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

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

Fungal infections are difficult to prevent and treat in large part due to heterogeneity in clinically relevant phenotypes. 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 generate genomic heterogeneity over timescales impacting 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 pathogenesis wherein *Starships* mediate variation in virulence-related gene content and expression. Together, our results demonstrate that *Starships* are a foundational mechanism generating disease-relevant genotypic and, in turn, phenotypic heterogeneity in a major human fungal pathogen.

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 ubiquitous human pathogen causing disease in an estimated 2 million people yearly[2]. Several infectious diseases are caused by *A. fumigatus*, including invasive pulmonary aspergillosis, which manifests primarily in immunocompromised individuals with mortality rates of upwards of 85%. The treatment of *A. fumigatus* infections are complicated by the remarkable variation strains display in virulence, resistance to antifungals, and other infection-relevant traits[3–7]. Strain heterogeneity confounds "one size fits all" therapies and poses a significant challenge for developing efficacious disease management strategies[8]. 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[9–11]. 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[12]. MGE transposition generates structural variation that directly impacts gene regulation and function[13–15], and many elements also modulate genome content by acquiring genes as "cargo" and transposing them within and between genomes, facilitating gene gain and loss[16,17]. The ability to generate contiguous genome assemblies with long-read sequencing technologies has dramatically enhanced our ability to find new lineages of MGEs[12,18]. For example, *Starships* are a recently discovered superfamily of MGEs found across hundreds of filamentous fungal taxa[19–21]. *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 recombinase at their 5' end that is both necessary and sufficient for transposition[20]. 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[22–25]. *Starships* are further capable of horizontal transfer, implicating them in the acquisition, dissemination and repeated evolution of diverse traits ranging from host interactions to stress resistance[19,20,23–25]. 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 MGEs are a source of clinically-relevant strain heterogeneity. We reveal that *Starships* are responsible for generating previously unexplainable variation in genome content and structure, including among different isolates of

what are purported to be the same laboratory strain. 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. 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 mechanism generating clinically-relevant 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 to detect in short-read assemblies due to their large size and frequent localization in regions with high repetitive content[19]. The wealth of publicly available short-read genomes are still of high value, however, due to the breadth of sampling they provide. We therefore deployed a hybrid sampling strategy that enabled an accurate and precise accounting of how *Starships* impact strain heterogeneity. We downloaded 507 short-read *A. fumigatus* assemblies isolated from environmental and clinical sources that span the known genetic diversity of this species and used Oxford Nanopore to sequence 12 additional strains with long-read technology, for a combined total of 519 assemblies[9,10]. 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[9]. We performed 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 for investigating the connection between *Starship*s and strain heterogeneity (Table S1).

At least 20 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[21]. 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)[9,10]. 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 our hybrid sampling approach (Fig. S1).

To determine how many different *Starships* 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. Using a conservative threshold that required observing the same *navis-haplotype* (i.e., *Starship*) at multiple 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).

Starships distributions are heterogeneous and only partially explained by strain relatedness

All high-confidence *Starships* show polymorphic presence/absence variation, representing a previously unexamined source of genetic heterogeneity across *A. fumigatus* strains (Fig. 1A, Table S6). For example, the two commonly used reference strains Af293 and CEA10 carry 12 different high-confidence *Starships*, but share only 4 in common (or only 2, using a more conservative threshold of not counting fragments or degraded elements; Fig. 1D, Table S7). The number of high-confidence *Starships* per long-read isolate ranges from 0-11 (median = 3, std. dev. = 3.5)[26,27]. To identify the underlying drivers of variation in *Starship* distributions, we investigated the relationship between *Starship* repertoires and strain relatedness by testing if phylogenetic signal underlies *Starship* distributions. We genotyped *Starship* presence/absence in the best sampled clade of *A. fumigatus* with well resolved phylogenetic relationships and good representation of environmental and clinical strains (Clade 1 *sensu* Lofgren et al., 2022, n = 150) and found that isolate relatedness is significantly but only weakly correlated with similarity in *Starship* repertoire (Fig. 1E, Table S8, Spearman's $\rho = 0.29$; $P < 2.2e^{-16}$, using the genotyping dataset). This indicates that *Starships* are a source of genetic heterogeneity among even closely related strains. While we are unable to unequivocally differentiate horizontal transfer events from sexual or parasexual recombination in *A. fumigatus*, we hypothesize that horizontal transfer within this species is at least partially responsible for generating the observed patchy distributions of *Starships*, since *Starship* transfer has occurred between fungal species of vastly different taxonomic ranks, and barriers to transfer would be lower within species.

Starships actively transpose in clones of reference laboratory strains

We determined the short-term potential for *Starships* to introduce unwanted variation into laboratory experiments by comparing lab-specific isolates of the same strain used by different research groups. First, we found that *Starship* insertions in the reference Af293 genome coincide

Family	Name	Predicted target site	Mean length in bp (SD)	Type element	Functions of interest in type element
Tardis	*Tardis h1*	TACGGAGTAG	81,743 (9,347)	W72310-lr_s00161	GH71, SAH
Prometheus	*Gnosis h1*	A(N_3)CTA(N_{17})T	82,313 (23,570)	A-fum-AFUG-100413-0667_s00261	GH18, lysM
Prometheus	*Gnosis h2*	-	93,292 (60)	CM2733_s04518	bafY, fumihopaside A BGC
Phoenix	*Janus h1*	-	62,365 (19,076)	F14513G_s04800	CHROMO
Phoenix	*Janus h2*	A(N_3)TTACTA(N_2)A(N_{15})CT	217,821 (102,982)	47-4_s00005	CHROMO, DHDPS, GH18, GH71
Phoenix	*Janus h3*	A(N_3)TTACTA(N_2)A(N_{15})CT	42,070 (37,412)	SF1S6_s05434	CHROMO
Phoenix	*Janus h4*	-	19,097 (5,047)	CM6458_s04616	CHROMO
Phoenix	*Janus h5*	-	34,546 (18,524)	47-10_s00011	CHROMO
Enterprise	*Osiris h1*	5S rDNA	37,633 (2,148)	A-fum-AFIS-13708-CDC-14_s00209	putative BGC
Enterprise	*Osiris h2*	5S rDNA	138,260 (2,424)	B5269_s04148	ARS, B-LAC, COP
Enterprise	*Osiris h3*	5S rDNA	70,476 (28,148)	Afu-218-E11_s00395	bafB, hrmB
Enterprise	*Osiris h4*	5S rDNA	51,342 (13,227)	Af293_s00032	putative BGC
Galactica	*Lamia h1*	A(N)TAGT	69,899 (24,972)	S02-30_s00148	GH18, lysM
Galactica	*Lamia h2*	A(N)TAGT	65,414 (6,136)	ATCC42202_s00060	fumigermin BGC
Galactica	*Lamia h3*	A(N)TAGT	236,680 (74,695)	ISSF-21_s05253	GH18, GH71
Galactica	*Lamia h4*	A(N)TAGT	291,189 (215,918)	F7763_s04695	bafX, fumigermin BGC, GH3, GH18, GH71, GT8, GT31
Prometheus	*Logos h1*	A(N_{10})TACTTATTA(N)A(N_6)A	69,008 (5,313)	CEA10-lr_s00105	bafZ
Prometheus	*Logos h2*	-	56,189 (31,755)	CEA10-lr_s00103	bafC
Hephaestus	*Neb h1*	TTACA(N_5)AAT	48,346 (5,585)	AF293_s00037	ARS, bafA, cgnA, hrmA
Hephaestus	*Neb h2*	TTACA(N_5)AAT	54,751 (7,226)	AF293_s00031	ARS

Table 1: High Confidence *Starships* in *Aspergillus fumigatus*. Abbreviations: GH: glycosyl hydrolase; GT: glycosyl transferase; SAH: salicylate hydroxylase; DHDPS: dihydrodipicolinate synthetase; CHROMO: chromatin modification domain; ARS: arsenic resistance cluster; COP: copper resistance; B-LAC: Beta-lactamase

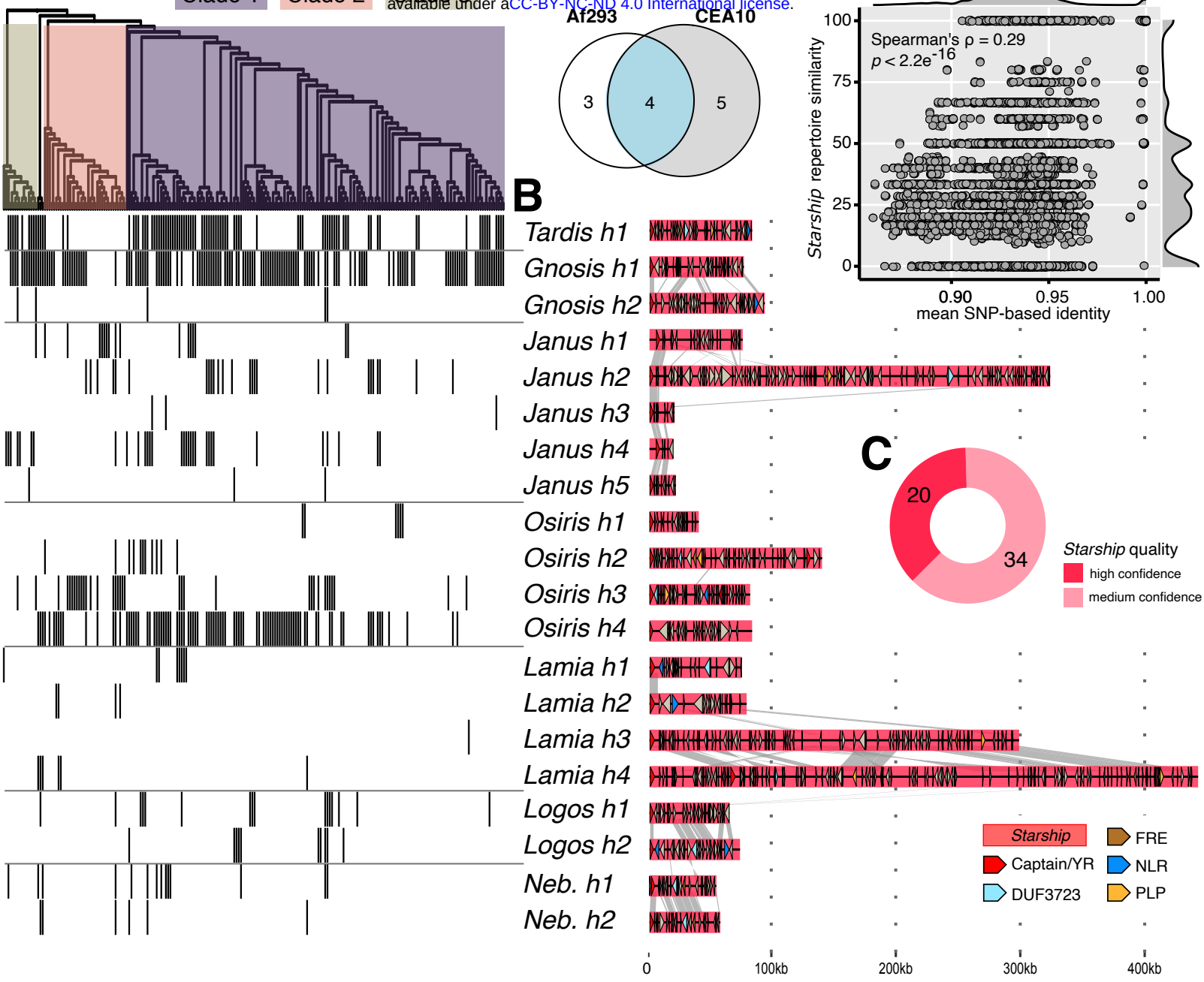

Figure 1: At least 20 distinct Starships carrying hundreds of protein-coding genes vary in their presence/absence across Aspergillus fumigatus strains. A) Top: a SNP-based maximum likelihood tree of 220 A. fumigatus strains, color-coded according to phylogenetic clade (sensu Lofgren et al., 2022)9. Bottom: 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).

with 4/8 known putative structural variants identified in other Af293 strains (Table S9), suggesting that *Starship* transposition and/or loss contributes substantially to structural variation observed among isolates of what should otherwise be clones of the same strain[5]. We complemented this literature-based approach with a genomics approach where we compared sequence data among isolates of the same strain. We compared the locations of *Starships* between the CEA10 assembly we generated here and a publically available CEA10 assembly (BioSample SAMN28487501), and found evidence that at least 1 *Starship* (*Nebuchadnezzar h1*) has jumped to a different chromosomal location, from chromosome 6 to chromosome 4 (Fig. S2). We also evaluated *Starship* heterogeneity in three additional strains with both publicly available short-read data and long-read data generated in this study. One of the strains (ATCC-46645-lr) showed signs of *Starship* heterogeneity among isolates used by different research groups. Specifically, a small number of reads (a single mapped read for *h2* and three for *h1*) supported precise excision events for *Nebuchadnezzar h1* and *h2*, suggesting these elements have transposed or have been lost in a subset of the nuclei within one of the sequenced isolates (Fig. S3). This suggests that *Starships* are active under laboratory conditions, which is directly relevant for experimental design in this case as *Nebuchadnezzar h1* carries the HAC gene cluster known to impact biofilm-associated virulence (see below)[28]. Additionally, one breakpoint of a large-scale reciprocal translocation between chromosomes 1 and 6 among CEA10 isolates is localized precisely within a *Starship*, suggesting that this structural variant was caused by the presence of the *Starship*, an association that has also been noted in at least one other species[19]. Together, *Starship*-mediated variation among isolates of the same *A. fumigatus* strain suggests these transposons cause genomic instability over short-enough timescales that they may affect routine laboratory work and experimental reproducibility.

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

A. fumigatus strains differ extensively in the combinations of genes found in their genomes, resulting in a large and diverse pangenome[9,10]. To quantify how much pangenomic variation is attributable to actively transposing *Starships,* we estimated the total proportion of genomic content present in the 20 high-confidence *Starships* (Fig. 2, Methods). We examined these *Starships* in the 13 reference-quality assemblies (Af293 plus the 12 newly sequenced long-read strains) and found that between 0-2.4% of genomic nucleotide sequence (and 0-2.1% of all genes 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 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

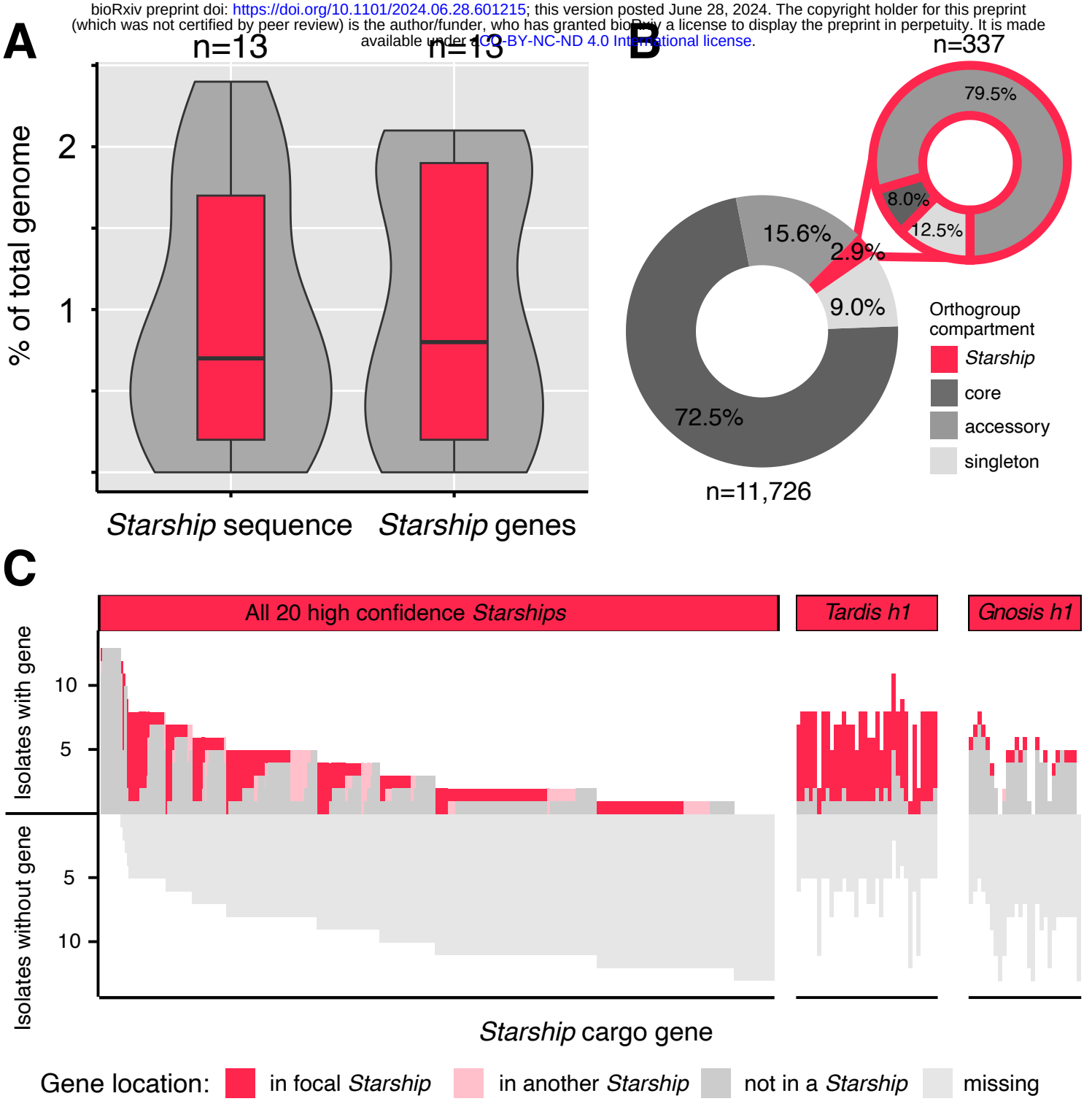

Figure 2: Starships are enriched in accessory genes whose presence/absence and genomic location vary across Aspergillus fumigatus strains. All panels visualize data derived from 13 reference-quality A. fumigatus assemblies and the 20 high-confidence Starships (n = 459 elements total). A) Box-and-whisker 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. S4).

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. S4, Table S11). 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. S5, 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 modulate accessory sequences among strains, highlighting the importance of investigating *Starships* at the individual element level.

Starships generate 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. S6; 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 tens 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*

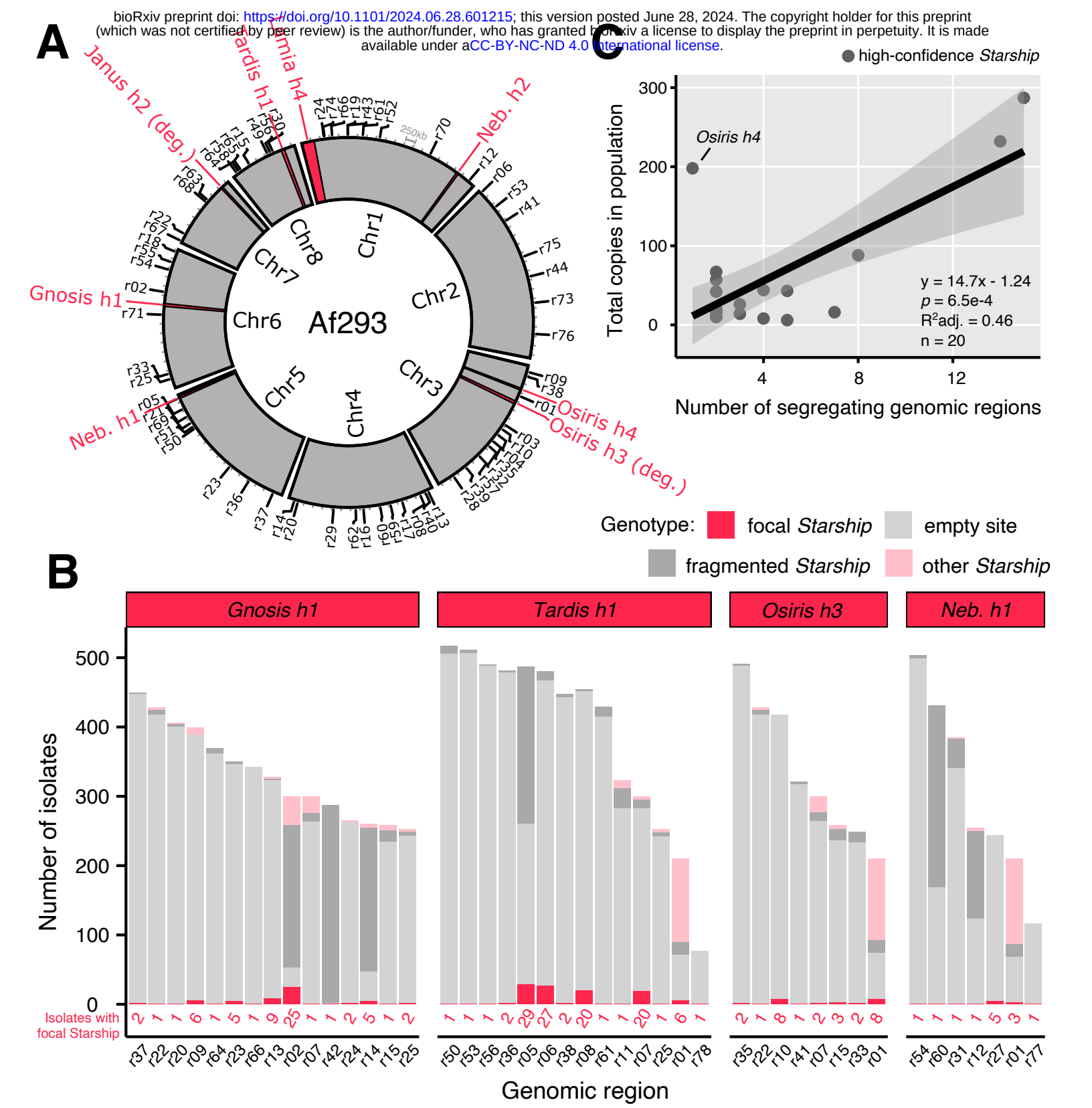

Figure 3: Starships and their insertion sites are distributed across all major chromosomes in the Aspergillus fumigatus genome. A) A Circos plot summarizing all inserted Starships (in red) and all genomic regions containing either an empty insertion site or a fragmented Starship (in black, along perimeter 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 A. fumigatus 519 strain population (Table S16; full dataset in Fig. S6). 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 A. fumigatus 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.

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).

Given that *Starships* are mobilized by a site-specific tyrosine recombinase[20], 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[29]. 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[20].

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[21]. 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; Fig. S7), indicating that this motif may represent some other type of highly conserved genomic sequence, such as a transcription factor binding site. Furthermore, this DR sequence is conserved among other *Eurotiomycetes*, which could correspond to deep conservation of this sequence among fungi[21]. This latter 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.

Starships modulate variation at an idiomorphic biosynthetic gene cluster

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

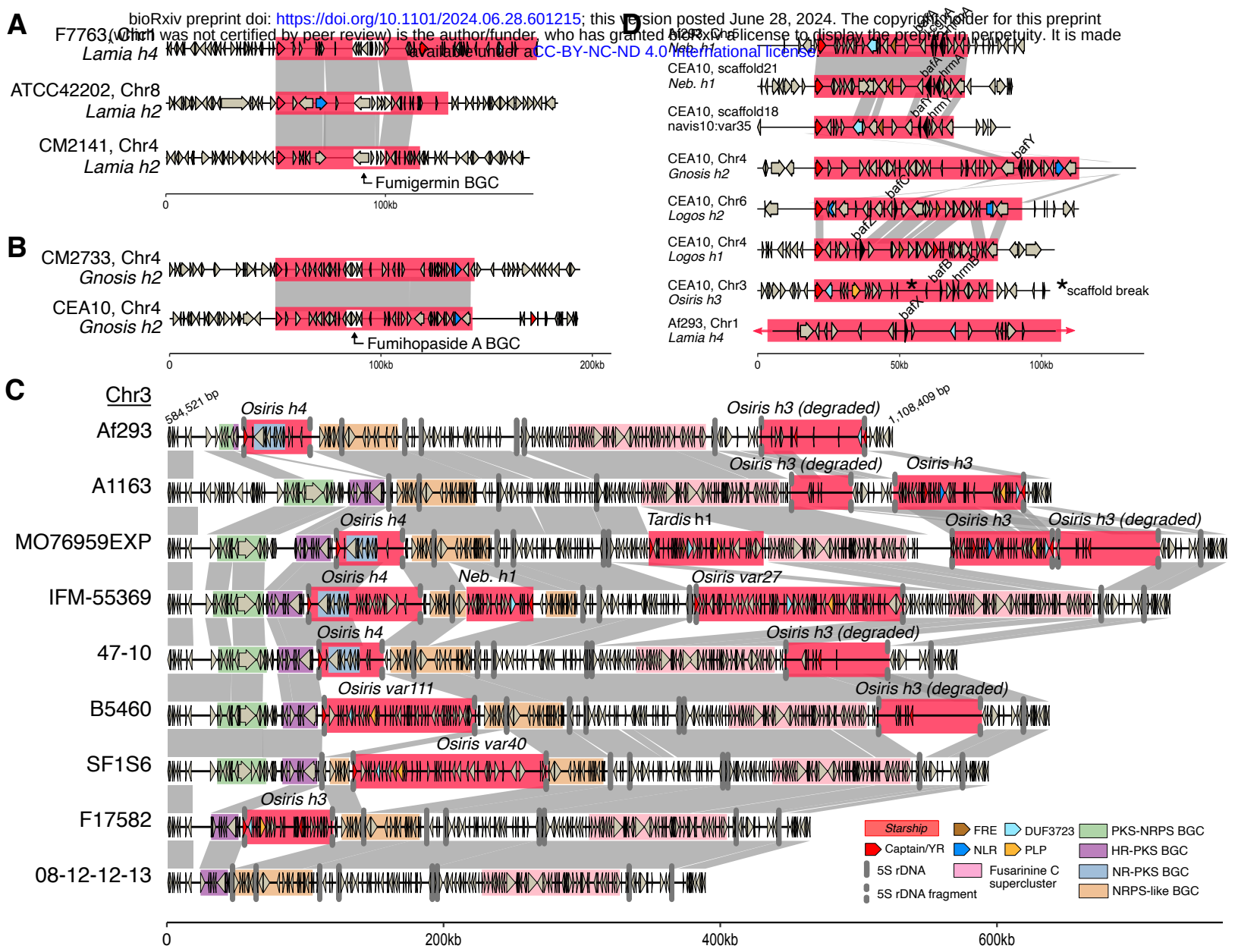

Figure 4: Starships mobilize adaptive traits and generate allelic diversity among Aspergillus fumigatus strains. Schematics and alignments of: A) Starship Lamia h2 and h4, which carry the biosynthetic gene cluster (BGC) encoding the polyketide secondary metabolite Fumigermin39, shown inserted at 3 independent sites. B) Starship Gnosis h2 carrying the BGC encoding the terpene secondary metabolite Fumihopaside A32, shown inserted at 2 independent sites. C) a large region on Chromosome 3 previously identified as an idiomorphic BGC30 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 (annotated by two vertical gray bars). D) Eight Starships in the reference strains Af293 and CEA10 that all carry homologs of biofilm architecture factor A (bafA)28. Starship Nebuchadnezzar h1 (Neb. h1) carries bafA as part of the HAAC (hrmA-associated gene cluster), while Starship navis10-var35 carries bafB as part of HBAC and Starship Osiris h3 carries bafC as part of HCAC28. 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).

this single locus, representing a ~10-fold enrichment relative to background expectations given the length of this region (using 5S rDNA frequencies in the Af293 genome). Thus, the enrichment of 5S rDNA at this locus potentiates *Starship*-mediated variation and the generation of a *Starship* hotspot.

We were surprised to find that in addition to multiple segregating *Starships*, the Chromosome 3 region contains an idiomorphic biosynthetic gene cluster (BGC) that was previously noted to be a recombination hotspot (Cluster 10 *sensu* Lind et al. 2017)[30,31]. The idiomorphic BGC locus is polymorphic for upwards of 4 smaller BGC modules and is upstream of a fifth BGC encoding the biosynthesis of Fusarinine C (BGC8 *sensu* Bignell et al. 2016) that itself is embedded within a larger stretch of sequence containing NRPS and KR-PKS core genes. While the idiomorphic modules were not predicted by antiSMASH, manual inspection revealed that each contains a different core SM biosynthesis gene and numerous other genes involved in metabolic processes, suggesting they are part of a larger BGC or represent different cryptic BGCs. The NR-PKS BGC module at this locus is carried by *Starship Osiris h4,* and its presence/absence is directly associated with the presence/absence of the element, implicating *Starships* as a mechanism generating idiomorphic BGCs. Furthermore, some BGC modules are disrupted by the insertion of a *Starship* (e.g., the NRPS-like BGC module). Together, these findings support the assertion that *Starship* activity generates selectable variation in BGC genotypes. The implication of this variation for the expression and diversification of natural products warrants further investigation.

Starships encode clinically-relevant phenotypes and are enriched in clinical strains

Starships encode genes with diverse functions associated with pathogen fitness, including carbohydrate active enzymes, BGCs and metal detoxification genes (Table 1, Table S5). To investigate broad trends in how *Starships* might contribute to pathogen phenotypes, 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. 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 pathogenicity and overall fitness, we examined granular classifications within the Metabolism category and found that *Starship* types differ specifically in their contributions to metabolic heterogeneity (Fig. S8). 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.

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)[9,10,30,32–38]. The BGC encoding Fumihopaside A (AFUA_5G00100-00135) 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[32]. Similarly, the BGC encoding the polyketide Fumigermin (AFUA_1G00970-01010), which inhibits bacterial spore germination, is carried by *Lamia h2* and *h4* (Fig. 4B)[39]. 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)[30]. Finally, a cluster of three genes (AFUA_5G14900/*hrmA*, AFUA_5G14910/*cgnA*, AFUA_5G14915/*bafA*, collectively referred to as the *hrmA*-associated cluster or HAC) carried by *Nebuchadnezzar h1* increases virulence and low oxygen growth, and its expression modulates colony level morphological changes associated with biofilm development (Fig. 4D)[28,40]. The genes *bafB* and *bafC*, which are homologs of *bafA,* also determine colony and submerged biofilm morphology and are each carried by up to 5 additional *Starships*, indicating a sustained association between this gene family and *Starships* (Fig. 4D, Fig. S9)[28]. Mobilization within and between genomes likely contributes to the rapid evolution of these survival and virulence traits through the processes of gene gain and loss, effectively increasing the capacity of these traits to evolve in populations. We propose that by mobilizing these genes, *Starships* drive heterogeneity in clinically-relevant phenotypes[28].

To identify more *Starships* that could be relevant in clinical settings, we tested for an association between the high-confidence *Starships* and strain isolation source (genotyping dataset; Methods). We found that 5 high-confidence *Starships* are significantly enriched (P_{adj} <0.05) in strains from either clinical or environmental isolation sources across the 475/479 genotyped strains for which source data exists (Table S2, Table S20, Fig. S10). *Lamia h3* (P_{adj} = 0.013) and *Janus h3* (P_{adj} = 0.034) are enriched in environmental strains, while *Osiris h4* (P_{adj} = 0.022), *Nebuchadnezzar h1* (P_{adj} = 0.015; carries *bafA*, see above) and J*anus h2* (P_{adj} = 0.013) are enriched in clinical strains. Although we can’t completely rule out biases in our sampling that would contribute to these enrichment patterns, strains with these 5 *Starships* were isolated from different countries (between 3-10) and belong to different phylogenetic groups (between 1-3 clades sensu Lofgren et al., 2022 and 1-6 phylogenetic clusters sensu Barber et al., 2021). Further investigation is necessary to test how these particular *Starships* contribute to fitness in non-clinical and clinical settings.

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

The revelation of *A. fumigatus' Starships* allowed us to re-examine RNA-seq datasets available for this pathogen to test the hypothesis that *Starship* cargo is expressed under clinically-relevant conditions. We analyzed patterns of differential *Starship* cargo gene expression using 14 transcriptomic studies with easily comparable treatments from three commonly used reference strains (Af293, CEA10, A1163), for a total of 177 samples. We identified 459 differentially expressed genes (DEGs) in *Starships* across the three reference strains (Supplementary Data), out of 596 *Starship* genes total. We did not find a significant enrichment of DEGs within *Starships* compared to the genomic background of any strain (Fisher's Exact Test *P*-values > 0.05). However, we found that *Starship* DEGs vary across element naves, fungal strains, and experimental conditions, indicative of pervasive strain-specific, treatment-specific and strain by treatment interactions impacting *Starship* cargo expression (Fig. 5B & Fig. S13).

To identify experimental conditions that may impact *Starship* transposition and the subsequent generation of *Starship*-mediated heterogeneity, we examined patterns of transcript abundance for captain genes responsible for *Starship* transposition within 12 high-confidence *Starships* from the reference strains Af293 (Fig. 5A), CEA10 and A1163 (Fig. S11). All captain genes have evidence for constitutive gene expression (a minimum median transcript coverage of 1 across the gene body for biological replicates) in at least one experimental condition (Fig. S12). Overall, 8 of the 12 captain genes from high-confidence *Starships* are differentially expressed in at least one study for one or more reference strains, revealing myriad conditions that create opportunities for *Starship* transposition (Supplemental information). Differences in experimental conditions and/or strain backgrounds have no consistent positive or negative influence on captain gene expression, indicating that the intricacies of captain expression remain to be elucidated (Fig. 5B).

We then examined patterns of differential expression for *Starship*-mobilized cargo genes in order to identify conditions where *Starships* have potential to impact organismal phenotypes. As with captain genes, the expression of cargo genes is context-dependent, with strain background, experimental condition and *Starship haplotype* impacting gene transcription (Fig. 5B). For example, we compared transcript abundance patterns for genes within the *hrmA*-asscociated cluster (HAC) between treatment categories and found that the transcription of genes in the HAC appear to be down-regulated in select infection environments (Fig. 5A & Figure S11). Together, strain identity, treatment conditions, and non-additive interactions between strains and treatments generate variation in *Starship* cargo expression.

To gain further insight into *Starship*-mediated heterogeneity in gene expression and regulation, we built weighted gene co-expression networks (WGCNA) that capture correlations between gene expression patterns to reveal the broader regulatory network of the cell. We identified five modules as being significantly associated with "antifungal"-, "infection"-, or "nutrient"-based

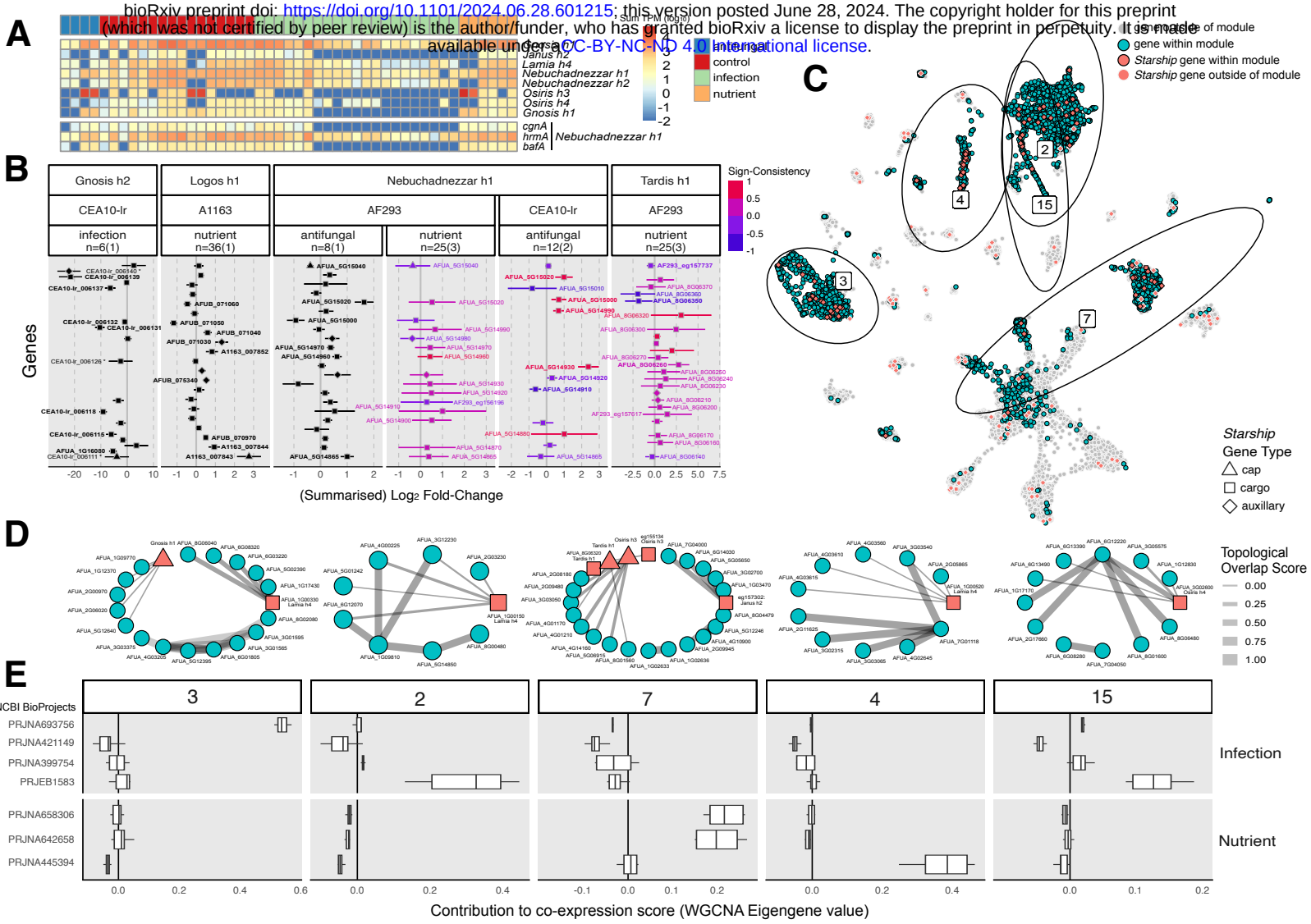

Figure 5: A) A heatmap of transcript abundances (log10TPM) of Starship captain tyrosine recombinase genes (top) and genes within the hrmA-associated cluster (HAC; bottom), collapsed across treatment replicates, for 11 RNAseq studies from A. fumigatus Af293. B) Results from differential expression (DE) 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 log2 fold-change (log2FC) in black with standard error bars, whereas DE genes identified across multiple studies are represented with summarized log2FC 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. S15).

studies (Fig. 5E). We identified two major trends in the transcriptional network properties of *Starship* cargo. First, we found multiple *Starship*-associated genes that integrate into modules containing numerous background genes, underscoring the possibility that phenotypes emerging from these networks are the product of regulatory interactions between *Starship* and non-*Starship* genes (Fig. 5D & Fig. S14). Connections involving captains in particular further highlight candidate loci involved in regulatory interactions between *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*. 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 little is known about the genetic basis of this variation. We hypothesized that a newly discovered group of unusual MGEs, *Starships*, would be major drivers of strain heterogeneity. We tested this hypothesis by systematically characterizing *Starship* activity and expression in *A. fumigatus*, an important human fungal pathogen of critically high research priority[1,41]. Our work provides foundational insight into the mode and tempo of fungal evolution by revealing that strain variation emerges not only from single nucleotide or copy number variation [9,10] 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[42]. Given our strict requirements that a high-confidence *Starship* be found in 2 different locations in 2 different strains, our estimates represent a 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 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.

Online 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 [43]. Biomass was collected by gravity filtration through Miracloth (Millipore). High molecular weight DNA extraction followed the Fungal CTAB DNA extraction protocol[44]. 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[9]. 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 v wrapped with AAFTF v0.3.0[45–47]. Summary statistics for each assembly were computed with AAFTF and BUSCO using eurotiomycetes_odb10 marker set[48]. 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[49,50].

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)[51]. 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)[52,53]. 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)[54,55]. On average, 35.7% of genes within *Starships* have no known PFAM or Interpro domain, GO term, COG category or eggNOG orthology (range = 0-67.6%, std. dev. = 14.8%) but approximately half of all mobilized genes could be assigned to

a COG (e.g., 50% of 11,900 genes total, across 459 high-confidence elements). To investigate associations between *Starships* and pathogen phenotypes, we searched for the presence of 798 genes known to be associated with virulence or infection-relevant phenotypes on *Starships* (Table S19, Methods)[9,10,30,32–38].

SNP analysis

We calculated pairwise kinship among strains by analyzing a previously generated VCF of all high-quality, filtered, biallelic single nucleotide polymorphisms (SNPs) from the Lofgren et al. 2022 population with plink (--distance square ibs flat-missing --double-id --allow-extra-chr; Fig. 1E)[56,57]. We calculated Jaccard similarity in *Starship* repertoire among all pairwise combinations of strains using genotyping data for the set of 20 high-confidence *Starships* and the following formula: *Starships* shared between strain 1 and 2 / (*Starships* unique to strain 1 + *Starships* unique to strain 2 + *Starships* shared between strain 1 and 2).

Pangenome construction

We constructed a pangenome for all 519 strains using Orthofinder (--only-groups -S diamond), and then extracted all orthogroups containing at least 1 of the 13 reference-quality strains for analysis (Fig. 2B)[58]. Core orthologs are defined as being present in all 13 strains; accessory orthologs are present in between 2-12 strains; singleton orthologs are present in 1 isolate. We included all 519 strains in the Orthofinder analysis to ensure that ortholog groups are consistent across different population subsets.

BLAST analysis

We determined the genomic locations of all best reciprocal blast hits to cargo genes carried by the 20 high-confidence *Starships* using BLASTp (Fig. 1C)[59]. For each cargo gene, we retrieved the highest scoring hit with ≥90% identity and ≥90% query coverage in each of the 13 reference-quality assemblies, and determined whether that hit was found in the same *Starship*, a different *Starship*, or not in a *Starship* at all. If no hit was retrieved, that gene was marked as missing.

Structural variant analysis

To evaluate the validity of putative structural rearrangements within our CEA10-lr assembly, we took two approaches. First, we aligned the raw nanopore reads to the CEA10-lr assembly using Mummer 4.0 with the nucmer command and default parameters[60]. Alignments were then visually inspected using the Genome Ribbon software[61]. This method confirmed that individual reads spanned the junctions of putative rearrangements and confirmed that no misassemblies were present. However, this approach suggested that two alternative chromosomal topologies existed: one that matches the chromosomal assemblage of the reference strain Af293, and another that exhibits a reciprocal translocation between Chromosomes 1 and 6, as reported for other assemblies of CEA10[62]. To resolve this paradox, we then mapped the raw nanopore reads to

both the CEA10-lr assembly, as well as a publicly available high-quality assembly (BioSample: SAMN28487501) with the minimap2 map-ont command [63]. These were then visually inspected using the IGV genome browser[64]. Multiple reads support both topologies, implying that the CEA10-lr isolate sequenced here is a heterogeneous mix of at least two lineages derived from the same CEA10 ancestor, likely obscuring estimates of the frequency with which *Starships* are mobilizing. Additionally, we examined the location of the annotated *Starship* elements, which for many elements revealed multiple reads where the given *Starship* was absent, suggesting that the element had either excised itself when mycelia was cultured for DNA extraction, or that additional derived lineages with variation at these loci were present in the sample. To resolve these possibilities, single conidial strains will have to be obtained and resequenced.

To further assess whether *Starships* transpose frequently enough to generate variation among isolates of the same strain, we evaluated three additional strains that we sequenced with Nanopore long-read technology that also had publicly available short-read data generated by other research groups: ATCC46645-lr, S02-30, and TP9. Short-read sequence data were mapped to our long-read assemblies and visually inspected in IGV to determine if any *Starships* were deleted/transposed in the public data. Both S02-30 and TP9 had no apparent variation at *Starship* loci. However, ATCC46645-lr showed polymorphism at three *Starships*, namely *Nebuchadnezzar* h1, *Nebuchadnezzar* h2, and *Gnosis* h1, indicative of active transposition. Both of the *Nebuchadnezzar haplotypes* show clean deletions, indicating that they may have jumped out of the genome of some nuclei within a particular isolate, while the deletions around *Gnosis* (5 in total) were more complex, which may be a signature of genomic instability rather than *Starship* mobilization. Note that neither S02-30 nor TP9 have *Nebuchadnezzar h1*.

Starship annotation

We systematically annotated *Starships* in the 12 newly sequenced long-read genomes plus 507 publicly available *A. fumigatus* genomes by applying the starfish workflow v1.0.0 (default settings) in conjunction with metaeuk, mummer4, CNEFinder and BLASTn (Table S2)[21,59,60,65,66]. Briefly, captain tyrosine recombinase genes were *de novo* predicted with starfish's Gene Finder Module and full length elements associated with captains were predicted with starfish's Element Finder Module using pairwise BLAST alignments to find empty and occupied insertion sites. We filtered out all elements that were <15kb in length, which we have found to correspond to indels of captain genes but never to the transposition of a full length *Starship*. We then manually examined alignments between each putative *Starship* and its corresponding insertion site, and filtered out all "low confidence" poorly supported alignments indicative of a false positive insertion. Poor alignments were observed most often when *Starships* inserted into smaller transposons; when an inversion breakpoint occurred within a putative *Starship*, resulting in an over-estimation of *Starship* length; or when the insertion site was on a very small contig resulting in small flanking region alignments.

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 was used to generate alignments with default parameters[67]. 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)[68]. *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"[69,70]. 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 conservative 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[21]. 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 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 paired-end short read Illumina sequencing data for all 466 strains for which these data were available and mapped them to the 20 high-confidence *Starships* using the command "starfish coverage" and the aligner strobealign[71], ensuring that we would correctly genotype the presence/absence of a *Starship* even if it is not present in an assembly or partially assembled. 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.

Starship datasets

We explored a variety of methods for predicting *Starships*, including manual and automated annotation, BLAST-based detection, and read-mapping to a known *Starship* library (see above). Certain detection methods were more appropriate for particular analyses, resulting in the partitioning of our findings into three datasets: the high-confidence dataset, expanded dataset, and genotyping dataset. To analyze *Starship* features, like gene content and pangenome distributions, we took a conservative approach by limiting 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 all 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 (i.e., with starfish 'extend' boundaries'), 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 presence/absence of an element regardless of where it is found in the genome, we took an even more conservative approach by limiting our analysis to the 13 reference quality assemblies plus 466 *A. fumigatus* strains with paired-end short-read Illumina sequencing data that were used to validate *Starship* presence/absence with short-read mapping (genotyping dataset; Table S21).

Meta-Analysis of *A. fumigatus* RNAseq Datasets

We collected transcriptomes from a total of 434 paired-end libraries of *A. fumigatus* strains Af293, A1163, and CEA10/CEA17. 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*[72], we created separate sets of BAM files to assess transcript coverage across *Starship* genes by mapping to the transcriptome with *STAR*[73]. 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*[74] and excluded BioProjects with inconsistent/unclear metadata, or those that remained as outliers in the PCA after batch-correction. *DESeq2*[75] was

used to perform a series of differential expression (DE) using the corrected transcript counts. These DE tests were performed individually for each BioProject, DE tests compared the differences in expression between all control and treatment samples in each BioProject, including treatment level 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*[76]. 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*[74] 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*[77]. 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*[78].

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[79,80]. All other figures were generated in R using ggplot2[81].

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[57]. 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. *WHO Fungal Priority Pathogens List to Guide Research, Development and Public Health Action.* (2022).
2. Denning, D. W. Global incidence and mortality of severe fungal disease. *Lancet Infect. Dis.* S1473-3099(23)00692–8 (2024) doi:10.1016/S1473-3099(23)00692-8.
3. Latgé, J.-P. & Chamilos, G. Aspergillus fumigatus and Aspergillosis in 2019. *Clin. Microbiol. Rev.* **33**, 10.1128/cmr.00140-18 (2019).
4. Kowalski, C. H. *et al.* Heterogeneity among Isolates Reveals that Fitness in Low Oxygen Correlates with Aspergillus fumigatus Virulence. *mBio* **7**, e01515-16 (2016).
5. Colabardini, A. C. *et al.* Heterogeneity in the transcriptional response of the human pathogen Aspergillus fumigatus to the antifungal agent caspofungin. *Genetics* **220**, iyab183 (2022).
6. Horta, M. A. C. *et al.* Examination of Genome-Wide Ortholog Variation in Clinical and Environmental Isolates of the Fungal Pathogen Aspergillus fumigatus. *mBio* **13**, e01519-22 (2022).
7. Keller, N. P. Heterogeneity Confounds Establishment of “a” Model Microbial Strain. *mBio* **8**, 10.1128/mbio.00135-17 (2017).
8. Steenwyk, J. L., Rokas, A. & Goldman, G. H. Know the enemy and know yourself: Addressing cryptic fungal pathogens of humans and beyond. *PLoS Pathog.* **19**, e1011704 (2023).
9. Lofgren, L. A., Ross, B. S., Cramer, R. A. & Stajich, J. E. The pan-genome of Aspergillus fumigatus provides a high-resolution view of its population structure revealing high levels of lineage-specific diversity driven by recombination. *PLOS Biol.* **20**, e3001890 (2022).
10. Barber, A. E. *et al.* Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection. *Nat. Microbiol.* **6**, 1526–1536 (2021).
11. Rhodes, J. *et al.* Population genomics confirms acquisition of drug-resistant Aspergillus fumigatus infection by humans from the environment. *Nat. Microbiol.* **7**, 663–674 (2022).
12. Hall, J. P. J., Harrison, E. & Baltrus, D. A. Introduction: the secret lives of microbial mobile genetic elements. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20200460 (2022).
13. Aubert, D., Naas, T., Héritier, C., Poirel, L. & Nordmann, P. Functional characterization of IS1999, an IS4 family element involved in mobilization and expression of beta-lactam resistance genes. *J. Bacteriol.* **188**, 6506–6514 (2006).
14. Krishnan, P. *et al.* Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. *BMC Biol.* **16**, 78 (2018).
15. Hottes, A. K. *et al.* Bacterial Adaptation through Loss of Function. *PLOS Genet.* **9**, e1003617 (2013).
16. Weisberg, A. J. & Chang, J. H. Mobile Genetic Element Flexibility as an Underlying Principle to Bacterial Evolution. *Annu. Rev. Microbiol.* **77**, null (2023).
17. Baquero, F. From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nat. Rev. Microbiol.* **2**, 510–518 (2004).
18. Benler, S. *et al.* Cargo Genes of Tn7-Like Transposons Comprise an Enormous Diversity of

Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *mBio* **12**, e02938-21 (2021).

19. Gluck-Thaler, E. *et al.* Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. *Mol. Biol. Evol.* **39**, msac109 (2022).
20. Urquhart, A. S., Vogan, A. A., Gardiner, D. M. & Idnurm, A. Starships are active eukaryotic transposable elements mobilized by a new family of tyrosine recombinases. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2214521120 (2023).
21. Gluck-Thaler, E. & Vogan, A. A. Systematic identification of cargo-mobilizing genetic elements reveals new dimensions of eukaryotic diversity. *Nucleic Acids Res.* gkae327 (2024) doi:10.1093/nar/gkae327.
22. Gourlie, R. *et al.* The pangenome of the wheat pathogen Pyrenophora tritici-repentis reveals novel transposons associated with necrotrophic effectors ToxA and ToxB. *BMC Biol.* **20**, 239 (2022).
23. Urquhart, A. S., Chong, N. F., Yang, Y. & Idnurm, A. A large transposable element mediates metal resistance in the fungus Paecilomyces variotii. *Curr. Biol.* **32**, 937-950.e5 (2022).
24. Urquhart, A. S., Gluck-Thaler, E. & Vogan, A. A. Gene acquisition by giant transposons primes eukaryotes for rapid evolution via horizontal gene transfer. 2023.11.22.568313 Preprint at https://doi.org/10.1101/2023.11.22.568313 (2023).
25. Bucknell, A. *et al.* Sanctuary: A Starship transposon facilitating the movement of the virulence factor ToxA in fungal wheat pathogens. 2024.03.04.583430 Preprint at https://doi.org/10.1101/2024.03.04.583430 (2024).
26. Pain, A. *et al.* 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 (2004).
27. Monod, M. *et al.* Virulence of alkaline protease-deficient mutants of Aspergillus fumigatus. *FEMS Microbiol. Lett.* **106**, 39–46 (1993).
28. Kowalski, C. H., Morelli, K. A., Stajich, J. E., Nadell, C. D. & Cramer, R. A. A Heterogeneously Expressed Gene Family Modulates the Biofilm Architecture and Hypoxic Growth of Aspergillus fumigatus. *mBio* **12**, e03579-20 (2021).
29. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
30. Lind, A. L. *et al.* Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol.* **15**, e2003583 (2017).
31. Abdolrasouli, A. *et al.* Genomic Context of Azole Resistance Mutations in Aspergillus fumigatus Determined Using Whole-Genome Sequencing. *mBio* **6**, e00536-15 (2015).
32. Ma, K. *et al.* Characterization and Biosynthesis of a Rare Fungal Hopane-Type Triterpenoid Glycoside Involved in the Antistress Property of Aspergillus fumigatus. *Org. Lett.* **21**, 3252–3256 (2019).
33. Brown, N. A. & Goldman, G. H. The contribution of Aspergillus fumigatus stress responses to virulence and antifungal resistance. *J. Microbiol.* **54**, 243–253 (2016).
34. Steenwyk, J. L. *et al.* Genomic and Phenotypic Analysis of COVID-19-Associated Pulmonary Aspergillosis Isolates of Aspergillus fumigatus. *Microbiol. Spectr.* **9**,

10.1128/spectrum.00010-21 (2021).
35. Raffa, N., Osherov, N. & Keller, N. P. Copper Utilization, Regulation, and Acquisition by Aspergillus fumigatus. *Int. J. Mol. Sci.* **20**, 1980 (2019).
36. Bignell, E., Cairns, T. C., Throckmorton, K., Nierman, W. C. & Keller, N. P. Secondary metabolite arsenal of an opportunistic pathogenic fungus. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20160023 (2016).
37. Abad, A. *et al.* What makes Aspergillus fumigatus a successful pathogen? Genes and molecules involved in invasive aspergillosis. *Rev. Iberoam. Micol.* **27**, 155–182 (2010).
38. Auxier, B. *et al.* Identification of heterokaryon incompatibility genes in Aspergillus fumigatus highlights a narrow footprint of ancient balancing selection. 2022.11.25.517501 Preprint at https://doi.org/10.1101/2022.11.25.517501 (2022).
39. Stroe, M. C. *et al.* Targeted induction of a silent fungal gene cluster encoding the bacteria-specific germination inhibitor fumigermin. *eLife* **9**, e52541 (2020).
40. Kowalski, C. H. *et al.* Fungal biofilm morphology impacts hypoxia fitness and disease progression. *Nat. Microbiol.* **4**, 2430–2441 (2019).
41. Rokas, A., Mead, M. E., Steenwyk, J. L., Oberlies, N. H. & Goldman, G. H. Evolving moldy murderers: Aspergillus section Fumigati as a model for studying the repeated evolution of fungal pathogenicity. *PLOS Pathog.* **16**, e1008315 (2020).
42. Shaw, L. P. *et al.* Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci. Adv.* **7**, eabe3868 (2021).
43. Shimizu, K. & Keller, N. P. 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 (2001).
44. Carter-House, D., Stajich, J., Unruh, S. & Kurbessoian, T. Fungal CTAB DNA Extraction. (2020).
45. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
46. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
47. Palmer, J. M. & Stajich, J. E. Automatic assembly for the fungi (AAFTF): genome assembly pipeline. Zenodo https://doi.org/10.5281/zenodo.7439973 (2022).
48. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. 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 (2021).
49. Palmer, J. M. & Stajich, J. Funannotate v1.8.1: Eukaryotic genome annotation. Zenodo https://doi.org/10.5281/zenodo.4054262 (2020).
50. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
51. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.

eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* (2021) doi:10.1093/molbev/msab293.

53. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).
54. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
55. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
56. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
57. Lofgren, L. A., Ross, B., Cramer Jr, R. A. & Stajich, J. E. Aspergillus fumigatus pangenome dataset. Zenodo https://doi.org/10.5281/zenodo.5775266 (2021).
58. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
59. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
60. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
61. Nattestad, M., Aboukhalil, R., Chin, C.-S. & Schatz, M. C. Ribbon: intuitive visualization for complex genomic variation. *Bioinformatics* **37**, 413–415 (2020).
62. Bowyer, P., Currin, A., Delneri, D. & Fraczek, M. G. Telomere-to-telomere genome sequence of the model mould pathogen Aspergillus fumigatus. *Nat. Commun.* **13**, 5394 (2022).
63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
64. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
65. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
66. Ayad, L. A. K., Pissis, S. P. & Polychronopoulos, D. CNEFinder: finding conserved non-coding elements in genomes. *Bioinforma. Oxf. Engl.* **34**, i743–i747 (2018).
67. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
68. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
69. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
70. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

71. Sahlin, K. Strobealign: flexible seed size enables ultra-fast and accurate read alignment. *Genome Biol.* **23**, 260 (2022).
72. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
73. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
74. Rezaie, N., Reese, F. & Mortazavi, A. PyWGCNA: a Python package for weighted gene co-expression network analysis. *Bioinformatics* **39**, btad415 (2023).
75. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
76. MetaVolcanoR. *Bioconductor* http://bioconductor.org/packages/MetaVolcanoR/.
77. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
78. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9**, ELIXIR-709 (2020).
79. Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* (2009) doi:10.1101/gr.092759.109.
80. Hackl T, Ankenbrand M, & van Adrichem B. gggenomes: A Grammar of Graphics for Comparative Genomics. (2023).
81. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-24277-4.
82. Filho, A. P. da C. *et al.* 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 (2020).

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 [20]). 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. S12A) have core gene transcript coverage greater than this threshold, compared to genes in A1163 (Fig. S12C) and CEA10 (Fig. S12B) *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. S11), suggesting that the transcriptional response of *Starship* captains within the infection environment differs between strains.

Specific treatments applied within a single study can have a contrasting influence on the prevailing expression patterns of captain genes. Expression of the *Tardis h1* captain gene in Af293 is generally decreased across nutrient supplementation studies compared to controls (Fig. 5B). Yet, supplementation of glucose and xylose has a positive and negative influence on captain expression, respectively, compared to controls with nutrient media [82]. Similarly, expression of CEA10 *Logos h1* and *h2* captain genes both decreased and increased, with supplementation of acetate (PRJNA668271), respectively, compared to nutrient media controls containing dilute acetate (0.1%) (Fig. 5D). Together, these gene expression datasets demonstrate the context-dependent expression of captain genes, which we speculate would lead to variation in *Starship* transposition rates across different environments and strain backgrounds.

Similar to captain genes, the expression patterns of cargo genes within *Starships* are heterogeneous. The transcriptional profile of *Starships* is punctuated with expression of cargo genes (Figure SH), which can have substantial transcript coverage even in the absence of captain expression, suggesting that the transcriptional regulation of these cargo genes are decoupled from transposition. Of particular note are the genes that form biosynthetic gene clusters (BGC) which reside within *Starships*. In Af293, 47 BGC genes across six Starships (*Gnosis h1, Lamia h4, Nebuchadnezzar h1, Nebuchadnezzar h2, Osiris h4, Tardis h1*) are DE in one or more of the treatment categories investigated in this study. At least one gene from the BGCs across these six *Starships* are DE in response to caspofungin exposure, from a specific Af293 study (PRJNA472460). Furthermore, multiple genes associated with a BGC present in Osiris h4 (AFUA_3G02580, AFUA_3G02600, AFUA_3G02620, AFUA_3G02630, AFUA_3G02640) have significantly increased expression compared to controls. Interestingly, generally fewer genes in BGCs are DE in Af293 “antifungal” or “nutrient” studies. Relatively fewer Starship BGC genes are DE in other strains. Notably, a collection of cargo genes from Fumihopaside A and terpene biosynthesis (CEA10-lr_006118) BGCs in CEA10 *Gnosis h2* have decreased expression during an *in vitro* infection of a human cell line. Cargo genes within *Osiris h4* that increased expression under both antifungal exposure and nutrient manipulation included multiple genes from the putative NR-PKS BGC (Fig. S13).

As a human pathogen, we were also interested in *A. fumigatus* genes that have specific relevance to virulence, such as genes within *hrmA*-associated cluster (HAC). Two HAC genes present in *Nebuchadnezzar h1*, *cgnA* and *bafA*, have near-zero transcript abundances within samples from an Af293 study of an experimental infection of a murine model (PRJNA421149). However, genes within the HAC cluster were only DE within certain nutrient and antifungal studies. Specifically, *cgnA* (AFUA_5G14910) and *hrmA* (AFUA_5G14900) both have increased expression within Af293 samples supplemented with calcium, and *bafA* (AFUA_3G04070) has decreased expression in CEA10 samples under caspofungin exposure (Fig. 5A).

Multiple DE cargo genes serve putative functions that are relevant for the metabolic or enzymatic activity of the cell. Collections of cargo genes within the Af293 *Starships Lamia h4* and *Osiris h4* have consistently increased expression from studies employing nutrient manipulation treatments

compared to controls (Fig. S13). This includes several genes in *Lamia h4* with a putative role in enzymatic secretion (e.g., glycosyl hydrolases, phosphotransferases, and dehydrogenases). Genes with domains relevant to dehydrogenases (NAD binding domains) were generally positive DEGs within antifungal and nutrient studies of Af293 and A1163, respectively.

Genes with a putative role in specific molecular functions were also consistently DE in *Starships*. We found that genes involved in zinc ion binding activity (IBR) (negative DEG in Af293 antifungal and positive DEG in CEA10 antifungal), dynamin-like GTPase domain and BTB (both highly negative DEGS in CEA10 *Logos h1* within infection studies). In Af293 *Tardis h1*, a gene with a putative role in structural molecular activity (domains for Cytochrome b5-like Heme/Steroid binding and flocculin; AFUA_8G06250) has increased expression in nutrient supplementation. Whereas HATPase_c/HisKA response regulator (AFUA_8G06140) and glycoside hydrolase/lipase (AFUA_8G06350) have decreased expression in nutrient supplementation in the same *Starship*.

Additional cargo genes of interest may convey resistance to environmental stressors. Two genes involved in arsenic detoxification, present within *Nebuchadnezzar h1*, were found to be DE in Af293 and CEA10 studies. Arsenate reductase (AFUA_5G15000) in *Nebuchadnezzar h1* was found to both have increased and decreased expression within the CEA10 and Af293 copies, respectively. Three additional genes within this pathway were found to be DE in response across “antifungal” or “infection” studies (Fig. 5B). Two of these genes in in Af293 *Nebuchadnezzar h1*, arsenate methyl-transferase (AFUA_5G15020) and arsenite efflux transporter (AFUA_5G15010), were found to have either increased or decreased expression levels under caspofungin exposure, respectively. The third, arsenite permease (AFUA_1G16100) has decreased expression in Af293 infection studies.

Starship co-expression networks

We performed weighted gene co-expression network analysis (WGCNA) using *PyWGCNA* to construct gene co-expression networks for each of the three *A. fumigatus* reference strains Af293, CEA10, and A1163. Overall, we found that genes within or between *Starships* are more strongly co-expressed together than with non-*Starship* genes, based on pairwise comparisons of the distribution of scores in the topological overlap matrices (TOM) (corrected p-values < 0.01 for all comparisons; Fig. S14). This suggests that a tighter transcriptional relationship exists between genes mobilized by *Starships* compared to co-expression with other genes in the genome.

Genes in the *A. fumigatus* Af293 co-expression network were resolved into 27 modules (Table S24). The majority of genes found within each module are not contained within a *Starship*, yet almost all modules contain one or more *Starship* genes, including those previously identified as DEGs (Table S24). Modules were selected for further analyses based on the correlation of their gene expression profiles (WGCNA eigengenes) and their association with a treatment category (“antifungal”, “infection”, or “nutrient”) or specific study (Fig. 5E & Fig. S15).

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 S25SV).

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 S25SV). 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 S25SV). 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.

Figure legends

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, color-coded according to phylogenetic clade (sensu Lofgren et al., 2022)[9]. 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. S4).

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 *A. fumigatus* 519 strain population (Table S16; full dataset in Fig. S6). 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 *A. fumigatus* 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.

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[39], shown inserted at 3 independent sites. B) *Starship Gnosis h2* carrying the BGC encoding the terpene secondary metabolite Fumihopaside A[32], shown inserted at 2 independent sites. C) a large region on Chromosome 3 previously identified as an idiomorphic BGC[30] 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 (annotated by two vertical gray bars). D) Eight *Starships* in the reference strains Af293 and CEA10 that all carry homologs of *biofilm architecture factor A* (*bafA*)[28]. *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[28]. 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_2FC) in black with standard error bars, whereas DE genes identified across multiple studies are represented with summarized log_2FC 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. S15).

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: 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.

Figure S3: Screenshots from the IGV genome browser showing deletions and/or movement of *Nebuchadnezzar h1* among different isolates of the same named strain sequenced by different research groups (regions depicted in red denote *Nebuchadnezzar h1*). A. Nanopore long-read sequences of strain CEA10-lr (generated in this study) mapped to the publicly available assembly of a different isolate of this same strain (accession: SAMN28487501; sequenced by a different research group). A zoom in of the genomic location of *Nebuchadnezzar h1* in the public genome is shown (contig accession: CP097568.1). Only a single long-read supports the presence of the *Starship* at this locus. All other reads indicate either a deletion (above) or presence elsewhere in the genome (below). B. Illumina short-reads of strain ATCC46645 sequenced by a different research group (accession: SRR7418935) mapped to the ATCC46645-lr genome sequenced in this study. A zoom in of the genomic location of *Nebuchadnezzar h1* in ATCC46645-lr is shown (contig accession: scaffold_15). Three short-read tracks indicate a deletion of the *Starship* (above). Note that more short reads are mapped than shown in the image. Track and color descriptions can be found in the IGV manual.

Figure S4: Donut charts summarizing the percentages of gene orthogroups in the core, accessory, singleton, and *Starship*-associated compartments of the *Aspergillus fumigatus* pangenome, derived from the 13 reference-quality *A. fumigatus* strains and the expanded set of 54 high and medium-confidence *Starships* (n = 1818 elements total; Table S11).

Figure S5: 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 the 20 high-confidence *Starships* in 519 *Aspergillus fumigatus* assemblies (Table S13). Each column represents a cargo gene. A) Results broken down by *Starship*, with columns arranged according to gene order within each *Starship*. B) Compiled results across all 20 *Starships*, with columns arranged according to the number of strains with BLASTp hits. Bars are colored according to the genomic location in which the BLASTp hits are found.

Figure S6: Barcharts summarizing the genotypes of segregating genomic regions associated with the 20 high-confidence elements in the *Aspergillus fumigatus* 519 strain population (Table S16). If an isolate did not have any Starships within a given region, it was assigned either an “empty” or “fragmented” genotype (Methods).

Figure S7: The putative target site of *Starship Tardis* (indicated with an arrow) occurs more often than you would expect by chance in the *Aspergillus fumigatus* genome. We estimated the copy numbers of all k-mers of length 10 in the Af293 reference genome and found that the k-mer of length 10 that corresponds to the putative target site of *Tardis* is present in high copy numbers that exceed the expected genome-wide frequency of k-mers of this length (>99.99th percentile).

Figure S8: *Starships* mobilize genes encoding diverse metabolic functions. Barcharts 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 S9: 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 S10: *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 S11: 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 S12: 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 S13: 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_2FC) in black with standard error bars, whereas DEGs identified across multiple studies are represented with summarized log_2FC 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 S14: 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"; $p < 0.05$ = "*", $p < 0.01$ = "**", $p < 0.001$ = "***", $p < 0.0001$ = "****").

Figure S15: Heatmap displaying the correlation values for eigengenes in each module and their association with samples from a specific study (labeled with BioProject IDs), or different levels of experimental treatment category across studies. Cells of the heatmap are coloured by the correlation value, as well as labeled in each cell. P-values from a pairwise comparison between observations (*pdist*) are shown in parentheses.

Supplemental table legends

Table S1: Genome assembly statistics for the 12 strains sequenced with Oxford Nanopore long-read technology

Table S2: Metadata for publically available *Aspergillus fumigatus* strains

Table S3: Manual annotation data of 86 Starship elements

Table S4: Metadata for all starfish-predicted Starships

Table S5: Sequence features of all starfish-predicted Starships in the 519 strain population

Table S6: Frequencies of the 20 high-confidence Starships in the 519 strain population using the expanded dataset

Table S7: Manual annotation of Starship elements in three *Aspergillus fumigatus* reference strains

Table S8: Pairwise comparisons between strains of SNP identity by state and Starship repertoire similarity

Table S9: Comparison of Starship coordinates with structural variants identified by Colabardini et al 2022 (doi: 10.1371/journal.pgen.1010001)

Table S10: Orthogroup frequencies in the 519 strain population

Table S11: Orthogroup frequencies in the 13 reference-quality strains

Table S12: BLAST recovery of Starship cargo genes from the 13 reference-quality strains

Table S13: BLAST recovery of Starship cargo genes from the 519 strain population

Table S14: Genotyping of genomic regions with segregating Starship insertions

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.

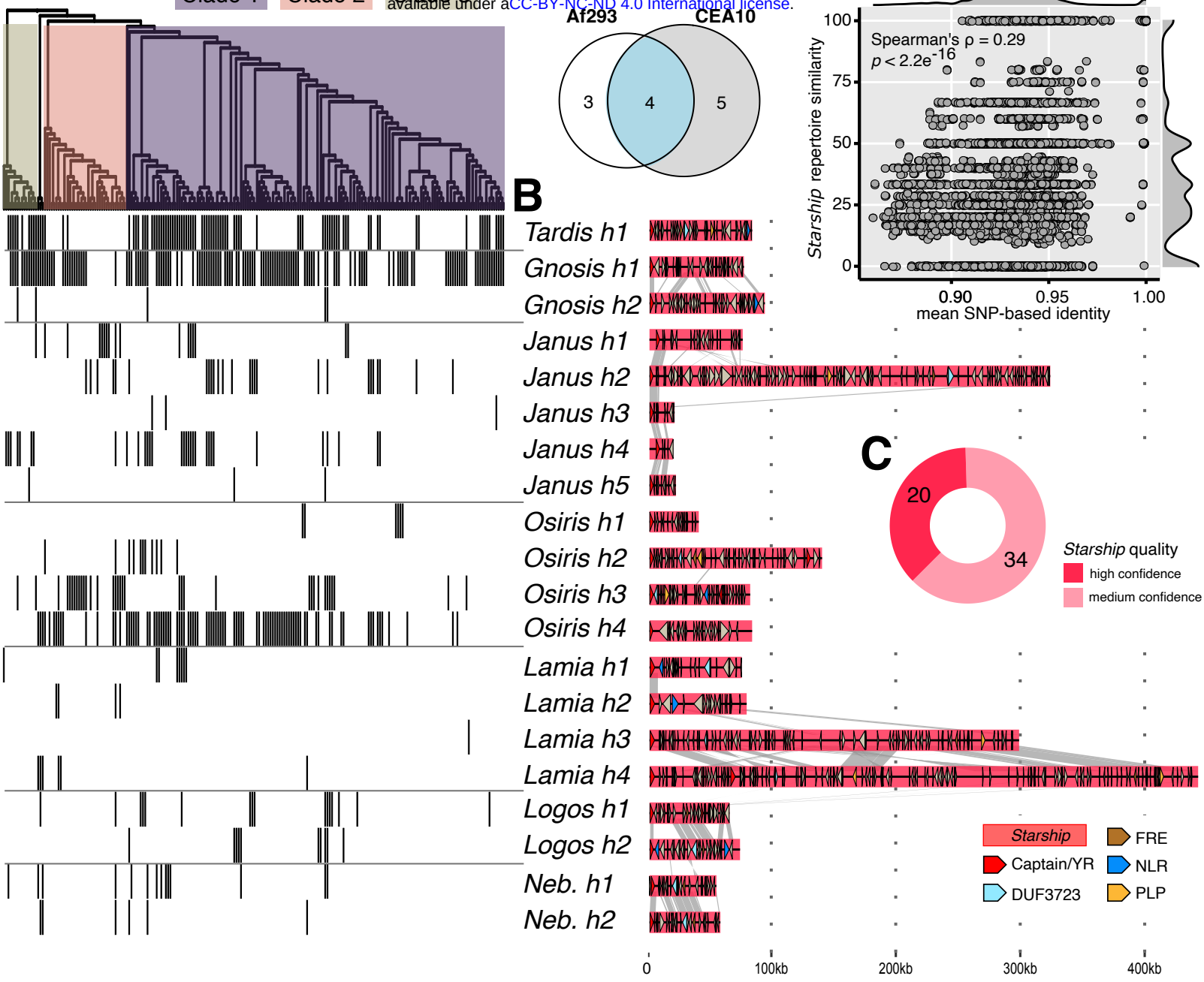

Figure 1: At least 20 distinct Starships carrying hundreds of protein-coding genes vary in their presence/absence across Aspergillus fumigatus strains. A) Top: a SNP-based maximum likelihood tree of 220 A. fumigatus strains, color-coded according to phylogenetic clade (sensu Lofgren et al., 2022)9. Bottom: 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).

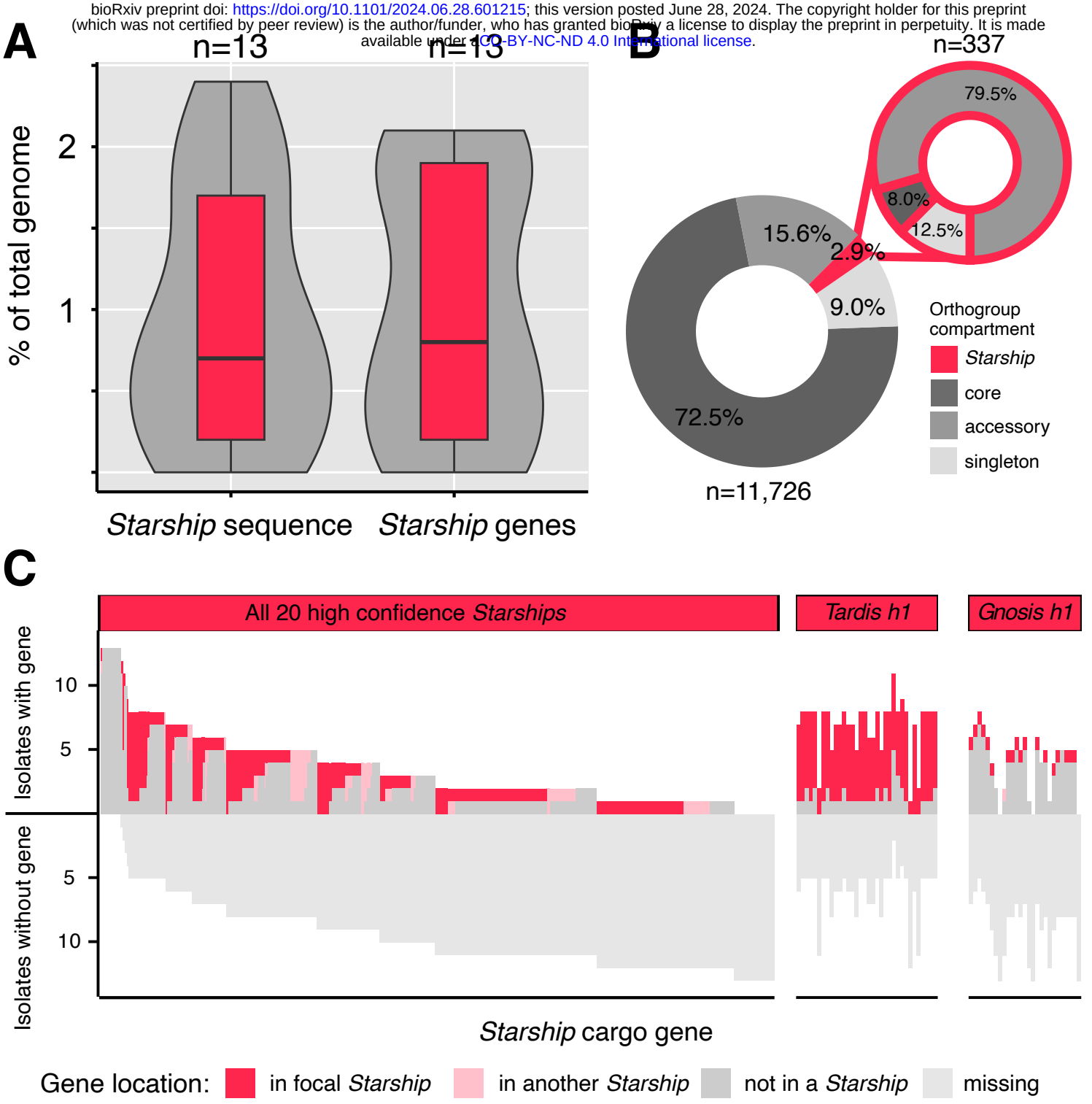

Figure 2: Starships are enriched in accessory genes whose presence/absence and genomic location vary across Aspergillus fumigatus strains. All panels visualize data derived from 13 reference-quality A. fumigatus assemblies and the 20 high-confidence Starships (n = 459 elements total). A) Box-and-whisker 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. S4).

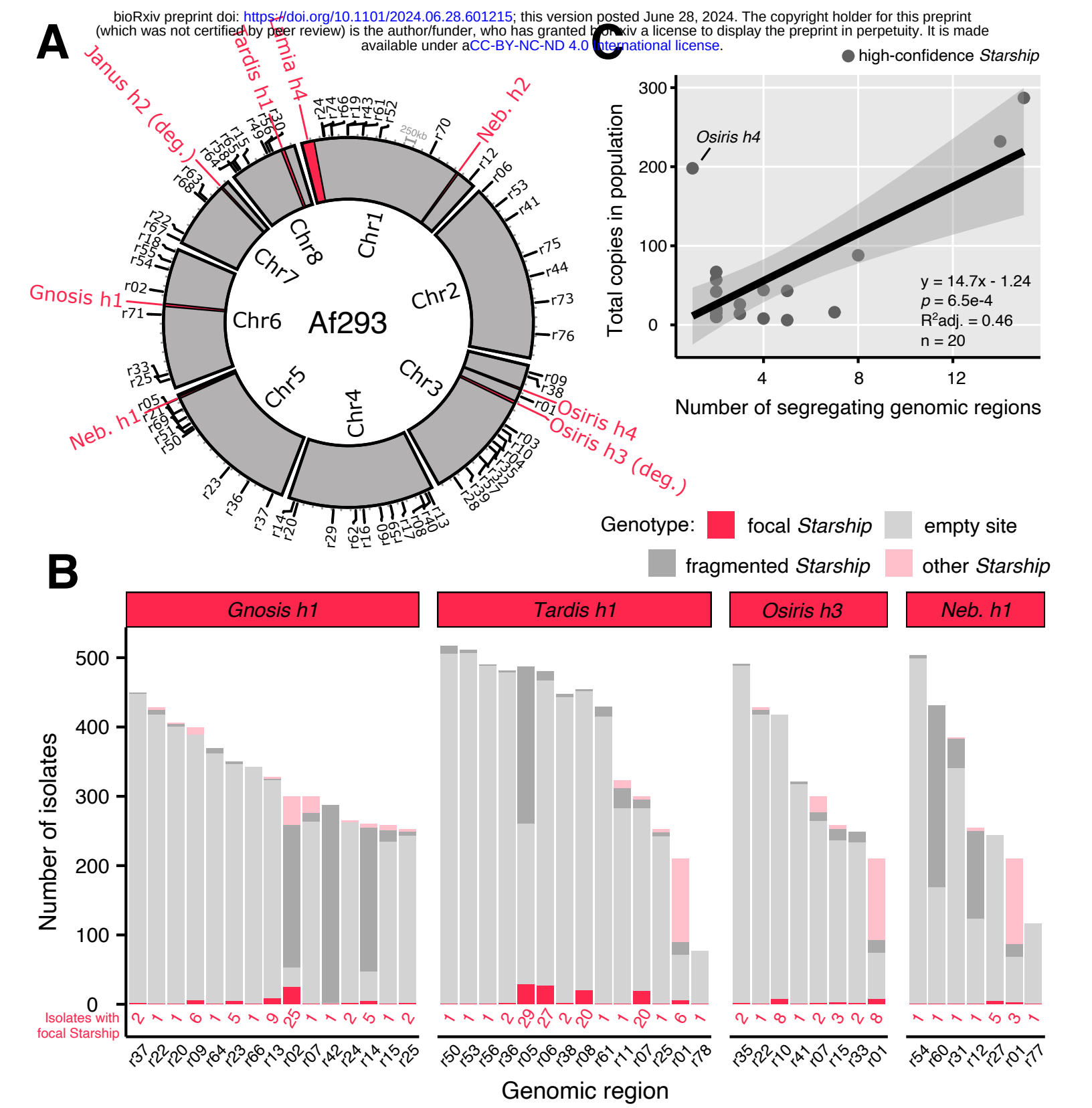

Figure 3: Starships and their insertion sites are distributed across all major chromosomes in the Aspergillus fumigatus genome. A) A Circos plot summarizing all inserted Starships (in red) and all genomic regions containing either an empty insertion site or a fragmented Starship (in black, along perimeter 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 A. fumigatus 519 strain population (Table S16; full dataset in Fig. S6). 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 A. fumigatus 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.

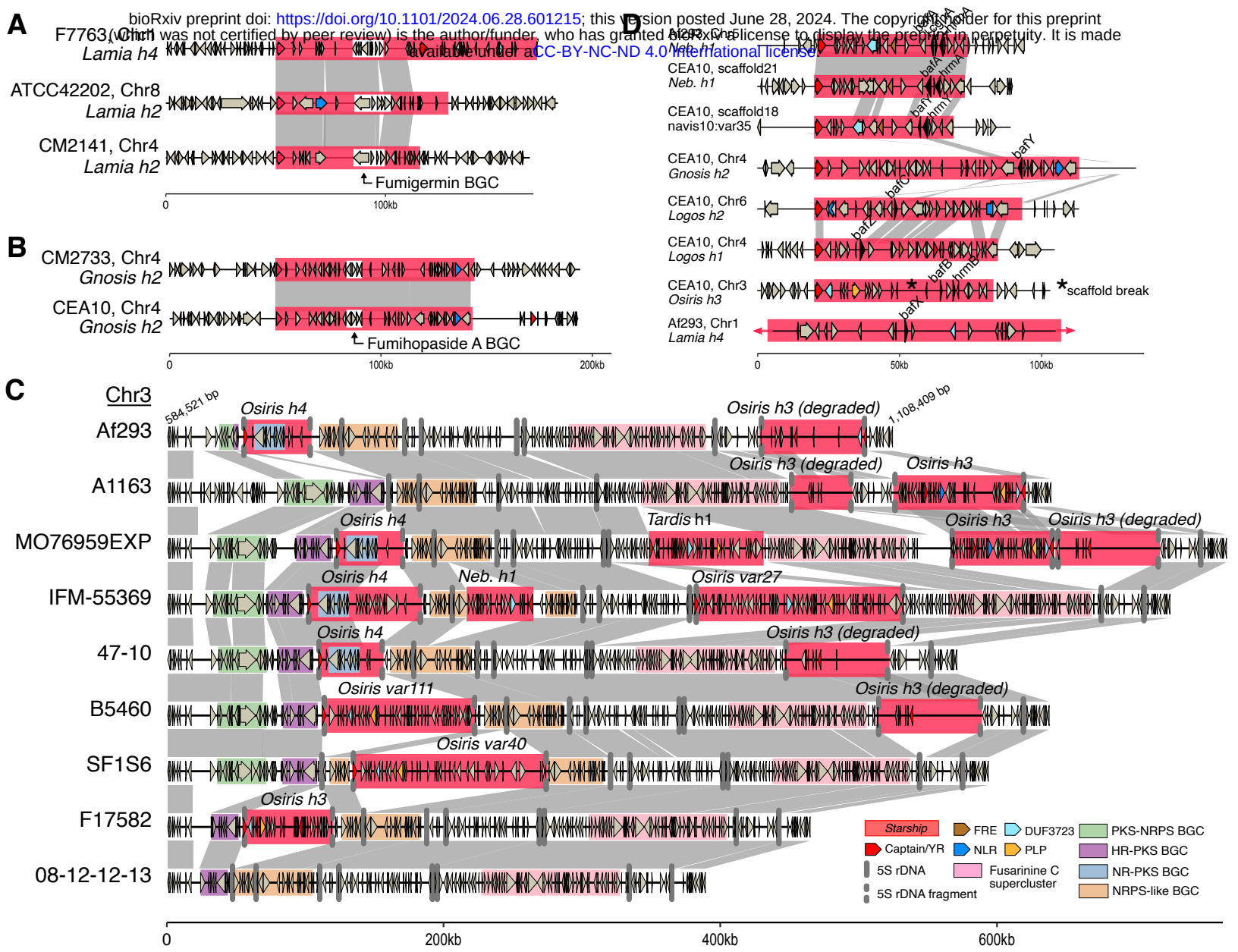

Figure 4: Starships mobilize adaptive traits and generate allelic diversity among Aspergillus fumigatus strains. Schematics and alignments of: A) Starship Lamia h2 and h4, which carry the biosynthetic gene cluster (BGC) encoding the polyketide secondary metabolite Fumigermin39, shown inserted at 3 independent sites. B) Starship Gnosis h2 carrying the BGC encoding the terpene secondary metabolite Fumihopaside A32, shown inserted at 2 independent sites. C) a large region on Chromosome 3 previously identified as an idiomorphic BGC30 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 (annotated by two vertical gray bars). D) Eight Starships in the reference strains Af293 and CEA10 that all carry homologs of biofilm architecture factor A (bafA)28. Starship Nebuchadnezzar h1 (Neb. h1) carries bafA as part of the HAAC (hrmA-associated gene cluster), while Starship navis10-var35 carries bafB as part of HBAC and Starship Osiris h3 carries bafC as part of HCAC28. 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).

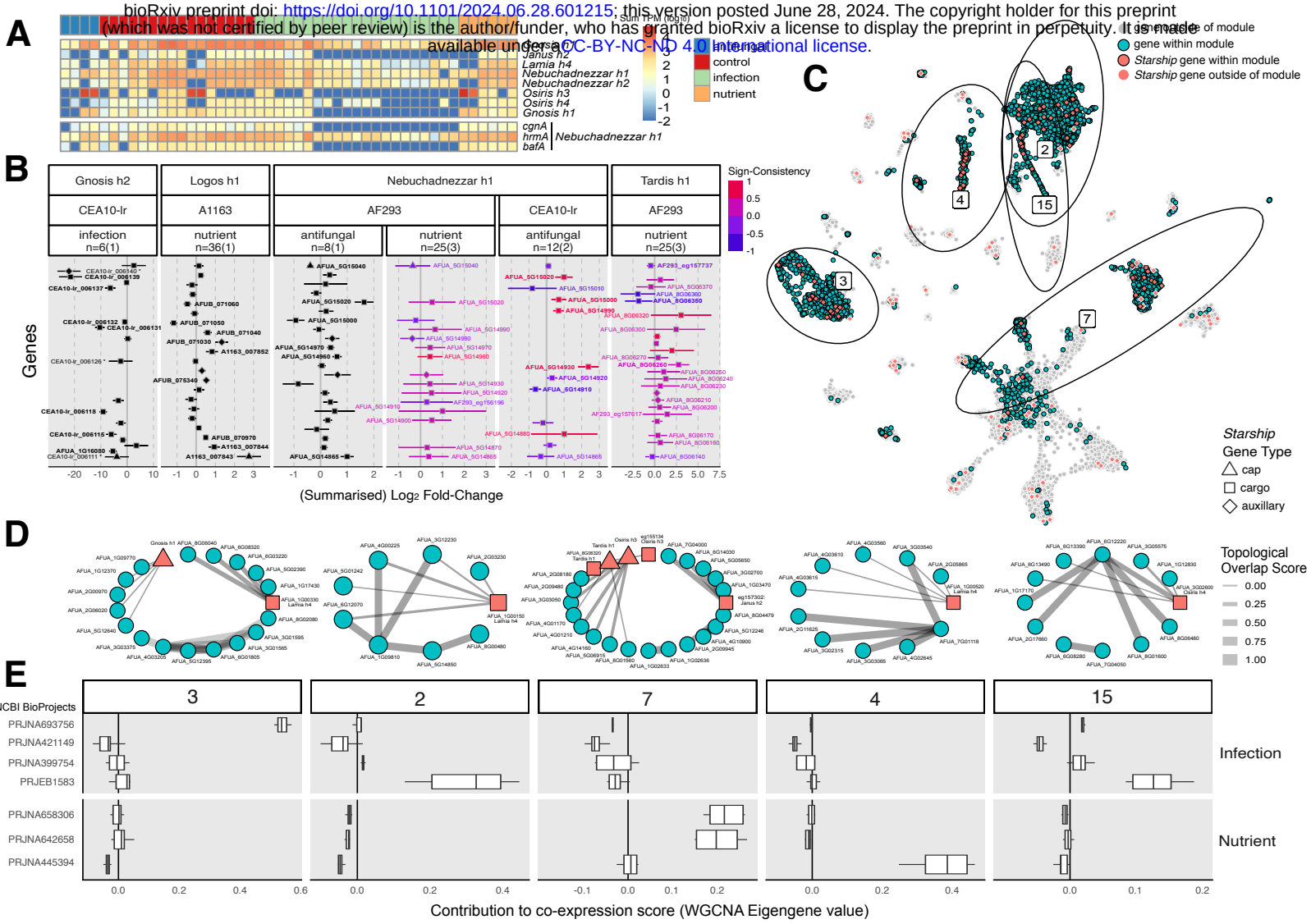

Figure 5: A) A heatmap of transcript abundances (log10TPM) of Starship captain tyrosine recombinase genes (top) and genes within the hrmA-associated cluster (HAC; bottom), collapsed across treatment replicates, for 11 RNAseq studies from A. fumigatus Af293. B) Results from differential expression (DE) 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 log2 fold-change (log2FC) in black with standard error bars, whereas DE genes identified across multiple studies are represented with summarized log2FC 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. S15).