ARTICLES
https://doi.org/10.1038/s41564-021-00993-x

nature microbiology

Check for updates

Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection

Amelia E. Barber[1,8,9], Tongta Sae-Ong[2,9], Kang Kang[2], Bastian Seelbinder[2], Jun Li[3,4], Grit Walther[5], Gianni Panagiotou[2,6] ✉ and Oliver Kurzai[1,5,7] ✉

Aspergillus fumigatus **is an environmental saprobe and opportunistic human fungal pathogen. Despite an estimated annual occurrence of more than 300,000 cases of invasive disease worldwide, a comprehensive survey of the genomic diversity present in *A. fumigatus*—including the relationship between clinical and environmental isolates and how this genetic diversity contributes to virulence and antifungal drug resistance—has been lacking. In this study we define the pan-genome of *A. fumigatus* using a collection of 300 globally sampled genomes (83 clinical and 217 environmental isolates). We found that 7,563 of the 10,907 unique orthogroups (69%) are core and present in all isolates and the remaining 3,344 show presence/absence of variation, representing 16–22% of the genome of each isolate. Using this large genomic dataset of environmental and clinical samples, we found an enrichment for clinical isolates in a genetic cluster whose genomes also contain more accessory genes, including genes coding for transmembrane transporters and proteins with iron-binding activity, and genes involved in both carbohydrate and amino-acid metabolism. Finally, we leverage the power of genome-wide association studies to identify genomic variation associated with clinical isolates and triazole resistance as well as characterize genetic variation in known virulence factors. This characterization of the genomic diversity of *A. fumigatus* allows us to move away from a single reference genome that does not necessarily represent the species as a whole and better understand its pathogenic versatility, ultimately leading to better management of these infections.**

Diseases caused by the mould *Aspergillus fumigatus* are a major cause of human morbidity and mortality[1,2]. Invasive aspergillosis is particularly problematic in immunocompromised patients, resulting in a mortality rate of up to 50%[3,4]. Treatment of infections caused by *A. fumigatus* relies on triazole antifungal drugs. However, resistance to these frontline therapies is increasing, and the mortality rate for resistant infections is 25% higher than susceptible infections[5,6]. Although the most frequently identified resistance mutations occur in the cellular target of the triazoles—that is, *cyp51a*—up to 30% of the resistant isolates have no identifiable resistance mechanisms[7], complicating the recognition and treatment of these problematic infections.

While the host immune status is an important determinant in the development of aspergillosis, the substantial phenotypic variability observed among *A. fumigatus* isolates indicates that intraspecies diversity also plays a role in the disease[8–14]. This includes marked differences in virulence in animal models[9,10,14], fitness under hypoxia[9], growth under chemical stress(es)[11], nutritional heterogeneity[12] and induction of host inflammatory mediators[8]. As an indicator of the genomic diversity underlying the phenotypic variability observed in *A. fumigatus*, genomic comparisons between the reference strains Af293 and A1163 reveal tracts of variable gene content between the two[15], and 7% of Af293 genes are not present in A1163 (FungiDB). Despite this variation, previous studies of *A. fumigatus* have largely only analysed genomic information in the context of the reference genome and were limited to the genetic material present in Af293 due to the technical challenges of de novo eukaryotic genome analysis[16–19]. In addition, most of the isolates that have been sequenced to date are of clinical origin, thereby obscuring the genomic relationship between environmental isolates and those causing human disease.

In this study we constructed de novo genome assemblies of 300 *A. fumigatus* genomes ($n=217$ environmental isolates and $n=83$ clinical isolates) and used them to define the pan-genome of this important human fungal pathogen as well as the relationship between environmental and clinical isolates. We also leveraged the power of genome-wide association studies (GWAS) to identify genomic variation associated with human infection and triazole resistance, revealing a new range of therapeutic targets to combat these life-threatening infections.

Results

De novo assembly of 300 *A. fumigatus* genomes. In this study we used reference-guided and de novo assembly methods to analyse

[1]Research Group Fungal Septomics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [2]Research Group Systems Biology and Bioinformatics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [3]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China. [4]School of Data Science, City University of Hong Kong, Hong Kong, China. [5]National Reference Center for Invasive Fungal Infections (NRZMyk), Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [6]Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China. [7]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany. [8]Present address: Junior Research Group Fungal Informatics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [9]These authors contributed equally: Amelia E. Barber, Tongta Sae-Ong. ✉e-mail: gianni.panagiotou@leibniz-hki.de; okurzai@hygiene.uni-wuerzburg.de

the genomes of 300 *A. fumigatus* isolates, representing environmental and clinical isolates from different locations across the globe. Among these, 188 samples were novel environmental and clinical isolates from Germany that were sequenced as part of this study. The remaining 112 isolates, including 64 isolates that were sequenced by us in a previous study[20], were pulled from public data repositories as raw sequence data. Our overall dataset was comprised of 217 environmental isolates and 83 clinical isolates from Europe, Asia, North America, South America and the International Space Station (Supplementary Data 1). Forty-three of 294 isolates were resistant to one or more medical triazoles, as determined by European Committee on Antimicrobial Susceptibility Testing (EUCAST) broth microdilution[21]. Azole susceptibility data were not available for six isolates. We generated de novo genome assemblies of these 300 isolates using paired-end Illumina sequencing to facilitate the unrestricted analysis of genomic diversity in *A. fumigatus*. The mean number of contigs in our assemblies was 948 and the mean N50, a marker of genome contiguity representing the weighted median contig length, was 145,494 base pairs (bp; Supplementary Table 1 and Supplementary Data 1). The mean genome size of our assemblies was 28.6 Mb (range, 26.9–30.8 Mb), with an average of 9,408 open reading frames (ORFs) per isolate and a range of 9,169 to 11,231. Using BUSCO as a measure of genome completeness, we found that an average of 97% of the expected single-copy orthologues were found and present as single copies in our genome assemblies.

To perform population genomic analyses, we aligned reads against the Af293 reference genome. We observed an average of 78,692 single nucleotide variants (SNVs) per isolate (range, 23,029–149,537) or approximately three SNVs per kilobase (Supplementary Data 1). We also detected an average of 7,383 short insertions or deletions (indels) per isolate (range, 2,528–16,134). Of the 329,405 non-redundant SNVs identified among our isolates, 33% (107,779) were not described in FungiDB, release 39. Together, this reveals a pronounced level of genetic diversity in *A. fumigatus* at the nucleotide level and considerably extends the previously recognised diversity.

The *A. fumigatus* pan-genome contains 7,563 core and 3,344 accessory genes. To examine the full genomic diversity of *A. fumigatus*, we used our de novo genome assemblies to define and characterize its pan-genome. The pan-genome is the collective gene set of a species and is composed of core genes found in all individuals and accessory genes that are not shared between all members of the species. We identified a total of 12,798 gene clusters that condensed into 10,907 non-redundant orthogroups. The *A. fumigatus* pan-genome was composed of a core genome of 7,563 orthogroups in all 300 isolates (69% of the pan-genome), 935 softcore orthogroups in >95% of the isolates (9% of the pan-genome), 1,367 shell genes in 5–95% of the isolates (13% of the pan-genome) and a cloud genome of 1,043 genes present in less than 5% of the isolates (10% of the pan-genome; Fig. 1a). Each isolate contained an average of 9,199 orthogroups (range, 8,987–9,629) and an average of 1,636 orthologous accessory-gene clusters (range, 1,424–2,066), corresponding to 16–22% of the total genome of the isolate. The pan-genome was closed—that is, the number of pan-genes did not substantially increase after the addition of approximately 250 genomes (Fig. 1b). Gene association analysis identified 53 co-occurring gene modules containing 2–251 genes (Fig. 1c).

The protein sequences of the core genes were significantly longer than the softcore or accessory genomes. The geometric mean of the length of the core genes was 436 amino acids compared with 310 amino acids for the softcore genes and 191 for the shell/cloud genes (Fig. 1d). To examine the evolutionary forces working on the core and accessory genomes, we calculated the rate of non-synonymous-to-synonymous substitutions (d_N/d_S). The geometric mean of the d_N/d_S ratio among all 10,907 pan-genes was 0.53, with significant differences between the genome compartments. The core genome showed the strongest evidence of negative or purifying selection ($d_N/d_S = 0.49$), whereas the softcore and accessory genomes had d_N/d_S ratios of 0.68 and 0.69, respectively (Fig. 1e). The lower d_N/d_S values for the core genes relative to the accessory genes indicate that they are under a higher degree of purifying selection—although neither genome compartment is evolving neutrally, as indicated by ratios of less than one.

The core genome contained a higher proportion of proteins with annotated domains, as 85% of the core genes contained at least one annotated InterPro domain compared with 71% of the softcore genes and 51% of the accessory genes (Fig. 1f). The core genome was enriched for 3,140 Pfam domains—including protein kinase domains, transcription factor domains and ABC transporters—whereas the accessory genome was enriched for 546 Pfam domains—including short-chain dehydrogenases and cytochrome P450 enzymes (Extended Data Fig. 1a). For Gene Ontology (GO) annotations, the core genome was enriched for protein binding, ATP binding, carbohydrate metabolic functions, signal transduction and 1,497 total annotations (Supplementary Data 2). The accessory genome was enriched for haem binding, response to oxidative stress and 244 total GO annotations (Supplementary Data 2).

Many of the shell and cloud genes were located on the subtelomeric ends of chromosomes 1 and 7, as measured by their position in Af293 (Extended Data Fig. 1b). Of the 10,907 orthologous gene clusters (homologous genes identified in different isolates) identified in the *A. fumigatus* pan-genome, 87% were present in Af293 (Supplementary Data 3). Overall, we identified an average of 494 genes per isolate that were absent in Af293 and a cumulative 1,934 unique ORFs were not present in Af293. In summary, the core genome of *A. fumigatus* represented 69% of the total identified orthogroups and was distinct from the accessory genome in length, function and the strength of purifying selection.

Chronic disease isolates are more genetically diverse than isolates from invasive disease and the environment. We examined the population genomics of isolates from the environment, invasive disease and chronic aspergillosis. Due to the lower number of isolates in the chronic disease group, the environmental and clinical samples were downsampled to match the number of chronic disease isolates ($n = 19$). Interestingly, the isolates from chronic disease group were significantly more diverse at the nucleotide level than isolates from invasive disease or the environment, as measured by the nucleotide diversity (π) calculated across overlapping 5 kb windows (Extended Data Fig. 2). In contrast, isolates from the invasive disease group showed less nucleotide diversity than isolates from the environment and chronic disease. The geometric mean of the genome-wide nucleotide diversity was 1.3×10^{-5} for the isolates from the chronic disease group, 8.3×10^{-6} for the environmental isolates and 6.9×10^{-6} for the isolates from the invasive disease group.

The Af293-containing genetic cluster is enriched for clinical isolates. In a phylogeny built from the coding nucleotide sequences of 5,380 single-copy orthologues, all 300 isolates formed a monophyletic group that was clearly distinct from the related outgroups of *Aspergillus oerlinghausenensis* and *Aspergillus fischeri* (Fig. 2 and Extended Data Fig. 3a). Isolates from Germany, collected and sequenced by us, intermixed with the globally sampled isolates from publicly available repositories, with no strong geographic clustering observed. We also found a high degree of congruence between the phylogeny built from the core genome sequence from de novo genome assemblies and phylogenies built using reference-guided SNV data from whole-genome SNVs and neutral loci (Extended Data Fig. 3b–d). Based on genome coverage at the MAT locus, we found an equal split of isolates of both mating types ($n = 148$ MAT1-1 isolates and $n = 149$ MAT1-2 isolates; Fig. 2).

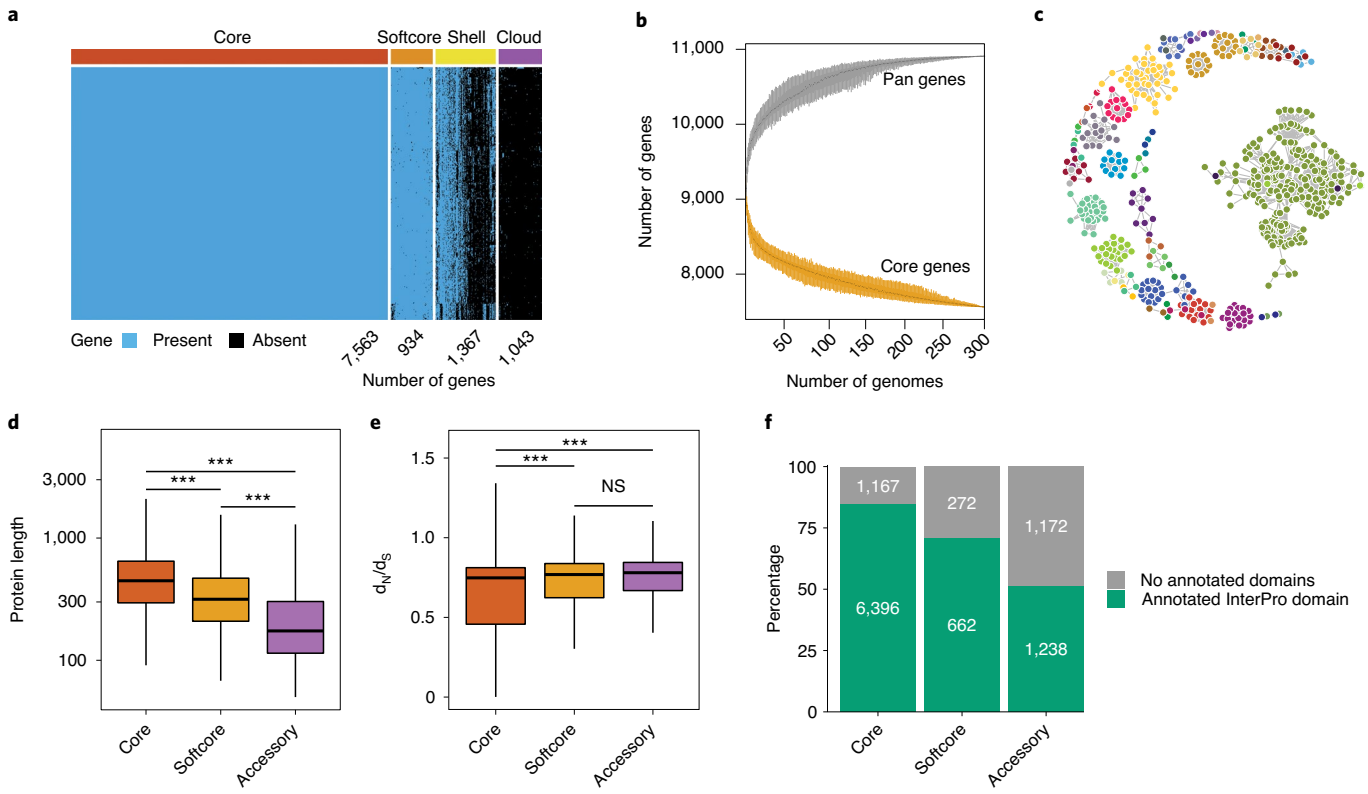


Fig. 1 | The pan-genome of *A. fumigatus*. **a**, Presence/absence matrix of 10,907 orthologous gene clusters identified from 300 *A. fumigatus* genomes. The pan-genome is subdivided into core (orthogroups present in all isolates), softcore (orthogroups present in >95% of the isolates), shell (orthogroups present in 5–95% of the isolates) and cloud (orthogroups present in less than 5% of the isolates) genomes. **b**, Pan and accessory (softcore, cloud and shell) genome size as the number of genomes included increases. Darker hues represent the 25th and 75 percentiles, while the lighter hues represent the range. **c**, Co-occurring gene modules ($n=53$) of the *A. fumigatus* accessory genome. Each circle indicates a gene and its association with other genes indicated by edges (lines). The module significance was identified using two-sided binomial exact tests with Bonferroni's correction ($P<0.05$). Only positive associations are illustrated. The colour indicates module membership. **d**, Amino-acid-sequence lengths of core, softcore and accessory (cloud and shell) genes. The exact P values were $<2\times10^{-16}$ for all of the indicated comparisons. **e**, Ratio of the non-synonymous substitutions to synonymous substitutions in core, softcore and accessory genes. Genes with ratios greater than one are under positive selection, whereas genes with ratios less than one are under purifying selection. The exact P values were: core versus softcore, $P=1.2\times10^{-7}$; core versus accessory, $P<2\times10^{-16}$; and softcore versus accessory, $P=0.15$. **c**–**e**, $n=7{,}563$ core, 935 softcore and 2,410 accessory orthogroups. **d**,**e**, 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 1.5× the interquartile range. Statistical significance was determined using a two-sided Mann–Whitney *U*-test with Bonferroni's correction; ***$P<0.001$ and NS, not significant. **f**, Number (indicated in the bars) and fraction of core, softcore and accessory genes containing an annotated InterPro domain.

To look for evidence of genomic recombination in *A. fumigatus*, we performed a neighbour-net analysis, a phylogenetic method that allows for the representation of conflicting genetic signals that result from sexual recombination or gene conversion. The neighbour-net tree built from core genes had a highly reticulated centre, which indicates a marked degree of conflicting genetic information in the phylogenetic network and is suggestive of abundant genetic recombination in the species (Fig. 3a).

Discriminant analysis of principle components[22] was used to identify seven as the best supported number of genetic clusters in our dataset based on our de novo, reference and pan-gene count-based approaches (Extended Data Fig. 4a–c). Cluster 6 had the largest number of isolates ($n=80$), followed by cluster 2 ($n=53$), cluster 5 ($n=48$), cluster 7 ($n=43$), cluster 3 ($n=35$), cluster 4 ($n=22$) and cluster 1 ($n=19$; Fig. 4a). Interestingly, cluster 5 was enriched for clinical isolates (Fisher's exact test with Benjamini–Hochberg correction, $P=0.02$). This cluster also contained the reference strain Af293, which is a clinical isolate from a patient who died of invasive aspergillosis[23]. Together, we observed an enrichment for clinical isolates in one cluster as well as evidence of abundant genetic recombination in *A. fumigatus*.

Genetic cluster 5 contains more accessory genes and a distinct genomic profile. As genetic cluster 5 was statistically enriched for clinical isolates, we examined the genomes of each cluster to identify differences that might predispose the genetic background of cluster 5 towards human infection as well as characterize potential functional differences between the genetic clusters. Interestingly, clusters 5 and 2 contained significantly more accessory genes than the other clusters (Fig. 4a). The median number of accessory genes for cluster 5 was 1,965 compared with 1,895, 1,842, 1,882, 1,814 and 1,790 for clusters 2, 3, 1, 7 and 6, respectively (Fig. 4a). Cluster 4 had the smallest number of accessory genes, with a median of 1,749.

To predict the functional differences between the clusters, we calculated the abundance of Pfam domains and the frequency of GO annotations in the different clusters and compared the variance between clusters. A total of 170 GO annotations showed significant variation in their relative frequency between clusters (Fig. 4b and

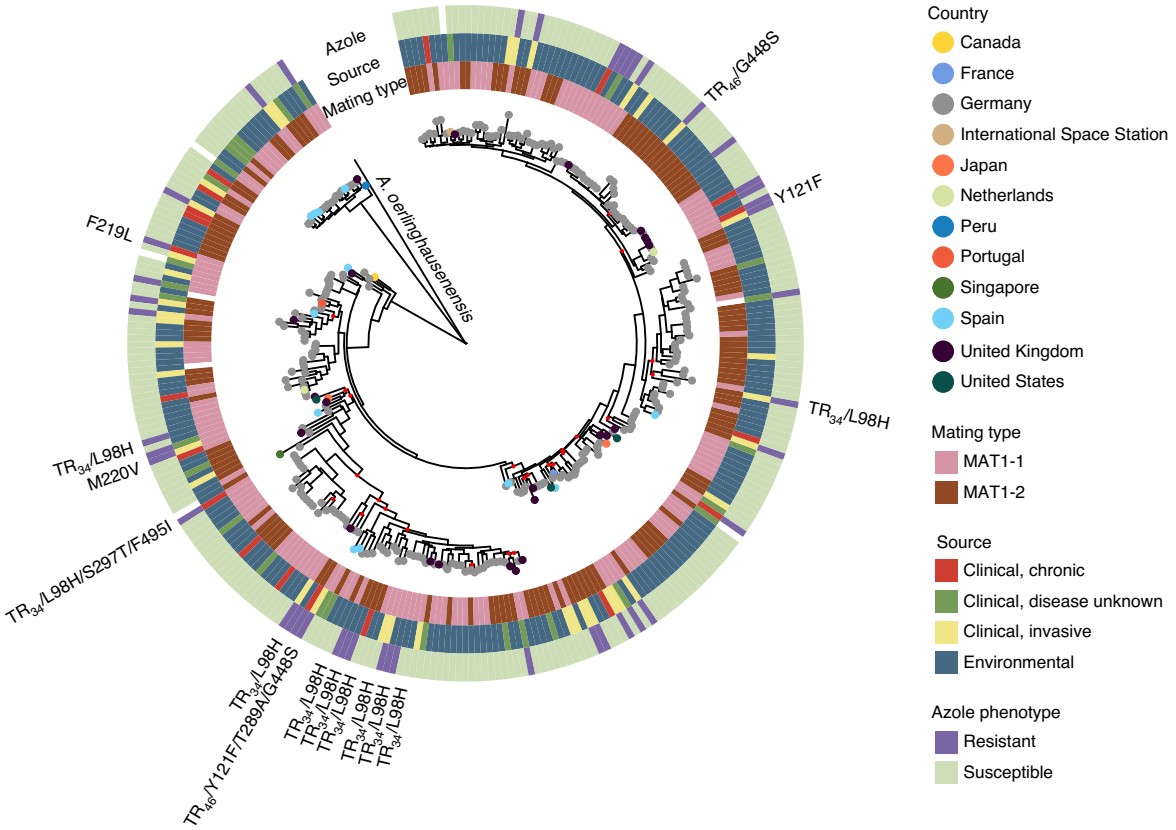


Fig. 2 | Whole-genome phylogeny of environmental and clinical *A. fumigatus*. Phylogenetic tree constructed from coding nucleotide sequences of 5,380 single-copy orthologues shared by *A. fumigatus*, *A. fischeri* and *A. oerlinghausenensis*. The phylogeny is rooted with *A. oerlinghausenensis* and the branch length was shortened for illustration. The coloured symbols at the end of branches represent the country where the sample was isolated. The red dots in the tree structure indicate nodes with ultrafast bootstrap values of less than 0.96. The metadata rings on the outside of the tree indicate the Azole phenotype (where resistance is defined as a minimum inhibitory concentration above the EUCAST breakpoint for one or more triazoles), source of the isolate and the mating type. Mutations in the *cyp51a* gene relative to the Af293 reference genome are also indicated on the outside of the tree.

Supplementary Data 4). Among these were an increased frequency of genes involved in oxidation–reduction processes, iron-ion binding, carbohydrate metabolic processes and proteolysis in cluster 5 (Fig. 4c and Supplementary Data 4). Significant variation between clusters was observed for the abundance of 269 Pfam domains (Supplementary Data 4). Cluster 5 had an increased abundance of major facilitator superfamily transporters, amino-acid permeases and chitin-recognition proteins (Fig. 4d and Supplementary Data 4).

For the GO categories and Pfam domains that did not show a significant difference in copy number between the genetic clusters, we reasoned that there could still be functional differences due to the presence of high-impact variants such as frameshifts or the gain/loss of stop codons. To examine this, we calculated the fraction of genes containing a high-impact variant(s) for each functional annotation and compared the incidence across the clusters. A total of 945 GO annotations contained significant differences in the incidence of high-impact variants between the clusters (Supplementary Data 4). Among these were a reduced number of high-impact variants in chromatin organization and mismatch repair-annotated genes in clusters 5 and 2 (Fig. 4e). We also quantified the incidence of high-impact variants in Pfam domain-containing genes and identified 482 domains with significant differences between the clusters (Supplementary Data 4). These included a reduced number of high-impact variants in cytochrome P450 enzymes and bZIP transcription factors in clusters 2 and 5. In summary, we observed distinct genomic profiles between the genetic clusters of *A. fumigatus*, including a larger number of accessory genes in clusters 2 and 5 in addition to copy-number variation and incidence of high-impact variants in functional annotations such as Pfam domains and GO categories.

***A. fumigatus* exhibits variation in virulence-associated genes.** Using a database of 360 virulence- or fitness-associated genes for *A. fumigatus*[19,24–28] (Supplementary Data 5), we examined our 300 genomes for the presence/absence of these genes, changes in copy number relative to Af293 and the incidence of high-impact variants (for example, frameshifts and nonsense mutations). This list includes genes involved in metabolism, signalling, cell-wall biology, secondary metabolism, stress responses and antifungal drug resistance. Overall, these virulence-associated factors were well conserved. No variation in copy number, presence/absence of genes or genetic alterations anticipated to have a high functional impact was detected in 57% (205/360) of the genes. The remaining 155 virulence-associated genes had some degree of genetic variation expected to affect gene function among our 300 genomes, which can be visualized in Fig. 5 (full summary in Supplementary Data 5). Underscoring the fundamental role of these genes for the fitness of *A. fumigatus* in the environment and the human host, most cases of gene loss or high-impact genetic variation were uncommon and observed in less than 5% of the isolates ($n = 121$ genes). However, the remaining 34 genes displayed more pervasive genetic variation, including 76% of the isolates (229/300) showing frameshifts in the

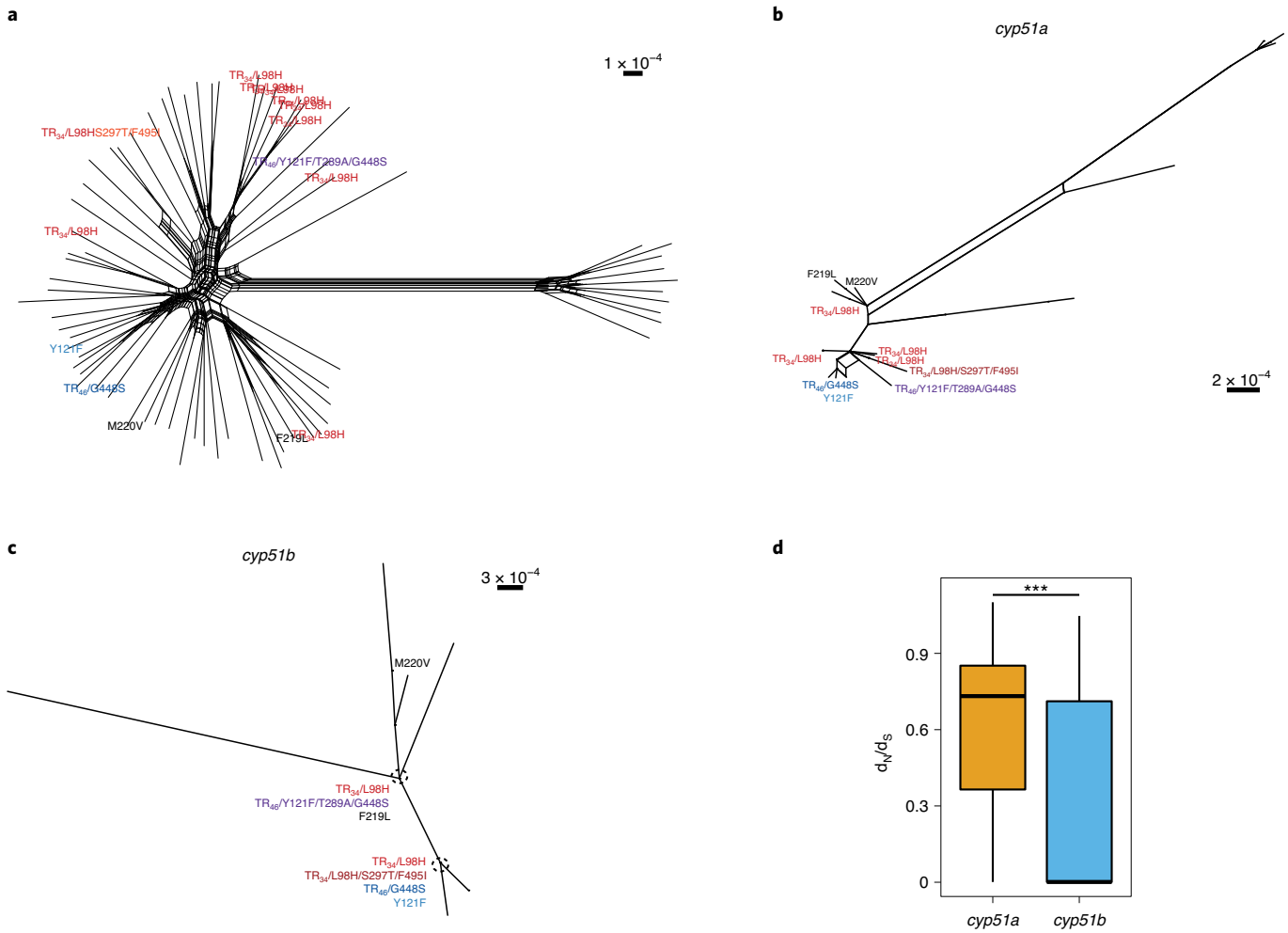


Fig. 3 | Neighbour-net trees from 70 *A. fumigatus* isolates. a, Phylogenetic network of the coding nucleotide sequence of 5,380 single-copy orthologous genes (whole-genome sequencing). The reticulated core indicates abundant conflicting genetic information among isolates, suggestive of recombination in the species. **b**,**c**, Phylogenetic networks of the coding nucleotide sequences of *cyp51a* (**b**) and *cyp51b* (**c**) plus 1,000 bp up- and downstream of the genes. Splits (parallel bands) indicate conflicting nucleotide patterns and their length is proportional to the number of bases supporting the split. The *cyp51a* genotypes of the isolates are indicated at the branch tips. Red-hued labels indicate TR_{34}-containing alleles. Blue-hued labels denote TR_{46} lineage-associated alleles. All additional *cyp51a* polymorphisms are indicated with black labels. Dashed circles in **c** denote the point on the phylogenetic network where all the indicated alleles localize. **a**–**c**, $n = 10$ isolates from each of the genetic clusters identified in the dataset. Scale bar indicates nucleotide substitutions per site. **d**, Ratio of non-synonymous-to-synonymous substitutions in *cyp51a* and *cyp51b*. The ratios were calculated from the genomes of 300 *A. fumigatus* isolates. Both genes have ratios of less than one and are under purifying selection. Statistical significance was determined using a two-sided Mann–Whitney *U*-test; $^{***}P < 0.001$; the exact *P* value was $P < 2.2 \times 10^{-16}$. The bold horizontal line indicates the 50th percentile and the filled box extends from the 25th to the 75th percentile. The whiskers encompass the lowest and highest values within 1.5× the interquartile range.

serine protease *pr1* (*Afu7g04930*) and 71% of the isolates (213/300) showing high-impact variants in the putative sensor histidine kinase *tcsB* (*Afu2g00660*).

Overall, secondary metabolism genes showed the highest variability among the virulence-associated genes, with 59 genes either being absent or showing a predicted loss of function among the 300 isolates. Interestingly, 97% of the isolates (292/300) possessed extensively degraded copies of the non-ribosomal peptide synthetase *nrps8* (also known as *pes3* or *Afu5g12730*), a gene whose deletion showed increased virulence in a murine model of invasive aspergillosis[29]. We also observed variability in the biosynthetic gene cluster encoding fumagillin, including absence of the fumagillin tailoring enzyme *fmaG* in 89% of the isolates (267/300), the absence of *fumR* in 49% of the isolates (146/300) and a complete loss of the cluster in three isolates (1%). Finally, we observed variants predicted to impact the biosynthesis of the immunosuppressive virulence factor gliotoxin in 6% of the isolates (17/300). These included high-impact variants in *gliZ* ($n = 11$ isolates); *gliA* ($n = 5$ isolates); *gliP* and *gliF* ($n = 3$ isolates each); *gliI*, *gliT* and *gliJ* ($n = 2$ isolates each); and *glicC* and *gliG* ($n = 1$ isolate each).

In addition to cases of gene loss, we observed cases of gene amplification in virulence-associated genes relative to Af293 and A1163. A total of 53 genes showed gene amplification, including 5% of the isolates ($n = 16$) with increased copy number of the putative catalase-peroxidase *cat2* (*Afu8g01670*), which is upregulated in response to neutrophils[30]. In addition, 5% of the isolates ($n = 14$) had increased copy numbers of the zinc transporter *zrfC* (*Afu4g09560*) and 3% ($n = 10$) had increased copy numbers of the putative ABC multidrug transporter *Afu5g12720*. In summary, although roughly half of the virulence-associated genes described to date were well conserved among the 300 genomes examined, we observed high-impact genetic variation in many virulence-associated genes, which could perhaps explain the wide range in virulence observed among *A. fumigatus* isolates.

GWAS-identified fungal genetic variation associated with clinical isolates. To better understand how the environmental saprobe *A. fumigatus* can cause disease in the non-native niche of the human lung, we performed a GWAS study to identify fungal variants associated with clinical isolates relative to environmental isolates as well as fungal variants associated with the specific disease states of invasive and chronic disease (Extended Data Fig. 5a). Using a linear mixed model and a minor allele frequency (MAF) > 0.05, we identified 68 genomic positions with genetic variants associated with clinical isolates relative to environmental isolates (Supplementary Data 6). These variants included hits in 27 protein-coding genes, comprising both genes with established roles in virulence as well as uncharacterized ORFs (Supplementary Table 2). Among the genes previously implicated in the virulence of *A. fumigatus* were the sterol regulatory element binding protein *srbA*, which is involved in both growth in hypoxia and iron homeostasis[31,32], the global transcriptional regulator *pacC* required for fungal invasion during pulmonary infection[33,34] and the transcription factor *acuK* that regulates gluconeogenesis and iron acquisition[35]. The analysis also identified variants in genes whose role in virulence is less established, including a microtubule spindle protein (*Afu2g16260*), a heat shock-responsive protein (*Afu4g04680*), a putative polyketide synthase (*Afu6g13930*) and histone H1 (*Afu3g06070*), which is upregulated in conidia exposed to neutrophils (AspDB).

The virulence potential of *A. fumigatus* is influenced by the host and its underlying disease status. The factors critical for the establishment of invasive infection in a neutropenic lung are probably not the same as those required for long-term survival in the human lung, as in the case of chronic diseases such as cystic fibrosis and allergic bronchopulmonary aspergillosis. We thus performed association analysis for genetic variants associated with isolates from both invasive (acute) and chronic aspergillosis. There was a high degree of overlap between the genetic variants identified in this analysis and those from the analysis of all clinical isolates, regardless of the disease status of the host, but fewer variants and genes were identified for each underlying clinical disease (Extended Data Fig. 5b). We identified 21 genomic positions with SNVs and short indels significantly associated with invasive aspergillosis (Supplementary Table 2 and Supplementary Data 6). Nine of the ten variants located in coding genes that were associated with invasive disease were shared with isolates from all clinical origins and included the transcription factors *acuK* and *pacC* as well as the tubulin beta-2 subunit *tub2*. Chronic disease had variants at five genomic positions, two of which were within coding genes: *Afu2g03540*, an orthologue of GPI-anchored cell protein *cspA* (*Afu3g08990*) and a L-cytosine transmembrane transporter (*Afu6g14530*; Supplementary Table 2 and Supplementary Data 6).

Triazole target genes display distinct phylogenetic networks and imbalanced levels of stabilizing selection. The paralogous genes *cyp51a* (*Afu4g06890*) and *cyp51b* (*Afu7g03740*) encode the molecular targets of the triazoles. Despite this, most resistance mutations and mechanisms have been described in *cyp51a*. Triazole-resistant isolates were distributed throughout the phylogeny (Fig. 2). However, most isolates carrying the TR_{34}/L98H allele of *cyp51a* were clustered near each other. The close genetic relationship between isolates carrying TR_{34}/L98H is in agreement with previous work suggesting a single origin of this allele[36].

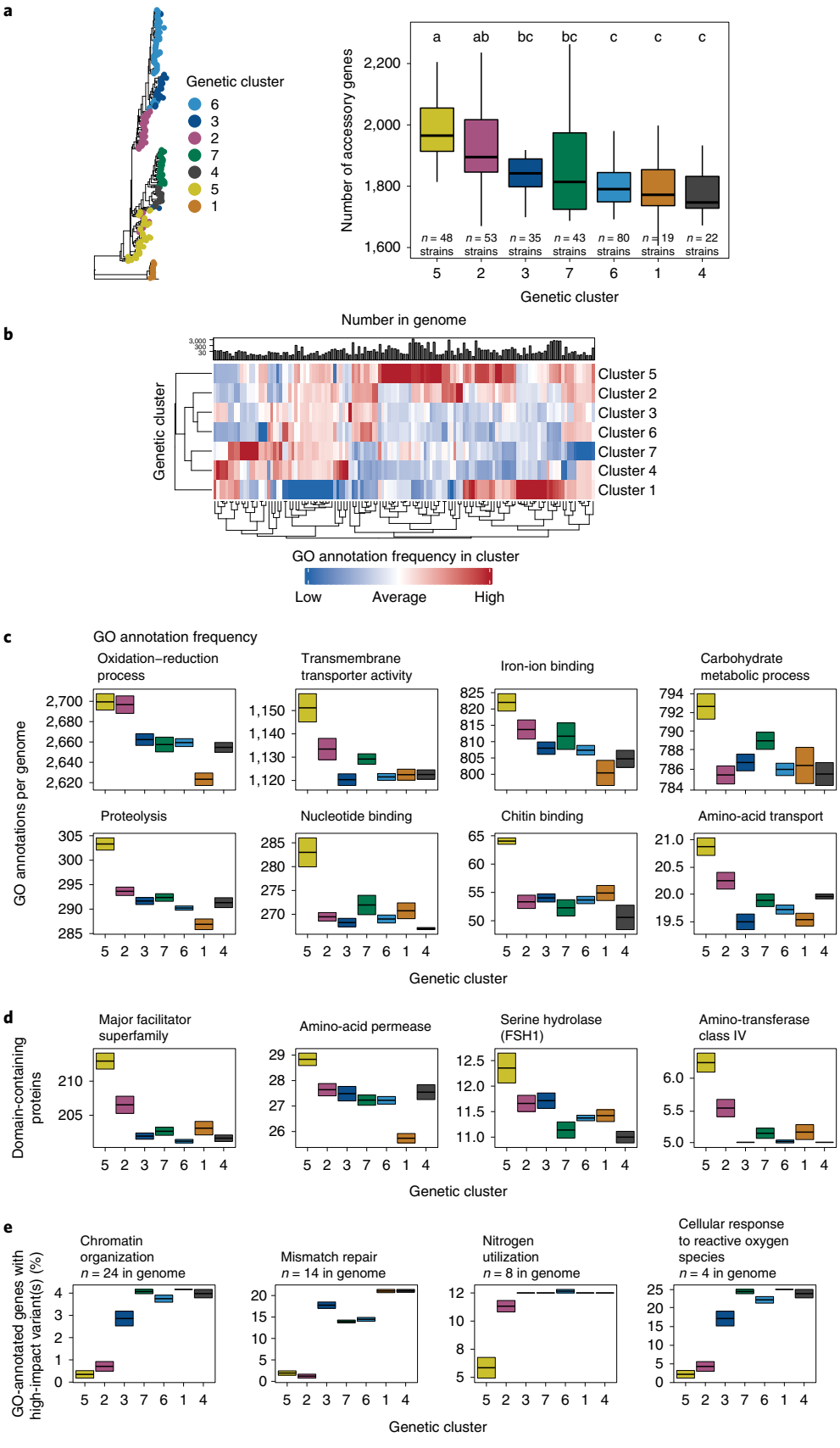
To investigate the evolutionary features of the triazole targets, we built neighbour-net trees from the coding sequence of *cyp51a* and *cyp51b* plus 1,000 bp of the up- and downstream flanking sequences. A phylogenetic network built from *cyp51a* sequences showed multiple splits (parallel bands), indicating conflicting genetic information among our isolates that could arise from recombination (Fig. 3b). Genetic recombination by isolates carrying the TR_{34}/L98H allele is supported by its presence in isolates of both mating types (Fig. 2). By comparison, a neighbour-net tree of *cyp51b*, which is located on a different chromosome, did not show any conflicting genetic information, as demonstrated by the lack of reticulation in the phylogenetic network (Fig. 3c). We also observed a reassortment of *cyp51a* genotypes in the tree constructed from *cyp51b* sequences relative to that constructed from *cyp51a* (Fig. 3b,c). In the tree constructed from *cyp51a* sequences, isolates carrying the TR_{34}/L98H allele of *cyp51a* were located at five distinct points on the phylogenetic network at positions that did not overlap with the positions of other *cyp51a* mutant alleles. In the network of *cyp51b* sequences, strains carrying the TR_{34}/L98H allele were found only at two positions in the network that also contained other *cyp51a* mutant alleles.

To assess the selective forces working on *cyp51a* and *cyp51b*, we examined the d_N/d_S ratios of each gene. The d_N/d_S ratios of *cyp51b* were significantly lower than *cyp51a* (mean value of 0.01 and 0.27 for *cyp51b* and *cyp51*, respectively), indicating that *cyp51b* is under a stronger degree of stabilization selection than *cyp51a* (Fig. 3d). Together, our results demonstrate higher levels of genetic disagreement in the isolate sequences of *cyp51a* compared with *cyp51b* and that *cyp51a* is under less stabilizing selection than *cyp51b*.

GWAS-identified genetic changes associated with triazole resistance. We subsequently performed variant-based GWAS to identify genomic changes associated with triazole resistance. Among the 294 samples with available susceptibility data, 44 were resistant to one or more triazole. Of these, 15 contained mutations in *cyp51a* that have been previously shown to confer triazole resistance (for example, TR_{34}/L98H and TR_{46}/Y121F/T289A/G448S) and 29 were resistant by unknown mechanisms. When we performed a linear mixed-model GWAS using a MAF > 0.01, we identified 16 genomic positions associated with triazole resistance (Supplementary Table 3 and Supplementary Data 6). These included the known TR_{34} and L98H variants in the triazole target enzyme *cyp51a*. However, we repeated our analysis using a MAF > 0.05 to give a more robust variant list with fewer false positives given that association studies with smaller datasets such as ours are underpowered for the detection of true associations with rare variants (Extended Data Fig. 5c). Using this more stringent criterion, we condensed our variant list

Fig. 4 | Pan-genomic differences between the clusters of *A. fumigatus*. **a**, Number of accessory genes (right) present in the genomes of isolates belonging to each genetic cluster (left). Statistical significance was determined using a one-way analysis of variance and 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; $P < 0.05$. The bold line in the box-and-whisker plot indicates the 50th percentile, and the box extends from the 25th to the 75th percentiles. The whiskers denote the lowest and highest values within 1.5× the interquartile range. **b**, Heatmap showing the normalized abundance of GO annotations exhibiting significant variance in frequency between the clusters (bottom; $n = 127$ GO annotations). Statistical significance was determined using one-way analysis of variance with Bonferroni's correction ($P < 0.05$). The mean number of genes containing each GO annotation across the 300 genomes is shown (top). Note the graph is on a $\log_{10}$ scale but the *y*-axis shows actual values. **c**, Genome copy number for select GO annotations from **b** across the clusters. **d**, Genome copy number for select Pfam annotations across the clusters. **e**, The incidence of high-impact variants (for example, frameshift and loss of start) relative to Af293 was analysed for GO annotations that did not contain significant copy-number variation between the clusters. A selected subset of GO categories with significant variation in the incidence of high-impact variants between the genetic clusters is shown. **c**–**e**, The boxes denote the mean (crossbar) ± s.e.m. for the isolates of each cluster.

to variants in three protein-coding genes (Extended Data Fig. 5d and Supplementary Table 3). These included a microtubule bundle protein (*Afu2g16260*), a FGGY-family kinase induced by heat shock (*Afu4g04680*) and its adjacent, uncharacterized ORF *Afu4g04690*. The role of these genes in triazole resistance is an exciting area to follow up on.



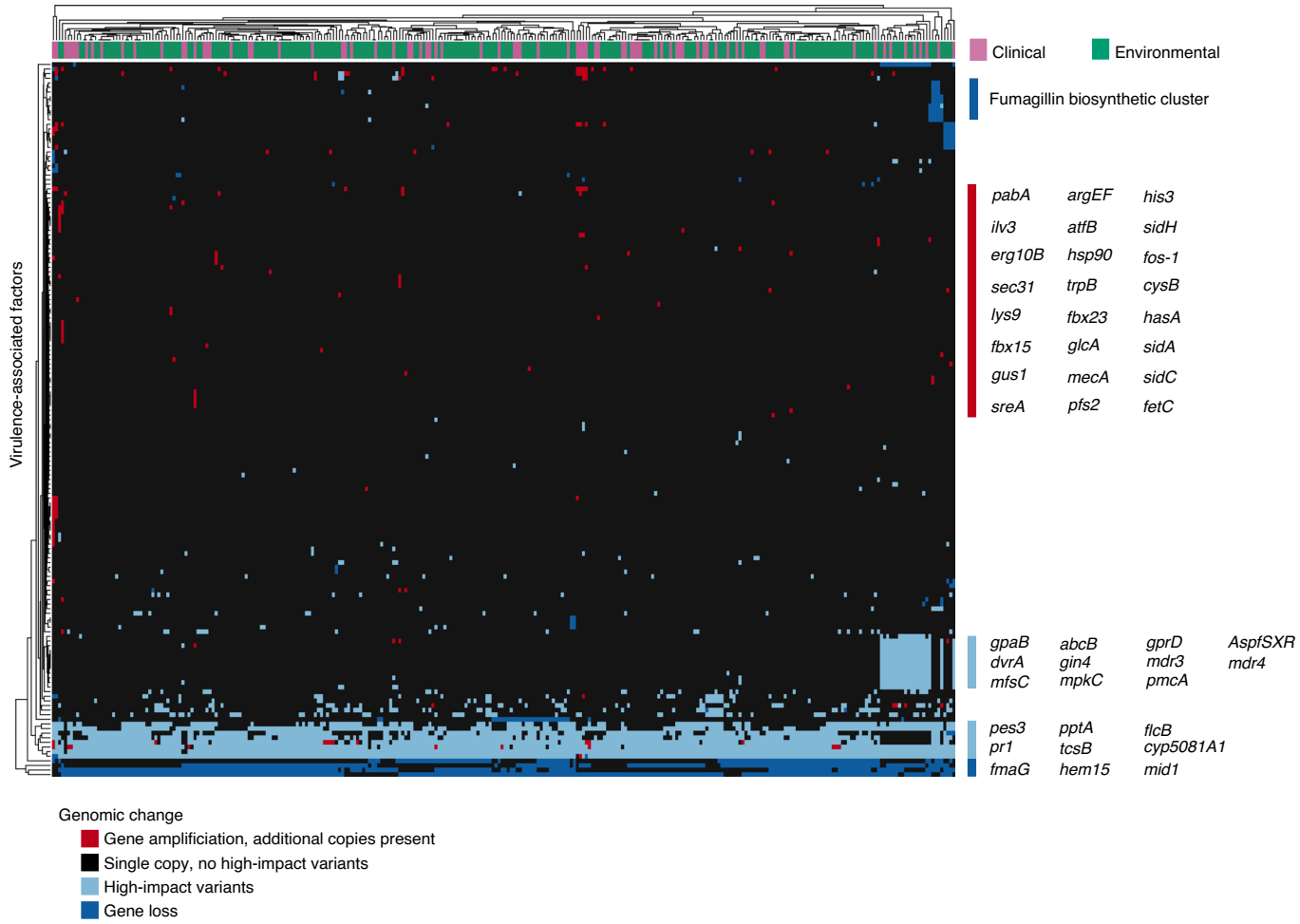


Fig. 5 | Genomic variation among known *A. fumigatus* virulence-associated factors. Heatmap of 155 virulence-associated genes where variation in copy number (gene loss or amplification) or the presence of high-impact variants (for example frameshift, loss or gain of stop codon) was observed relative to Af293 (bottom). The source of the isolate is indicated (top). Gene names for select virulence-associated factors are annotated (right).

Discussion

In this study we defined the pan-genome of *A. fumigatus* using 300 genomes, including a large number of environmental isolates largely absent from previous analyses[15–19]. Compared with *A. fumigatus*, the human commensal and opportunistic pathogen *Candida albicans* was shown to have a lower level of pan-genomic diversity, with 91% of pan-genes present in all isolates[37]. In the same study, a proof-of-concept pan-genome for *A. fumigatus* was also built using genomic data from 12 isolates and 83% of the pan-genome was found to be conserved in all isolates[37]. By contrast, our findings indicate that *A. fumigatus* has a much larger pan-genome and only 69% of the genes identified are present in all isolates; this discrepancy in results is likely to be due to the limited number of isolates included in the former study. In addition, the average BUSCO genome completeness of the assemblies used for their analysis was below 85%, suggesting that notable genetic content was unaccounted for[37]. Future work utilizing chromosome-level assemblies of *A. fumigatus* isolates will allow for a finalized pan-genome of the species with additional information on the evolutionary dynamics of chromosomal organization.

Through pan-genomic analyses we discovered notable genetic variation in virulence factors that have largely only been studied in one or two reference strains. Although most of these cases were infrequent and observed in fewer than 5% of the isolates, some, such as pseudogenization of the non-ribosomal peptide synthetase *nrps8* (or *pes3*), was observed in 97% of the isolates. The largest virulence-associated genetic variation was in secondary metabolism genes, an observation in line with a previous study of 66 isolates[38]. Both studies observed low-incidence variation in the gliotoxin and fumagillin/pseurotin biosynthetic gene clusters as well as high-incidence variation in the fumigermin biosynthetic gene cluster. In addition, although our analysis quantified high-impact genetic changes in virulence determinants, there is almost certainly additional genetic variation that impacts fungal virulence that is difficult to predict on the global scale. The genomes generated here provide a valuable resource for addressing how intraspecies variability in virulence determinants affects infection.

We observed an enrichment for clinical isolates in genetic cluster 5, suggesting that this genetic background might be more fit in the human environment. However, clinical isolates were distributed throughout the phylogeny, highlighting the overall fitness of *A. fumigatus*. In addition, this organism can take advantage of numerous, diverging clinical diseases to establish an infection. We thus performed a genome-wide association study (GWAS) to identify fungal variants associated with clinical disease in general as well as acute and chronic disease, and identified largely overlapping gene sets. However, information on the underlying clinical disease was not available for all samples and the isolates from chronic disease represented a small fraction of the dataset. Future genomic analyses including additional samples from specific underlying diseases will

further illuminate the complex interplay between *A. fumigatus* and specific host disease environments.

The rising incidence of resistance to first-choice antifungals, the triazoles, is a major challenge for the management of *A. fumigatus* infections. This problem is further complicated by up to 30% of the isolates having no identifiable resistance mechanism. We performed GWAS and identified 12 genes associated with triazole resistance. These hits included previously identified variants in the triazole target gene *cyp51a* as well as genes that had not been previously linked to triazole resistance. As a caveat, association studies are underpowered at detecting associations with rare variants. Accordingly, we only screened for association of genetic variants present in at least 1% and 5% of samples. Thus, there are potentially additional resistance-associated variants that were not identified by our analysis. This is perhaps the case for the HMG-CoA demethylase *hmg1*. Clinically observed mutations in this gene conferred triazole resistance to *A. fumigatus* following reconstruction in an isogenic background[39]. Although our analysis did not identify any variants of this gene associated with triazole resistance, manual examination uncovered three triazole-resistant isolates containing single non-synonymous substitutions in *hmg1* (E306D, P309L and C369R). These variants were not considered in the GWAS due to their low prevalence in the dataset. However, the exact role these substitutions play in the resistance of this isolate is unclear, particularly for one isolate that also contained *cyp51a* alterations associated with resistance (TR_{46}/G448S). No variants were observed in *hapE* and *cyp51b*, two additional genes linked to triazole resistance.

In summary, this study provides a comprehensive view of the genetic diversity in this important human fungal pathogen. Characterization of the intraspecies diversity and moving away from a single reference genome that does not necessarily represent *A. fumigatus* as a whole will ultimately help us understand its metabolic and pathogenic versatility.

Methods

***A. fumigatus* isolates analysed in this study.** Of the 300 isolates analysed, 188 (49 clinical and 139 environmental isolates) were newly sequenced as part of this study. The 49 clinical isolates sequenced were collected by the German National Reference Center for Invasive Fungal Infections between 2014 and 2018. Bronchial alveolar lavage was the most frequent form of sample collection, representing 31% (15/49) of clinical isolates. The remaining clinical samples were isolated from other pulmonary sources, such as sputum or bronchial secretions, and the exact site of isolation was unavailable for 10% (5/49) of the samples. The 139 environmental isolates sequenced as part of this study were obtained from soil sampling of 11 farms in Germany between 2016 and 2018. Sixty-four of the remaining 112 isolates had been previously sequenced by us as part of a previous study[20] (BioProject PRJNA595552), while 48 had been previously sequenced by other groups and data downloaded from the NCBI Sequence Read Archive. In total, the dataset was comprised of 213 environmental isolates and 87 clinical isolates from Europe, Asia, North America, South America and the International Space Station. A detailed list of the isolates and their metadata characteristics can be found in Extended Data Fig. 1.

Antifungal susceptibility testing. The 188 novel isolates included in this study were screened for azole resistance using the agar-based VIPcheck Assay (Mediaproducts BV) based on EUCAST E.DEF 10.1 following the manufacturer's protocol. Isolates that showed distinguishable germination and hyphal growth on any of the azole-containing wells were subjected to EUCAST broth microdilution (protocol E.DEF 9.3.2; ref. [21]) to define the minimum inhibitory concentrations. Antifungal susceptibility in the clinical isolates was also assessed following EUCAST protocol E.DEF 9.3.2 and resistance was defined in both isolate sets using the EUCAST-established clinical breakpoints. Antifungal susceptibility data from published isolates were obtained from the source publications detailed in Supplementary Data 1.

Genome sequencing and quality assessment. Genomic DNA was extracted from isolates (cultured in Sabouraud Glucose broth at 37 °C with shaking) using a Quick-DNA fungal/bacterial miniprep kit (Zymo Research) following the manufacturer's suggested protocol. Library preparation and Illumina 2 × 150 bp paired-end sequencing were performed on a NextSeq 500 v.2 by LGC Genomics (environmental isolates) and GeneWiz (clinical isolates). Raw FASTQ files were filtered for quality using the following steps: adaptor sequences were removed, bases with an overall quality score of <20 were trimmed and reads shorter than 30 bp were removed. The remaining sequences were verified for quality using FastQC v.0.11.5 (Babraham Institute).

Reference-guided genome analysis. High-quality sequencing reads were aligned to the *A. fumigatus* Af293 reference genome v.2015-09-27 using BWA-MEM v.0.7.8-r779-dirty[40]. PCR duplicates were marked using MarkDuplicate from Picard v.2.18.25. Variant calling to detect SNVs and short indels was performed using the GATK Toolkit (v.4.1.0.0)[41]. Briefly, before variant calling, BAM files were recalibrated using GATK BaseRecalibrator, ApplyBSQ and an in-house dataset of known SNVs generated from the Af293 reference genome and SNVs present in FungiDB, release 39, with ≥80% read frequency and base call ≥ 20. Variant detection was performed using HaplotypeCaller and high-quality variants were identified using GATK best practices (SNP: QD < 2.0 || MQ < 40.0 || FS > 60.0 || MQ RankSum < −12.5 || ReadPosRankSum < −8.0; indel: QD < 2.0 || FS > 200.0 || ReadPosRankSum < −20.0). For downstream analyses, individual VCF files were combined into a single file using bcftools v.01.1.1. Variant function was predicted using SnpEff v.4.3t[42] and 1,000 bp as the cutoff for upstream and downstream flanking of the ORFs. To balance the analysis of high-quality variants with the potential bias introduced by true variants being discarded due to insufficient support, individual variants that failed the quality filter in a sample were included in the variant dataset if at least 95% of the total samples with a variant at that position passed the quality control. This hybrid-filtered variant dataset was used as the input for GWAS and genetic diversity analyses.

A. fumigatus has two mating types (MAT1-1 and MAT1-2), which are encoded within idiomorphic loci on chromosome 3 (refs. [43,44]). The mating type was assigned by calculating the genomic coverage at *Afu3g06160* and *Afu3g06170* using the knowledge that MAT1-2 isolates, including the reference strain Af293, contain a truncated copy of the HMG box mating-type transcription factor (*Afu3g06170*) and an additional gene (*MAT1-2-4*; also known as *Afu3g06160*) that are absent in MAT1-1 isolates. Isolates showing zero coverage in the genomic region of *Afu3g06160* following alignment to Af293 were assigned the mating type MAT1-1. Samples that were not assigned the mating type MAT1-1 were confirmed to be the mating type MAT1-2 through calculation of the genomic coverage at *MAT1-2-1* and *MAT1-2-4*. The ratio of coverage for *MAT1-2-1* and *MAT1-2-4* relative to the genome-wide depth of coverage was between 0.75 and 1.25 for all samples that were assigned the mating type MAT1-2.

Analysis of genomic diversity. Genomic diversity statistics were calculated based on SNV data generated as described earlier. The nucleotide diversity (π) was also calculated using VCFtools[45] with a window size of 5,000 bp and a 500 bp step size. To ensure that differences in sample sizes between the isolate populations did not skew the results, environmental and clinical samples from the acute disease group were downsampled to match the number of isolates from chronic disease in the dataset.

De novo genome assembly and annotation. Genomes were assembled de novo using IDBA-hybrid v.1.1.3 with the Af293 reference genome as a guide[46]. The quality of the genome assembly was assessed using QUAST v.5.0.2 (ref. [47]). Contigs that were shorter than 500 bp or possessing >95% identity and coverage overlap with other contigs were removed. Gene prediction and functional annotation were performed using Funannotate pipeline v.1.5.2-4cfc7f8 (ref. [48]), integrating the following steps. Assemblies were masked for repetitive elements using RepeatMasker (v.4.0.8)[49] using Dfam and RepBase repeat libraries[50]. Gene prediction was performed using EvidenceModeler v.1.1.1 (ref. [51]), incorporating evidence data generated using GeneMark-ES[52] (minimum gene length, 120 bp; and maximum intron length, 3,000 bp) and Augustus[53] (training set, *A. fumigatus*). Gene models predicted to encode peptides shorter than 50 amino acids or transposable elements, or to include span gaps were removed. Transfer-RNA prediction was performed using tRNAscan-SE v.2.0 (ref. [54]). Functional annotation was predicted using PFAM v.43 (ref. [55]), MEROPS v.12 (ref. [56]), dbCAN2 release 7.0 (ref. [57]) and BUSCO v.4.1.4 (ref. [58]). KofamScan v.1.2.0-0 (ref. [59]) was used to assign Kyoto Encyclopedia of Genes and Genomes orthologues to predicted protein sequences and InterProScan v.5.19 (ref. [60]) was used to identify the protein families.

Pan-genome analysis. OrthoFinder was used to identify and cluster orthologous genes[61]. Clustering was performed on the protein sequences of the 300 *A. fumigatus* genomes analysed in this study. In addition, protein sequences from the reference strains Af293 and A1163 were added to improve the identification of the cluster functions. Orthologous gene clusters were assigned a gene identifier from Af293 if they grouped with a single sequence of Af293. If a cluster was not assigned a Af293 gene identifier, but a single A1163 sequence was present, the cluster was assigned the gene identifier from A1163. Orthologous clusters that could not be grouped with a single Af293 or A1163 gene were queried against the NCBI RefSeq non-redundant protein database using DIAMOND using the following criteria: *E*-value cutoff of 1×10^5, percent identity > 70%, minimum query coverage > 50% and minimum subject coverage > 50%. If at least 70% of the protein sequences in the cluster were assigned to any protein in the NCBI non-redundant protein database, the cluster name was assigned to the name of the RefSeq with the highest contribution. If only 50–70% of the protein sequences in a cluster were assigned to the same protein, the matching sequences were assigned the name of the RefSeq match and the

remaining sequences were left unassigned. The remaining clusters without a match in Af293, A1163 or the non-redundant database were considered novel clusters and had putative functions assigned based on their Funannotate (KofamScan and InterProScan) prediction. For these clusters that were not present in Af293 or the non-redundant database, only clusters present in at least 5% of samples were included to limit false gene predictions. The pan-genome was defined based on gene presence/absence variation in the approved cluster meeting the above criteria. Enrichment analysis was performed using a Fisher's exact test with Bonferroni's correction.

Whole-genome phylogeny. The core genome phylogeny (Fig. 2) was inferred from 5,380 single-copy orthologous genes shared by the two reference strains Af293 and A1163, the 300 *A. fumigatus* genomes analysed in this study, the related species *A. oerlinghausenensis* and *A. fischeri*, which was used to root the tree. Orthologues were identified and clustered using OrthoFinder[61]. Cluster peptide sequences were aligned using MUSCLE v.3.8.1551 (ref. [62]). The resulting peptide alignment was back-translated to a nucleotide sequence using PAL2NAL[63] and concatenated. The phylogeny was inferred from this core nucleotide alignment using IQ-TREE 2 (ref. [64]). The ModelFinder Plus module of IQ-TREE 2 was used to identify GTR + F + R8 as the best fitting substitution and site heterogeneity models for phylogeny construction. Branch support was computed using UFBoot2 ultrafast bootstraps[65]. ClonalFrameML[66] was then used to account for recombination in the phylogeny and rescale branch lengths accordingly.

The SNV-based phylogenies (Extended Data Fig. 4c,d) were constructed by first filtering out loci that showed zero coverage in any sample. For the phylogeny constructed from neutral loci, fourfold degenerate sites were used. For both the non-zero coverage and neutral loci phylogenies, SNVs were concatenated and used as the input for IQ-TREE 2. As with the core nucleotide phylogeny, ModelFinder was employed and identified GTR + F + ASC + R8 as the best fitting model for the non-zero coverage phylogeny and TVM + F + ASC + R8 as the best fitting model for the neutral loci phylogeny. Branch supports were calculated using UFBoot2.

Genetic clusters were identified using discriminant analysis of principle components[22]. To create phylogenetic network trees with clearly visible branches and network structure, the genomes were downsampled by randomly selecting ten genomes per cluster, resulting in a total of 70 samples. Neighbour-net phylogenies were inferred and visualized using the R package phangorn (v.2.5.5)[67] based on a similarity matrix of core nucleotide sequences for the whole-genome network phylogeny and nucleotide sequence alignment for the *cyp51a* and *cyp51b* genes with network phylogenies. The phylogenies were visualized using the R package Ggtree[68].

Estimation of d_N/d_S. The protein-coding sequences of each gene cluster were aligned using MUSCLE v.3.8.1551 (ref. [62]). PAL2NAL[63] was then used to convert the resulting amino-acid alignment to a nucleotide alignment that records whether a base-pair substitution resulted in a synonymous or non-synonymous change. Finally, the CODEML package of PAML[69] was used to calculate the d_N/d_S value of each orthogroup. Median values were used for comparison.

Gene co-occurrence in the pan-genome. Gene co-occurrence networks were computed using Coinfinder[70] using a presence/absence matrix of the pan-genome and a significance cutoff of 0.05 by binomial exact test with Bonferroni's correction. Networks were visualized using the R package igraph.

SNV-based GWAS and pan-GWAS. Before analysis, variant classes were assigned as follows: C, SNVs; G, insertions; D, deletions; and A, reference base. VCF files were converted to plink format using VCFtools[45] and filtered using a MAF of 0.05, which resulted in 352,306 SNVs and 24,726 indels for analysis. Positions with a missingness, or the number of individuals where there was SNV information was available, of >1% were removed from the analysis. The GWAS was performed using the EMMA eXpedited (EMMAX) software package[71], applying a linear mixed model with azole resistance (susceptible/resistant), source (environmental/clinical) or clinical disease (chronic/acute infection) as the phenotypic traits. The GEMMA, treeWAS and ECAT software packages were also tested in the framework of this project. EMMAX was ultimately selected over these tools because it accounted for sample structure the best, providing the least-inflated Q–Q plots (Extended Data Fig. 5a,c). Significant variants were determined using a cutoff of $P < 0.01$ with false-discovery-rate correction. The pan-GWAS was performed using a presence/absence matrix of the orthologous gene clusters, where zero denoted absent gene clusters and one represented gene clusters that were present in the genome. Associations between pan-gene presence/absence, isolate source and azole resistance were calculated using Scoary v.1.6.16 (ref. [72]).

Availability of isolates. The isolates that were sequenced in this study were submitted to and are publicly available in the Jena Microbial Resource Collection.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw FASTQ files for the isolates sequenced in this study were uploaded to the NCBI Sequence Read Archive and are publicly available under BioProject PRJNA697844. The accession numbers for the publicly available sequence data are listed in Extended Data Fig. 1. Annotated genome assemblies for sequence data generated in this study and for 64 isolates sequenced by us in a previous study[20] were submitted to NCBI GenBank and are available under the NCBI BioSample numbers listed in Extended Data Fig. 1. Datasets from FungiDB, release 39, are available at https://fungidb.org/fungidb/app/downloads/release-39/. The NCBI RefSeq non-redundant protein database v.22.01.08 is accessible at https://ftp.ncbi.nlm.nih.gov/blast/db/cloud/2018-01-22/. Source data are provided with this paper.

Received: 5 March 2021; Accepted: 8 October 2021;
Published online: 24 November 2021

References

1. Latgé, J. P. and Chamilos, G. *Aspergillus fumigatus* and Aspergillosis in 2019. *Clin. Microbiol. Rev.* https://doi.org/10.1128/CMR.00140-18 (2019).
2. *Invasive Aspergillosis. LIFE* http://www.life-worldwide.org/fungal-diseases/invasive-aspergillosis (2020).
3. Harrison, N. et al. Incidence and characteristics of invasive fungal diseases in allogeneic hematopoietic stem cell transplant recipients: a retrospective cohort study. *BMC Infect. Dis.* **15**, 584 (2015).
4. Kuster, S. et al. Incidence and outcome of invasive fungal diseases after allogeneic hematopoietic stem cell transplantation: a Swiss transplant cohort study. *Transpl. Infect. Dis.* **20**, e12981 (2018).
5. Heo, S. T. et al. Changes in in vitro susceptibility patterns of *Aspergillus* to triazoles and correlation with aspergillosis outcome in a tertiary care cancer center, 1999–2015. *Clin. Infect. Dis.* **65**, 216–225 (2017).
6. Lestrade, P. P. et al. Voriconazole resistance and mortality in invasive aspergillosis: a multicenter retrospective cohort study. *Clin. Infect. Dis.* **68**, 1463–1471 (2019).
7. Snelders, E. et al. Emergence of azole resistance in *Aspergillus fumigatus* and spread of a single resistance mechanism. *PLoS Med.* **5**, 1629–1637 (2008).
8. Rizzetto, L. et al. Strain dependent variation of immune responses to *A. fumigatus*: definition of pathogenic species. *PLoS ONE* **8**, 2–14 (2013).
9. Kowalski, C. H. et al. Heterogeneity among isolates reveals that fitness in low oxygen correlates with *Aspergillus fumigatus* virulence. *mBio* **7**, e01515-16 (2016).
10. Alshareef, F. & Robson, G. D. Genetic and virulence variation in an environmental population of the opportunistic pathogen *Aspergillus fumigatus*. *Microbiology* **160**, 742–751 (2014).
11. Knox, B. P. et al. Characterization of *Aspergillus fumigatus* isolates from air and surfaces of the International Space Station. *mSphere* **1**, e00227-16.
12. Ries, L. N. A. et al. Nutritional heterogeneity among *Aspergillus fumigatus* strains has consequences for virulence in a strain- and host-dependent manner. *Front. Microbiol.* **10**, 854 (2019).
13. Steenwyk, J. L. et al. Variation among biosynthetic gene clusters, secondary metabolite profiles, and cards of virulence across *Aspergillus* species. *Genetics* **216**, 481–497 (2020).
14. Dos Santos, R. A. C. et al. Genomic and phenotypic heterogeneity of clinical isolates of the human pathogens *Aspergillus fumigatus*, *Aspergillus lentulus*, and *Aspergillus fumigatiaffinis*. *Front. Genet.* **11**, 459 (2020).
15. Fedorova, N. D. et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet.* **4**, e1000046 (2008).
16. Abdolrasouli, A. et al. Genomic context of azole-resistance mutations in *Aspergillus fumigatus* using whole-genome sequencing. *mBio* **6**, e00536 (2015).
17. Garcia-Rubio, R. et al. Genome-wide comparative analysis of *Aspergillus fumigatus* strains: the reference genome as a matter of concern. *Genes* **9**, 363 (2018).
18. Fan, Y., Wang Y. and Xu, J. Comparative genome sequence analyses of geographic samples of *Aspergillus fumigatus*—relevance for amphotericin B resistance. *Microorganisms* **8**, 1673 (2020).
19. Puértolas-Balint, F. et al. Revealing the virulence potential of clinical and environmental *Aspergillus fumigatus* isolates using whole-genome sequencing. *Front. Microbiol.* **10**, 1970 (2019).
20. Barber, A. E. et al. Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure. *mBio* **11**, e02213-20 (2020).
21. Arendrup, M. C. et al. Method for the determination of broth dilution minimum inhibitory concentrations of antifungal agents for conidia forming moulds. E.DEF 9.3.2. *EUCAST* https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/AFST/Files/EUCAST_E_Def_9.3.2_Mould_testing_definitive_revised_2020.pdf (2020).
22. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
23. Nierman, W. C. et al. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* **438**, 1151–1156 (2005).
24. Steenwyk, J. L. et al. Genomic and phenotypic analysis of COVID-19-associated pulmonary aspergillosis isolates of *Aspergillus fumigatus*. *Microbiol. Spectr.* **9**, e0001021 (2021).

25. Abad, A. et al. What makes *Aspergillus fumigatus* a successful pathogen? Genes and molecules involved in invasive aspergillosis. *Rev. Iberoam. Micol.* **27**, 155–82 (2010).
26. Bignell, E., et al. Secondary metabolite arsenal of an opportunistic pathogenic fungus. *Philos. Trans. R Soc. B* **371**, 20160023 (2016).
27. Kjaerbolling, I. et al. Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species. *Proc. Natl Acad. Sci. USA* **115**, E753–E761 (2018).
28. Urban, M. et al. PHI-base: the pathogen-host interactions database. *Nucleic Acids Res.* **48**, D613–D620 (2020).
29. O'Hanlon, K. A. et al. Targeted disruption of nonribosomal peptide synthetase pes3 augments the virulence of *Aspergillus fumigatus*. *Infect. Immun.* **79**, 3978–3992 (2011).
30. Sugui, J. A. et al. Genes differentially expressed in conidia and hyphae of *Aspergillus fumigatus* upon exposure to human neutrophils. *PLoS ONE* **3**, e2655 (2008).
31. Willger, S. D. et al. A sterol-regulatory element binding protein is required for cell polarity, hypoxia adaptation, azole drug resistance, and virulence in *Aspergillus fumigatus*. *PLoS Pathog.* **4**, e1000200 (2008).
32. Blatzer, M., et al. SREBP coordinates iron and ergosterol homeostasis to mediate triazole drug and hypoxia responses in the human fungal pathogen *Aspergillus fumigatus*. *PLoS Genet.* **7**, e1002374 (2011).
33. Bertuzzi, M. et al. The pH-responsive PacC transcription factor of *Aspergillus fumigatus* governs epithelial entry and tissue invasion during pulmonary aspergillosis. *PLoS Pathog.* **10**, e1004413 (2014).
34. Bignell, E. et al. The *Aspergillus* pH-responsive transcription factor PacC regulates virulence. *Mol. Microbiol.* **55**, 1072–1084 (2005).
35. Pongpom, M. et al. Divergent targets of *Aspergillus fumigatus* AcuK and AcuM transcription factors during growth in vitro versus invasive disease. *Infect. Immun.* **83**, 923–933 (2015).
36. Camps, S. M. T. et al. Molecular epidemiology of *Aspergillus fumigatus* isolates harboring the TR34/L98H azole resistance mechanism. *J. Clin. Microbiol.* **50**, 2674–2680 (2012).
37. McCarthy, C. G. P. & Fitzpatrick, D. A. Pan-genome analyses of model fungal species. *Microb. Genom.* **5**, e000243 (2019).
38. Lind, A. L. et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol.* **15**, e2003583 (2017).
39. Rybak, J. Mutations in *hmg1*, challenging the paradigm of clinical triazole resistance in *Aspergillus fumigatus*. *mBio* **10**, e00437-19 (2019).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
43. Varga, J. Mating type gene homologues in *Aspergillus fumigatus*. *Microbiology* **149**, 816–819 (2003).
44. Paoletti, M. et al. Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Curr. Biol.* **15**, 1242–1248 (2005).
45. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
46. Peng, Y., et al. IDBA—a practical iterative de Bruijn graph de novo assembler. In *Proc. Research in Computational Molecular Biology. RECOMB 2010.* (ed. Berger, B.) 426–440 (Springer, 2010).
47. Gurevich, A. et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
48. Palmer, J & Stajich J. Funannotate v.1.5.3. *Zenodo* https://zenodo.org/record/2604804 (2019).
49. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013–2015); http://www.repeatmasker.org
50. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
51. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
52. Ter-Hovhannisyan, V. et al. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
53. Stanke, M. et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
54. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
55. Finn, R. D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
56. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **40**, D343–D350 (2012).
57. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
58. Simao, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
59. Aramaki, T. et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
60. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
61. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
62. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
63. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
64. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
65. Hoang, D. T. et al. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
66. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
67. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
68. Yu, G. et al. Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
69. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
70. Whelan, F. J., Rusilowicz, M. and McInerney, J. O. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genom.* **6**, e000338 (2020).
71. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
72. Brynildsrud, O. et al. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).

Acknowledgements

This project was supported by The Federal Ministry for Education and Science (Bundesministerium für Bildung und Forschung) within the framework of InfectControl 2020 projects FINAR and FINAR 2.0 (grant nos 03ZZ0809 and 03ZZ0834 to O.K.). The NRZMyk is supported by the Robert Koch Institute with funds provided by the German Ministry of Health (grant no. 1369-240 to O.K.). A.E.B. and G.P. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy (EXC 2051, Project ID 390713860). G.P. thanks the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction', subproject INF (project number 210879364) for intellectual input. The authors thank M. Blango for thoughtful discussions on this manuscript.

Author contributions

A.E.B., G.P. and O.K. conceptualized and designed the study. The experimental work was performed by A.E.B. and G.W. A.E.B., B.S., G.P., K.K., J.L., O.K., T.S.-O. and G.W. analysed the data and interpreted results. A.E.B. wrote the primary manuscript and all of the listed authors were involved in the editing and review of the manuscript. O.K. acquired the primary funding for this work.

Competing interests

The authors declare no competing interests.

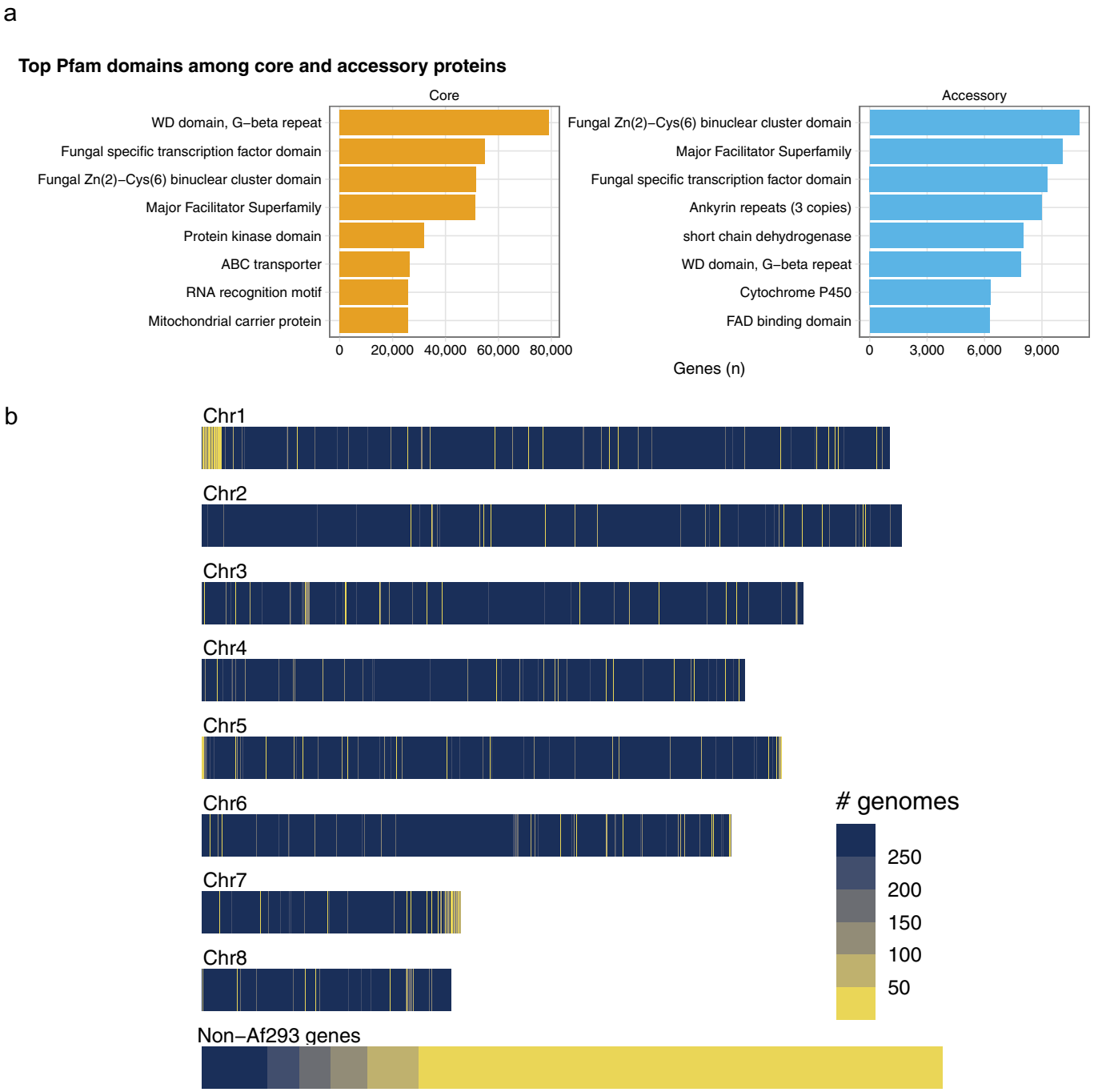
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-021-00993-x.

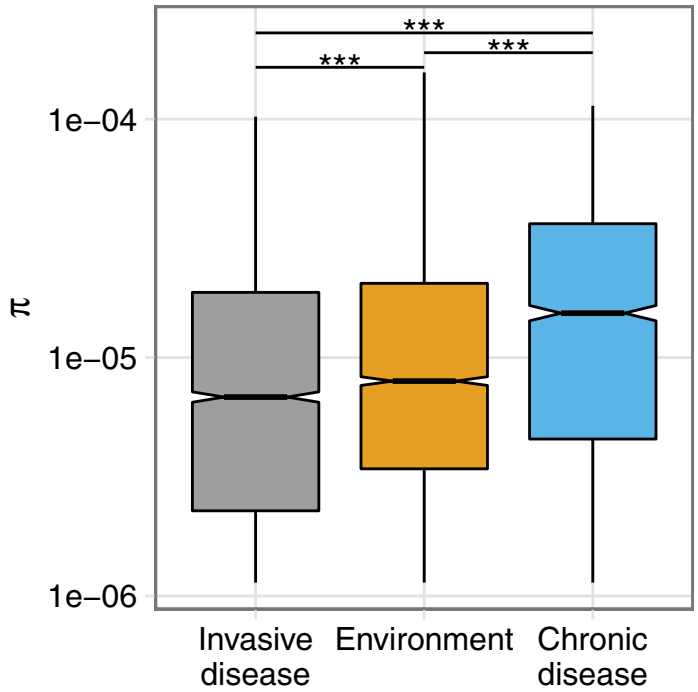
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-00993-x.

Correspondence and requests for materials should be addressed to Gianni Panagiotou or Oliver Kurzai.

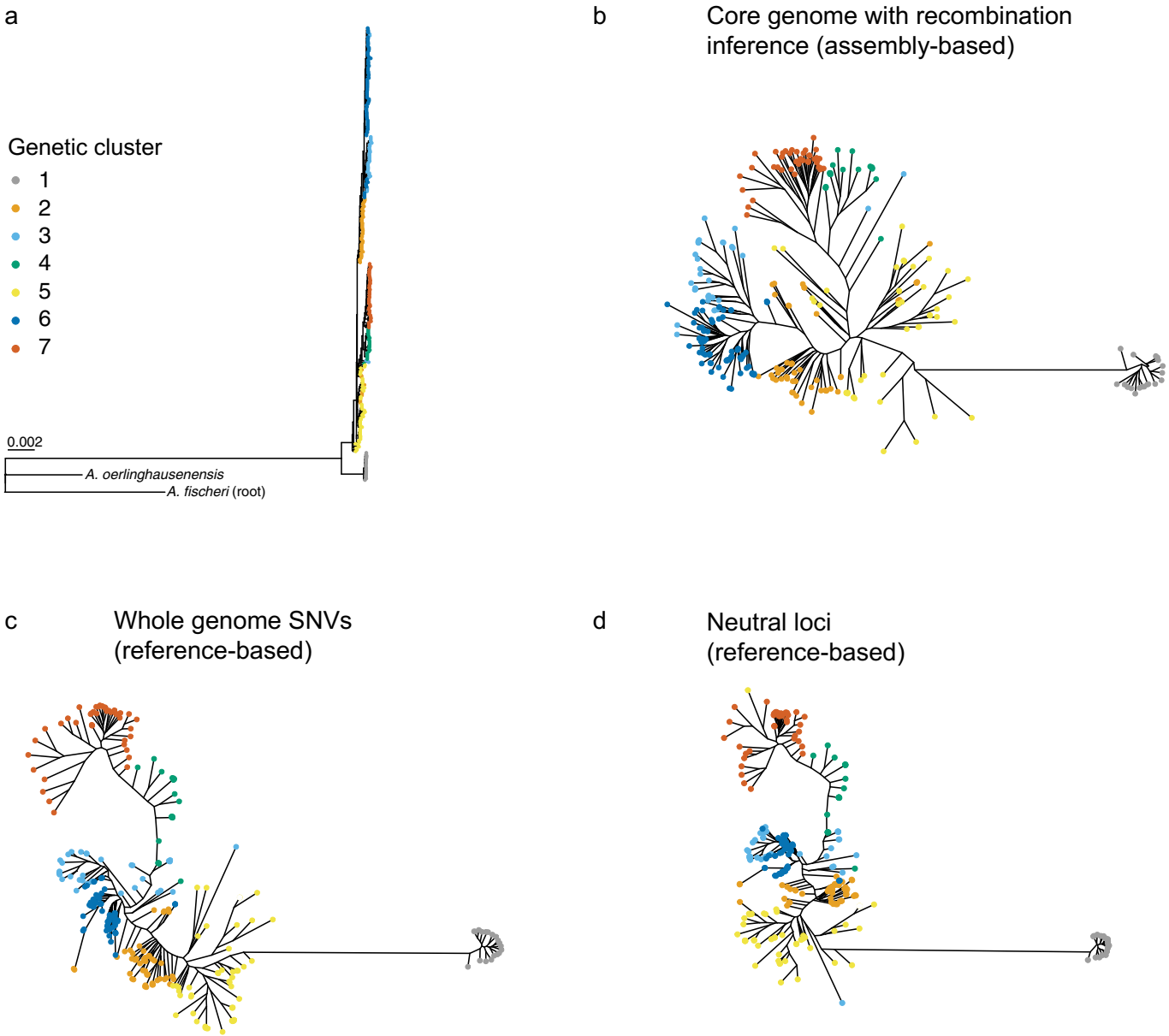
Peer review information *Nature Microbiology* thanks Nancy Keller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



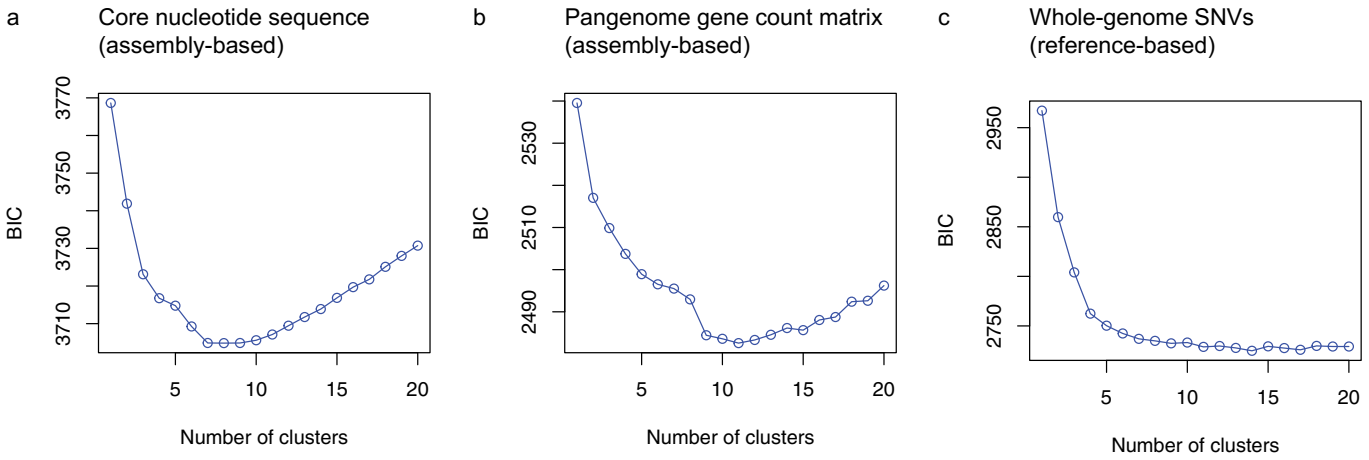
Extended Data Fig. 1 | The pan-genome of *A. fumigatus*. (A) Most frequently occurring Pfam domains among the core and accessory genomes. Values represent the total sum of domain-containing proteins among all 300 genomes. (B) Conservation of Af923 genes in the *A. fumigatus* pan-genome, arranged by chromosomal location in Af293. Each gene in Af293 is represented by a uniform-sized band that is coloured according to its prevalence among the 300 isolates analysed. Genes not in Af293 and their relative frequency are depicted at the bottom.



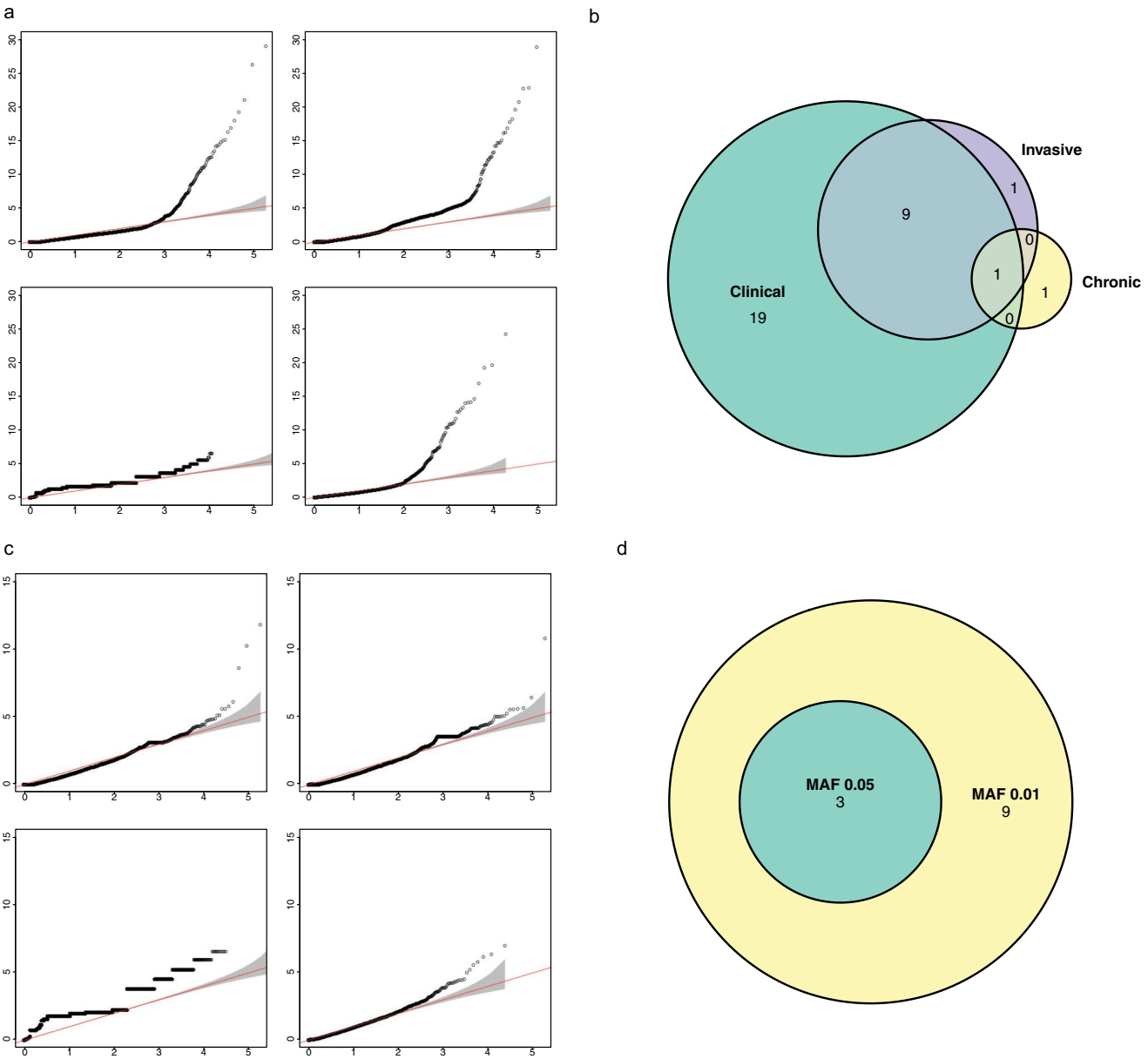
Extended Data Fig. 2 | Nucleotide diversity (π) of *A. fumigatus* isolates from the environment, invasive disease and chronic disease. π was calculated using 5 kb sliding windows across with genome with a 500 bp step size. Due to the underrepresentation of isolates from chronic disease in the dataset, isolates from the environment and invasive disease were downsampled to match the number of isolates from chronic disease (n = 19 isolates per group). The bold line in the box-and-whisker plot indicates the 50th percentile, and the box extends from the 25th to the 75th percentiles. The whiskers denote the lowest and highest values within 1.5 interquartile range. Statistical significance determined by two-sided Mann–Whitney *U* test with Bonferroni correction. *** represents $P < 0.001$. Exact P-values are: chronic vs. environmental: $P = 97e\text{-}39$; chronic vs invasive: $P = 1.6e\text{-}78$; invasive vs. environmental: $P = 5.6e\text{-}14$.



Extended Data Fig. 3 | Phylogenies constructed from the genomes of 300 *A. fumigatus* using de novo assembled genomes and reference-base analyses. (a-b) Core genome phylogeny built from nucleotide coding sequence of 5,380 single-copy orthologous genes shared by all 300 *A. fumigatus* isolates, *A. oerlinghausenensis* and *A. fischeri* (alignment length = 9,178,893 bp). Panel a shows the phylogeny rooted with *A. fischeri* and depicts the scaled relationship between the two outgroups and the *A. fumigatus* samples. Panel b depicts this phylogeny unrooted and with outgroups removed for comparison to the other phylogenies. (c) Phylogeny from concatenated SNVs following read alignment to Af293 and variant calling (n = 341,031 base pair). Genomic positions with zero coverage in any sample were removed from the alignment. (d) SNV-based phylogeny constructed from 4-fold degenerate (neutral) loci (n = 35,052 base pair).



Extended Data Fig. 4 | Discriminant analysis of principle components of 300 *A. fumigatus* isolates. (a-c) Number of clusters vs. Bayesian information criteria (BIC) was used to assess the best supported number of genetic clusters for the dataset. The input for analysis was either (a) a non-gapped core nucleotide alignment from 5,830 single-copy orthologous genes (b) a gene count matrix from orthogroup-based clustering of pansequences or (c) or whole-genome SNV data with of positions with zero genomic coverage in any isolate in the dataset excluded.



Extended Data Fig. 5 | GWAS for variants associated with clinical isolates and triazole resistance. (a & c) Q-Q plots for association with isolate source (a; clinical vs. environmental) and triazole resistance (c; resistance to one or more triazole vs. susceptible to all examined; c). Four software were utilized: EMMAX (top, left), GEMMA (top, right), treeWAS (bottom, left) and ECAT (bottom, right). The resulting Q-Q plots were used to identify the tool that produced outputs where the expected p-value distribution (x-axis) best matched the observed p-values (y-axis). (b) Venn diagram showing the gene overlap between GWAS for all clinical strains relative to environmental and significant genes specific to acute and chronic disease. (d) Venn diagram showing the gene overlap for association with triazole resistance when minor allele frequencies (MAF) of 0.01 and MAF 0.05 were used.

nature research

Corresponding author(s): Oliver Kurzai, Gianni Panagiotou

Last updated by author(s): Oct 3, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software or code was used for data collection
Data analysis	The R programming language (version 4.0.3), FastQC (version 0.11.5), BWA-MEM (version 0.7.9-r779-dirty), Picard (version 2.18.25), GATK Toolkit (version 4.1.0.0), bcftools (version 01.1.1), VCFTools (version 0.1.16), SnpEff (version 4.3t), IDBA-hybrid (version 1.1.3), QUAST (version 5.0.2), Funannotate (version 1.5.2-4cfc7f8), RepeatMasker (version 1.332), EvidenceModeler (version 1.1.1), GeneMark-ES (version 4.38), tRNAscan-SE (version 1.4), BUSCO (version 4.1.4), PFAM (32.0), MEROPS (12.0), dbCAN2 (7.0), KofamScan (version 1.2.0), InterProScan (version 5.42-78.0), Orthofinder (version 2.5.2), DIAMOND (version 0.9.24), MUSCLE (version 3.8.31), IQ-TREE (version 2.1.2), phangorn (version 2.5.5), PAL2NAL (version 14), PAML (version 4.9), CoinFinder (version 1.1), EMMAX (version beta-07Mar2010), TreeWAS (version 1.0), GEMMA (0.98.3), ECAT (version 1), Scoary (version 1.6.16), igraph (version 1.2.6)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw FASTQ files for samples sequenced in this study were uploaded to the NCBI Sequence Read Archive and are publicly available under BioProject PRJNA697844. Annotated genome assemblies were submitted to NCBI GenBank and are available under the NCBI BioSample accession numbers listed in Supp. File 1. FungiDB

release 39 datasets are available at https://fungidb.org/fungidb/app/downloads/release-39/. The NCBI RefSeq non-redundant (nr) protein database v22.01.08 is accessible at https://ftp.ncbi.nlm.nih.gov/blast/db/cloud/2018-01-22/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The primary dataset of this study is the genomic sequences of 300 Aspergillus fumigatus isolates. The number of isolates needed to capture the genomic diversity of this organism was determined by monitoring the size of the pangenome as the number of genomes analyzed increased (Figure 1b). The plateau observed in the number of pangenes identified after ~250 genomes demonstrates that the number of isolates analyzed in this study capture the majority of genetic diversity of A. fumigatus and the analysis of additional isolates is unlikely to provide substantially more information.
Data exclusions	No data was excluded from this study.
Replication	The only experimental work in this study involves the antifungal susceptibility testing of the samples analyzed genomically. For this, replication was handled through an initial screening test for increased minimum inhibitory concentration with the clinical VIPCheck diagnostic and then confirmed via broth microdilution following standard (EUCAST) protocol. All isolates that screened positive for azole resistance via the VIPCheck assay were found to have MICs above the clinical breakpoints upon subsequent broth microdilution.
Randomization	As this was a genomic survey of 300 Aspergillus fumigatus isolates, no group randomization was necessary or applied.
Blinding	As this was a genomic survey of the intraspecies diversity of Aspergillus fumigatus, blinding was not necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging