Giant transposons promote strain heterogeneity in a major fungal pathogen

Emile Gluck-Thaler[1,2,3]*, Adrian Forsythe[4], Charles Puerner[5], Jason E. Stajich[6], Daniel Croll[1], Robert A. Cramer[5], Aaron A. Vogan[4]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Neuchâtel, Neuchâtel 2000, Switzerland
[2] Department of Plant Pathology, University of Wisconsin-Madison, Madison, WI 53706, USA
[3] Wisconsin Institute for Discovery, Madison, WI 53706, USA
[4] Systematic Biology, Department of Organismal Biology, Uppsala University, Uppsala, 752 36, Sweden
[5] Department of Microbiology and Immunology, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755 USA
[6] Department of Microbiology and Plant Pathology, University of California-Riverside, Riverside, CA 92521, USA

* Corresponding author: Tel: +1 608 264 3239; Email: gluckthaler@wisc.edu

Running title: *Starships* generate variation in a fungal pathogen

ORCIDs
Emile Gluck-Thaler: https://orcid.org/0000-0003-0438-7495
Adrian Forsythe: https://orcid.org/0000-0003-1966-9946
Charles Puerner: https://orcid.org/0000-0002-4402-0295
Jason E. Stajich: https://orcid.org/0000-0002-7591-0020
Daniel Croll: https://orcid.org/0000-0002-2072-380X
Robert A. Cramer: https://orcid.org/0000-0001-5503-5006
Aaron A. Vogan https://orcid.org/0000-0003-2013-7445

Key words:
transposable element, transposon, Aspergillosis, Aspergillus fumigatus, strain heterogeneity, secondary metabolism

Abstract and Importance word count: 299
Main text word count: 9,094

Abstract

Fungal infections are difficult to prevent and treat in large part due to strain heterogeneity. However, the genetic mechanisms driving pathogen variation remain poorly understood. Here, we determined the extent to which *Starships*—giant transposons capable of mobilizing numerous fungal genes—generate genetic and phenotypic variability in the human pathogen *Aspergillus fumigatus*. We analyzed 519 diverse strains, including 12 newly sequenced with long-read technology, to reveal 20 distinct *Starships* that are generating genomic heterogeneity over timescales potentially relevant for experimental reproducibility. *Starship*-mobilized genes encode diverse functions, including biofilm-related virulence factors and biosynthetic gene clusters, and many are differentially expressed during infection and antifungal exposure in a strain-specific manner. These findings support a new model of fungal evolution wherein *Starships* help generate variation in gene content and expression among fungal strains. Together, our results demonstrate that *Starships* are a previously hidden mechanism generating genotypic and, in turn, phenotypic heterogeneity in a major human fungal pathogen.

Importance

No “one size fits all” option exists for treating fungal infections in large part due to genetic and phenotypic variation among strains. Accounting for strain heterogeneity is thus fundamental for developing efficacious treatments and strategies for safeguarding human health. Here, we report significant progress towards achieving this goal by uncovering a previously hidden mechanism generating heterogeneity in the major human fungal pathogen *Aspergillus fumigatus*: giant transposons called *Starships* that span dozens of kilobases and mobilize fungal genes as cargo. By conducting the first systematic investigation of these unusual transposons in a single fungal species, we demonstrate their contributions to population-level variation at the genome, pangenome and transcriptome levels. The *Starship* atlas we developed will not only help account for variation introduced by these elements in laboratory experiments but will serve as a foundational resource for determining how *Starships* shape clinically-relevant phenotypes, such as antifungal resistance and pathogenicity.

Introduction

Infectious diseases caused by fungi pose a grave threat to human health. The World Health Organization recently coordinated a global effort to prioritize research among fungal pathogens based on unmet research needs and public health importance (1). Among the pathogens deemed most important for research include *Aspergillus fumigatus*, a globally distributed human pathogen causing disease in an estimated 3 million people yearly (2, 3). Several infectious diseases are caused by *A. fumigatus*, including invasive pulmonary aspergillosis, which manifests primarily in immunocompromised individuals with mortality rates of up to 88% (4). The treatment of *A. fumigatus* infections are complicated by the remarkable variation strains display in virulence, resistance to antifungals, metabolism, and other infection-relevant traits (3, 5–8). Strain heterogeneity confounds "one size fits all" therapies and poses a significant challenge for developing efficacious disease management strategies (9). Recent evaluations of the *A. fumigatus* pangenome have revealed extensive genetic variability underlying variation in clinically-relevant traits such as antifungal resistance and virulence, yet in many cases, the origins of such variation remain unexplained (10–12). Determining the genetic drivers of strain heterogeneity will help accelerate the development of strain-specific diagnostics and targeted therapies.

Mobile genetic elements (MGEs) are ubiquitous among microbial genomes and their activities profoundly shape phenotypic variation (13). MGE transposition generates structural variation that directly impacts gene regulation and function (14–16), and many elements also modulate genome content by acquiring genes as "cargo" and transposing them within and between genomes, facilitating gene gain and loss (17, 18). The ability to generate contiguous genome assemblies with long-read sequencing technologies has dramatically enhanced our ability to find new lineages of MGEs (13, 19). For example, *Starships* are a recently discovered superfamily of MGEs found across hundreds of filamentous fungal taxa (20–22). *Starships* are fundamentally different from other fungal MGEs because they are typically 1-2 orders of magnitude larger (~20-700 kb versus 1-8kb) and carry dozens of protein-coding genes encoding fungal phenotypes. In addition to flanking short direct repeats, all *Starships* possess a "captain" tyrosine site-specific recombinase at their 5' end that is both necessary and sufficient for transposition (21). Many *Starships* make key contributions to adaptive phenotypes: for example, *Horizon* and *Sanctuary* carry the ToxA virulence factor that facilitates the infection of wheat, and *Hephaestus* and *Mithridate* encode resistance against heavy metals and formaldehyde, respectively (23–26). *Starships* are capable of horizontal transfer, implicating them in the acquisition, dissemination and repeated evolution of diverse traits ranging from host interactions to stress resistance (20, 21, 24–26). *Starships* may also fail to re-integrate into the genome during a transposition event, leading to the loss of the element, its cargo and its encoded phenotypes (21). Through horizontal transfer and failed re-integration events, *Starships* contribute directly to the generation of rapid gene gain and loss polymorphisms across individuals. However, we know little about how *Starships* drive genetic and phenotypic variation in species relevant for human disease.

Here, we conduct the first systematic assessment of *Starship* activity and expression in a human fungal pathogen to test the hypothesis that these unusual transposons are a source of strain heterogeneity. We reveal that *Starships* are responsible in part for generating key instances of previously unexplainable variation in genome content and structure. Our interrogation of 507 diverse clinical and environmental strains of *A. fumigatus,* combined with highly contiguous assemblies of 12 newly sequenced strains, enabled an unprecedented quantification of *Starship* diversity within a single species. We leveraged the wealth of functional data available for *A. fumigatus* to draw causal links between *Starship*-mediated genetic variation and phenotypic heterogeneity in secondary metabolite and biofilm production contributing to pathogen survival and virulence. We analyzed multiple transcriptomic studies and determined that variation in *Starship* cargo expression arises from strain- and treatment-specific effects. By revealing *Starships* as a previously unexamined mechanism generating phenotypic variation, our work sheds light on the origins of strain heterogeneity and establishes a predictive framework to decipher the intertwining impacts of transposons on fungal pathogenesis and human health.

Results

Data curation and long-read sequencing

Starships are difficult (but not impossible) to detect in short-read assemblies due to their large size and frequent localization in regions with high repetitive content (20). The wealth of publicly available short-read genomes are still of high value, however, due to the breadth of sampling they provide combined with the still relevant possibility of finding *Starships*, especially in higher quality short-read assemblies such as those available for *A. fumigatus*. We therefore deployed a hybrid sampling strategy that leveraged both short- and long-read genome assemblies, combined with short-read mapping to genotype *Starship* presence/absence, to enable an accurate and precise accounting of how *Starships* impact strain heterogeneity. We downloaded 507 publicly available short-read *A. fumigatus* assemblies that span the known genetic diversity of this species isolated from environmental and clinical sources (51.7% clinical, 47.9% environmental, 0.4% unknown isolation source) and used Oxford Nanopore to sequence 12 additional strains with long-read technology (66.7% clinical, 25% environmental, 8.3% unknown), for a combined total of 519 assemblies (10, 11). The 12 isolates selected for long-read sequencing were chosen because they represent commonly utilized laboratory strains, clinical isolates and environmental strains from three major *A. fumigatus* clades (10). We performed additional short-read Illumina sequencing of the 12 strains to provide additional support and error-correct the long-read assemblies. The long-read assemblies are of reference quality, representing nearly full chromosome assemblies with an L50 range of 4-5 and a N50 range of 1.9-4.7 Mb, and are thus ideal complements to the available short-read assemblies for investigating the connection between *Starship*s and strain heterogeneity (Table S1).

At least 20 active and distinct *Starships* vary among *A. fumigatus* strains

We began evaluating the impact of *Starships* on strain heterogeneity by systematically annotating them in the 519 assemblies using the starfish workflow (22). We identified a total of 787 individual *Starships* and validated these predictions by manually annotating a subset of 86 elements (Table S2, Table S3, Methods) (10, 11). As expected, more *Starships* were recovered in total from short-read assemblies, but on average more *Starships* were recovered from each reference-quality assembly, highlighting the utility of sampling both short- and long-read assemblies (Fig. S1). Importantly, unlike many other fungal transposons, we found that with few exceptions, if a *Starship* is present in a strain, it is present at most in a single copy. Some other species appear to readily accumulate multiple copies of the same *Starship* (20), but this is not the case for *A. fumigatus*.

To determine how many different *Starships* actively generate variability in *A. fumigatus*, we assigned each *Starship* element to a family (based on similarity to a reference library of captain tyrosine recombinases), a *navis* (Latin for “ship”; based on orthology of the *A. fumigatus* captains) and a *haplotype* (based on k-mer similarity scores of the cargo sequences; Methods), following Gluck-Thaler and Vogan 2024. For example, *Starship Osiris h4* belongs to *haplotype* 4 within the *Osiris navis*, which is part of the Enterprise-family. The most reasonable conclusion when observing the same transposon bounded by identical direct repeats at two different loci in two individuals from the same species is that this transposon is presently or recently active (21). Using a threshold that required observing the same *navis-haplotype* (i.e., *Starship*) at two or more sites, we identified 20 high-confidence *Starships* that met these criteria and 34 medium-confidence *Starships* that did not, revealing a phylogenetically and compositionally diverse set of *Starships* actively transposing within *A. fumigatus* (Methods, Table 1, Fig. 1, Table S4, Table S5). For each of the 20 high-confidence elements, we aligned representative copies present in different genomic regions to highlight how precisely and repeatedly element boundaries are conserved across transposition events (Fig. S2).

After using all available genome assembly data to build the *Starship* library with automated methods, we explored a variety of other approaches for augmenting and validating our predictions, including manual annotation, BLAST-based detection, and short-read mapping (Methods). Certain detection methods were more appropriate for particular analyses, resulting in the partitioning of our findings into the high-confidence, the expanded, and the genotyping datasets. To analyze *Starship* features, like gene content and pangenome distributions (see below), we limited our analyses to the set of 459 high-confidence elements for which we have end-to-end boundaries and evidence of transposition (high-confidence dataset). For analyses that examine the presence/absence of elements across genomic regions, we used an expanded set of 1818 elements that consists of the 787 high and medium-confidence elements as well as 1031 elements whose presence was detected through BLAST searches, because these methods maximize our ability to correctly predict the presence/absence of an element in a given region (expanded dataset; Table S4, Table S5). For genotyping analyses that focus on detecting the

148 presence/absence of an element regardless of where it is found in the genome, we took a
149 conservative approach by limiting our analysis to the 13 reference quality assemblies plus 466 *A.*
150 *fumigatus* strains with publicly available paired-end short-read Illumina data that were used to
151 validate *Starship* presence/absence by short-read mapping to a reference *Starship* library
152 (genotyping dataset; Table S21).

153 *Starships* distributions are heterogeneous and partially explained by strain 154 relatedness

155

156 All high-confidence *Starships* show polymorphic presence/absence variation, representing a
157 previously unexamined source of genetic heterogeneity across *A. fumigatus* strains (Fig. 1A,
158 Table S6). For example, the two commonly used reference strains Af293 and CEA10 carry 12
159 different high-confidence *Starships*, but share only 4 in common (or only 2, using a more
160 conservative threshold of not counting fragments or degraded elements; Fig. 1D, Table S7). The
161 number of high-confidence *Starships* per long-read isolate ranges from 0-11 (median = 3, std.
162 dev. = 3.5)(27, 28). To identify the underlying drivers of variation in *Starship* distributions, we
163 investigated the relationship between *Starship* repertoires and strain relatedness by testing if
164 phylogenetic signal underlies *Starship* distributions. We genotyped *Starship* presence/absence
165 using short-read mapping in the best sampled clade of *A. fumigatus* with well resolved
166 phylogenetic relationships (Clade 1 *sensu* Lofgren et al., 2022, n = 150) and found that isolate
167 relatedness is significantly but only weakly correlated with similarity in *Starship* repertoire (Fig.
168 1E, Table S8, Spearman's $\rho = 0.29$; $P < 2.2e^{-16}$, using the genotyping dataset). This indicates that
169 *Starships* may be a source of genetic heterogeneity among even closely related strains. While no
170 sequence-based or phylogenetic method to our knowledge would enable us to unequivocally
171 differentiate horizontal transfer events from incomplete lineage sorting, loss or sexual and
172 parasexual recombination at the within species level, we hypothesize that horizontal transfer
173 within *A. fumigatus* is at least partially responsible for generating the observed patchy distributions
174 of *Starships*, since *Starship* transfer has occurred between fungal species of vastly different
175 taxonomic ranks, and barriers to transfer are likely lower within species (24, 25). The hypothesis
176 of horizontal transfer within species remains an exciting possibility that must be confirmed with
177 laboratory experiments.

178 *Starships* may actively transpose in clones of reference laboratory strains

179

180 We examined the short-term potential for *Starships* to introduce unwanted variation into laboratory
181 experiments by comparing lab-specific isolates of the same strain used by different research
182 groups. First, we found that *Starship* insertions in the reference Af293 genome coincide with 4/8
183 known putative structural variants identified in other Af293 strains by Collabardini et al. 2022
184 (Table S9), suggesting that in addition to other mechanisms, *Starship* transposition and/or loss
185 contributes to structural variation observed among isolates of what should otherwise be clones of

186 the same strain (6). We complemented this literature-based approach with a genomics approach
187 where we compared sequence data among isolates of the same strain. We compared the
188 locations of *Starships* between the CEA10 assembly we generated here and a publically available
189 CEA10 assembly (BioSample SAMN28487501), and found evidence that at least 1 *Starship*
190 (*Nebuchadnezzar h1*) has jumped to a different chromosomal location, from chromosome 6 to
191 chromosome 4 (Fig. S3). While this specific translocation appears robust, other results from these
192 analyses suggest that the strain of CEA10 sequenced here may have been derived from a sample
193 with mixed lineages (See methods). As a result we cannot currently rule out other more exotic
194 explanations for the movement of *Nebuchadnezzar h1* in this isolate. To further address the
195 possibility of *Starship* movement under short timescales relevant to laboratory experiments, we
196 evaluated *Starship* heterogeneity in three additional strains with both publicly available short-read
197 data and long-read data generated and found that one of the strains (ATCC-46645-lr) showed
198 signs of *Starship* heterogeneity among isolates used by different research groups. Specifically, a
199 small number of reads (a single mapped read for *h2* and three for *h1*) supported precise excision
200 events for *Nebuchadnezzar h1* and *h2*, suggesting these elements have transposed or have been
201 lost in a subset of the nuclei within one of the sequenced isolates (Fig. S4). This suggests that
202 some but not all *Starships* appear to be active under laboratory conditions in some strains, which
203 is directly relevant for experimental design in this case as *Nebuchadnezzar h1* carries the HAC
204 gene cluster known to impact biofilm-associated virulence (see below) and genes on *Starships*
205 may be lost through failed transposition events (29). Additionally, one breakpoint of a large-scale
206 reciprocal translocation between chromosomes 1 and 6 among CEA10 isolates is localized
207 precisely within a *Starship*, suggesting that this structural variant may have been caused by the
208 presence of the *Starship*, an association that has also been noted in at least one other species
209 (20). Together, *Starship*-mediated variation among isolates of the same *A. fumigatus* strain
210 suggests these transposons may cause genomic instability over short-enough timescales to
211 potentially impact routine laboratory work and experimental reproducibility. It is interesting to note
212 that not all strains or closely related strains demonstrate intraspecific variation: for example,
213 similar to Bowyer et al. 2022, we did not detect large-scale variation in *Starship* presence/absence
214 between CEA10 and its derivative A1163 (30), which have been separated for decades, indicating
215 that much remains to be known about the factors promoting within strain *Starship* variability.

216 *Starships* mobilize upwards of 16% of accessory genes that differ across 217 strains

218

219 *A. fumigatus* strains differ extensively in the combinations of genes found in their genomes,
220 resulting in a large and diverse pangenome (10, 11). To quantify how much pangenomic variation
221 is attributable to actively transposing *Starships,* we estimated the total proportion of genomic
222 content present in the 20 high-confidence *Starships* (Fig. 2, Methods). We examined these
223 *Starships* in the 13 reference-quality assemblies (Af293 plus the 12 newly sequenced long-read
224 strains) and found that between 0-2.4% of genomic nucleotide sequence (and 0-2.1% of all genes
225 per genome) is mobilized as *Starship* cargo (Fig. 2A). We then built a pangenome with all 519

strains and extracted all orthogroups (i.e., genes) found in the 13 reference-quality genomes to gain insight into the distribution of *Starship*-associated gene content (Table S10). Across the 13 isolate pangenome, 2.9% of all genes, which corresponds to 9.7% of all accessory and singleton genes, have at least 1 member carried as *Starship* cargo (Fig. 2B). Accessory and singleton genes are overrepresented >3 fold in *Starships*, with ~92% of *Starship* genes being either accessory or singleton compared with 24.6% of non-*Starship* associated genes. We determined the upper bounds of this conservative estimate by examining the 13 isolate pangenome in the context of all 54 high- and medium-confidence *Starships*, and found that 4.8% of all genes, representing 16% of all accessory and singleton genes, have at least 1 member carried as *Starship* cargo (Fig. S5, Table S11). Drawing parallels from decades of observations of bacterial MGEs (17), we hypothesize that localization on a *Starship* promotes a gene's rate of gain and loss through *Starship*-mediated horizontal transfer and failed re-integration events (21). Together, these data reveal a previously hidden association between *Starships* and the making of *A. fumigatus'* accessory pangenome.

Starships differ in their potential to mediate gain and loss of unique sequence

We further determined the potential for *Starships* to introduce variation into *A. fumigatus* strains by testing if *Starship* genes are uniquely found in those elements or found elsewhere in the genome. We calculated the degree of association between a given cargo gene and a *Starship* by identifying the genomic locations of best reciprocal BLAST hits to genes in the 20 high-confidence type elements (Fig. 2C, Fig. S6, Table S12, Table S13, Methods). Across the 13 reference-quality assemblies, we found that the vast majority of *Starship* genes display presence/absence variation between strains (i.e., are accessory or singleton) although several genes are present in conserved regions in these particular strains (while seemingly counter-intuitive, these genes do vary in their presence/absence and are *Starship*-associated when examining the larger 519 strain population, as expected). We found that many genes are almost always carried as cargo in active *Starships* when present in a given genome (e.g., the majority of cargo on *Tardis h1*), while others have weaker associations and are found in both *Starships* and non-*Starship* regions across different strains (e.g., *Gnosis h1*). Variation in the degree to which genes associate with active *Starships* suggests elements differ in capacity to generate accessory sequence variation among strains, highlighting the importance of investigating *Starships* at the individual element level.

Starship activity generates structural variation at a genome-wide scale

We next asked where *Starship*-mediated variation occurs in the genome to better understand the implications of *Starship* activity for genome organization. We identified the genomic locations where *Starships* introduce structural variation by sorting all 787 high and medium-confidence elements, along with elements detected by BLASTn, into homologous genomic regions (Table S14, Table S15, Table S16; expanded dataset; Methods). This enabled us to genotype individuals in the 519 strain population for segregating *Starship* insertions that are polymorphic across strains

(Fig. 3, Fig. S7; Methods). Across all strains and chromosomes, we found 79 regions that contain at least 1 segregating "empty" insertion site, for a total of 154 sites distributed across all eight chromosomes (a single region can have >1 insertion site). The average number of empty sites per genome is 44.5 (range = 9-63, std. dev. = 13.88), indicating that each strain harbors dozens of sites with structural variation introduced by *Starships*. For example, the Af293 reference strain has a total of 6 full length *Starships* with annotated boundaries and 2 *Starship* fragments, and a total of 56 segregating empty sites (Fig. 3A). We found a significant linear relationship between the number of genomic regions a given *Starship* is present in (a proxy for transposition activity) and the total copy number of that element in the 519 strain population, suggesting active *Starships* contribute more to strain heterogeneity compared with less active elements (Fig. 3C; $y = 14.7x - 1.24$; $P = 6.5e^{-4}$; R^2 adj = 0.46; expanded dataset). Predicted *Starship* boundaries are precisely conserved across copies present in different genomic regions, suggesting transposition, and not translocation or segmental duplication, is the major mechanism responsible for generating variation in *Starship* location (Fig. 3D).

Given that *Starships* are mobilized by a tyrosine site-specific recombinase (21), we attempted to predict each *Starship*'s target site to gain insight into the genomic sequences that are susceptible to *Starship* insertions. Among high-confidence *Starships*, 19/20 have identifiable direct repeats (DRs) ranging from 1-12 bp in length and 18/20 have terminal inverted repeats (TIRs; Table 1, Table S3). DRs typically reflect a portion of the element's target site, while TIRs are predicted to facilitate transposition (31). While precise target site motifs must be confirmed experimentally, the target site TTACA(N_5)AAT that we recovered for *Nebuchadnezzar* elements, which belong to the Hephaestus-family, resembles the canonical *Hephaestus* target site TTAC(N_7)A, demonstrating the utility of our approach as a first step towards understanding what sequence features promote the gain and loss of *Starship*-mediated variation (21).

Generally, DRs >6 bp are associated with targeting of the 5S rDNA gene (e.g. *Osiris*), presumably because this is a relatively stable target site (22). In contrast to this, *Tardis* has a highly conserved 9 - 11 bp DR that does not correspond to the 5S gene. A k-mer analysis shows that the 10 bp consensus motif CTACGGAGTA is strongly overrepresented in the genome of Af293 (>99.99th percentile of k-mers; Fig. S8), indicating that this motif may represent some other type of highly conserved genomic sequence, such as a transcription factor binding site. Although inconclusive, this motif does overlap with predicted motifs associated with C6 zinc cluster factors in *A. nidulans* (https://jaspar.elixir.no/matrix/UN0291.2/) (32). This DR sequence is further conserved among other *Eurotiomycetes* (22). This result highlights the fact that studying *Starship* elements in detail can reveal other important aspects of a species' biology, and characterizing the *Tardis* motif should be of interest for future research.

302

303 ## *Starships* modulate variation at an idiomorphic biosynthetic gene cluster

304

305 Several genomic regions harbor multiple types of *Starships*, raising the possibility that strain
306 heterogeneity arises in part from the formation of *Starship* insertion hotspots. We investigated the
307 genomic region with the highest density of *Starship* insertions to define the upper bounds of
308 *Starship*-mediated structural variation at a single locus (Fig. 4C, Methods). This region spans an
309 average of 498.76 Kb (range = 158.07-781.54 kb; std. dev. = 83.02 kb) across the n = 210 strains
310 for which we could detect it, and ranges from position 584,521-1,108,409 on Chromosome 3 in
311 the Af293 reference assembly (representing 12.84% of the entire chromosome). By
312 supplementing starfish's automated genotyping with manual annotation of nine strains with
313 distinct alleles at this region, we found this region contains at least 7 distinct *Starships* with
314 identifiable DRs and 1 degraded *Starship* that range from 47.52-151.62 kb long and are inserted
315 into 6 independent segregating sites (Fig. 4C). The majority (75%) of the *Starships* in this region
316 are inserted into 5S rDNA coding sequences, which effectively fragments that copy of the 5S
317 rDNA gene. Total 5S rDNA copy number varies between 28-33 in the 13 reference quality
318 assemblies (median = 31, std. dev. = 1.4; Table S17), but between 6-8 intact copies are typically
319 present at this single locus, representing a ~10-fold enrichment of the 5S rDNA sequence relative
320 to background expectations given the length of this region (using 5S rDNA frequencies in the
321 Af293 genome). Thus, the enrichment of 5S rDNA at this locus potentiates *Starship*-mediated
322 variation and the generation of a *Starship* hotspot.

323

324 We were surprised to find that in addition to multiple segregating *Starships*, the Chromosome 3
325 region contains an idiomorphic biosynthetic gene cluster (BGC) that was previously noted to be
326 a recombination hotspot (Cluster 10 *sensu* Lind et al. 2017) (33, 34). The idiomorphic BGC locus
327 is polymorphic for upwards of 4 smaller BGC modules and is upstream of a fifth BGC encoding
328 the biosynthesis of Fusarinine C (BGC8 *sensu* Bignell et al. 2016) that itself is embedded within
329 a larger stretch of sequence containing NRPS and KR-PKS core genes. While the idiomorphic
330 modules were not predicted by antiSMASH, manual inspection revealed that each contains a
331 different core SM biosynthesis gene and numerous other genes involved in metabolic processes,
332 suggesting they are part of a larger BGC or represent different cryptic BGCs. The NR-PKS BGC
333 module at this locus is carried by *Starship Osiris h4,* and its presence/absence is directly
334 associated with the presence/absence of the element, implicating *Starships* as a mechanism
335 generating idiomorphic BGCs. Furthermore, some BGC modules are disrupted by the insertion of
336 a *Starship* (e.g., the NRPS-like BGC module). Together, these findings support the assertion that
337 *Starship* activity helps generate selectable variation in BGC genotypes. The implication of this
338 variation for the expression and diversification of natural products warrants further investigation.

Starships encode clinically-relevant phenotypes and are enriched in strains isolated from clinical and environmental sources

Starships encode genes with diverse functions typically associated with fungal fitness, including carbohydrate active enzymes, BGCs and metal detoxification genes (Table 1, Table S5). To investigate broad trends in how *Starships* might contribute to phenotypic heterogeneity, we compared the predicted cargo functions among the 20 high-confidence *Starships* (Table S18; high-confidence dataset; Methods). As expected, *Starship* types differ from each other in their predicted functional content, but functional diversity also varies to some extent within elements from the same type, indicating that both inter- and intra-specific *Starship* variation may contribute to functional heterogeneity among *A. fumigatus* strains. In this sense, different copies of the "same" *Starship* in different individuals have much greater potential to differ from each other compared with much smaller transposons. We found that 49.7% of COG-annotated genes are "Poorly Categorized"; 28.8% belong to the "Metabolism" category, 14.7% belong to "Cellular Processes and Signaling" and 6.8% belong to "Information Storage and Processing".

Given the importance of metabolic processes for diverse fungal traits, we examined granular classifications within the Metabolism category and found that *Starship* types differ specifically in their contributions to metabolic heterogeneity (Fig. S9). For example, nearly all 75 copies of *Gnosis h1* carry at least 1 gene with a predicted role in "Coenzyme transport and metabolism," and none carry genes with predicted roles in "Lipid transport and metabolism", while the exact opposite is true for the 119 copies of *Tardis h1*. The most frequently mobilized COG metabolism categories are "Carbohydrate transport and metabolism" (present in 371 individual elements belonging to 13 high-confidence *Starships*), "Secondary metabolite transport and metabolism" (present in 329 elements belonging to 16 high-confidence *Starship* types) and "Energy production and conversion" (present in 281 elements belonging to 9 high-confidence *Starship* types), each of which has known associations with clinically-relevant pathogen phenotypes. We found no evidence that any metabolism-associated COG category was enriched in *Starships* compared with background frequencies in Af293 (data not shown).

While investigating associations between *Starships* and characterized pathogen phenotypes, we found three notable examples of cargo genes that encode known traits important for pathogen survival and virulence (Table S19, Methods) (10, 11, 33, 35–41). The BGC encoding Fumihopaside A (AFUA_5G00100-AFUA_5G00135) is carried by *Starship Gnosis h2* (Fig. 4A). Fumihopaside A is a triterpenoid glycoside that increases fungal spore survival under heat and UV stress exposure (35). Similarly, the BGC encoding the polyketide Fumigermin (AFUA_1G00970-AFUA_1G01010), which inhibits bacterial spore germination, is carried by *Lamia h2* and *h4* (Fig. 4B) (42). Both the Fumigermin and Fumihopaside A BGCs were previously identified as "mobile gene clusters" based on their presence/absence at different chromosomal locations in *A. fumigatus* strains, and our results reveal that *Starships* are the mechanism underpinning their mobility (Cluster 1 and 33, respectively from Lind et al., 2017) (33). Finally, a

380 cluster of three genes (AFUA_5G14900/*hrmA*, AFUA_5G14910/*cgnA*, AFUA_5G14915/*bafA*,
381 collectively referred to as the *hrmA*-associated cluster or HAC) carried by *Nebuchadnezzar h1*
382 increases virulence and low oxygen growth, and its expression modulates colony level
383 morphological changes associated with biofilm development (Fig. 4D) (29, 43). The genes *bafB*
384 and *bafC*, which are homologs of *bafA,* also mediate colony and submerged biofilm morphology
385 and are each carried by up to 5 additional *Starships*, indicating a sustained association between
386 this gene family and *Starships* (Fig. 4D, Fig. S10) (29). Mobilization of biosynthetic gene clusters
387 and biofilm-related loci by *Starships* likely contributes to the rapid evolution of these survival and
388 virulence traits by mediating gene gain and loss through, we speculate, horizontal *Starship*
389 transfer (which has yet to be experimentally demonstrated) and *Starship* loss (which has been
390 experimentally demonstrated to occur through failed re-integration events (21)), effectively
391 increasing the capacity of these traits to evolve in populations. We propose that by mobilizing
392 genes contributing to biofilm formation, *Starships* help drive heterogeneity in clinically-relevant
393 phenotypes at a population level (29).
394
395 To identify *Starships* that could be relevant in either environmental or clinical settings, we tested
396 for an association between the high-confidence *Starships* and strain isolation source (genotyping
397 dataset; Methods). We found that 5 high-confidence *Starships* are significantly enriched (P_{adj}
398 <0.05) in strains from either clinical or environmental isolation sources across the 475/479
399 genotyped strains for which source data exists (Table S2, Table S20, Table S21, Fig. S11). *Lamia*
400 *h3* (P_{adj} = 0.013) and *Janus h3* (P_{adj} = 0.034) are enriched in environmental strains. *Osiris h4* (P_{adj}
401 = 0.022), *Nebuchadnezzar h1* (P_{adj} = 0.015; carries *bafA*, see above) and J*anus h2* (P_{adj} = 0.013)
402 are enriched in clinical strains, providing complementary evidence to recently published analyses
403 of *Starship* captain enrichment in clinical isolates from several fungal species (44). Although
404 clinical and environmental strains were sampled in roughly equal proportions, we can't completely
405 rule out unknown biases in our sampling that would contribute to these enrichment patterns.
406 Nevertheless, strains with these 5 *Starships* were isolated from different countries (between 3-10
407 countries) and belong to different phylogenetic groups (between 1-3 clades *sensu* Lofgren et al.,
408 2022 and 1-6 phylogenetic clusters *sensu* Barber et al., 2021), suggesting these *Starships* are
409 excellent candidates to test how giant transposons contribute to fitness in environmental and
410 clinical settings.

411 *Starship* expression contributes to heterogeneity in a strain-, treatment- and 412 strain by treatment-dependent manner

413
414 The revelation of *A. fumigatus' Starships* allowed us to test the hypothesis that *Starship* cargo is
415 expressed under clinically-relevant conditions through re-examination of the RNA-seq datasets
416 available for this pathogen. We analyzed patterns of differential *Starship* cargo gene expression
417 in 14 transcriptomic studies from the three commonly used reference strains (Af293, CEA10,
418 A1163). In total, 177 samples were included in this meta-analysis. Samples were split into broad
419 treatment categories and corrected for batch effects to allow for comparison of cargo gene

420 expression between studies (Methods). Out of 596 *Starship* genes total (carried by the high-
421 confidence *Starships* present in the three reference strains), we identified 459 differentially
422 expressed genes (DEGs; Supplementary Data). We did not find a significant enrichment of DEGs
423 within *Starships* compared to the genomic background of any strain (Fisher's Exact Test *P*-values
424 > 0.05). However, the expression and significance of *Starship* DEGs vary across *Starship* naves,
425 fungal strains, and experimental conditions, indicative of pervasive strain-specific, treatment-
426 specific and strain by treatment interactions impacting *Starship* cargo expression (Fig. 5B & Fig.
427 S14).
428
429 To identify experimental conditions that may impact *Starship* transposition and the subsequent
430 generation of *Starship*-mediated heterogeneity, we examined patterns of transcript abundance
431 for captain genes responsible for *Starship* transposition within the 12 high-confidence *Starships*
432 from the reference strains Af293 (Fig. 5A), CEA10 and A1163 (Fig. S12). All captain genes have
433 evidence for constitutive gene expression (a minimum median transcript coverage of 1 across the
434 gene body for biological replicates) in at least one experimental condition (Fig. S13). Overall, 8 of
435 the 12 captain genes from high-confidence *Starships* are differentially expressed in at least one
436 study for one or more reference strains, revealing myriad conditions that create opportunities for
437 *Starship* transposition (Supplemental information). Differences in experimental conditions and/or
438 strain backgrounds have no consistent positive or negative influence on captain gene expression,
439 indicating that the intricacies of captain expression remain to be elucidated (Fig. 5B).
440
441 In order to identify conditions where *Starships* have potential to impact phenotypes, we examined
442 patterns of differential expression for *Starship*-mobilized cargo genes. As with captain genes, the
443 expression of cargo genes is context-dependent, within each strain background, with
444 experimental conditions and *Starship* haplotypes impacting levels of gene expression (Fig. 5B).
445 For example, we compared transcript abundance patterns for genes within the *hrmA*-asscociated
446 cluster (HAC) between treatment categories and found that the transcription of *cgnA* and *bafA*
447 genes within HAC appear to be down-regulated with a select group of infection samples (Fig. 5A
448 & Fig. S12). Together, strain identity, treatment conditions, and non-additive interactions between
449 strains and treatments generate variation in *Starship* cargo expression.
450
451 To gain further insight into *Starship*-mediated heterogeneity in gene expression and regulation,
452 we constructed weighted gene co-expression networks (WGCNA) which visualize correlations in
453 gene expression. Genes are grouped into modules based on the level of co-expression
454 (Methods). We identified five modules as being significantly associated with "antifungal"-,
455 "infection"-, or "nutrient"-based studies (Fig. 5E). Two major trends in the transcriptional network
456 properties of *Starship* cargo are apparent. First, multiple *Starship*-associated genes are integrated
457 into modules containing many other genes from the genome at large, underscoring the possibility
458 that phenotypes emerging from these networks are the product of regulatory interactions between
459 *Starship* cargo and non-*Starship* genes (Fig. 5D & Fig. S15). In particular, connections involving
460 captain genes further highlight candidate loci involved in regulatory interactions between
461 *Starships* and the *A. fumigatus* genome, and are prime targets for future investigations of the

gene networks involved in *Starship* transposition (Fig. 5E). Second, we also identified modules composed of mainly *Starship* cargo, indicative of modular networks specific to particular *Starships* (Table S24). These network analyses paint a nuanced picture of how *Starships* interact with the broader transcriptional networks of the cell, and implicate the expression of *Starship* cargo in the generation of transcriptional variation.

Discussion

Phenotypic heterogeneity among fungal pathogen strains poses a major challenge for combating infectious diseases, yet we often know little about the genetic basis of this variation. We hypothesized that a newly discovered group of unusual transposons, the *Starships*, make important contributions to strain heterogeneity. We tested this hypothesis by systematically characterizing *Starship* presence/absence and expression in *A. fumigatus*, an important human fungal pathogen of critically high research priority (1, 45). Our work provides fundamental insight into the mode and tempo of fungal evolution by revealing that strain variation emerges not only from expected genetic mechanisms (e.g., single nucleotide polymorphisms, copy number variation, chromosomal aneuploidies, and other types of structural variants) (10, 11) but from the previously hidden activity of giant transposons.

Although much remains unknown about the fundamentals of *Starship* biology, *Starships* in *A. fumigatus* likely generate variation on a scale approaching MGE-mediated evolution in bacteria, which is a major mechanism driving bacterial adaptation. The median number of *Starship*-borne genes per *A. fumigatus* isolate is 0.8% (range = 0 - 2.1%), while between 2.9-4.8% of the pangenome (corresponding to between 9.7-16% of the accessory genome) is *Starship*-associated. These observations approach analogous measurements of bacterial plasmids in the Enterobacteriaceae, where the median number of plasmid-borne genes per genome is estimated at 3.3% (range = 0 - 16.5%) and where between 12.3 - 21.5% of the pangenome is plasmid-associated (46). Given our requirements that a high-confidence *Starship* be found in 2 different locations in 2 different strains, our estimates represent a relatively conservative assessment of the mobile fraction of *A. fumigatus'* genome. Mobility underpins prokaryotic genome dynamics, yet current models of fungal pathogen physiology and evolution do not take gene mobility through transposition into account. These data highlight the need to understand the population level impacts of *Starships* across a larger number of plant and animal pathogen species and to integrate these findings into a new predictive framework for fungal pathogenesis. Our *Starship* atlas thus provides practical insights for elucidating *A. fumigatus* pathobiology and establishes a roadmap for deciphering the origins of strain heterogeneity across the fungal tree of life.

Materials and Methods

Long-read genome sequencing, assembly and annotation

For DNA extraction strains were grown at 37 °C for 24 hours in shaking liquid cultures using 1% glucose minimal media (47). Biomass was collected by gravity filtration through Miracloth (Millipore). High molecular weight DNA extraction followed the Fungal CTAB DNA extraction protocol (48). Biomass was ground in liquid nitrogen and incubated at 65 °C in a lysis buffer composed of 650 µl of Buffer A (0.35 M sorbitol, 0.1 M Tris-HCl pH 9, 5 mM EDTA pH 8), 650 µl Buffer B (0.2 M Tris-HCl pH 9, 0.05 M EDTA pH8, 2M NaCl, 2% CTAB), 260 µl of Buffer C (5% Sarkosyl), 0.01% PVP, 0.2 mg proteinase K for 30 minutes. Potassium acetate (280 µl of 5M solution) was added and incubated on ice for 5 minutes. DNA was extracted by phenol:chloroform:isoamyl alcohol extraction followed by a secondary extraction with chloroform:isoamyl alcohol. The supernatant was RNAse treated (2.5µg), and precipitated using sodium-acetate (0.3M) and isopropanol. DNA was finally purified using 70% ethanol, dried and rehydrated in TE (pH 9). To preserve long strands of DNA suitable for long-read sequencing samples were gently mixed by inversion and transferred using large bore pipette tips. DNA quality was assessed using NanoDrop and concentration quantified using Qubit dsDNA BR assay kit. Finally DNA fragment size was assessed on 1% agarose gel, looking for large quantities of DNA to remain in the well after 1 hour of running at 100v. DNA was sent to SeqCoast for Oxford Nanopore Sequencing. We used previously sequenced Illumina short read data for polishing the assemblies with Pilon (10). Two strains not previously sequenced were sequenced with Illumina at SeqCoast.

The long-read genome assemblies were assembled with Canu v2.2 and polished with five rounds of Pilon v1.24 wrapped with AAFTF v0.3.0 (49–51). Summary statistics for each assembly were computed with AAFTF and BUSCO v5.7 using eurotiomycetes_odb10 marker set (52). Genome annotation was performed using Funannotate v1.8.10 which trained gene predictors with PASA using RNA-Seq generated for each strain (Puerner et al., forthcoming), and predicted genes de novo in each assembly and produced a consensus gene prediction set for each strain (53, 54).

Functional predictions

We annotated all predicted genes with Pfam domains, InterPro domains and GO terms using InterproScan v5.61-93.0 (-appl Pfam --iprlookup --goterms)(55). We annotated carbohydrate active enzymes using dbcan v3.0.4 (--tools all) and EggNOG orthogroups and COG categories using eggnog-mapper v2.1.7 (--sensmode very-sensitive --tax_scope Fungi)(56, 57). We annotated biosynthetic gene clusters using antiSMASH v6.0.1 (--taxon fungi --clusterhmmer --tigrfam) and 5S rDNA using infernal v1.1.4 (--rfam -E 0.001)(58, 59). On average, 35.7% of genes within *Starships* have no known PFAM or Interpro domain, GO term, COG category or eggNOG

535 orthology (range = 0-67.6%, std. dev. = 14.8%) but approximately half of all mobilized genes could
536 be assigned to a COG (e.g., 50% of 11,900 genes total, across 459 high-confidence elements).
537 To investigate associations between *Starships* and pathogen phenotypes, we searched for the
538 presence of 798 genes known to be associated with virulence or infection-relevant phenotypes
539 on *Starships* (Table S19, Methods)(10, 11, 33, 35–41).
540

541 SNP analysis

542 We calculated pairwise kinship among strains by analyzing a previously generated VCF of all
543 high-quality, filtered, biallelic single nucleotide polymorphisms (SNPs) from the Lofgren et al. 2022
544 population with plink v1.9 (--distance square ibs flat-missing --double-id --allow-extra-chr; Fig.
545 1E)(60, 61). We calculated Jaccard similarity in *Starship* repertoire among all pairwise
546 combinations of strains using genotyping data for the set of 20 high-confidence *Starships* and the
547 following formula: *Starships* shared between strain 1 and 2 / (*Starships* unique to strain 1 +
548 *Starships* unique to strain 2 + *Starships* shared between strain 1 and 2).
549

550 Pangenome construction

551 We constructed a pangenome for the 13 reference-quality strains (our 12 long-read assemblies
552 plus Af293) combined with the 506 publicly available short-read assemblies (10, 11) using
553 Orthofinder v2.5.4 (--only-groups -S diamond), and then extracted all orthogroups with a
554 sequence from at least 1 of the 13 reference-quality strains (Fig. 2B)(62). Core orthologs are
555 defined as being present in all 13 strains; accessory orthologs are present in between 2-12 strains;
556 singleton orthologs are present in 1 isolate. We included all 519 strains in the Orthofinder analysis
557 to ensure that ortholog groups are consistent across different population subsets.

558 BLAST analysis

559 We determined the genomic locations of all best reciprocal blast hits to cargo genes carried by
560 the 20 high-confidence *Starships* using BLASTp v2.13 (Fig. 1C)(63). For each cargo gene, we
561 retrieved the highest scoring hit with ≥90% identity and ≥90% query coverage in each of the 13
562 reference-quality assemblies, and determined whether that hit was found in the same *Starship*, a
563 different *Starship*, or not in a *Starship* at all. If no hit was retrieved, that gene was marked as
564 missing.
565

566 Structural variant analysis

567 To evaluate the validity of putative structural rearrangements within our CEA10-lr assembly, we
568 took two approaches. First, we aligned the raw nanopore reads to the CEA10-lr assembly using
569 Mummer 4.0 with the nucmer command and default parameters (64). Alignments were then
570 visually inspected using the Genome Ribbon software (65). This method confirmed that individual
571 reads spanned the junctions of putative rearrangements and confirmed that no misassemblies
572 were present. However, this approach suggested that two alternative chromosomal topologies
573 existed: one that matches the chromosomal assemblage of the reference strain Af293, and
574 another that exhibits a reciprocal translocation between Chromosomes 1 and 6, as reported for

575 other assemblies of CEA10 (30). To resolve this paradox, we then mapped the raw nanopore
576 reads to both the CEA10-lr assembly, as well as a publicly available high-quality assembly
577 (BioSample: SAMN28487501) with the minimap2 v2.24 map-ont command (66). These were then
578 visually inspected using the IGV genome browser (67). Multiple reads support both topologies but
579 genome-wide variation at single nucleotide polymorphisms is extremely low, implying that the
580 CEA10-lr isolate sequenced here is a heterogeneous mix of at least two lineages derived from
581 the same CEA10 ancestor. The heterogeneous mixture was subsequently confirmed by
582 morphological phenotyping, so single conidial strains of the mixed culture will have to be obtained
583 and resequenced. Despite the mixed CEA10 culture, annotations of *Starships* in our CEA10
584 assembly are still valid for this strain, as all of its *Starships* are present at the same loci in the
585 other publicly available CEA10 long-read assembly (30).
586
587 To further assess whether *Starships* transpose frequently enough to generate variation among
588 isolates of the same strain, we evaluated three additional strains that we sequenced with
589 Nanopore long-read technology that also had publicly available short-read data generated by
590 other research groups: ATCC46645-lr, S02-30, and TP9. Short-read sequence data were mapped
591 to our long-read assemblies and visually inspected in IGV to determine if any *Starships* were
592 deleted/transposed in the public data. Both S02-30 and TP9 had no apparent variation at *Starship*
593 loci. However, ATCC46645-lr showed polymorphism at three *Starships*, namely *Nebuchadnezzar*
594 h1, *Nebuchadnezzar* h2, and *Gnosis* h1, indicative of active transposition. Both of the
595 *Nebuchadnezzar haplotypes* show clean deletions, indicating that they may have jumped out of
596 the genome of some nuclei within a particular isolate, while the deletions around *Gnosis* (5 in
597 total) were more complex, which may be a signature of genomic instability rather than *Starship*
598 mobilization. Note that neither S02-30 nor TP9 have *Nebuchadnezzar h1*.

599 *Starship* annotation

600 We systematically annotated *Starships* in the 12 newly sequenced long-read genomes plus 507
601 publicly available *A. fumigatus* genomes by applying the starfish workflow v1.0.0 (default settings)
602 in conjunction with metaeuk v14.7e284, Mummer v4.0, CNEFinder and BLASTn (Table S2)(22,
603 63, 64, 68, 69). Briefly, captain tyrosine recombinase genes were *de novo* predicted with starfish's
604 Gene Finder Module and full length elements associated with captains were predicted with
605 starfish's Element Finder Module using pairwise BLAST alignments to find empty and occupied
606 insertion sites. We filtered out all elements that were <15kb in length, which we have found to
607 correspond to indels of captain genes but never to the transposition of a full length *Starship* (22).
608 We then manually examined alignments between each putative *Starship* and its corresponding
609 insertion site, and filtered out all "low confidence" poorly supported alignments indicative of a false
610 positive insertion. Poor alignments were observed most often when *Starships* inserted into smaller
611 transposons; when an inversion breakpoint occurred within a putative *Starship*, resulting in an
612 over-estimation of *Starship* length; or when the insertion site was on a very small contig resulting
613 in small flanking region alignments.
614

We verified and supplemented these automated *Starship* predictions with manual annotations of the 3 reference strains Af293, CEA10 and A1163, along with a set of 86 additional elements (Table S7, Table S3). Insertion sites and DRs were manually verified by generating alignments of the elements plus the 50kb flanks to a corresponding putative insertion site, as determined by starfish (Table S4). MAFFT v7.4 was used to generate alignments with default parameters (70). The DRs were visually assessed and the insertion regions were examined for the presence of TEs, which could lead to erroneous target site, DR and TIR determinations. For *Starships* in the three reference genomes, the set of reference *Starships* was used as a query with BLAST to verify their start and end coordinates, and insertion sites. Additionally, the starfish output was examined for additional elements which did not meet our initial strict cutoffs to attempt to capture the entire *Starship* repertoire of these strains.

Starship classification

Starships were then grouped into *naves* (singular: *navis*, latin for “ship”) by clustering captain sequences in ortholog groups using mmseqs easy-cluster (--min-seq-id 0.5 -c 0.25 --alignment-mode 3 --cov-mode 0 --cluster-reassign)(71). *Starship* sequences were then grouped into *haplotypes* using the MCL clustering algorithm in conjunction with sourmash sketch (-p k=510 scaled=100 noabund) that calculated pairwise k-mer similarities over the entire sequence length, as implemented in the commands “starfish sim” and “starfish group”(72, 73). These automated and systematic predictions yielded a core set of 787 individual elements grouped into 54 distinct *navis-haplotype* combinations.

We identified segregating insertion sites associated with the 787 elements across all 519 strains using the command starfish dereplicate (--restrict --flanking 6 --mismatching 2) in conjunction with the Orthofinder orthogroups file (see above) that was filtered to contain groups absent in at most 517 strains and present in at most 8 copies per isolate. Starfish dereplicate enables the identification of independently segregating insertions by grouping *Starships* and their insertion sites into homologous genomic regions using conserved combinations of orthogroups between individuals. We genotyped the presence of each *navis-haplotype* combination within each region for each isolate. If an isolate did not have any *Starships* within a given region, it was assigned either an “empty” genotype (if the set of orthogroups that define the upstream region flank were adjacent to the set of orthogroups that define the downstream region flank) or a “fragmented” genotype (if additional orthogroups were in between the upstream and downstream region flanks).

To define high-confidence *Starships,* we used a threshold that required observing the same *navis-haplotype* in ≥2 independent segregating sites in different strains, which ensured an accurate and precise determination of each element’s boundaries and the sequences therein (22). We found 18 *navis-haplotype* combinations meeting these criteria. We considered 2 additional *navis-haplotype* combinations of interest (*Osiris h4* and *Lamia h4*) to be high-confidence after manually annotating their boundaries and finding evidence for flanking direct repeats that are signatures of *Starship* boundaries. Each high-confidence *Starship* is represented by a type element (similar in

concept to a type specimen for defining a biological species) that constitutes the longest element assigned to that *navis-haplotype*. The type elements of the reference *Starships* range in size from 16,357-443,866 bp, and collectively represent 459 individual elements with an average length of 71,515 bp (Table S4, Table S5). The remaining 328 elements have an average length of 50,249 bp and are represented by 34 "medium-confidence" *navis-haplotypes* that were only observed at a single segregating site. We predicted the target sites of each high-confidence *Starship* by manually aligning multiple empty insertion site sequences to each other and identifying columns within 25bp of the core motif that had conserved nucleotides in at least 80% of sequences (Table 1).

Each *navis* associated with a high-confidence element was assigned a charismatic name (e.g., *Tardis*) and a new, sequentially named *haplotype* (*h1, h2*, etc.) to distinguish it from the automatically generated *haplotype* codes (*var14, var09*, etc.). All other 34 *navis-haplotype* combinations with evidence of only a single segregating site were classified as "medium-confidence". Medium-confidence *Starships* with captains not belonging to the high-confidence naves kept their automatically generated *navis* and *haplotype* codes (e.g., *navis01-var09*), while medium-confidence *Starships* with captains belonging to the high-confidence naves were assigned to that *navis* but kept their automatically assigned *haplotype* (e.g., *Osiris-var04*).

Starship genotyping

Starfish requires a well-resolved empty insertion site in order to annotate the boundaries of a contiguous *Starship* element. It will therefore not annotate any element whose corresponding empty site is missing, or any element that is partially assembled or not assembled at all. We therefore supplemented starfish output using a combination of BLASTn and short-read mapping to the library of reference *Starship* sequences to decrease the false negative rate for presence/absence genotyping. First, we recovered full length and partial elements using the set of high-confidence *Starships* as input to the command "starfish extend". This command uses BLASTn to align full length *Starship* elements to the sequence downstream from all tyrosine recombinase genes not affiliated with a full length *Starship* element.

Separately, we downloaded publicly available paired-end short read Illumina sequencing data for all 466 strains for which these data were available (10, 11) and mapped them to the library of 20 high-confidence *Starships* using the command "starfish coverage" and the aligner strobealign (74), ensuring that we would correctly genotype the presence/absence of a *Starship* even if it is not present in an assembly or partially assembled. By mapping short reads directly to a library of known *Starship* elements, we overcame some of the drawbacks associated with working with short-read assemblies. We considered any *Starship* with a minimum of 5 mapped reads at each position across 95% of its entire length to be present. We considered a *Starship* to be present in a given individual if either the main starfish workflow, BLASTn extension, or short-read mapping identified it as present; otherwise we considered it absent.

Meta-Analysis of *A. fumigatus* RNAseq Datasets

We collected transcriptomes from a total of 434 publicly available paired-end libraries of *A. fumigatus* strains Af293, A1163, and CEA10/CEA17 (Table S22). We surveyed the available metadata from each BioProject and binned the samples into general categories based on the type of treatment applied in each study. The categories that were the most well represented across multiple BioProjects include exposure to antifungals ("antifungal"), *in vivo* and *in vitro* infection experiments ("infection"), or supplemented growth media with specific nutrients ("nutrient"; Table S22, Table S23).

RNAseq reads were retrieved from the NCBI SRA using fasterq-dump from the SRA toolkit (last accessed June 26, 2024: https://github.com/ncbi/sra-tools) and were trimmed for quality and sequencing artifacts using TrimGalore (last accessed June 26, 2024: https://github.com/FelixKrueger/TrimGalore). Transcripts were quantified using salmon quant, which employs a reference-free pseudo-mapping based approach to quantify transcripts. In addition to the pseudo-BAM files created by salmon v1.10.1 (75), we created separate sets of BAM files to assess transcript coverage across Starship genes by mapping to the transcriptome with STAR v2.7.11a (76). We assessed transcript coverage across the core of Starship genes using the bamsignals R package (last accessed June 26, 2024: https://github.com/lamortenera/bamsignals. Genes with a median core-transcript coverage less than 1 were considered to have insufficient evidence of being expressed. We applied an abundance filter on transcript abundances, removing genes represented by fewer than 10 transcripts in at least 3 samples. We corrected for batch effects between BioProjects using CombatSeq (77) and excluded BioProjects with inconsistent/unclear metadata, or those that remained as outliers in the PCA after batch-correction. DESeq2 v1.45.0 (78) was used to perform a series of differential expression (DE) tests using the corrected transcript counts. These DE tests were performed individually for each BioProject, with each test comparing the differences in expression between all control and treatment samples. Differences in treatment level were included as a cofactor in the model, where applicable.

To summarize expression patterns across BioProjects which tested similar conditions, we employed a random effects model (REM) using the R package metaVolcanoR (79). This method identifies differentially expressed genes (DEGs) as genes that are significantly differentially expressed in all studies and accounts for the variation in the expression for each DEG observed in multiple BioProjects. The output from the meta-volcano REM are summarized log2 fold-change, a summary p-value, and confidence intervals for each DEG. In addition, a measure of "sign-consistency" is used to evaluate the consistency of DEG, which is expressed as a count of the number of BioProjects where a DEG was observed with the same directional change (+/-), centered around 0.

We performed weighted gene co-expression network analysis (WGCNA) using PyWGCNA v1.72-5 (80) to construct gene co-expression networks for *A. fumigatus* reference strain Af293. A single

network was constructed using the batch-corrected TPM values from the collection of samples belonging to “antifungal”, “infection”, and “nutrient” treatment categories. PyWGCNA automatically estimates an appropriate soft-power threshold based on the lowest power for fitting scale-free topology and identifies modules of co-expressed genes through hierarchical clustering of the network and performing a dynamic tree cut based on 99% of the dendrogram height range. The topological overlap matrix is then computed and a correlation matrix is constructed to produce the final network. Distributions of TOM scores were generated based on edges in the network between genes within *Starships*, between *Starships*, between *Starships* and the rest of the genome, and between non-*Starship* genes in the genome. We compared these distributions using a Wilcoxon test and anova, adjusting p-values with the Holm–Bonferroni method. Modules in the network that were significantly associated with samples from a specific treatment category were identified using the pairwise distances between observations using pdist from the python module scipy (81). To highlight the genes within each module that have the most strongly correlated expression profiles, sub-networks were constructed using only the edges 10 highest TOM scores made between either *Starship* captain genes or any non-*Starship* gene that was present in the module. The collections of genes within these modules were also tested for functional enrichment using gprofiler2 (82).

Statistical tests and data visualization

Enrichment tests were performed with either the Binomial test (binom.test; alternative = “greater”) or the Fisher’s exact test (fisher.test; alternative = “greater”) implemented in base R, using the p.adjust function (method = Benjamini-Hochberg) to correct for multiple comparisons. We visualized all element alignments and insertion site distributions using circos and gggenomes (83, 84). All other figures were generated in R using ggplot2 (85).

Data availability

Sequencing reads and genomic assemblies for the 12 isolates sequenced with Oxford Nanopore have been deposited at NCBI (Table S1). We downloaded the Af293 reference assembly (accession: GCF_000002655.1) as well as 252 assemblies and annotations from Barber et al. 2021 from NCBI, and 256 assemblies and annotations from Lofgren et al. 2022 from the associated Zenodo repository (61). We subsequently reformatted the contig and gene identifiers to include genome codes generated as part of this study (Table S2). All scripts (including raw data for figure generation), genomic data used for this study, and results from differential expression tests are available through the following Figshare repository DOI: 10.6084/m9.figshare.26049703.

Acknowledgements

We thank Jacob Steenwyk for insightful feedback on an earlier draft of this manuscript.

Funding

This work was supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation to EGT; by funding from start-up funds from the Department of Plant Pathology at the University of Wisconsin-Madison and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement to EGT (grant number 890630); by the Swedish Research Council Formas (grant number 2019-01227) and the Swedish Research Council VR (grant number 2021-04290) to AAV. JES, CP, and RAC are supported by funding from the National Institutes of Health (grant no. 2R01AI130128), to RAC. JES is a CIFAR Fellow in the program Fungal Kingdom: Threats and Opportunities. Analyses and data management at UC Riverside was performed on the High-Performance Computing Center at the University of California, Riverside, in the Institute of Integrative Genome Biology, supported by grants from NSF (DBI-2215705) and NIH (S10-OD016290).

References

1. 2022. WHO fungal priority pathogens list to guide research, development and public health action. World Health Organization, Geneva.
2. Denning DW. 2024. Global incidence and mortality of severe fungal disease. Lancet Infect Dis S1473-3099(23)00692–8.
3. Latgé J-P, Chamilos G. 2019. Aspergillus fumigatus and Aspergillosis in 2019. Clin Microbiol Rev 33:10.1128/cmr.00140-18.
4. Lin S-J, Schranz J, Teutsch SM. 2001. Aspergillosis Case-Fatality Rate: Systematic Review of the Literature. Clin Infect Dis 32:358–366.
5. Kowalski CH, Beattie SR, Fuller KK, McGurk EA, Tang Y-W, Hohl TM, Obar JJ, Cramer RA. 2016. Heterogeneity among Isolates Reveals that Fitness in Low Oxygen Correlates with Aspergillus fumigatus Virulence. mBio 7:e01515-16.
6. Colabardini AC, Wang F, Dong Z, Pardeshi L, Rocha MC, Costa JH, dos Reis TF, Brown A, Jaber QZ, Fridman M, Fill T, Rokas A, Malavazi I, Wong KH, Goldman GH. 2022. Heterogeneity in the transcriptional response of the human pathogen Aspergillus fumigatus to the antifungal agent caspofungin. Genetics 220:iyab183.
7. Horta MAC, Steenwyk JL, Mead ME, dos Santos LHB, Zhao S, Gibbons JG, Marcet-Houben M, Gabaldón T, Rokas A, Goldman GH. 2022. Examination of Genome-Wide Ortholog Variation in Clinical and Environmental Isolates of the Fungal Pathogen Aspergillus fumigatus. mBio 13:e01519-22.
8. Keller NP. 2017. Heterogeneity Confounds Establishment of “a” Model Microbial Strain. mBio 8:10.1128/mbio.00135-17.
9. Steenwyk JL, Rokas A, Goldman GH. 2023. Know the enemy and know yourself: Addressing cryptic fungal pathogens of humans and beyond. PLoS Pathog 19:e1011704.
10. Lofgren LA, Ross BS, Cramer RA, Stajich JE. 2022. 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.
11. Barber AE, Sae-Ong T, Kang K, Seelbinder B, Li J, Walther G, Panagiotou G, Kurzai O. 2021. Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection. 12. Nat Microbiol 6:1526–1536.
12. Rhodes J, Abdolrasouli A, Dunne K, Sewell TR, Zhang Y, Ballard E, Brackin AP, van Rhijn N, Chown H, Tsitsopoulou A, Posso RB, Chotirmall SH, McElvaney NG, Murphy PG, Talento AF, Renwick J, Dyer PS, Szekely A, Bowyer P, Bromley MJ, Johnson EM, Lewis White P, Warris A, Barton RC, Schelenz S, Rogers TR, Armstrong-James D, Fisher MC. 2022. Population genomics confirms acquisition of drug-resistant Aspergillus fumigatus infection by humans from the environment. Nat Microbiol 7:663–674.
13. Hall JPJ, Harrison E, Baltrus DA. 2022. Introduction: the secret lives of microbial mobile genetic elements. Philos Trans R Soc B Biol Sci 377:20200460.
14. Aubert D, Naas T, Héritier C, Poirel L, Nordmann P. 2006. Functional characterization of IS1999, an IS4 family element involved in mobilization and expression of beta-lactam resistance genes. J Bacteriol 188:6506–6514.

15. Krishnan P, Meile L, Plissonneau C, Ma X, Hartmann FE, Croll D, McDonald BA, Sánchez-Vallet A. 2018. Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. BMC Biol 16:78.
16. Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. 2013. Bacterial Adaptation through Loss of Function. PLOS Genet 9:e1003617.
17. Weisberg AJ, Chang JH. 2023. Mobile Genetic Element Flexibility as an Underlying Principle to Bacterial Evolution. Annu Rev Microbiol 77:null.
18. Baquero F. 2004. From pieces to patterns: evolutionary engineering in bacterial pathogens. Nat Rev Microbiol 2:510–518.
19. Benler S, Faure G, Altae-Tran H, Shmakov S, Zhang F, Koonin E. 2021. Cargo Genes of Tn7-Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. mBio 12:e02938-21.
20. Gluck-Thaler E, Ralston T, Konkel Z, Ocampos CG, Ganeshan VD, Dorrance AE, Niblack TL, Wood CW, Slot JC, Lopez-Nicora HD, Vogan AA. 2022. Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. Mol Biol Evol 39:msac109.
21. Urquhart AS, Vogan AA, Gardiner DM, Idnurm A. 2023. Starships are active eukaryotic transposable elements mobilized by a new family of tyrosine recombinases. Proc Natl Acad Sci U S A 120:e2214521120.
22. Gluck-Thaler E, Vogan AA. 2024. Systematic identification of cargo-mobilizing genetic elements reveals new dimensions of eukaryotic diversity. Nucleic Acids Res gkae327.
23. Gourlie R, McDonald M, Hafez M, Ortega-Polo R, Low KE, Abbott DW, Strelkov SE, Daayf F, Aboukhaddour R. 2022. The pangenome of the wheat pathogen Pyrenophora tritici-repentis reveals novel transposons associated with necrotrophic effectors ToxA and ToxB. BMC Biol 20:239.
24. Urquhart AS, Chong NF, Yang Y, Idnurm A. 2022. A large transposable element mediates metal resistance in the fungus Paecilomyces variotii. Curr Biol 32:937-950.e5.
25. Urquhart AS, Gluck-Thaler E, Vogan AA. 2023. Gene acquisition by giant transposons primes eukaryotes for rapid evolution via horizontal gene transfer. bioRxiv https://doi.org/10.1101/2023.11.22.568313.
26. Bucknell A, Wilson HM, Santos KCG do, Simpfendorfer S, Milgate A, Germain H, Solomon PS, Bentham A, McDonald MC. 2024. Sanctuary: A Starship transposon facilitating the movement of the virulence factor ToxA in fungal wheat pathogens. bioRxiv https://doi.org/10.1101/2024.03.04.583430.
27. Pain A, Woodward J, Quail MA, Anderson MJ, Clark R, Collins M, Fosker N, Fraser A, Harris D, Larke N, Murphy L, Humphray S, O'Neil S, Pertea M, Price C, Rabbinowitsch E, Rajandream M-A, Salzberg S, Saunders D, Seeger K, Sharp S, Warren T, Denning DW, Barrell B, Hall N. 2004. Insight into the genome of Aspergillus fumigatus: analysis of a 922 kb region encompassing the nitrate assimilation gene cluster. Fungal Genet Biol FG B 41:443–453.
28. Monod M, Paris S, Sarfati J, Jaton-Ogay K, Ave P, Latgé JP. 1993. Virulence of alkaline protease-deficient mutants of Aspergillus fumigatus. FEMS Microbiol Lett 106:39–46.
29. Kowalski CH, Morelli KA, Stajich JE, Nadell CD, Cramer RA. 2021. A Heterogeneously

867 Expressed Gene Family Modulates the Biofilm Architecture and Hypoxic Growth of
868 Aspergillus fumigatus. mBio 12:e03579-20.
869 30. Bowyer P, Currin A, Delneri D, Fraczek MG. 2022. Telomere-to-telomere genome
870 sequence of the model mould pathogen Aspergillus fumigatus. Nat Commun 13:5394.
871 31. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
872 Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification
873 system for eukaryotic transposable elements. Nat Rev Genet 8:973–982.
874 32. Lambert SA, Yang AWH, Sasse A, Cowley G, Albu M, Caddick MX, Morris QD, Weirauch
875 MT, Hughes TR. 2019. Similarity regression predicts evolution of transcription factor
876 sequence specificity. Nat Genet 51:981–989.
877 33. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, Rodrigues F,
878 Goldman GH, Rokas A. 2017. Drivers of genetic diversity in secondary metabolic gene
879 clusters within a fungal species. PLoS Biol 15:e2003583.
880 34. Abdolrasouli A, Rhodes J, Beale MA, Hagen F, Rogers TR, Chowdhary A, Meis JF,
881 Armstrong-James D, Fisher MC. 2015. Genomic Context of Azole Resistance Mutations in
882 Aspergillus fumigatus Determined Using Whole-Genome Sequencing. mBio 6:e00536-15.
883 35. Ma K, Zhang P, Tao Q, Keller NP, Yang Y, Yin W-B, Liu H. 2019. Characterization and
884 Biosynthesis of a Rare Fungal Hopane-Type Triterpenoid Glycoside Involved in the
885 Antistress Property of Aspergillus fumigatus. Org Lett 21:3252–3256.
886 36. Brown NA, Goldman GH. 2016. The contribution of Aspergillus fumigatus stress responses
887 to virulence and antifungal resistance. J Microbiol 54:243–253.
888 37. Steenwyk JL, Mead ME, de Castro PA, Valero C, Damasio A, dos Santos RAC, Labella
889 AL, Li Y, Knowles SL, Raja HA, Oberlies NH, Zhou X, Cornely OA, Fuchs F, Koehler P,
890 Goldman GH, Rokas A. 2021. Genomic and Phenotypic Analysis of COVID-19-Associated
891 Pulmonary Aspergillosis Isolates of Aspergillus fumigatus. Microbiol Spectr
892 9:10.1128/spectrum.00010-21.
893 38. Raffa N, Osherov N, Keller NP. 2019. Copper Utilization, Regulation, and Acquisition by
894 Aspergillus fumigatus. 8. Int J Mol Sci 20:1980.
895 39. Bignell E, Cairns TC, Throckmorton K, Nierman WC, Keller NP. 2016. Secondary
896 metabolite arsenal of an opportunistic pathogenic fungus. Philos Trans R Soc B Biol Sci
897 371:20160023.
898 40. Abad A, Fernández-Molina JV, Bikandi J, Ramírez A, Margareto J, Sendino J, Hernando
899 FL, Pontón J, Garaizar J, Rementeria A. 2010. What makes Aspergillus fumigatus a
900 successful pathogen? Genes and molecules involved in invasive aspergillosis. Rev
901 Iberoam Micol 27:155–182.
902 41. Auxier B, Zhang J, Marquez FR, Senden K, Heuvel J van den, Aanen DK, Snelders E,
903 Debets AJM. 2022. Identification of heterokaryon incompatibility genes in Aspergillus
904 fumigatus highlights a narrow footprint of ancient balancing selection. bioRxiv
905 https://doi.org/10.1101/2022.11.25.517501.
906 42. Stroe MC, Netzker T, Scherlach K, Krüger T, Hertweck C, Valiante V, Brakhage AA. 2020.
907 Targeted induction of a silent fungal gene cluster encoding the bacteria-specific
908 germination inhibitor fumigermin. eLife 9:e52541.

43. Kowalski CH, Kerkaert JD, Liu K-W, Bond MC, Hartmann R, Nadell CD, Stajich JE, Cramer RA. 2019. Fungal biofilm morphology impacts hypoxia fitness and disease progression. Nat Microbiol 4:2430–2441.
44. O'Donnell S, Rezende G, Vernadet J-P, Snirc A, Ropars J. 2024. Harbouring Starships: The accumulation of large Horizontal Gene Transfers in Domesticated and Pathogenic Fungi. bioRxiv https://doi.org/10.1101/2024.07.03.601904.
45. Rokas A, Mead ME, Steenwyk JL, Oberlies NH, Goldman GH. 2020. Evolving moldy murderers: Aspergillus section Fumigati as a model for studying the repeated evolution of fungal pathogenicity. PLOS Pathog 16:e1008315.
46. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, Barker L, Rodger G, Bowes MJ, Hubbard ATM, Pickford H, Swann J, Gilson D, Smith RP, Hoosdally SJ, Sebra R, Brett H, Peto TEA, Bailey MJ, Crook DW, Read DS, Anjum MF, Walker AS, Stoesser N, REHAB CONSORTIUM. 2021. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. Sci Adv 7:eabe3868.
47. Shimizu K, Keller NP. 2001. Genetic involvement of a cAMP-dependent protein kinase in a G protein signaling pathway regulating morphological and chemical transitions in Aspergillus nidulans. Genetics 157:591–600.
48. Carter-House D, Stajich J, Unruh S, Kurbessoian T. 2020. Fungal CTAB DNA Extraction.
49. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736.
50. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9:e112963.
51. Palmer JM, Stajich JE. 2022. Automatic assembly for the fungi (AAFTF): genome assembly pipeline (v0.4.1). Zenodo.
52. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol 38:4647–4654.
53. Palmer JM, Stajich J. 2020. Funannotate v1.8.1: Eukaryotic genome annotation (v1.8.1). Zenodo.
54. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666.
55. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
56. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the

951 Metagenomic Scale. Mol Biol Evol https://doi.org/10.1093/molbev/msab293.
952 57. Zheng J, Ge Q, Yan Y, Zhang X, Huang L, Yin Y. 2023. dbCAN3: automated
953 carbohydrate-active enzyme and substrate annotation. Nucleic Acids Res 51:W115–W121.
954 58. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber
955 T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic
956 Acids Res 49:W29–W35.
957 59. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches.
958 Bioinformatics 29:2933–2935.
959 60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P,
960 de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome
961 Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575.
962 61. Lofgren LA, Ross B, Cramer Jr RA, Stajich JE. 2021. Aspergillus fumigatus pangenome
963 dataset (v2021-12-12). Zenodo.
964 62. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
965 comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157.
966 63. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
967 BLAST+: architecture and applications. BMC Bioinformatics 10:421.
968 64. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A
969 fast and versatile genome alignment system. PLOS Comput Biol 14:e1005944.
970 65. Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. 2020. Ribbon: intuitive visualization for
971 complex genomic variation. Bioinformatics 37:413–415.
972 66. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
973 34:3094–3100.
974 67. Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV):
975 high-performance genomics data visualization and exploration. Brief Bioinform 14:178–
976 192.
977 68. Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene
978 discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 8:48.
979 69. Ayad LAK, Pissis SP, Polychronopoulos D. 2018. CNEFinder: finding conserved non-
980 coding elements in genomes. Bioinforma Oxf Engl 34:i743–i747.
981 70. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
982 Improvements in Performance and Usability. Mol Biol Evol 30:772–780.
983 71. Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep
984 clustering and searching of large protein sequence sets. Bioinformatics 32:1323–1330.
985 72. Brown CT, Irber L. 2016. sourmash: a library for MinHash sketching of DNA. J Open
986 Source Softw 1:27.
987 73. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale
988 detection of protein families. Nucleic Acids Res 30:1575–1584.
989 74. Sahlin K. 2022. Strobealign: flexible seed size enables ultra-fast and accurate read
990 alignment. Genome Biol 23:260.
991 75. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and
992 bias-aware quantification of transcript expression. Nat Methods 14:417–419.

76. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.
77. Zhang Y, Parmigiani G, Johnson WE. 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma 2:lqaa078.
78. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
79. MetaVolcanoR. Bioconductor. http://bioconductor.org/packages/MetaVolcanoR/. Retrieved 26 June 2024.
80. Rezaie N, Reese F, Mortazavi A. 2023. PyWGCNA: a Python package for weighted gene co-expression network analysis. Bioinformatics 39:btad415.
81. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272.
82. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. 2020. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Research 9:ELIXIR-709.
83. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res https://doi.org/10.1101/gr.092759.109.
84. Hackl T, Ankenbrand M, van Adrichem B. 2023. gggenomes: A Grammar of Graphics for Comparative Genomics (R package version 0.9.9.9000).
85. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis, 3rd ed. Springer International Publishing, Cham. http://link.springer.com/10.1007/978-3-319-24277-4. Retrieved 10 October 2023.
86. Filho AP da C, Brancini GTP, de Castro PA, Valero C, Ferreira Filho JA, Silva LP, Rocha MC, Malavazi I, Pontes JG de M, Fill T, Silva RN, Almeida F, Steenwyk JL, Rokas A, dos Reis TF, Ries LNA, Goldman GH. 2020. Aspergillus fumigatus G-Protein Coupled Receptors GprM and GprJ Are Important for the Regulation of the Cell Wall Integrity Pathway, Secondary Metabolite Production, and Virulence. mBio 11:10.1128/mbio.02458-20.
87. McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 3:861.

Supplemental information and results

Starship frequencies

The population-level frequency of high-confidence *Starships* ranges from 1-60%, with 13 of the high-confidence *Starships* present in <10% of strains and no *Starship* present in more than 60% of strains (Table S6). Each high-confidence *Starship* is present across multiple individuals; however, any given *Starship* element is only present at most once in a genome, with only 3 exceptions that likely involve recently degraded copies, suggesting some mechanism for controlling their copy number (e.g., repeat-induced point mutation (21)). Individual elements typically segregate at low frequencies, with each segregating site containing on average an inserted element in just 0.4% of strains across the 20 high-confidence *Starships* (range = 0.2-52.3%, std. dev. = 6.5%, n = 154; Fig. 3B). However, multiple insertion sites often exist for a given element such that each high-confidence *Starship* is present on average in 4.35 different regions (range = 1-15, std. dev. = 3.92).

Differential expression of *Starship* captains and cargo

Patterns of transcript coverage, transcript abundance, and differential expression of *Starship*-associated genes were generally heterogeneous, with various strain-, treatment- and strain by treatment interactions. The transcriptional profile for many *Starships* is patchy, and not all *Starship*-mobilized genes are expressed in response to experimental conditions. Only certain cargo genes have the minimum required evidence for constitutive expression: a median transcript coverage of 1 across the gene body. A greater number of genes present in *Starships* from *A. fumigatus* Af293 (Fig. S13A) have core gene transcript coverage greater than this threshold, compared to genes in A1163 (Fig. S13C) and CEA10 (Fig. S13B) *Starships*. Interestingly, specific studies of Af293 or CEA10 which tested exposure to heatshock (PRJDB6203), hypoxia (PRJNA144647), *in vitro* infections (PRJEB1583, PRJNA399754, PRJNA399754). and various gene knock-out experiments (PRJNA390719, PRJNA396210, and PRJNA601094) have consistently reported expression of all genes across the *Starship* element.

Captain genes of multiple *Starships* from Af293 and CEA10 had decreased transcript abundance associated with “infection” studies. Samples of a study performing an experimental infection using a murine model (PRJNA421149) had near-zero transcript abundances of captain genes from Af293 *Starships Janus h2*, *Osiris h3*, and *Gnosis h1*. However, this was not found within samples of a different study of an Af293 infection of human pneumocyte cell lines (PRJNA399754). Near-zero abundances were also found for a different set of captain genes, within CEA10 *Starships Tardis h1, Logos h2,* and *Gnosis h3*, from an *in vitro* infection study using human cell lines (Fig. S12), suggesting that the transcriptional response of *Starship* captains within the infection environment differs between strains.

1066 Specific treatments applied within a single study can have a contrasting influence on the prevailing
1067 expression patterns of captain genes. Expression of the *Tardis h1* captain gene in Af293 is
1068 generally decreased across nutrient supplementation studies compared to controls (Fig. 5B). Yet,
1069 supplementation of glucose and xylose has a positive and negative influence on captain
1070 expression, respectively, compared to controls with nutrient media (86). Similarly, expression of
1071 CEA10 *Logos h1* and *h2* captain genes both decreased and increased, with supplementation of
1072 acetate (PRJNA668271), respectively, compared to nutrient media controls containing dilute
1073 acetate (0.1%) (Fig. 5D). Together, these gene expression datasets demonstrate the context-
1074 dependent expression of captain genes, which we speculate would lead to variation in *Starship*
1075 transposition rates across different environments and strain backgrounds.
1076
1077 Similar to captain genes, the expression patterns of cargo genes within *Starships* are
1078 heterogeneous. The transcriptional profile of *Starships* is punctuated with expression of cargo
1079 genes (Fig. SH), which can have substantial transcript coverage even in the absence of captain
1080 expression, suggesting that the transcriptional regulation of these cargo genes are decoupled
1081 from transposition. Of particular note are the genes that form biosynthetic gene clusters (BGC)
1082 which reside within *Starships*. In Af293, 47 BGC genes across six Starships (*Gnosis h1, Lamia*
1083 *h4, Nebuchadnezzar h1, Nebuchadnezzar h2, Osiris h4, Tardis h1*) are DE in one or more of the
1084 treatment categories investigated in this study. At least one gene from the BGCs across these six
1085 *Starships* are DE in response to caspofungin exposure, from a specific Af293 study
1086 (PRJNA472460). Furthermore, multiple genes associated with a BGC present in Osiris h4
1087 (AFUA_3G02580, AFUA_3G02600, AFUA_3G02620, AFUA_3G02630, AFUA_3G02640) have
1088 significantly increased expression compared to controls. Interestingly, generally fewer genes in
1089 BGCs are DE in Af293 “antifungal” or “nutrient” studies. Relatively fewer Starship BGC genes are
1090 DE in other strains. Notably, a collection of cargo genes from Fumihopaside A and terpene
1091 biosynthesis (CEA10-lr_006118) BGCs in CEA10 *Gnosis h2* have decreased expression during
1092 an *in vitro* infection of a human cell line. Cargo genes within *Osiris h4* that increased expression
1093 under both antifungal exposure and nutrient manipulation included multiple genes from the
1094 putative NR-PKS BGC (Fig. S14).
1095
1096 As a human pathogen, we were also interested in *A. fumigatus* genes that have specific relevance
1097 to virulence, such as genes within *hrmA*-associated cluster (HAC). Two HAC genes present in
1098 *Nebuchadnezzar h1*, *cgnA* and *bafA*, have near-zero transcript abundances within samples from
1099 an Af293 study of an experimental infection of a murine model (PRJNA421149). However, genes
1100 within the HAC cluster were only DE within certain nutrient and antifungal studies. Specifically,
1101 *cgnA* (AFUA_5G14910) and *hrmA* (AFUA_5G14900) both have increased expression within
1102 Af293 samples supplemented with calcium, and *bafA* has decreased expression in CEA10
1103 samples under caspofungin exposure (Fig. 5A).
1104
1105 Multiple DE cargo genes serve putative functions that are relevant for the metabolic or enzymatic
1106 activity of the cell. Collections of cargo genes within the Af293 *Starships Lamia h4* and *Osiris h4*
1107 have consistently increased expression from studies employing nutrient manipulation treatments

1108 compared to controls (Fig. S14). This includes several genes in *Lamia h4* with a putative role in
1109 enzymatic secretion (e.g., glycosyl hydrolases, phosphotransferases, and dehydrogenases).
1110 Genes with domains relevant to dehydrogenases (NAD binding domains) were generally positive
1111 DEGs within antifungal and nutrient studies of Af293 and A1163, respectively.
1112
1113 Genes with a putative role in specific molecular functions were also consistently DE in *Starships*.
1114 We found that genes involved in zinc ion binding activity (IBR) (negative DEG in Af293 antifungal
1115 and positive DEG in CEA10 antifungal), dynamin-like GTPase domain and BTB (both highly
1116 negative DEGS in CEA10 *Logos h1* within infection studies). In Af293 *Tardis h1*, a gene with a
1117 putative role in structural molecular activity (domains for Cytochrome b5-like Heme/Steroid
1118 binding and flocculin; AFUA_8G06250) has increased expression in nutrient supplementation.
1119 Whereas HATPase_c/HisKA response regulator (AFUA_8G06140) and glycoside
1120 hydrolase/lipase (AFUA_8G06350) have decreased expression in nutrient supplementation in the
1121 same *Starship*.
1122
1123 Additional cargo genes of interest may convey resistance to environmental stressors. Two genes
1124 involved in arsenic detoxification, present within *Nebuchadnezzar h1*, were found to be DE in
1125 Af293 and CEA10 studies. Arsenate reductase (AFUA_5G15000) in *Nebuchadnezzar h1* was
1126 found to both have increased and decreased expression within the CEA10 and Af293 copies,
1127 respectively. Three additional genes within this pathway were found to be DE in response across
1128 "antifungal" or "infection" studies (Fig. 5B). Two of these genes in in Af293 *Nebuchadnezzar h1*,
1129 arsenate methyl-transferase (AFUA_5G15020) and arsenite efflux transporter (AFUA_5G15010),
1130 were found to have either increased or decreased expression levels under caspofungin exposure,
1131 respectively. The third, arsenite permease (AFUA_1G16100) has decreased expression in Af293
1132 infection studies.

1133 *Starship* co-expression networks

1134 We performed weighted gene co-expression network analysis (WGCNA) using *PyWGCNA* to
1135 construct gene co-expression networks for each of the three *A. fumigatus* reference strains Af293,
1136 CEA10, and A1163. Overall, we found that genes within or between *Starships* are more strongly
1137 co-expressed together than with non-*Starship* genes, based on pairwise comparisons of the
1138 distribution of scores in the topological overlap matrices (TOM) (corrected p-values < 0.01 for all
1139 comparisons; Fig. S15). This suggests that a tighter transcriptional relationship exists between
1140 genes mobilized by *Starships* compared to co-expression with other genes in the genome.
1141
1142 Genes in the *A. fumigatus* Af293 co-expression network were resolved into 27 modules (Table
1143 S24). The majority of genes found within each module are not contained within a *Starship*, yet
1144 almost all modules contain one or more *Starship* genes, including those previously identified as
1145 DEGs (Table S24). Modules were selected for further analyses based on the correlation of their
1146 gene expression profiles (WGCNA eigengenes) and their association with a treatment category
1147 ("antifungal", "infection", or "nutrient") or specific study (Fig. 5E & Fig. S16).

To investigate functional compartmentalization of WGCNA modules, we performed enrichment tests and identified which GO/KEGG terms are significantly overrepresented within each module. Genes in module “2” are significantly enriched in functions including the production of antimicrobial secondary metabolites (fumagillin), the secretion of *A. fumigatus* mycotoxins (helvolic acid), proteins for heme binding, and synthesis of immunosuppressive compounds (endocrcoin) (Table S25).

To identify which genes are most strongly co-expressed with *Starship* cargo or captain genes, we subsetted the co-expression network to keep only the top 10 edges that were made between any pair of genes or any gene and a *Starship* captain. Two modules within the Af293 co-expression network (“2” and “15”) are significantly more commonly co-expressed within samples from a single infection study of Af293 (PRJEB1583; Fig. 5E). The connections with the highest TOM scores in module “2” include those between an IBR finger domain protein within the *Starship Lamia h4* (AFUA_1G00150) and genes involved in RNA binding pathways (AFUA_6G12070), as well as genes for alpha-amylase (AFUA_2G03230) and anthrone oxygenase (AFUA_4G00225) (Fig. 5D). Genes within the module “15” include those tightly connected to the expression of F-box domain protein within *Lamia h4*. Genes within module “15” are enriched in glycerophospholipid metabolism (Table S25). These co-expression relationships provide insight into how *Starships*-mobilized genes integrate into the existing transcriptional network in *A. fumigatus* strains.

Module “3” is significantly associated with increased expression within a single study of an experimental infection in mice (PRJNA693756) (Fig. 5E & S15). Genes within module “3” are enriched for genes in ribosome biogenesis in eukaryotes, carbon metabolism, and pyruvate metabolism (Table S25). The connections with the highest TOM scores in module “3” include genes co-expressed with *Gnosis h1* captain gene: genes with predicted catalytic activity (AFUA_1G12370), zinc-containing alcohol dehydrogenase (AFUA_2G00970), mitochondrial respiration (AFUA_2G06020), and exonuclease activity/DNA-directed DNA polymerase activity/role in mitochondrial DNA replication and mitochondrion localization (AFUA_5G12640).

Module “7” is significantly associated with two studies testing supplementation with 5,8-dihydroxyoctadecadienoic acid (PRJNA658306) and lipo-chitooligosaccharides (PRJNA642658) in Af293 (Fig. 5E & S15). The focus of both of these studies is to understand how these supplementations impact the regulation of fungal growth and development. The connections with the highest TOM scores in the module “7”, include genes co-expressed with captain genes of *Tardis h1* and *Osiris h3. Osiris h3* captain co-expressed with predicted RNA binding, ribonuclease III activity and role in RNA processing (AFUA_3G03050). The connections made in this module with *Starship* genes present themselves as good candidates for future research to understand which genes are expressed along with transposition.

1187 # Figure legends

1188 Figure 1: At least 20 distinct *Starships* carrying hundreds of protein-coding genes vary in their
1189 presence/absence across *Aspergillus fumigatus* strains. A) Top: a SNP-based maximum
1190 likelihood tree of 220 *A. fumigatus* strains from Lofgren et al. 2022 (visualized as a cladogram for
1191 clarity), color-coded according to phylogenetic clade as defined by Lofgren et al. 2022 (10).
1192 Bottom: A heatmap depicting the presence (gray) and absence (white) of 20 high-confidence
1193 *Starships* in each isolate. B) Schematics and sequence alignments of the type elements from 20
1194 high-confidence *Starships*, where links between schematics represent alignable regions ≥500bp
1195 and ≥80% nucleotide sequence identity and arrows represent predicted coding sequences. C) A
1196 donut chart summarizing the number of distinct types of *Starships* by the quality of their prediction.
1197 D) A Venn diagram indicating the number of shared and unique high-confidence *Starships* in the
1198 reference strains Af293 and CEA10 (Table S7). E) A scatterplot depicting pairwise comparisons
1199 of SNP-based Identity by State (IBS) and Jaccard similarity in high-confidence *Starship*
1200 presence/absence profiles between 150 Clade 1 strains (Table S8).
1201
1202 Figure 2: *Starships* are enriched in accessory genes whose presence/absence and genomic
1203 location vary across *Aspergillus fumigatus* strains. All panels visualize data derived from 13
1204 reference-quality *A. fumigatus* assemblies and the 20 high-confidence *Starships* (n = 459
1205 elements total). A) Box-and-whisker plots summarizing the total percentage of nucleotide
1206 sequence and predicted genes carried by *Starships* per genome. B) Donut charts summarizing
1207 the percentages of gene orthogroups in the core, accessory, singleton, and *Starship*-associated
1208 compartments of the *A. fumigatus* pangenome (Table S10). C) Iceberg plots summarizing the
1209 genomic locations of the single best BLASTp hits (≥90% identity, ≥33% query coverage) to the
1210 cargo genes from the type elements of all 20 high-confidence *Starships* (left) and two individual
1211 *Starships* (right; Table S12). Each column in the iceberg plot represents a cargo gene and is
1212 color-coded according to the genomic location of hits (full dataset in Fig. S5).
1213
1214 Figure 3: *Starships* and their insertion sites are distributed across all major chromosomes in the
1215 *Aspergillus fumigatus* genome. A) A Circos plot summarizing all inserted *Starships* (in red) and
1216 all genomic regions containing either an empty insertion site or a fragmented *Starship* (in black,
1217 along perimeter and labeled with the *r* prefix) in the 8 chromosomes of the *A. fumigatus* reference
1218 strain Af293 (Table S14). All genomic regions contain a *Starship* insertion in some other individual
1219 from the 519 strain population. B) Barcharts summarizing the genotypes of segregating genomic
1220 regions associated with the four most active high-confidence elements in the 519 strain population
1221 (Table S16; full dataset in Fig. S7). If an isolate did not have any *Starships* within a given region,
1222 it was assigned either an "empty" or "fragmented" genotype (Methods). C) A scatterplot
1223 summarizing the relationship between the number of genomic regions containing a given *Starship*
1224 and the total number of copies of that *Starship* in the 519 strain population, where each point
1225 represents one of the 20 high-confidence *Starships.* A line derived from a linear regression model
1226 is superimposed, with shaded 95% confidence intervals drawn in gray. D) Alignments of *Tardis*

1227 *h1* copies +/- 50kb of flanking sequence across 14 genomic regions. Links between schematics
1228 represent alignable regions ≥1000bp and ≥90% nucleotide sequence identity.
1229
1230 Figure 4: *Starships* mobilize adaptive traits and generate allelic diversity among *Aspergillus*
1231 *fumigatus* strains. Schematics and alignments of: A) *Starship Lamia h2* and *h4*, which carry the
1232 biosynthetic gene cluster (BGC) encoding the polyketide secondary metabolite Fumigermin (42),
1233 shown inserted at 3 independent sites. B) *Starship Gnosis h2* carrying the BGC encoding the
1234 terpene secondary metabolite Fumihopaside A (35), shown inserted at 2 independent sites. C) a
1235 large region on Chromosome 3 previously identified as an idiomorphic BGC (33) containing
1236 multiple segregating *Starship* insertions and various combinations of putative BGCs, including a
1237 a non-reducing polyketide synthase (NR-PKS) BGC carried by *Starship Osiris h4*. *Starships* from
1238 the *Osiris navis* specifically insert in 5S rDNA sequence and are predicted to fragment it. D) Eight
1239 *Starships* in the reference strains Af293 and CEA10 that all carry homologs of *biofilm architecture*
1240 *factor A* (*bafA)*(29). *Starship Nebuchadnezzar h1 (Neb. h1)* carries *bafA* as part of the H_AAC
1241 (*hrmA*-associated gene cluster), while *Starship navis10-var35* carries *bafB* as part of H_BAC and
1242 *Starship Osiris h3* carries *bafC* as part of H_CAC (29). Only a portion of *Starship Lamia h4* is
1243 visualized for figure legibility. All data for A-D was collected from the 519 *A. fumigatus* strain
1244 population (Table S5). Links between schematics represent alignable regions ≥5000bp and ≥95%
1245 nucleotide sequence identity and arrows represent predicted coding sequences. Abbreviations:
1246 polyketide synthase (PKS); non-ribosomal peptide synthetase (NRPS); highly reducing (HR).
1247
1248 Figure 5: A) A heatmap of transcript abundances ($\log_{10}$TPM) of *Starship* captain tyrosine
1249 recombinase genes (top) and genes within the *hrmA*-associated cluster (HAC; bottom), collapsed
1250 across treatment replicates, for 11 RNAseq studies from *A. fumigatus* Af293. B) Results from
1251 differential expression (DE) tests for select *Starships*, treatment categories, and strain
1252 combinations. Only genes with valid output from DE tests are shown, and only genes with
1253 significantly DE are labeled with gene codes. Genes with extreme count outliers, as determined
1254 by Cook's Distance in DESeq2, do not have valid p-values and are labeled with an asterisk ("*").
1255 DE based on a single study are shown as $\log_2$ fold-change ($\log_2$FC) in black with standard error
1256 bars, whereas DE genes identified across multiple studies are represented with summarized
1257 $\log_2$FC values from a random effects model (REM) and coloured by "sign-consistency", the
1258 number of studies that reported DE in the positive (+1) or negative (-1) direction, centered around
1259 0. Labels in bold font represent DEGs that are significantly DE in more than one study. C) The
1260 results of a weighted gene co-expression network analysis (WGCNA) constructed from 14
1261 "antifungal", "infection", and "nutrient" RNAseq studies represented using UMAP clustering (87)
1262 based on the co-expression eigengene values for all genes in the *A. fumigatus* Af293 genome.
1263 Non-*Starship* genes present within modules of interest are shown in blue, while all *Starship* genes
1264 are shown in red, with those genes within modules of interest having a black outline. D) Genes
1265 within modules that were significantly associated with samples from specific treatment categories
1266 or studies were used to construct sub-networks which contained the top 10 edges in the network
1267 that were made between any pair of genes or any gene and a *Starship* captain. The connections
1268 between genes (edges) are based on the topological overlap matrix (TOM) for each module, and

have been 0-1 scaled. E) Boxplots of eigengene values from WGCNA, akin to a weighted average expression profile, indicate the extent of co-expression (correlation) of the genes present within each module. Pairwise comparisons of module eigengene values determined if a module was significantly associated with samples from specific treatment categories or studies (Fig. S16).

Supplemental figure legends

Figure S1: Counts of starfish-predicted *Starships* per *Aspergillus fumigatus* strain (n = 519), broken up by assembly project (either the 12 Oxford Nanopore Assemblies generated by study plus the AF293 reference genome, or from Lofgren et al 2022 or Barber et al 2021; Table S2). A) Counts of *Starships* with either 'insert' or 'flank' boundaries, which are derived directly from pairwise genome alignments against a putative insertion site. B) Counts of *Starships* with 'insert' or 'flank' boundaries plus those with 'extend' boundaries, which are derived from aligning genomic sequences to known *Starship* sequences. Insert and flank boundaries are associated with full-length *Starship* elements, while "extend" boundaries may be associated with either full-length *Starship* elements or element fragments.

Figure S2: Nucleotide alignments of representative *Starship* copies, +/- 100kb of flanking sequence, across all genomic regions where that *Starship* is found. Predicted functions of interest are annotated above the corresponding gene. Links between schematics represent alignable regions ≥1000bp and ≥90% nucleotide sequence identity. A) *Tardis*. B) *Gnosis*. C) *Janus*. D) *Osiris*. E) *Lamia*. F) *Logos*. G) *Nebuchadnezzar.*

Figure S3: Pairwise genome alignments of the CEA10 isolate sequenced by this study and the CEA10 isolate from BioSample SAMN28487501 demonstrating *Starship* movement. *Nebuchadnezzar h1* is present on Chromosome 6 in the SAMN28487501 isolate and absent at the corresponding locus in this study's sequenced isolate. Conversely, *Nebuchadnezzar h1* is present on Chromosome 4 in this study's isolate but absent from the corresponding locus in the SAMN28487501 isolate.

1296 Figure S4: Screenshots from the IGV genome browser showing deletions and/or movement of
1297 *Nebuchadnezzar h1* among different isolates of the same named strain sequenced by different
1298 research groups (regions depicted in red denote *Nebuchadnezzar h1*). A. Nanopore long-read
1299 sequences of strain CEA10-lr (generated in this study) mapped to the publicly available assembly
1300 of a different isolate of this same strain (accession: SAMN28487501; sequenced by a different
1301 research group). A zoom in of the genomic location of *Nebuchadnezzar h1* in the public genome
1302 is shown (contig accession: CP097568.1). Only a single long-read supports the presence of the
1303 *Starship* at this locus. All other reads indicate either a deletion (above) or presence elsewhere in
1304 the genome (below). B. Illumina short-reads of strain ATCC46645 sequenced by a different
1305 research group (accession: SRR7418935) mapped to the ATCC46645-lr genome sequenced in
1306 this study. A zoom in of the genomic location of *Nebuchadnezzar h1* in ATCC46645-lr is shown
1307 (contig accession: scaffold_15). Three short-read tracks indicate a deletion of the *Starship*
1308 (above). Note that more short reads are mapped than shown in the image. Track and color
1309 descriptions can be found in the IGV manual.
1310
1311 Figure S5: Donut charts summarizing the percentages of gene orthogroups in the core,
1312 accessory, singleton, and *Starship*-associated compartments of the *Aspergillus fumigatus*
1313 pangenome, derived from the 13 reference-quality *A. fumigatus* strains and the expanded set of
1314 54 high and medium-confidence *Starships* (n = 1818 elements total; Table S11).
1315
1316 Figure S6: Iceberg plots summarizing the genomic locations of the single best BLASTp hits (≥90%
1317 identity, ≥33% query coverage) to the cargo genes from the type elements of the 20 high-
1318 confidence *Starships* in 519 *Aspergillus fumigatus* assemblies (Table S13). Each column
1319 represents a cargo gene. A) Results broken down by *Starship*, with columns arranged according
1320 to gene order within each *Starship*. B) Compiled results across all 20 *Starships*, with columns
1321 arranged according to the number of strains with BLASTp hits. Bars are colored according to the
1322 genomic location in which the BLASTp hits are found.
1323
1324 Figure S7: Barcharts summarizing the genotypes of segregating genomic regions associated with
1325 the 20 high-confidence elements in the *Aspergillus fumigatus* 519 strain population (Table S16).
1326 If an isolate did not have any Starships within a given region, it was assigned either an "empty"
1327 or "fragmented" genotype (Methods).

1328 Figure S8: The putative target site of *Starship Tardis* (indicated with an arrow) occurs more often
1329 than you would expect by chance in the *Aspergillus fumigatus* genome. We estimated the copy
1330 numbers of all k-mers of length 10 in the Af293 reference genome and found that the k-mer of
1331 length 10 that corresponds to the putative target site of *Tardis* is present in high copy numbers
1332 that exceed the expected genome-wide frequency of k-mers of this length (>99.99th percentile).
1333
1334 Figure S9: *Starships* mobilize genes encoding diverse metabolic functions. Barcharts
1335 summarizing the presence of genes with metabolism-related COG (Clusters of Orthologous

Groups) annotations in the 20 high-confidence *Starships* in the 519 *Aspergillus fumigatus* strain population (Table S18). The X-axis measures the percentage of copies of a given *Starship* that carry at least 1 gene with an annotation of interest.

Figure S10: The Biofilm Architecture Factor (*baf*) gene family is closely associated with diverse *Starships* in *Aspergillus fumigatus*. A maximum likelihood tree of *baf* sequences from the 519 *Aspergillus fumigatus* strain population. Branches with ≥80% SH-ALRT and ≥95% ultrafast bootstrap support are in bold. *Baf* sequences found in *Starships* have the corresponding *Starship* identification number appended to their right (Table S5). Sequences are color-coded according to 6 corresponding *baf* clades of interest, and a summary of all the *Starship* types found associated with each clade is printed on the right.

Figure S11: *Starships* are enriched in environmental and clinical strains. Barcharts summarizing the proportion of strains from 475/479 *Aspergillus fumigatus* genotyped strains with known isolation sources that have the 20 high-confidence *Starships*, broken up by isolation source. Fisher's exact test P values for *Starship* enrichment across isolation source categories are shown above each bar (adjusted for multiple comparisons using the Benjamini Hochberg procedure; Table S20).

Figure S12: A heatmap of transcript abundances ($\log_{10}$ TPM) of *Starship* captain genes, collapsed across treatment replicates, for RNAseq studies from *A. fumigatus* A1163 (A) and CEA10 (B).

Figure S13: Heatmap of binary core transcript coverage across genes in *Starships* in strains Af293 (A), CEA10 (B) and A1163 (C). Genes that have a median core transcript coverage greater than 1 are shown here in blue, and values below 1 are shown in red.

Figure S14: Differentially expressed *Starship*-mobilized genes (DEGs) displayed in volcano plots across combinations of *A. fumigatus* strains and treatment categories. Differential expression based on singleton studies are shown as $\log_2$ fold-change ($\log_2$FC) in black with standard error bars, whereas DEGs identified across multiple studies are represented with summarized $\log_2$FC values from a random effects model (REM) and coloured by values "sign-consistency", the number of studies that reported differential gene expression in the positive (+1) or negative (-1) direction, centered around 0.

Figure S15: Distributions of TOM scores from the WGCNA constructed from Af293 *A. fumigatus* studies. TOM scores were separated into distinct categories for edges that connected any two genes within a *Starships*, between different *Starships*, between *Starships* and the rest of the genome, or between any two non-*Starship* genes in the genome were compared. These distributions were compared using a Wilcoxon test ($p > 0.05$ = "ns"; $p < 0.05$ = "*", $p < 0.01$* = "**", $p < 0.001$ = "***", $p < 0.0001$ = "****").

1377 Figure S16: Heatmap displaying the correlation values for eigengenes in each module and their
1378 association with samples from a specific study (labeled with BioProject IDs), or different levels of
1379 experimental treatment category across studies. Cells of the heatmap are coloured by the
1380 correlation value, as well as labeled in each cell. P-values from a pairwise comparison between
1381 observations (*pdist*) are shown in parentheses.

1382 Supplemental table legends

1383
1384 Table S1: Genome assembly statistics for the 12 strains sequenced with Oxford Nanopore long-
1385 read technology
1386
1387 Table S2: Metadata for publically available *Aspergillus fumigatus* strains
1388
1389 Table S3: Manual annotation data of 86 Starship elements
1390
1391 Table S4: Metadata for all starfish-predicted Starships
1392
1393 Table S5: Sequence features of all starfish-predicted Starships in the 519 strain population
1394
1395 Table S6: Frequencies of the 20 high-confidence Starships in the 519 strain population using the
1396 expanded dataset
1397
1398 Table S7: Manual annotation of Starship elements in three *Aspergillus fumigatus* reference strains
1399
1400 Table S8: Pairwise comparisons between strains of SNP identity by state and Starship repertoire
1401 similarity
1402
1403 Table S9: Comparison of Starship coordinates with structural variants identified by Colabardini et
1404 al 2022 (doi: 10.1371/journal.pgen.1010001)
1405
1406 Table S10: Orthogroup frequencies in the 519 strain population
1407
1408 Table S11: Orthogroup frequencies in the 13 reference-quality strains
1409
1410 Table S12: BLAST recovery of Starship cargo genes from the 13 reference-quality strains
1411
1412 Table S13: BLAST recovery of Starship cargo genes from the 519 strain population
1413
1414 Table S14: Genotyping of genomic regions with segregating Starship insertions
1415
1416 Table S15: Starships in genotyped genomic regions with segregating Starship insertions

Table S16: Summary of genotyping data for genomic regions with segregating Starship insertions

Table S17: Coordinates of predicted 5s rDNA sequences in the 519 strain population

Table S18: Proportion of elements carrying at least one gene annotated with various COG categories

Table S19: List of putative and published virulence and stress resistance genotypes in *Aspergillus fumigatus*

Table S20: Fisher's exact test statistics for Starship enrichment by strain isolation source

Table S21: Presence/absence genotyping data for the 20 high-confidence Starships in the 519 strain population using the genotyping dataset

Table S22: Metadata from RNASeq studies collected from NCBI used in the meta-analyses of *Starship* gene expression.

Table S23: Summary of RNASeq studies collected from NCBI used in the meta-analyses of *Starship* gene expression.

Table S24: Summary of modules assigned, and genes within them, from the WGCNA constructed from Af293 samples.

Table S25: Significantly enriched functional terms from a series of enrichment tests (Fisher’s Exact Test) conducted on the genes within WGCNA modules.

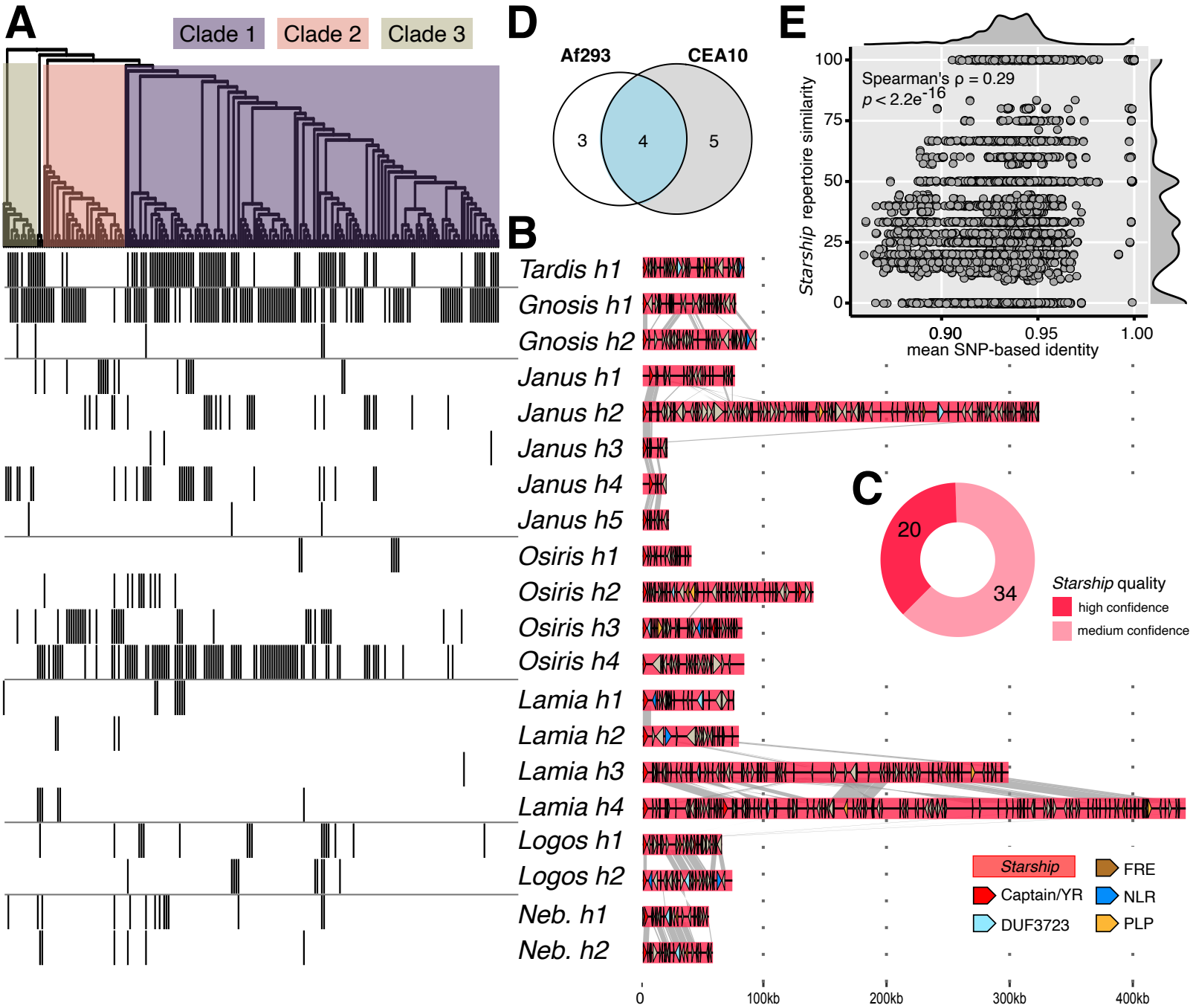

Figure 1: At least 20 distinct *Starships* carrying hundreds of protein-coding genes vary in their presence/absence across *Aspergillus fumigatus* strains. A) Top: a SNP-based maximum likelihood tree of 220 *A. fumigatus* strains from Lofgren et al. 2022 (visualized as a cladogram for clarity), color-coded according to phylogenetic clade as defined by Lofgren et al. 2022. Bottom: A heatmap depicting the presence (gray) and absence (white) of 20 high-confidence *Starships* in each isolate. B) Schematics and sequence alignments of the type elements from 20 high-confidence *Starships*, where links between schematics represent alignable regions ≥500bp and ≥80% nucleotide sequence identity and arrows represent predicted coding sequences. C) A donut chart summarizing the number of distinct types of *Starships* by the quality of their prediction. D) A Venn diagram indicating the number of shared and unique high-confidence *Starships* in the reference strains Af293 and CEA10 (Table S7). E) A scatterplot depicting pairwise comparisons of SNP-based Identity by State (IBS) and Jaccard similarity in high-confidence Starship presence/absence profiles between 150 Clade 1 strains (Table S8).

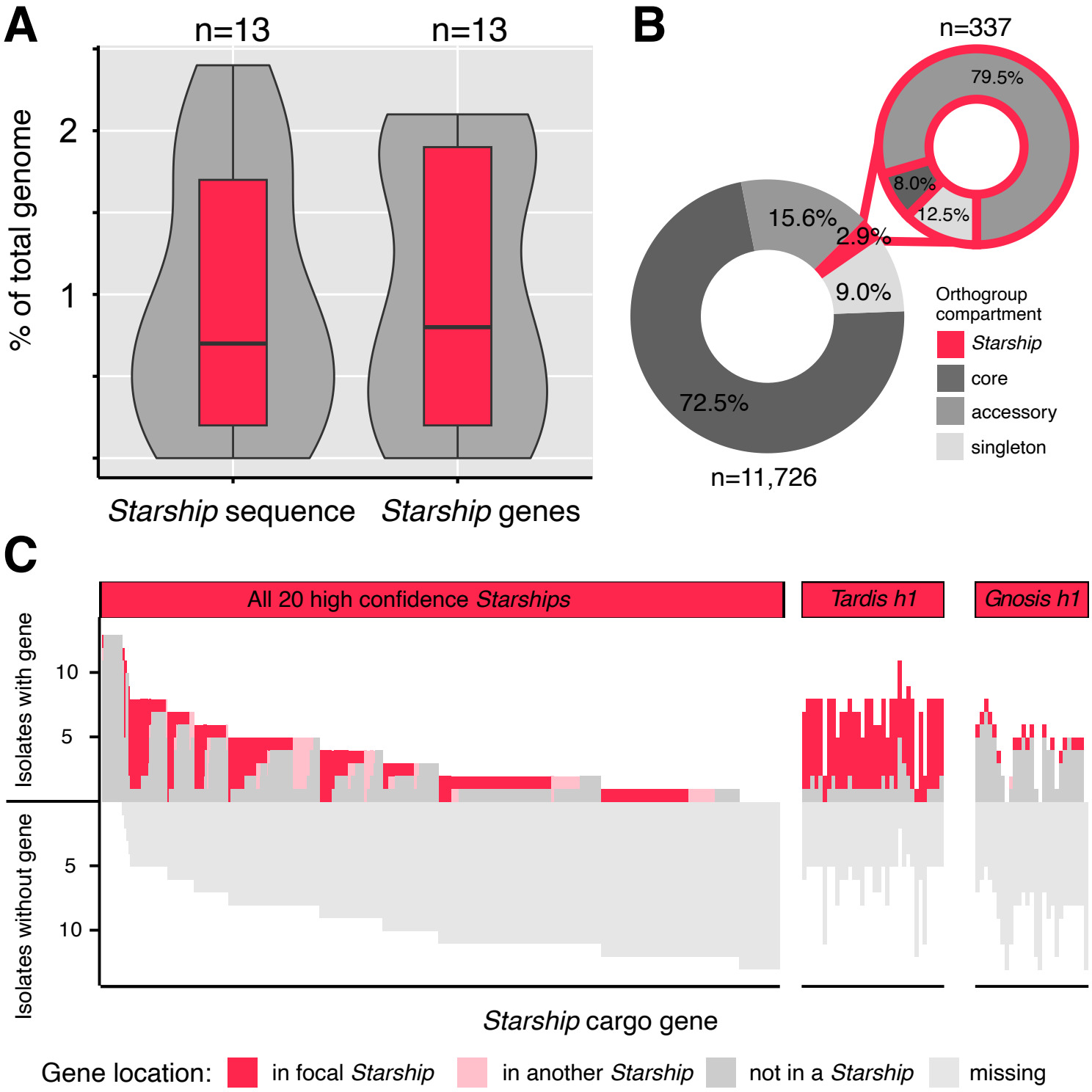

Figure 2: *Starships* are enriched in accessory genes whose presence/absence and genomic location vary across *Aspergillus fumigatus* strains. All panels visualize data derived from 13 reference-quality *A. fumigatus* assemblies and the 20 high-confidence *Starships* (n = 459 elements total). A) Box-and-whisker plots summarizing the total percentage of nucleotide sequence and predicted genes carried by *Starships* per genome. B) Donut charts summarizing the percentages of gene orthogroups in the core, accessory, singleton, and *Starship*-associated compartments of the *A. fumigatus* pangenome (Table S10). C) Iceberg plots summarizing the genomic locations of the single best BLASTp hits (≥90% identity, ≥33% query coverage) to the cargo genes from the type elements of all 20 high-confidence *Starships* (left) and two individual *Starships* (right; Table S12). Each column in the iceberg plot represents a cargo gene and is color-coded according to the genomic location of hits (full dataset in Fig. S5).

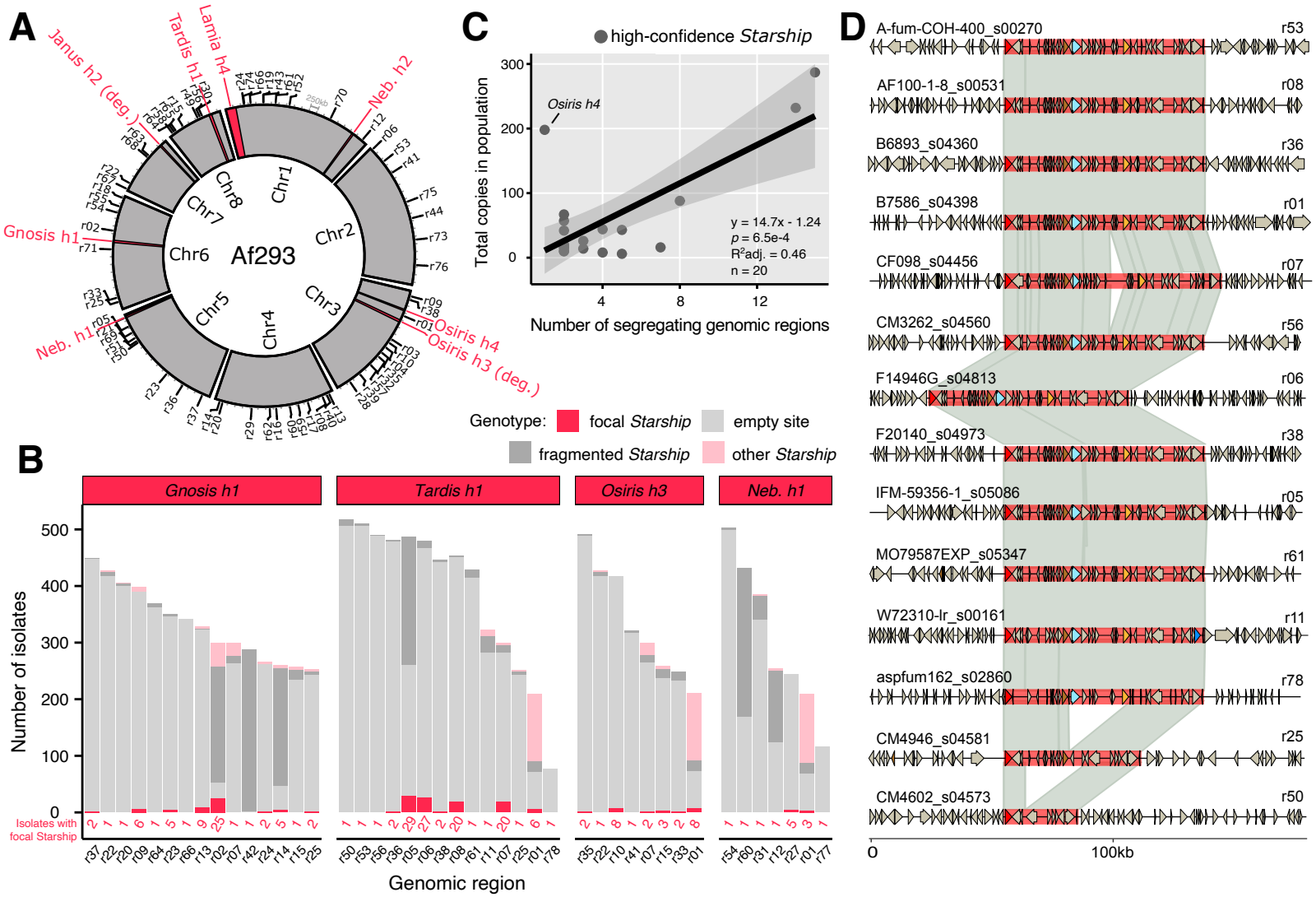

Figure 3: *Starships* and their insertion sites are distributed across all major chromosomes in the *Aspergillus fumigatus* genome. A) A Circos plot summarizing all inserted *Starships* (in red) and all genomic regions containing either an empty insertion site or a fragmented *Starship* (in black, along perimeter and labeled with the r prefix) in the 8 chromosomes of the *A. fumigatus* reference strain Af293 (Table S14). All genomic regions contain a *Starship* insertion in some other individual from the 519 strain population. B) Barcharts summarizing the genotypes of segregating genomic regions associated with the four most active high-confidence elements in the 519 strain population (Table S16; full dataset in Fig. S7). If an isolate did not have any *Starships* within a given region, it was assigned either an "empty" or "fragmented" genotype (Methods). C) A scatterplot summarizing the relationship between the number of genomic regions containing a given *Starship* and the total number of copies of that *Starship* in the 519 strain population, where each point represents one of the 20 high-confidence *Starships*. A line derived from a linear regression model is superimposed, with shaded 95% confidence intervals drawn in gray. D) Alignments of *Tardis h1* copies +/- 50kb of flanking sequence across 14 genomic regions. Links between schematics represent alignable regions ≥1000bp and ≥90% nucleotide sequence identity.

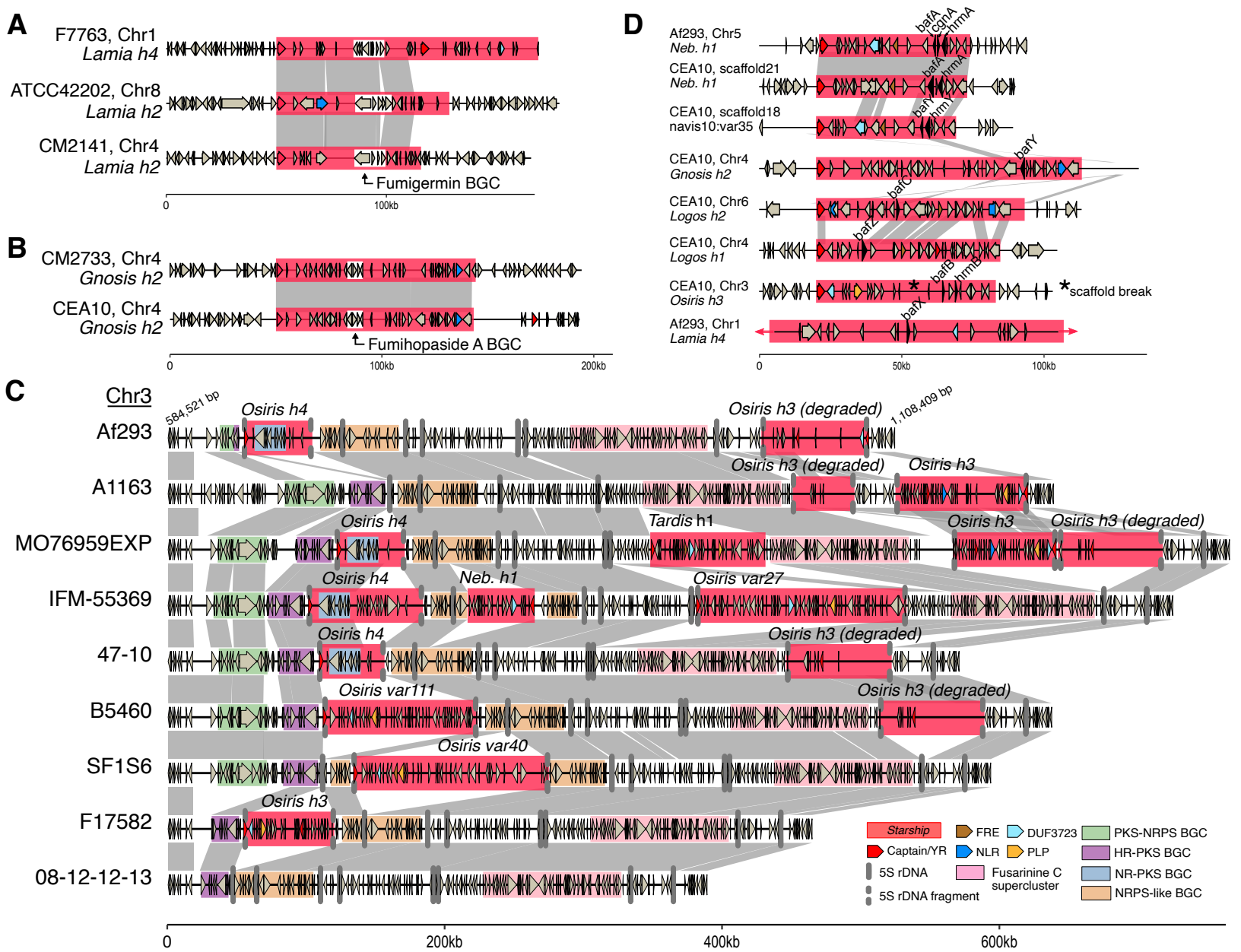

Figure 4: *Starships* mobilize adaptive traits and generate allelic diversity among *Aspergillus fumigatus* strains. Schematics and alignments of: A) *Starship Lamia h2* and *h4*, which carry the biosynthetic gene cluster (BGC) encoding the polyketide secondary metabolite Fumigermin, shown inserted at 3 independent sites. B) Starship *Gnosis h2* carrying the BGC encoding the terpene secondary metabolite Fumihopaside A, shown inserted at 2 independent sites. C) a large region on Chromosome 3 previously identified as an idiomorphic BGC containing multiple segregating *Starship* insertions and various combinations of putative BGCs, including a a non-reducing polyketide synthase (NR-PKS) BGC carried by *Starship Osiris h4*. *Starships* from the *Osiris* navis specifically insert in 5S rDNA sequence and are predicted to fragment it. D) Eight *Starships* in the reference strains Af293 and CEA10 that all carry homologs of biofilm architecture factor A (bafA). *Starship Nebuchadnezzar h1* (*Neb. h1*) carries bafA as part of the H_AAC (hrmA-associated gene cluster), while *Starship navis10-var35* carries bafB as part of H_BAC and *Starship Osiris h3* carries bafC as part of H_CAC. Only a portion of *Starship Lamia h4* is visualized for figure legibility. All data for A-D was collected from the 519 *A. fumigatus* strain population (Table S5). Links between schematics represent alignable regions ≥5000bp and ≥95% nucleotide sequence identity and arrows represent predicted coding sequences. Abbreviations: polyketide synthase (PKS); non-ribosomal peptide synthetase (NRPS); highly reducing (HR).

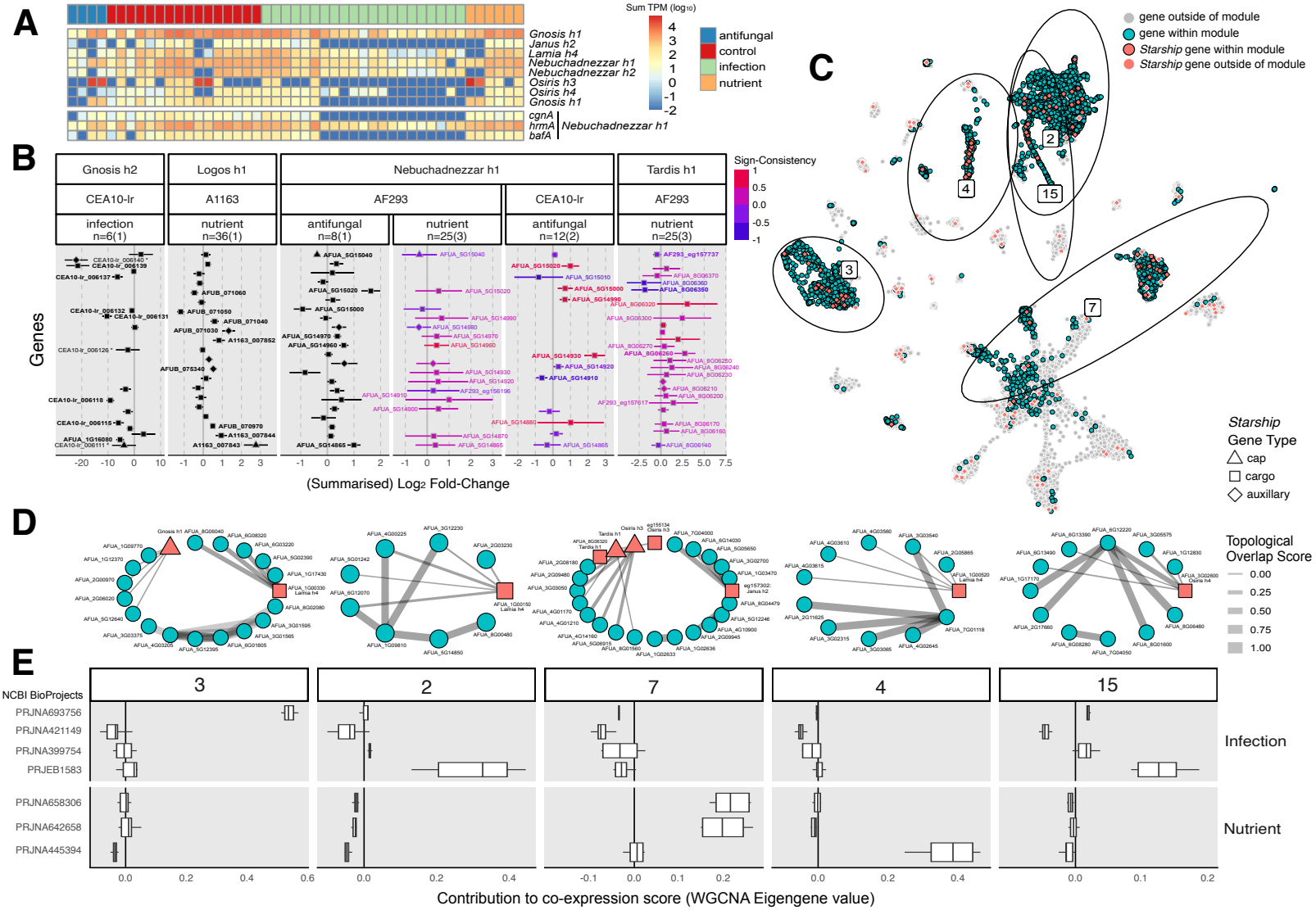

Figure 5: A) A heatmap of transcript abundances ($\log_{10}$TPM) of *Starship* captain tyrosine recombinase genes (top) and genes within the hrmA-associated cluster (HAC; bottom), collapsed across treatment replicates, for 11 RNAseq studies from *A. fumigatus* Af293. B) Results from differential expression (DE) tests for select *Starships*, treatment categories, and strain combinations. Only genes with valid output from DE tests are shown, and only genes with significantly DE are labeled with gene codes. Genes with extreme count outliers, as determined by Cook's Distance in DESeq2, do not have valid p-values and are labeled with an asterisk ("*"). DE based on a single study are shown as $\log_2$ fold-change ($\log_2$FC) in black with standard error bars, whereas DE genes identified across multiple studies are represented with summarized $\log_2$FC values from a random effects model (REM) and coloured by "sign-consistency", the number of studies that reported DE in the positive (+1) or negative (-1) direction, centered around 0. Labels in bold font represent DEGs that are significantly DE in more than one study. C) The results of a weighted gene co-expression network analysis (WGCNA) constructed from 14 "antifungal", "infection", and "nutrient" RNAseq studies represented using UMAP clustering based on the co-expression eigengene values for all genes in the *A. fumigatus* Af293 genome. Non-*Starship* genes present within modules of interest are shown in blue, while all *Starship* genes are shown in red, with those genes within modules of interest having a black outline. D) Genes within modules that were significantly associated with samples from specific treatment categories or studies were used to construct sub-networks which contained the top 10 edges in the network that were made between any pair of genes or any gene and a *Starship* captain. The connections between genes (edges) are based on the topological overlap matrix (TOM) for each module, and have been 0-1 scaled. E) Boxplots of eigengene values from WGCNA, akin to a weighted average expression profile, indicate the extent of co-expression (correlation) of the genes present within each module. Pairwise comparisons of module eigengene values determined if a module was significantly associated with samples from specific treatment categories or studies (Fig. S16).