEA v3.10.2⁶⁷ was used to estimate ancestry coefficients⁶⁸, from $K = 2$ to $K = 15$. Optimal K (number of populations) was determined using the entropy coefficient method⁶⁹.

To determine if molecular variation among populations was larger than variation within specific populations of *A. flavus*, we used Nei’s genetic distance⁷⁰ to implement an AMOVA (Analysis of Molecular Variance) in R v4.3.1 using the R package pegas v1.3.71 with 1000 permutations. A two-way ANOVA implemented in PRISM 10 (Graphpad) was used to establish the relationship between isolate sources (soil, plant-associated, and human-associated) and genetic populations identified by DAPC.

In population genomics, correlations between physical and genetic distance of isolates can impact population structure. As such, we tested for isolation by distance⁴⁰ using a Mantel test⁷² implemented in the R package dartR v2.9.73. Geographic location for clinical isolates was conservatively estimated by using the coordinates of each isolate’s culture collection. Nei’s genetic distance⁷⁰ was used for the genetic distance matrix. For the geographic distance matrix, latitude and longitude was either (1) obtained from previously published data or public metadata from NCBI or (2) estimated from listed hospital location for patient-derived isolates or city of sampling for environmental isolates. Isolates without latitude or longitude locations were considered missing data and coded as “NA” in the table.

Phylogenomics

Using the SNPs called and filtered as described above, we reconstructed a phylogeny of all 300 *A. flavus* isolates, with the close relative *Aspergillus minisclerotigenes* (SRA: SRR12001146) as an outgroup. Only loci present in at least eight isolates were included. We used Lewis ascertainment bias correction to include only variable (non-constant) characters from our SNP data⁷⁴. The phylogeny was constructed using IQ-Tree v2.2.2.6⁷⁵ using the ModelFinder Plus⁷⁶ (-m MFP) option and 1000 ultra-fast replicates for bootstrapping. The F81 + F model was chosen by IQ-Tree as the best-fit model according to BIC. The consensus tree was used for visualization. We used iTOL v6 to visualize and annotate the phylogeny⁷⁷. We also constructed a phylogenetic network using SplitsTreeCE v6.3.0⁷⁸ with default parameters to examine the relationships among isolates in a neighbor-net network.

To test whether clinical isolates were randomly distributed across the phylogeny or were more likely to be clustered (that is, whether clinical isolates had a phylogenetic signal), we calculated Fritz and Purvis’s D statistic for binary traits⁴¹ using the phylo.D command within the R package caper v1.0.3⁷⁹ with 1000 permutations.

Genome assembly and annotation

Draft genome assemblies were available from NCBI for 152 isolates (9 clinical and 143 environmental)^{25,32,42–44,80}, as of June 2021. For clinical isolates sequenced in this study, 16 clinical isolates with public data, and a further 12 environmental isolates from NCBI, genomes were assembled using trimmed reads (described above). Each de novo assembly was performed using SPAdes v3.15.0⁸¹ with default parameters except for k-mer count (set to 21, 33, 55, 77, 99, and 127). For all assemblies, scaffolds were filtered using Funannotate v1.8.10⁶⁹ to remove duplicate sequences and those under 500 bp in length. Scaffolds were masked for repeats using RepeatMasker⁸² within Funannotate v1.8.10. Mitochondrial sequences and any bacterial, primate, or viral contaminants identified through routine screening upon submission to NCBI were removed from the genome assemblies. All genomes were evaluated for completeness using the BUSCO v4.04 Eurotiales database of 4,191 single-copy genes⁸³. Several NCBI datasets did not have adequate sequencing depth for genome assembly, had contaminating reads, or had BUSCO scores below 95% and were therefore not included in the pan-genome analyses; we also removed clones as calculated using genetic distance based on SNPs. As such, we included 243 genomes in the pan-genome.

Gene predictions were generated by Funannotate v1.8.10 using the built-in gene models of *Aspergillus oryzae* (section *Flavi*) as predicted by Evidence Modeler⁸⁴, with additional evidence provided in the form of predicted amino acid sequences for proteins from the *A. flavus* NRRL 3357 annotation⁵⁹. To validate the gene-prediction procedure, we compared the new annotation of NRRL 3357 to two recent in-depth annotations^{48,59} using OrthoVenn2⁸⁵. Additional functional annotations were obtained through the “annotate” option within Funannotate that uses InterProScan v5.61.93⁸⁶ with default parameters. Twenty-one genes associated with plant pathogenicity in *A. flavus* were previously identified from the literature²⁵ and located in each genome via BLASTp⁸⁷. Global alignments were conducted using the Needleman-Wunsch algorithm implemented in EMBOSS v6.6.0⁸⁸. Predicted biosynthetic gene clusters (BGCs) were identified using the fungal version of antiSMASH v6.0⁸⁹, with default parameters, and collated into a table format. For specific clusters of interest, we used BLASTn to confirm the presence or absence of backbone genes as defined by antiSMASH (core biosynthetic genes), e.g., querying the *pksA* nucleotide sequence from the *A. flavus* reference strain NRRL 3357 against all genomes to confirm presence.

Pan-genome analysis

We identified orthologous proteins using OrthoFinder v2.5.4⁹⁰ in all *A. flavus* genomes that had $\geq 95\%$ completeness of the 4,191 genes in the

BUSCO Eurotiales gene set⁸³, resulting in a dataset of 247 isolates. The core genome was defined as in Lofgren et al.³⁸ as the set of genes that were present in at least 95% of isolates (in our dataset 236 or more); all other genes were considered part of the accessory genome. A subset of the accessory genome, the “cloud” genome includes orthogroups present in <5% of isolates. The presence/absence matrix of the accessory genome was visualized using the R package Complex Heatmap v2.16.0⁹¹. We created a gene accumulation curve and gene frequency histogram using the R packages vegan v2.6-10^{92,93}, phylentropy v0.9.0⁹⁴, and ggplot2 v3.5.1⁹⁵. Using vegan v2.6-10⁹², we also calculated a distance matrix for the presence or absence of orthogroups within the accessory genome using Jaccard distance. The distance matrix was then used as input for a principal coordinates analysis (weighted classical multidimensional scaling) to visualize population-level differences in accessory genome content. The accessory genome principal coordinates analysis was visualized using ggplot2 v3.5.1⁹⁵. Also using vegan v2.6-10⁹², the gene accumulation curve was calculated using the “random” method and 100 permutations. The alpha for Heap’s law was calculated using the R package micropan v2.1⁹⁶. Orthogroups were considered population-specific when absent in all isolates of a particular population but present in >90% of isolates in other populations, consistent with definitions from Lofgren et al.³⁸.

Orthogroups were associated with locus tags and functional annotations using a custom Python v3.9 script. Analysis of functional annotation differences among the populations was performed by ANOVA in R v4.3.1 using the number of genes in each isolate’s genome that contained each annotation as input. Tukey’s HSD was used for *post hoc* testing of gene ontology term abundance with a significance cutoff of $p < 0.05$. Statistics and Bonferroni false discovery rate correction were performed using base R v4.3.1. Heatmaps were constructed using the R package Complex Heatmap v2.16.0, and box plots were made using ggplot v3.4.4.

We used a phylogenetic generalized least squares (PGLS) analysis as conducted in the R package caper v1.0.3⁷⁹ to evaluate whether traits were more likely to be shared by closer relatives in accordance with a Brownian motion model of evolution. PGLS analyses incorporate the phylogenetic relationships between individual data points when examining linear regression of variables⁹⁷. We fit a model of genome size against the number of predicted genes, number of predicted tRNAs, and the number of predicted BGCs using a maximum likelihood estimate of lambda. We also fit a model to explain source (clinical or environmental) using the 10 orthogroups with the most variation in number of genes included in the orthogroup.

Data availability

Data associated with genomes from clinical isolates sequenced in this study, including paired-end reads and draft genome assemblies, are available under BioProject PRJNA836245. Accession numbers for previously published data used as part of this study are available in Supplementary Table 1. Predicted genes and annotations are available on FigShare at <https://doi.org/10.6084/m9.figshare.28830200>. Source data for Figs. 1 and 2 are provided in the source data file, source data for Fig. 3 in the form of predicted proteomes are provided in the Figshare repository (<https://doi.org/10.6084/m9.figshare.28830200>) and in Supplementary Data 6, source data for Fig. 4 is provided in Supplementary Data 6. Source data are provided with this paper.

References

- Rudramurthy, S. M., Paul, R. A., Chakrabarti, A., Mouton, J. W. & Meis, J. F. Invasive aspergillosis by *Aspergillus flavus*: epidemiology, diagnosis, antifungal resistance, and management. *J. Fungi* **5**, 55 (2019).
- Bongomin, F., Gago, S., Oladele, R. O. & Denning, D. W. Global and multi-national prevalence of fungal diseases—estimate precision. *J. Fungi* **3**, 57 (2017).
- Chong, W. H. & Neu, K. P. Incidence, diagnosis and outcomes of COVID-19-associated pulmonary aspergillosis (CAPA): a systematic review. *J. Hosp. Infect.* **113**, 115–129 (2021).
- Nasir, N., Farooqi, J., Mahmood, S. F. & Jabeen, K. COVID-19-associated pulmonary aspergillosis (CAPA) in patients admitted with severe COVID-19 pneumonia: an observational study from Pakistan. *Mycoses* **63**, 766–770 (2020).
- Schauwvlieghe, A. F. A. D. et al. Invasive aspergillosis in patients admitted to the intensive care unit with severe influenza: a retrospective cohort study. *Lancet Respir. Med.* **6**, 782–792 (2018).
- Devoto, T. B. et al. Molecular epidemiology of *Aspergillus* species and other moulds in respiratory samples from Argentinean patients with cystic fibrosis. *Med. Mycol.* **58**, 867–873 (2020).
- Truda, V. S. S. et al. A contemporary investigation of burden and natural history of aspergillosis in people living with HIV/AIDS. *Mycoses* **66**, 632–638 (2023).
- Denning, D. W. Global incidence and mortality of severe fungal disease. *Lancet Infect. Dis.* **24**, e428–e438 (2024).
- Ghosh, A. K. et al. Fungal keratitis in North India: spectrum of agents, risk factors and treatment. *Mycopathologia* **181**, 843–850 (2016).
- Brown, L., Leck, A. K., Gichangi, M., Burton, M. J. & Denning, D. W. The global incidence and diagnosis of fungal keratitis. *Lancet Infect. Dis.* **21**, e49–e57 (2021).
- Gheith, S. et al. Characteristics of invasive aspergillosis in neutropenic haematology patients (Sousse, Tunisia). *Mycopathologia* **177**, 281–289 (2014).
- Sarigüzél, F. M. et al. Molecular epidemiology and antifungal susceptibilities of *Aspergillus* species isolated from patients with invasive aspergillosis. *Rev. Assoc. Med. Bras.* **69**, 44–50 (2023).
- Dabas, Y. et al. Epidemiology and antifungal susceptibility patterns of invasive fungal infections (IFIs) in India: a prospective observational study. *J. Fungi* **8**, 33 (2021).
- Fedorova, N. D. et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet* **4**, e10000046 (2008).
- Rokas, A. Evolution of the human pathogenic lifestyle in fungi. *Nat. Microbiol.* **7**, 607–619 (2022).
- Steinbach, W. J. et al. Clinical epidemiology of 960 patients with invasive aspergillosis from the PATH alliance registry. *J. Infect.* **65**, 453–464 (2012).
- Walther, G. et al. Eye infections caused by filamentous fungi: spectrum and antifungal susceptibility of the prevailing agents in Germany. *J. fungi* **7**, 511 (2021).
- Hedayati, M. T., Pasqualotto, A. C., Warn, P. A., Bowyer, P. & Denning, D. W. *Aspergillus flavus*: human pathogen, allergen and mycotoxin producer. *Microbiology* **153**, 1677–1692 (2007).
- Eaton, D. L. & Gallagher, E. P. Mechanisms of aflatoxin carcinogenesis. *Annu. Rev. Pharm. Toxicol.* **34**, 135–172 (1994).
- Drott, M. T. et al. Microevolution in the pansecondary metabolome of *Aspergillus flavus* and its potential macroevolutionary implications for filamentous fungi. *Proc. Natl. Acad. Sci. USA* **118**, e2021683118 (2021).
- Raffa, N. & Keller, N. P. A call to arms: Mustering secondary metabolites for success and survival of an opportunistic pathogen. *PLoS Pathog.* **15**, e1007606 (2019).
- Vidal-Garcia, M. et al. Production of the invasive aspergillosis biomarker bis (methylthio) gliotoxin within the genus *Aspergillus*: in vitro and in vivo metabolite quantification and genomic analysis. *Front Microbiol.* **9**, 1246 (2018).
- Dowd, P. F. Synergism of aflatoxin B1 toxicity with the co-occurring fungal metabolite kojic acid to two caterpillars. *Entomol. Exp. Appl.* **47**, 69–71 (1988).
- Lan, H. et al. Investigation of *Aspergillus flavus* in animal virulence. *Toxicon* **145**, 40–47 (2018).

25. Hatmaker, E. A. et al. Genomic and phenotypic trait variation of the opportunistic human pathogen *Aspergillus flavus* and its close relatives. *Microbiol. Spectr.* **10**, e03069-22 (2022).
26. Chang, P.-K., Ehrlich, K. C. & Hua, S.-S. T. Cladal relatedness among *Aspergillus oryzae* isolates and *Aspergillus flavus* S and L morphotype isolates. *Int. J. Food Microbiol.* **108**, 172–177 (2006).
27. Singh, P., Mehl, H. L., Orbach, M. J., Callicott, K. A. & Cotty, P. J. Genetic diversity of *Aspergillus flavus* associated with chili in Nigeria and identification of haplotypes with potential in aflatoxin mitigation. *Plant Dis.* **106**, 1818–1825 (2022).
28. Acur, A. et al. Genetic diversity of aflatoxin-producing *Aspergillus flavus* isolated from groundnuts in selected agro-ecological zones of Uganda. *BMC Microbiol.* **20**, 252 (2019).
29. Drott, M. T., Fessler, L. M. & Milgroom, M. G. Population subdivision and the frequency of aflatoxigenic isolates in *Aspergillus flavus* in the United States. *Phytopathology* **109**, 878–886 (2019).
30. Cherif, G. et al. *Aspergillus flavus* genetic structure at a turkey farm. *Vet. Med Sci.* **9**, 234–241 (2023).
31. Choi, M. J. et al. Microsatellite typing and resistance mechanism analysis of voriconazole-resistant *Aspergillus flavus* isolates in South Korean hospitals. *Antimicrob. Agents Chemother.* **63**, 10–1128 (2019).
32. Drott, M. T. et al. The frequency of sex: population genomics reveals differences in recombination and population structure of the aflatoxin-producing fungus *Aspergillus flavus*. *mBio* **11**, 1–13 (2020).
33. Buil, J. B. et al. Genetic and phenotypic characterization of in-host developed azole-resistant *Aspergillus flavus* isolates. *J. Fungi* **7**, 164 (2021).
34. Toyotome, T. et al. Comparative genome analysis of *Aspergillus flavus* clinically isolated in Japan. *DNA Res.* **26**, 95–103 (2019).
35. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, e1002608 (2012).
36. McCarthy, C. G. P. & Fitzpatrick, D. A. Pan-genome analyses of model fungal species. *Micro. Genom.* **5**, e000243 (2019).
37. Barber, A. E. et al. *Aspergillus fumigatus* pan-genome analysis identifies genetic variants associated with human infection. *Nat. Microbiol.* **6**, 1526–1536 (2021).
38. Lofgren, L. A., Ross, B. S., Cramer, R. A. & Stajich, J. E. The pan-genome of *Aspergillus fumigatus* provides a high-resolution view of its population structure revealing high levels of lineage-specific diversity driven by recombination. *PLoS Biol.* **20**, e3001890 (2022).
39. Horta, M. A. C. et al. Examination of genome-wide ortholog variation in clinical and environmental isolates of the fungal pathogen *Aspergillus fumigatus*. *mBio* **13**, e01519-22 (2022).
40. Wright, S. Isolation by distance. *Genetics* **28**, 114 (1943).
41. Fritz, S. A. & Purvis, A. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24**, 1042–1051 (2010).
42. Weaver, M. A., Mack, B. M. & Gilbert, M. K. Genome sequences of 20 georeferenced *Aspergillus flavus* isolates. *Microbiol. Resour. Announc.* **8**, e01718–e01718 (2019).
43. Yin, G. et al. Genome sequence and comparative analyses of atoxigenic *Aspergillus flavus* WRR1 1519. *Mycologia* **110**, 482–493 (2018).
44. Arias, R. S. et al. Sixteen draft genome sequences representing the genetic diversity of *Aspergillus flavus* and *Aspergillus parasiticus* colonizing peanut seeds in Ethiopia. *Microbiol. Resour. Announc.* **9**, 10–1128 (2020).
45. Ajmal, M. et al. Characterization of 260 isolates of *Aspergillus* section Flavi obtained from sesame seeds in Punjab, Pakistan. *Toxins* **14**, 117 (2022).
46. Pennerman, K. K., Yin, G., Bennett, J. W. & Hua, S.-S. T. *Aspergillus flavus* NRRL 35739, a poor biocontrol agent, may have increased relative expression of stress response genes. *J. Fungi* **5**, 53 (2019).
47. Palmer, J. M. & Stajich, J. E. Funannotate v1. 8.1: eukaryotic genome annotation. Zenodo <https://doi.org/10.3389/fmicb.2019.01013> (2020).
48. Hatmaker, E. A. et al. Revised transcriptome-based gene annotation for *Aspergillus flavus* strain NRRL 3357. *Microbiol. Resour. Announc.* **9**, e01155–20 (2020).
49. Fassler, J. S. & West, A. H. Fungal Skn7 stress responses and their relationship to virulence. *Eukaryot. Cell* **10**, 156–167 (2011).
50. St Leger, R. J., Screen, S. E. & Shams-Pirzadeh, B. Lack of host specialization in *Aspergillus flavus*. *Appl Environ. Microbiol.* **66**, 320–324 (2000).
51. Desjardins, C. A. et al. Population genomics and the evolution of virulence in the fungal pathogen *Cryptococcus neoformans*. *Genome Res.* **27**, 1207–1219 (2017).
52. Paulussen, C. et al. Ecology of aspergillosis: insights into the pathogenic potency of *Aspergillus fumigatus* and some other *Aspergillus* species. *Micro. Biotechnol.* **10**, 296–322 (2017).
53. Freese, J. & Beyhan, S. Genetic diversity of human fungal pathogens. *Curr. Clin. Microbiol. Rep.* **10**, 17–28 (2023).
54. de Vries, R. P. & Visser, J. *Aspergillus* enzymes involved in degradation of plant cell wall polysaccharides. *Microbiol. Mol. Biol. Rev.* **65**, 497–522 (2001).
55. Crawford, A. & Wilson, D. Essential metals at the host–pathogen interface: nutritional immunity and micronutrient assimilation by human fungal pathogens. *FEMS Yeast Res.* **15**, fov071 (2015).
56. Spikes, S. et al. Gliotoxin production in *Aspergillus fumigatus* contributes to host-specific differences in virulence. *J. Infect. Dis.* **197**, 479–486 (2008).
57. Yu, J. et al. Tight control of mycotoxin biosynthesis gene expression in *Aspergillus flavus* by temperature as revealed by RNA-Seq. *FEMS Microbiol. Lett.* **322**, 145–149 (2011).
58. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
59. Skerker, J. M. et al. Chromosome assembled and annotated genome sequence of *Aspergillus flavus* NRRL 3357. *G3* **11**, jkab213 (2021).
60. Ben, L. & Salzberg, S. L. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).
61. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. McKenna, A. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
63. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).
64. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
65. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 1–15 (2010).
66. de Jong, M. J., de Jong, J. F., Hoelzel, A. R. & Janke, A. SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. *Mol. Ecol. Resour.* **21**, 1369–1379 (2021).
67. Frichot, E. & François, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925–929 (2015).
68. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12**, 1–6 (2011).
69. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).

70. Nei, M. Genetic distance between populations. *Am. Nat.* **106**, 283–292 (1972).
71. Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419–420 (2010).
72. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
73. Gruber, B., Unmack, P. J., Berry, O. F. & Georges, A. dartr: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* **18**, 691–699 (2018).
74. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
75. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
76. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
77. Letunic, I. & Bork, P. Interactive tree Of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
78. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
79. Orme, D. et al. The caper package: comparative analysis of phylogenetics and evolution in R. *R. Package Version* **5**, 1–36 (2013).
80. Gebru, S. T. et al. Draft genome sequences of 20 *Aspergillus flavus* isolates from corn kernels and cornfield soils in Louisiana. *Microbiol. Resour. Announc.* **9**, e00826–20 (2022).
81. Bankevich, A. et al. SPAdes: A New genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
82. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. <https://www.repeatmasker.org/faq.html> (2025).
83. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
84. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
85. Xu, L. et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, W52–W58 (2019).
86. Jones, P. et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–40 (2014).
87. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
88. Rice, P., Longden, I. & Bleasby, A. EMBL: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
89. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
90. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
91. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
92. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
93. Oksanen, J. et al. Package ‘Vegan’. *Community Ecology Package, Version*. <https://cran.r-project.org/web/packages/vegan/vegan.pdf> (2019).
94. Drost, H.-G. Philentropy: information theory and distance quantification with R. *J. Open Source Softw.* **3**, 765 (2018).
95. Ginestet, C. ggplot2: Elegant graphics for data analysis. *J. R. Stat. Soc. Ser. A Stat. Soc.* **174**, 245–246 (2011).
96. Snipen, L. & Liland, K. H. micropan: an R-package for microbial pan-genomics. *BMC Bioinforma.* **16**, 1–8 (2015).
97. Symonds, M. R. E. & Blomberg, S. P. A primer on phylogenetic generalised least squares. In *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice* (Garamszegi, L.) 105–130 (Springer, 2014).
98. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

Acknowledgements

This work was partially funded by the National Institutes of Health/ National Eye Institute (F31 EY033235 to E.A.H.) and the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356 to A.R.). Research in A.R.’s lab is also supported by the National Science Foundation (DEB-2110404) and the Burroughs Wellcome Fund. A.E.B. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research 358 Foundation) under Germany’s Excellence Strategy—EXC 20151—Project-ID 390813860. M.T.D. is supported by the United States Department of Agriculture, Agricultural Research Service. Work of the German NRZMyk is supported by the Robert Koch Institute from funds provided by the German Ministry of Health (grant-No. 1369-240). A. A.-I. is supported by Fondo de Investigaciones Sanitarias from Instituto de Salud Carlos III (grant-No. PI20CIII/00043). N.P.K. and J.L.E. are supported by the National Institutes of Health/National Institutes of General Medical Science (R01 GM112739 and R35 GM156119 to N.P.K.).

Author contributions

E.A.H. and A.R. conceptualized and designed the research and acquired funding. E.A.H., A.E.B., M.T.D., A.G., A.A.-I., D.G.-H. and J.L.E. performed experiments. E.A.H., A.E.B., M.T.D., and T.J.C.S. analyzed data. A.E.B., A.A.-I., D.G.-H., D.P.K., and O.K. provided key materials. A.R. and N.P.K. supervised the research. E.A.H. wrote and prepared the original draft. E.A.H., A.E.B., M.T.D., T.J.C.S., A.G., A.A.-I., D.G.-H., J.L.E., N.P.K., D.P.K., O.K., and A.R. reviewed and edited the manuscript.

Competing interests

A.R. is a scientific consultant for LifeMine Therapeutics, Inc. The other authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62777-9>.

Correspondence and requests for materials should be addressed to Antonis Rokas.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

¹Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA. ²Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN, USA.

³Institute for Microbiology, Friedrich Schiller University, Jena, Germany. ⁴Cereal Disease Laboratory, Agricultural Research Service, USDA, Saint Paul, MN, USA. ⁵Mycology Reference Laboratory, National Center for Microbiology, Instituto de Salud Carlos III, Madrid, Spain. ⁶Center for Biomedical Research in Network in Infectious Diseases (CIBERINFEC), Carlos III Health Institute, Madrid, Spain. ⁷Institut Pasteur, Université Paris Cité, National Reference Center for Invasive Mycoses and Antifungals, Translational Mycology Research Group, Mycology Department, Paris, France. ⁸Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI, USA. ⁹Department of Plant Pathology, University of Wisconsin-Madison, Madison, WI, USA.

¹⁰Department of Infectious Diseases, University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹¹National Reference Center for Invasive Fungal Infections NRZMyk, Leibniz Institute for Natural Product Research and Infection Biology—Hans-Knoell-Institute, Jena, Germany. ¹²Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany. ✉ e-mail: antonis.rokas@vanderbilt.edu