



UPPSALA  
UNIVERSITET

UPTEC X 25029

Examensarbete 30 hp

Juni 2025

# Finding Starship Genetic Elements in Histoplasma Fungus

---

Gabriel Pettersson



UPPSALA  
UNIVERSITET

## Finding Starship Genetic Elements in Histoplasma Fungus

---

Gabriel Pettersson

### Abstract

*Starships* are a class of giant cargo mobilizing element found in fungi that are known to be capable of horizontal gene transfer. Due to the limited availability of high quality fungal genomes and the relative recency of their discovery, *Starships* have not yet been extensively documented in many fungal species known to contain them.

This project sought to address that knowledge gap in the human pathogen *Histoplasma capsulatum*. 79 assemblies from publicly available *H. capsulatum* DNA, including finished assemblies and reads that were assembled during the project, were analyzed using the Starfish pipeline, developed for identifying *Starships*. A total of 50 *Starships* were found in the *H. capsulatum* genomes and classified alongside previously known *Starships* from other fungi. While this study found no cargo genes that could be linked to the ability for *H. capsulatum* to infect and survive within a human host, a disproportionate number of the *Starships* identified contained several retrotransposon domains.

These domains are not commonly found in other *Starships*, and along with additional proof, such as an abundance of *Starships* with high AT content, suggest *H. capsulatum* *Starships* carry many smaller transposons and repetitive elements. This is not a common feature in the *Starships* currently documented in other fungal species and could provide future insights into the behavior of both *H.capsulatum* and *Starships* in general.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Aaron Vogan Ämnesgranskare: Fabien Burki

Examinator: Siv Andersson





# Själviska "Starships" av DNA kan vara nyckeln till sjukdomsorsakande svampar

Populärvetenskaplig sammanfattning

Gabriel Pettersson

All DNA kan anses vara "självisk" eftersom endast DNA som dupliceras och överlever kommer föras vidare i generationerna. Det finns dock DNA-fragment som anses vara "själviskt DNA" av forskare eftersom de inte kräver resten av sin organisms genetiska kod för att överleva. Exempelvis kan sådant själviskt DNA kopiera sig självt från en cell till en annan, vilket gör att den kan överleva vidare självständigt från sin ursprungliga cell. Majoriteten av DNA inom människor och djur är uppbyggt av dessa själviska fragment, som ofta är mycket små. Inom djur, svampar och växter har dessa fragment historiskt ansetts ha en negativ effekt på organismens överlevnad.

Nyligen har en ny variant av dessa fragment, "Starships", upptäckts inom svampar. *Starships* är speciella eftersom de är väldigt stora fragment, som ofta innehåller fungerande gener som gynnar organismen, inte bara fragmentet i sig. Exempelvis kan dessa fragment snabbt sprida gener som gör en svamp tolerant mot tungmetaller. Via *Starships* kan dessa svampar därför dela med sig gener som har en gynnsam effekt med andra svampar under en kort tid.

I denna studie undersöktes en av dessa svampar, *Histoplasma*, som vid inandning kan orsaka sjukdomen histoplasmos hos människor, något som kan vara livshotande för de som saknar ett starkt immunförsvar. Eftersom *Starships* tidigare visats innehålla viktiga gener för *Aspergillus*, en relaterad svamp som också kan smitta människor, handlade denna studie om att undersöka *Starships* inom *Histoplasma*. Via datorbaserade metoder söktes offentligt *Histoplasma* DNA för att hitta nya *Starships*. Dessa undersöktes sedan för att hitta vilka gener som fanns och hur de gick att relatera till *Starships* från andra svamparter.

I slutändan gick det inte att hitta någon koppling mellan *Starships* och förmågan att smitta människor inom *Histoplasma*. Dock hittades 50 nya *Starships*, varav många visade sig innehålla en stor mängd mindre själviska DNA fragment inuti sig. Detta är en intressant upptäckt, eftersom *Starships* inom andra svampar inte har lika mycket själviskt DNA inuti sig. Mitt projekt har därför bidragit till att utöka informationen om dessa själviska DNA fragment inom både svampen *Histoplasma* men också angående hur dessa fragment fungerar. Vidare forskning kan fortsätta leta efter gener inom dessa 50 *Starships* men också efter förklaringar till varför just dessa har så mycket själviskt DNA inuti sig.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>9</b>
1.1	Transposable elements	9
1.2	<i>Starships</i>	10
1.3	<i>Starfish</i>	12
1.4	<i>Starships</i> in Fungal Pathogens	14
1.4.1	<i>Aspergillus fumigatus</i>	14
1.4.2	<i>Histoplasma capsulatum</i>	14
1.5	Project Goals	15
<b>2</b>	<b>Materials and methods</b>	<b>16</b>
2.1	Materials	16
2.2	Assembly of SRA reads	16
2.3	Annotation	17
2.4	Finding <i>Starships</i>	17
2.5	Phylogenetics	18
2.6	Annotation of <i>Starship Cargo</i>	19
<b>3</b>	<b>Results</b>	<b>19</b>
3.1	50 <i>Starships</i> Were Identified from 79 <i>H. capsulatum</i> Assemblies	19
3.2	Phylogenetics	22
3.2.1	Phylogenomics of assemblies	22
3.2.2	Captains of novel <i>Starships</i> can be classified into distinct element families	22
3.3	Synteny	25
3.3.1	The “Enterprise” <i>Starships</i> are highly interconnected	25
3.4	Annotations	26
3.4.1	Lift-over annotations	26
3.4.2	InterProScan	26
3.4.3	Several elements have low GC-content	27
<b>4</b>	<b>Discussion</b>	<b>28</b>
4.1	Identified Elements	28
4.2	Phylogenies	29
4.3	Annotations	30
<b>5</b>	<b>Future Perspective and Conclusion</b>	<b>32</b>
5.1	Future Perspective	32
5.2	Conclusion	32

<b>6</b>	<b>Ethical Aspects and Conflicts of Interest.....</b>	<b>33</b>
<b>7</b>	<b>Acknowledgements .....</b>	<b>33</b>
	<b>References .....</b>	<b>34</b>
	<b>Appendix A – Figures.....</b>	<b>41</b>
	<b>Appendix B – Supplemental Material .....</b>	<b>43</b>

# Abbreviations

List of abbreviations:

CME(s)	Cargo mobilizing element(s)
DR(s)	Direct repeat(s)
HGT	Horizontal gene transfer
HMM	Hidden-markov model
MGE(s)	Mobile genetic element(s)
NCBI	National Center for Biotechnology Information
RIP	Repeat-induced point mutation
SRA	Sequence-read archive
TE(s)	Transposable element(s)
TIR(s)	Terminal-inverted repeat(s)
YR(s)	Tyrosine site-specific recombinase(s)

# 1 Introduction

## 1.1 Transposable elements

Mobile genetic elements (MGEs) are segments of DNA which encode enzymes and other proteins that mediate the transfer of genetic material, ranging from transfers within the same chromosome to transfers between different species (Hall *et al.* 2021). They are present in all life and in many forms. Examples of MGEs include plasmids, viruses and transposons (Frost *et al.* 2005). MGEs are the main source of horizontal gene transfer (HGT) throughout life and are therefore important both to our understanding of an organism's evolutionary history, but also our knowledge of pathogens, which often rely on MGEs to quickly transfer beneficial genes when present in a new environment (Frost *et al.* 2005, Hall *et al.* 2021).

One such class of MGEs are transposable elements (TEs). TEs are mobile segments in a genome which can reproduce and move to a different genomic locus once transcribed (Frost *et al.* 2005). They are generally considered “selfish”, since their ability to move between different loci and organisms means their evolutionary success is not linked to a single genetic lineage. Furthermore, their transposition can often be harmful toward the host's fitness, due to TEs being able to transpose inside an active gene, disabling it – or by increasing the genome's overall size, causing greater burden during replication. Eukaryotic TEs are generally divided into two main classes: retrotransposons (Class I) and DNA transposons (Class II). These are distinguished by their mechanism of operation, with Class I operating through a “copy and paste” mechanism, while Class II operates through a “cut and paste” mechanism (Makałowski *et al.* 2019). Retrotransposons transpose using a similar mechanism to retroviruses. When transcribed into RNA, an RNA-intermediate is produced, along with a reverse-transcriptase protein. The reverse transcriptase is then able to reverse-transcribe the RNA-intermediate back into DNA, integrating it into a different part of the genome. DNA transposons contain terminal inverted repeats (TIRs) at their flanks, which allows a transposase enzyme encoded by the transposon to bind to and excise the genetic element from its current location. These excised transposons can then reintegrate elsewhere in the genome. In total; TEs make up a significant portion of most eukaryotic genomes, with 85% of the maize genome composed of TEs (Schnable *et al.* 2009), and the human genome consisting of ~50% transposable elements (Lander *et al.* 2001) constituting much of what is commonly labeled as “junk DNA”. Most of these elements are vestigial, their ability to transpose eliminated by point mutations, other transposons or active host defenses such as repeat-induced point-mutations in fungi (RIP) and epigenetic barriers (Cambareri *et al.* 1989, Fouché *et al.* 2022).

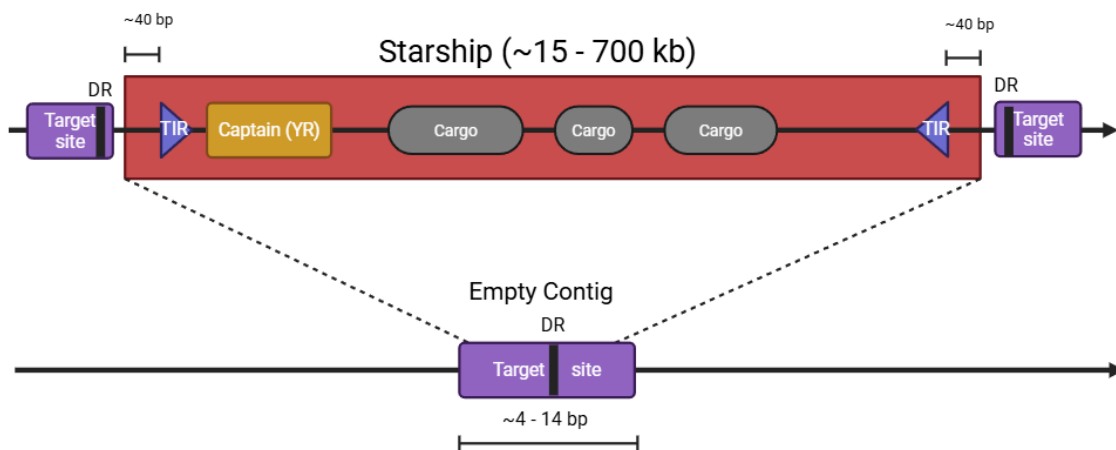
While TEs are selfish genetic elements, they are not always parasitic, with many providing benefits to the host genome along with being an important contributor to HGT. This is because TEs can transport genes which are not associated with their own transposition. If otherwise unrelated fragments of genomic DNA are positioned within the boundary of a TE, such as between the TIRs of a DNA transposon, those fragments will also be transposed through the TE. When such fragments are intact genes, they are considered “cargo”, with the TEs transporting them known as cargo-mobilizing elements (CMEs) (Arkhipova & Yushenova 2019). In bacteria, a well-documented variant of CMEs are integrative and conjugative elements (ICEs). These ICEs have been shown to contribute antibiotic resistance, genes which allow alternative metabolism and pathogenicity genes to their host bacteria (Johnson & Grossman 2015). In eukaryotes, such elements have historically proven rarer, with common examples being smaller retrotransposons (1-10kb) transducing genes to different hosts, or in DNA transposons such as “Pack-MULEs” in plants (<6kb) (Jiang *et al.* 2004, Arkhipova & Yushenova 2019). One of the reasons for this discrepancy is that the genomes of eukaryotes with a germline, such as humans, are believed to be much less susceptible to HGT by MGEs from other genomes, since only a tiny number of cells are evolutionarily relevant in such organisms (Andersson *et al.* 2001). Furthermore, the lack of high-quality genome assemblies, i.e. those generated with long-read sequencing technology, also contributed towards the difficulty in finding large eukaryotic ICEs. Genomes assembled using short-read, i.e. Illumina, while high coverage, are often not contiguous enough to allow for identification of larger TEs in eukaryotes. With greater availability of long-read sequencing technologies such as PacBio and Oxford nanopore, several new CMEs have been uncovered in eukaryotes, along with proof of their HGT (Arkhipova & Yushenova 2019).

## 1.2 *Starships*

*Starships* are a new class of DNA transposons which are present in several species of fungi, namely the Pezizomycotina subphylum and the Basidiomycetes phylum (Gluck-Thaler & Vogan 2024, Urquhart A *et al.* 2024). The first two *Starships*; *Enterprise* and *HEPHAESTUS*, were discovered independently in *Podospira anserina* and *Paecilomyces variotii*, respectively (Vogan *et al.* 2021, Urquhart *et al.* 2022). The *Enterprise* in *P. anserina* carries a block of selfish spore killing (*Spok*) genes, which produce a protein which kills any spore not carrying the *Spok* genes (Vogan *et al.* 2019). By comparison, *Starship HEPHAESTUS* found within *P. variotii*, carries a package of cargo-genes for zinc, lead, cadmium, and copper resistance, allowing strains of *P. variotii* to quickly colonize environments which would otherwise be toxic to them (Urquhart *et al.* 2022, Urquhart *et al.* 2023).

The genomic structure of a basic *Starship* can be seen illustrated in Figure 1. *Starships* are characterized by their massive size - which generally ranges between 15kb to 700 kb and are on average ~100kb long (Gluck-Thaler *et al.* 2022). It is this massive size that allows the *Starships* to transport entire packages of functioning genes, such as *Starship HEPHAESTUS*

in *P. variotii* (Urquhart *et al.* 2022). However, size is not the main distinguishing feature of a *Starship*, that is the “captain” gene. All *Starships* transpose with the help of a gene which encodes a tyrosine site-specific recombinase protein (YR), which in *Starships* is referred to as a “captain”. YRs are involved in the movement of several MGEs, such as many plasmids and other eukaryotic TEs such as *Cryptons*. In *Starships*, all YRs contain a DUF3435 protein domain and are encoded at the 5’-end of the *Starship*. This is the main method of identifying the captains, since captains are otherwise very diverse, with the median amino acid identity within the same *Starship* family only ~35% (Gluck-Thaler & Vogan 2024). The previously mentioned *Enterprise* and *HEPHAESTUS* have only 15% amino acid identity between their captains despite the captains being the most conserved portion of the *Starship* (Urquhart *et al.* 2023). DUF3435 is the most conserved YR domain and is therefore used to identify captains. Furthermore, all *Starships* contain direct repeats (DRs) at both flanks which are in the 4-14 bp range and can be used to help identify the boundaries of the element. Many *Starships* also possess, in addition to the flanking DRs, short terminal inverted repeats (TIRs) at the flanks (Urquhart A *et al.* 2024).



**Figure 1:** A typical embedded *Starship*. It contains a captain (YR) gene, several cargo genes, flanking TIRs and DRs. The red rectangle outlines the *Starship* boundary, with the target site flanking it on either side. Also illustrated is an empty contig containing a target site, which typically falls within the 4-14 bp range. The figure is not to scale, only the areas marked with a scale bar should be considered to have a specific size.

More recent studies on *Starships* have developed a method of systematically describing *Starship* diversity through three taxonomic ranks (Gluck-Thaler & Vogan 2024). All *Starships* belong to the *Starship* superfamily, themselves a part of the Class II transposons. Each *Starship* is first assigned one of 11 families based on the similarity of its YR against a library of known *Starship* YRs. These 11 families were themselves constructed by clustering known YRs into monophyletic families based on phylogenetic similarity. The phylogenetic comparison was done through alignment of the two main functional domains in the YR, the

core binding and catalytic domains. Secondly, since *Starship* captains have high diversity even within the same families (~35% aa similarity), they are further grouped into separate “naves” (ships in latin, “navis” singular) based on the orthologous relationship between the captain and other captains in the same family. Finally, since *Starships* can belong to the same navis, but contain different cargo genes, each *Starship* is assigned a unique haplotype based on sequence similarity scores. Ideally, each unique family, navis and haplotype combination would therefore describe a specific TE with a specific captain, mobilizing a specific combination of cargo genes. If two *Starships* of the same family, navis and haplotype are found at different loci, whether in the same organism or not, it is likely that it has transposed.

Considerable bioinformatic evidence of past *Starship* HGT between different fungal species exists (Urquhart *et al.* 2023, Gluck-Thaler & Vogan 2024). For example: BLAST searches between a strain of *Aspergillus fumigatus* and *P. variotii* produced a 96% nucleotide identity match between genes present in the *Galactica* family *Starship* contained in both strains (the average nucleotide identity between genes in the two strains was 72.5%). However, the mechanisms that trigger and regulate *Starship* transposition are currently unknown and are currently being researched by the Vogan Lab, with only the *Starships* HEPHAESTUS and Pegasus proven to transpose *in vitro* (Urquhart *et al.* 2023, Urquhart *et al.* 2025).

### 1.3 *Starfish*

Because *Starships* lack many of the traditional markers distinguishing them as TEs, it is difficult for traditional annotation software to correctly identify them. In response, the *Starfish* pipeline was created, which combines several existing software and methods to obtain a *de novo* annotated list of candidate captains and *Starships* when provided with multiple genome assemblies (Gluck-Thaler & Vogan 2024). The pipeline contains several commands which adapt existing bioinformatic software to the purpose of finding *Starships* and captains.

The first is *Starfish annotate*, which conducts a targeted *de novo* annotation to find protein coding sequences based on a preconstructed HMM file. It first runs *Metaeuk easy-predict* (Levy Karin *et al.* 2020) to annotate fungal amino acid sequences. From the identified sequences, it then runs *HMMsearch* (Eddy 2011) to find only the predicted sequences which match an existing HMM of known *Starship* YR sequences – or other cargo genes with an existing HMM profile. These *de novo* annotations can then be matched to existing annotations provided through a gff3 file using the *Starfish consolidate* command. The obtained coordinates of putative YR genes are then passed to *Starfish sketch*, which identifies genomic neighborhoods containing YR genes. These neighborhoods are mutually exclusive and help prevent false-positives and overlapping *Starships* by grouping nearby captains into the same neighborhood, instead of producing two separate elements because there are two captains. A neighborhood can therefore be considered a “proto-*Starship*” and it can contain several YRs within.

A BED file containing the coordinates of each neighborhood is parsed to *Starfish insert*, which conducts BLASTn alignments between the captains in the identified neighborhoods and the existing genome assemblies, to find the boundary for each element (Camacho *et al.* 2009). *Starfish insert* outputs another BED file, containing the coordinates for each putative *Starship* and its captain. These coordinates can then be input to *Starfish flank*, which uses *CNEFinder* to find *Starship* DRs and TIRs around the upstream and downstream sequences of each putative *Starship* identified through *Starfish insert* (Ayad *et al.* 2018). Since *insert* can often produce several different element boundaries for a single YR, *Starfish summarize* is used to consolidate the results from *insert* and *flank*, selecting the downstream and upstream boundary that produces the longest *Starship* and best corresponds with the identified DRs and TIRs. Finally, it outputs a BED file containing these *Starships* and annotations for all putative genes within.

*Starfish dereplicate* categorizes putative elements into homologous regions, based on genes found upstream and downstream of the *Starship* boundary using *MMseqs2 easy-cluster* (Hauser *et al.* 2016). These homologous regions are then used to find empty and/or fragmented regions which contain flanking genes orthologous to those in the homologous region, but without the *Starship*. This allows both a greater degree of certainty that an identified *Starship* is real and helps identify if the *Starship* is unique, and not a duplicate. It outputs a list of homologous regions and the elements and captains belonging to each region, it also filters captains based on their navis and haplotype. *Dereplicate* requires information on ortholog groups, obtained through software such as *Orthofinder* and *EggNOG-mapper*.

Finally, there are several auxiliary commands which aid in both the management of files and visualization of results. *Pair-viz* aligns *Starships* against potential empty insertion sites present in the input assemblies using *Nucmer* (Marçais *et al.* 2018) before running *dereplicate*. These alignments are then visualized using *Circos*, which allows the user to manually compare a contig containing a putative *Starship* with a contig containing an empty candidate insertion site (Krzywinski *et al.* 2009). If the alignment is poor, then it is likely the *Starship* is a false positive and can be manually filtered out. *Starfish sim* works similarly to *dereplicate* but does not require ortholog groups. It employs *Sourmash* to compare *Starship* sequences based on k-mer similarity (Pierce *et al.* 2019). The k-mer similarity scores can then be used to group *Starships* into unique navis-haplotype groups using *Starfish group*, which uses *mcl* (Enright *et al.* 2002). While these commands do help filter out duplicates and assign each *Starship* a unique navis-haplotype, *sim* and *group* are unable to output a list of homologous regions containing *Starships*, as there is no ortholog information input. *Locus-viz* takes as input either the regions output by *dereplicate* or a custom list of elements. It aligns the input elements using *Nucmer* and produces a list of inputs required to run *gggenomes* on R, which produces a synteny plot comparing elements (Hackl *et al.* 2024).

## 1.4 *Starships* in Fungal Pathogens

There are several fungal pathogens inside the peizizomycotina subkingdom, which is where the majority of currently documented *Starships* are present (Gluck-Thaler & Vogan 2024). Examples include *Aspergillus fumigatus* and *Histoplasma capsulatum*, which target humans and are both considered critical and high priority fungal pathogens by the World Health Organization (WHO) (World Health Organization 2022).

### 1.4.1 *Aspergillus fumigatus*

*A. fumigatus* infects humans through its spores, causing aspergillosis. This infection targets the respiratory system but can also target the nervous system. It is especially dangerous to immunocompromised individuals or those with previously existing lung conditions, such as asthma (WHO 2022). A recent study conducted on 519 strains of *A. fumigatus* uncovered 20 new high confidence *Starships* which contained several cargo genes encoding traits known for pathogen survival and virulence in fungi (Gluck-Thaler *et al.* 2024). Examples of such genes include the *HAC*-cluster, genes which together increase virulence, growth in low oxygen environments and assists in biofilm development, which is crucial for infection. Furthermore, the study found evidence of *Starships* being a major contributor of strain heterogeneity in *A. fumigatus*. Strain heterogeneity is a major confounding factor for combatting *A. fumigatus*, as strains display high levels of variation in traits such as virulence, antifungal resistance and metabolism. Antifungal resistance being especially concerning to the WHO due to *A. fumigatus* strains becoming increasingly resistant to Azoles, the main antifungal used to treat aspergillosis, especially in low-income countries (WHO 2022). Approximately 16% of the accessory genes which differ between *A. fumigatus* strains are transported as cargo by *Starships*, with 92% of *Starship* genes in *A. fumigatus* estimated as accessory, compared to the 24.6% rate in the rest of *A. fumigatus*. These findings suggest that *Starships* could be a notable source of strain heterogeneity in fungal pathogens (Gluck-Thaler *et al.* 2024), which is comparable to that of MGEs in bacteria which are well-known and documented. Furthermore, they show that genes involved in pathogenicity and infection are present in *Starships* and could therefore be equally present in *Starships* of other fungal pathogens.

### 1.4.2 *Histoplasma capsulatum*

*H. capsulatum* is an opportunistic pathogen, which means it does not depend on infecting humans to spread and is generally not adapted toward defeating the human immune system. *H. capsulatum* is a thermally dimorphic fungus which can infect the lungs through inhalation of spores or mycelial fragments, which causes histoplasmosis in humans (WHO 2022). In immunocompromised patients, histoplasmosis can often prove fatal, with a mortality rate of 21 to 53% in HIV/AIDS patients and 9 to 11% in immunosuppressed patients (WHO 2022).

Furthermore, *H. capsulatum* strains are also becoming increasingly resistant to antifungals, though the severity is lower compared to that in *A. fumigatus*, making it more difficult to protect patients with acute histoplasmosis.

Once inside the lungs, *H. capsulatum* transitions to its yeast phase at 37°C, which allows rapid growth and expression of crucial virulence factors. When *H. capsulatum* yeasts are phagocytized by macrophages, they can survive destruction due to proteins which prevent lysosomal fusion and disable several macrophage signals, preventing lysis and further immune responses (Valdez *et al.* 2022). Safe inside a macrophage, *H. capsulatum* multiplies and eventually kills the macrophage, allowing it to spread. Furthermore, through macrophages, *H. capsulatum* can reach the tissues and organs across the body, including bypassing the blood-brain barrier and infecting the brain. However, *H. capsulatum* is unable to survive against other phagocytes, such as neutrophils and dendritic cells, which instead destroy the yeasts within hours. The reason for this difference is that non-macrophage phagocytes employ different chemicals to destroy phagocytized cells which *H. capsulatum* is not resistant towards. Furthermore, neutrophils have been shown to create extracellular traps, which creates an environment hostile to fungal infection (Valdez *et al.* 2022). Therefore, in most immunocompetent individuals, *H. capsulatum* is neutralized before it can infect many macrophages and manifest serious symptoms, usually resulting in lenient pneumonia requiring little medication.

Currently, 37 *Starships* have been bioinformatically identified in *H. capsulatum* and are available in Starbase (<https://starbase.serve.scilifelab.se>), with 7 of those having been manually curated. These ships were obtained from a limited sample of *H. capsulatum* assemblies, making it likely more *Starships* exist in the wider *H. capsulatum* pangenome. Furthermore, their cargo genes have not yet been analyzed, meaning there currently exists little information concerning the contribution of *Starships* to *H. capsulatum* strains.

## 1.5 Project Goals

The objectives of this project are to identify novel *Starship* elements inside the pangenome of *H. capsulatum* with the aim of expanding the known *Starship* tree of life and identifying putative cargo genes which could be involved in the infection of humans by *H. capsulatum*. This will be accomplished by bioinformatically analyzing a large set of publicly available *H. capsulatum* genomes and assembling new genomes from short-read samples from the SRA, with the aim of covering as much of the *H. capsulatum* pangenome as reasonable within the available timeframe. The *Starships* will be identified using the *Starfish* pipeline. To identify their placement within the *Starship* tree of life, putatively identified *H. capsulatum* captains will be phylogenetically compared with captains of known high-confidence *Starships*. Furthermore, any novel *Starship* will be searched against fungal genomes within the pezizomycotina, with the aim of finding proof of past horizontal gene transfer. Finally, *de*

*novo* annotations will be made on the identified elements to find any genes or domains which could potentially assist *H. capsulatum* in infecting humans.

These analyses will provide an increased understanding of *Starships* within *H. capsulatum* and will ideally also assist future research in identifying how *H. capsulatum* infectivity works on a genetic level, providing clues on how to prevent and cure future infections.

## 2 Materials and methods

Python and bash scripts were written with assistance from Chatgpt (OpenAI, 2023, default free model versions for January – April 2025). All code has been manually verified and all parameters for software such as BLASTn was chosen manually without input from Chatgpt.

### 2.1 Materials

Genetic material from *H. capsulatum* was sampled from public databases with the aim of obtaining a diverse library of *H. capsulatum* samples to estimate its pangenome. 14 finished *H. capsulatum* genomic assemblies were downloaded from the NCBI genome database, constituting all available *H. capsulatum* assemblies in the NCBI as of the writing of this report (May 2025). These are referred to as the “NCBI assemblies”. 9 of those assemblies contained existing annotations, which were also downloaded. Additionally, 344 samples of raw DNA nucleotide reads were downloaded from the sequence read archive (SRA) using the Nextflow nf-core pipeline fetchngs 1.12.0 to streamline the downloading process and obtain all metadata from the samples (Ewels PA *et al.* 2020, Harshil Patel *et al.* 2024). These samples included 7 long-read samples and 337 short read samples. 6 of the long-read samples were sequenced using Oxford Nanopore and 1 was sequenced using PacBio technology. The short read samples included reads sequenced using Illumina and Capillary technologies. Because the project prioritized genetic diversity, transcriptomic material was not used, as most of the available DNA samples lacked a corresponding transcriptome.

### 2.2 Assembly of SRA reads

The quality of the 344 read samples was assessed using FastQC 0.11.8 and MultiQC 1.27 (Andrews 2010, Ewels P *et al.* 2016). All samples sequenced using capillary technologies were removed, as their coverage was determined to be too short. Furthermore, several other samples were removed, as they belonged to modified or mutant strains which do not reflect the wildtype *H. capsulatum* pangenome. The remaining short read samples were trimmed using trimmomatic 0.39 (Bolger *et al.* 2014) with the TruSeq3 and TruSeq2 primers selected, along with a sliding-window of 4:20 and a minimum length of 50 bp. None of the unpaired reads were kept after trimming, as they were considered too short. Before assembly,

additional filtering was done on the trimmed short reads to reduce assembly duration. To remove as many samples as possible, while retaining much of the genetic diversity present, each sample was sorted by its experiment and strain. If more than 5 samples shared the same experiment title and strain, only the five longest samples were kept. This was done to remove many of the samples which were simply separate runs of the same experiment, using the same strain. Finally, 7 short-read samples were not kept for assembly as they shared a corresponding long-read sample, making their assembly superfluous.

The remaining 59 short-read samples were assembled with SPAdes 4.0.0 using the default parameters for paired-end assembly (Prjibelski *et al.* 2020). The 7 long-read samples were assembled using Flye 2.9.5, with the parameters *pacbio-raw* and *nano-raw* used for assembling the PacBio and Nanopore reads, respectively (Kolmogorov *et al.* 2019). The long-read assemblies were then polished with their corresponding trimmed illumina short-reads using Pilon 1.24 (Walker *et al.* 2014). Both short-read and polished long-read assemblies were evaluated using QUAST 5.3.0 and summarized with MultiQC (Gurevich *et al.* 2013). From the assembly evaluations, it was determined that the PacBio assembly was too contaminated to be of use, as the assembly was 287 Mbp long, while the genome of *H. capsulatum* is on average 35 Mbp long.

## 2.3 Annotation

Annotation of the 65 *de novo* assembled genomes and the 5 NCBI assemblies lacking annotations was done using lift-over annotation instead of *de novo* annotation, which utilize existing annotations from similar assemblies to annotate. This was done since *de novo* annotation of the genomes was estimated to take too much time and computational resources, when there already existed several annotations for *H. capsulatum* which could be used for lift-over annotation. Lift-over annotation was done using Liftoff 1.6.3 with default parameters (Shumate & Salzberg 2021). The reference genome employed for lift-over annotation had the GenBank ID GCA\_017310585.1, known “hiscap2” in the dataset, and was chosen for its recent assembly, good assembly quality, and good annotation coverage. EggNOG mapper 2.1.12 was used on all 79 assemblies to annotate orthologous genes, with default genomic parameters used (Huerta-Cepas *et al.* 2019, Cantalapiedra *et al.* 2021).

## 2.4 Finding *Starships*

The *Starfish* auxiliary command “format” was used to reformat the FASTA assembly files and the gff3 liftoff annotation files to conform to a consistent naming format that can be used by *Starfish*. The pipeline was then run on the 79 assemblies to find candidate YR genes and elements. In total, 200 putative elements were found after *Starfish summarize* was finished. The *Starship* candidates were aligned to putative insertion sites present in the 79 input genomes using *Starfish pair-viz*, with the alignment visualized using Circos (Krzywinski *et al.*

2009, Katoh & Standley 2013). The *circos* plots for each *Starship* were manually inspected to filter out potential false-positive *Starships*. The final part of the *Starfish* pipeline, *dereplicate*, was unable to run alongside the liftoff annotations and *EggNOG* orthologs, with attempts to reformat the annotation and *EggNOG* files proving insufficient to address the issues. Instead, the *Starfish* auxiliary command “sim” was used. This allowed for the manual removal of duplicate *Starships* found earlier without using *dereplicate* and ortholog group information. The remaining dereplicated elements were visualized using *Starfish locus-viz*, with the synteny plots manually constructed to compare elements which belonged to the same family, or with phylogenetic similarity between captains. These were visualized using *gggenomes* 1.0.1 (Hackl *et al.* 2024).

To determine if any of the identified *Starships* had historically done horizontal gene transfer, a limited BLASTn 2.16.0 search was set up (Camacho *et al.* 2009). A custom database was constructed, containing the 241 *Starships* which had been manually curated and dereplicated from the Starbase database, accessed April 29 2025. This included 7 *Starships* from *H. capsulatum* which had been previously identified. A minimum percent identity of 95 was chosen for the search, along with a minimum word size of 100, to minimize the number of matches which were too short to be considered significant.

## 2.5 Phylogenetics

Phylogenetic analyses were done to both investigate the relationship between the 79 assemblies used, and to place the 50 identified captains within the *Starship* tree of life.

BUSCO 5.8.3 (Manni *et al.* 2021) was used with the *onygenales\_odb10* dataset on the 79 *H. capsulatum* assemblies used in *Starfish* along with 6 *Blastomyces dermatidis* assemblies downloaded from NCBI, to serve as an outgroup. *B. dermatidis* was chosen to be the outgroup due to being a closely related species with available long-read assemblies. The BUSCO results were directly input to the BUSCO phylogenomics pipeline (McGowan 2024) which identifies the BUSCO genes shared by all assemblies and produces a concatenated supermatrix of genes which can be used to generate a phylogenomic tree. BUSCO phylogenomics was used with default parameters, with trimming set to “trimAl gappyout”. Gappyout was chosen due to it being a conservative trimming approach that preserves much of the alignment information for tree construction (Tan *et al.* 2015). A maximum likelihood tree was constructed from the concatenated alignment using IQ-TREE2 2.4.0 with 1500 iterations of ultra-fast bootstrap with the *bnni* parameter, also known as UFBoot2+NNI (Hoang *et al.* 2018, Minh *et al.* 2020). ModelFinder was used with IQ-TREE to find the best tree model for each partition in the concatenated alignment (Chernomor *et al.* 2016, Kalyaanamoorthy *et al.* 2017).

The captain amino-acid sequences for the identified *H. capsulatum* *Starships* were aligned together with 35 known high-confidence captains sampled from a diverse range of *Starship*

families and fungal orders using MUSCLE 5.3 (Edgar 2022). No trimming was done, since the alignment file was considered small enough for performance not to be an issue. Furthermore, the risk of automatic trimming removing phylogenetically relevant sequences was considered higher than the potential gains in tree construction time (Tan *et al.* 2015). A phylogenetic tree was constructed using IQ-TREE2 2.4.0 with 2000 iterations of ultra-fast bootstrap with NNI enabled. ModelFinder was used with IQ-TREE to find the optimal tree model, which was determined to be JTT+F+R5.

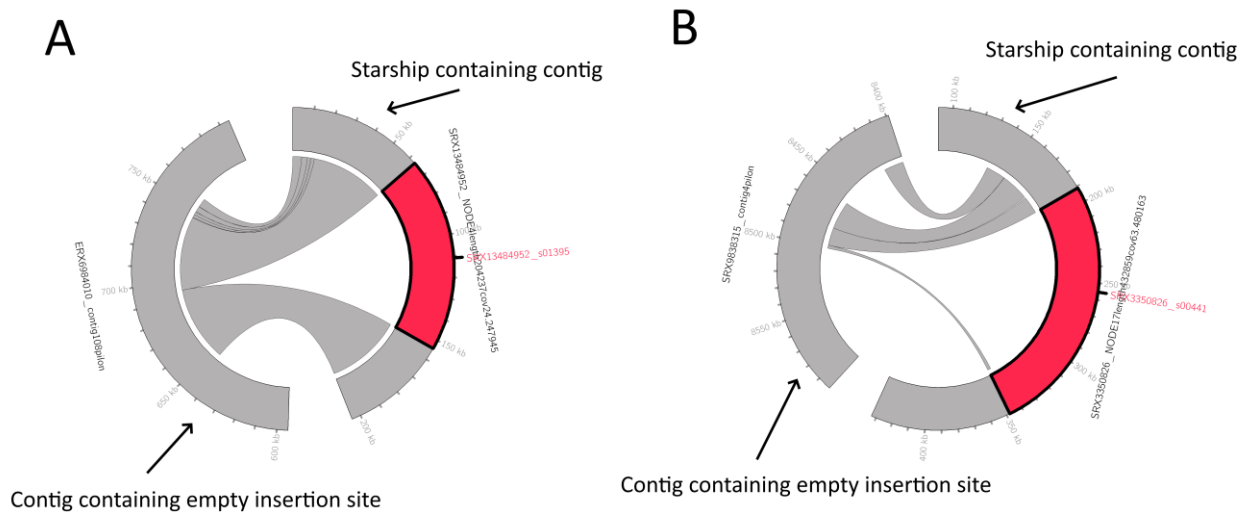
## 2.6 Annotation of *Starship Cargo*

The existing liftoff annotations of the *Starships* were complemented by a functional annotation of protein domains using InterProScan 5.73-104.0 (Blum *et al.* 2025).

# 3 Results

## 3.1 50 *Starships* Were Identified from 79 *H. capsulatum* Assemblies

A total of 65 *H. capsulatum* assemblies were constructed from publicly available DNA reads on the sequence read archive (SRA) (Supplemental Table 1). Combined with the 14 publicly available genome assemblies present in the NCBI genome database, a total of 79 *H. capsulatum* assemblies were searched for *Starships*. The *Starfish* pipeline found a total of 200 *Starships* (Supplemental Table 2), with 150 elements manually removed for either being duplicates or false positives (Figure 2) (Supplemental Table 3). Of the 50 unique *Starships* found, 28 belonged to the Prometheus family, 10 to the Tardis family, 9 to the Phoenix family and 3 to the Enterprise family. This corresponds with the 37 *Starships* documented within the Starbase database, where 17 belong to the Prometheus family, 8 to the Tardis family, 8 to the Phoenix family and 2 to the Enterprise family. With the remaining 2 Starbase *H. capsulatum* elements belonging to the Voyager and Galactica families.



**Figure 2: Examples of typical Starfish Pair-viz output, showing a *Circos* plot for the best alignment between the contig of a putative *Starship* (in red) and a contig containing a potential insertion site for the *Starship*. These plots are used to assist in manually filtering potential false-positive and low-confidence *Starships*. A) The alignment for the identified *Starship* “s01395”, which was interpreted as a true-positive, due to long and consistent synteny between the flanking regions of the *Starship* and the contig containing the insertion site. B) The alignment for the identified *Starship* “s00441” which was interpreted as a false-positive/low confidence, due to one of the *Starship* flanks having very few alignments to the contig containing the insertion site and therefore unlikely to be a genuine insertion site. Since no better insertion site alignments exist for “s00441”, the *Starship* was considered to have insufficient proof and was therefore removed.**

Furthermore, of the 50 identified *Starships*, only 12 belonged to one of the 65 assemblies constructed for this project, with the remainder identified within the 14 publicly available genomes assemblies downloaded. Of those 12 *Starships*, 3 belonged to assemblies constructed from Oxford Nanopore reads, while the other 9 were from Illumina assemblies. This discrepancy further extends to the 38 elements found within the NCBI assemblies, where 20 of the elements were found within the same assembly, labelled “hiscap2” in this project, with the GenBank accession “GCA\_017310585.1”. The explanation for this is that there was a bias toward hiscap2 during the manual removal of duplicates, since the hiscap2 elements are among the first in alphabetical order, meaning that for most duplicate elements, the element within hiscap2 would be considered the default. Comparing the number of elements present before (Supplemental Table 2) and after (Supplemental Table 3) deduplication shows the number of hiscap2 *Starships* remained consistent, which confirms that the manual process was biased.

Past HGT events for the *H. capsulatum* *Starships* were investigated through a BLASTn search against a database of 241 known high-quality *Starships*. The database contained 6 *H. capsulatum* *Starships*, with four of these producing >95% sequence identity alignments against the putative *Starships* that covered most of their length (Table 1). This suggests that at

least four of the 50 identified *Starships* have been identified previously, with a larger amount likely if the BLAST database had included the entire Starbase database, which contains 37 total *H. capsulatum* elements. The matches with *Starships* from other fungal species were largely with SBS000416 of *Coccidioides immitis* and SBS000310 of *Paracoccidioides brasiliensis* (Supplemental Table 4), both of which have several alignments >4000 bp with *H. capsulatum* elements. However, it is difficult to determine if these alignments are proof of past HGT between the *Starships*, or the result of highly conserved regions and other HGT events not involving the *Starships* themselves, due to the relatively small fraction of the elements being aligned. For instance, the best alignment is 6226 bp long, constituting 17% of the *H. capsulatum* element, while only constituting 4.7% of the *C. immitis* element.

**Table 1: The 10 longest alignments produced by a BLASTn comparison between the 50 identified *Starships* and 241 verified *Starships* from the Starbase database. All alignments have >95% sequence identity. The full table can be found in Supplementary Table 4.**

<b>Histoplasma <i>Starship</i></b>	<b>Starbase <i>Starship</i></b>	<b>Alignment length (bp)</b>	<b><i>Starship</i> length</b>	<b>Starbase species</b>
<b>hiscap1_s00042 -</b>	SBS000185	254753	254753	<i>H. capsulatum</i>
<b>hiscap2_s00113 -</b>	SBS000469	40380	40443	<i>H. capsulatum</i>
<b>hismis2_s00369 -</b>	SBS000468	19587	36226	<i>H. capsulatum</i>
<b>hiscap2_s00084 +</b>	SBS000301	16921	16931	<i>H. capsulatum</i>
<b>hismis2_s00369 -</b>	SBS000468	14733	36226	<i>H. capsulatum</i>
<b>hismis2_s00369 -</b>	SBS000416	6226	36226	<i>C. immitis</i>
<b>hiscap2_s00105 -</b>	SBS000310	5832	61968	<i>P. brasiliensis</i>
<b>hiscap2_s00101 -</b>	SBS000310	4799	55474	<i>P. brasiliensis</i>
<b>hismis1_s00269 -</b>	SBS000310	4799	55787	<i>P. brasiliensis</i>
<b>hiscap2_s00066 -</b>	SBS000310	4788	71200	<i>P. brasiliensis</i>

## 3.2 Phylogenetics

Phylogenetic trees were constructed for the 79 assemblies used and the 50 identified *Starship* captains. The assembly tree was constructed from a BUSCO list of conserved genes as a phylogenomic tree, with the purpose of verifying that the *H. capsulatum* assemblies belong to the correct species. The captain tree was constructed with the purpose of verifying that the *Starship* families assigned by *Starfish* to each captain was correct.

### 3.2.1 Phylogenomics of assemblies

The phylogenomic tree constructed for the 79 assemblies shows a clear separation between the *H. capsulatum* assemblies and the *B. dermatidis* assemblies used as an outgroup (Appendix, Figure A1). The internal phylogeny for the *H. capsulatum* assemblies shows three major geographical clades, and several smaller and more uncertain clades (Appendix, Figure A2). One clade for all samples collected in India, two North American clades centered on “hisohi” and “hismis” respectively – often referred to as “NA1” and “NA2” in the literature (Sepúlveda *et al.* 2017, Jofre *et al.* 2022). Many of the assemblies not part of the three first clades were difficult to geographically place, as the metadata often only included information on where the sequencing occurred, not where the sample originated. Furthermore, several reads were sequenced within a medical context, the geographic location of infection not directly documented. However, there exists enough evidence to classify several smaller clades based on geographic location, such as Africa (H88 strain), Panama (G186AR strain) and Brazil (Supplemental Table 5). This rough phylogenomic tree bears some resemblance to more thorough trees constructed in other studies, sharing a similar geographic distribution of clades (Sepúlveda *et al.* 2017, Jofre *et al.* 2022). Finally, it is likely that several of the assemblies used were identical, due to having little to no distance within the phylogeny and, in the case of “hiscap7” and “hiscap8”, having the same Genbank accession “GCA\_000150115.1” and “GCF\_000150115.1”, respectively.

### 3.2.2 Captains of novel *Starships* can be classified into distinct element families

While *Starfish* conducts its own phylogenetic similarity searches to place a putative *Starship* within one of the 11 pre-defined families, based on several conserved YR domains, a phylogenetic tree was also constructed with the complete amino-acid sequences of all 50 identified captains, and 35 high-quality verified captains from known *Starships* to help anchor the new *Starships* (Figure 3). The 4 element families where *H. capsulatum* *Starships* were identified through *Starfish* were clearly distinguished clades on the phylogram, with all *H. capsulatum* captains belonging to a specific family sharing the same clade as the reference captains which are known to belong to that family.

One family that does not share a homologous clade is the Tardis family (Figure 3), which has two separate clades that are separated by a clade containing captains from the Arwing family

and another clade from the Hephaestus family. However, both Tardis clades contain only YRs from the Tardis family, and they contain at least one Tardis reference captain, making it likely that both clades belong to the Tardis family. While the phylogram does indicate that the two Tardis clades share a closer common ancestor with either the Arwing or Hephaestus family, the bootstrap supports for the branches separating the four clades are below 95%, it is therefore equally possible that the two Tardis clades are in fact closest, which is supported by the existing literature (Gluck-Thaler & Vogan 2024). Finally, the phylogram further confirms that at least the captain sequences for the 50 putative *H. capsulatum* *Starships* are likely to be authentic captains, due to their consistent phylogenetic similarity with captains from other fungi.

# Phylogram of Starship Captains

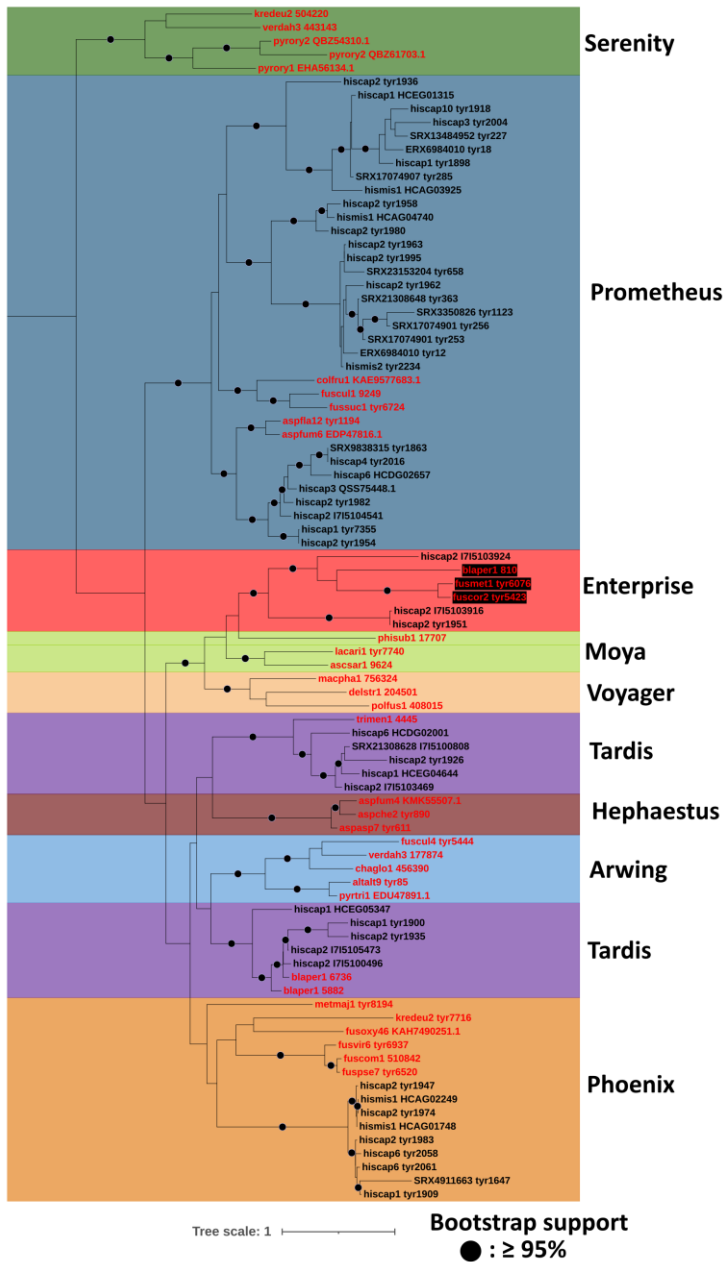


Figure 3: A maximum likelihood phylogram of the 50 putative *H. capsulatum* Starship captains, together with 35 reference captains (red text) sampled from known Starships, representing 9 of the currently defined Starship families. 2000 iterations of ultra-fast bootstrap were run during tree construction with all branches containing adequate (95%) bootstrap support marked in the phylogram. The 50 putative captains were assigned a family based on the family of nearby reference captains. The tree was midpoint rooted.

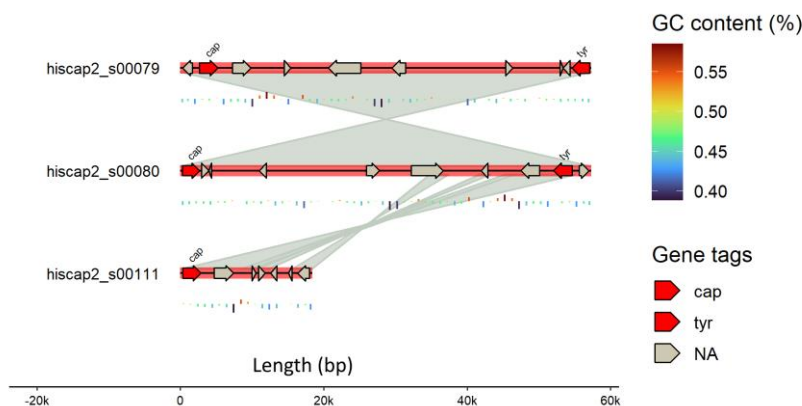
### 3.3 Synteny

The *Starfish* module *locus-viz* was used to inspect the synteny between the obtained elements. Because the lift-over annotation format clashed with the *EggNOG* annotation format, *Starfish dereplicate* was not able to run on the dataset. This meant that the *Starships* for each synteny comparison had to be selected manually, instead of being automatically assigned by *dereplicate*, reducing the number of alignments investigated.

#### 3.3.1 The “Enterprise” *Starships* are highly interconnected

Two of the *Starships* in the Enterprise family, hiscap2\_s00079 and hiscap2\_s00080, show a complete inverse match on the synteny plot, and have the exact same boundary coordinates and contig location (Figure 4). This suggests that the two *Starships* are inversions, with the only difference being which of the two YR genes present in the element is considered the “captain”. Furthermore, hiscap2\_s00111 shows strong synteny to the rear portion of hiscap2\_s00080, although there are several gaps in the synteny which suggest it is not a perfect match. This, combined with the relatively shorter length of hiscap2\_s00111, provides some evidence that hiscap2\_s00111 could historically have transposed to the precursor of hiscap2\_s00079/s00080 and become embedded, forming the double *Starship*.

Since no transcriptomic information was analyzed, it is impossible to assess if having two opposite captains has any effect on the fungus, i.e. if both variants of the element are transcribed. Furthermore, it is possible that only one of the YRs is capable of transposition, with the other having become inert. It is also possible that with both strands containing the *Starship*, it is more likely to transpose itself, granting a selective advantage to the element.



**Figure 4: Synteny comparison and graphic illustrating the 3 Enterprise *Starships*. Synteny lines are present where there is a >95% confidence nucleotide alignment with a neighboring element. "Gene tags" are assigned from existing *Starfish* and *liftoff* annotations. In these plots, “cap” and “tyr” are YRs identified by *Starfish*, where cap is considered the *Starships* captain.**

## 3.4 Annotations

The *Starships* were annotated using two annotation approaches, lift-over annotation on the 79 genome assemblies using the existing hiscap2 annotations as a reference (Supplemental Table 6), and a functional annotation on the 50 identified elements through InterProScan (Supplemental Table 7).

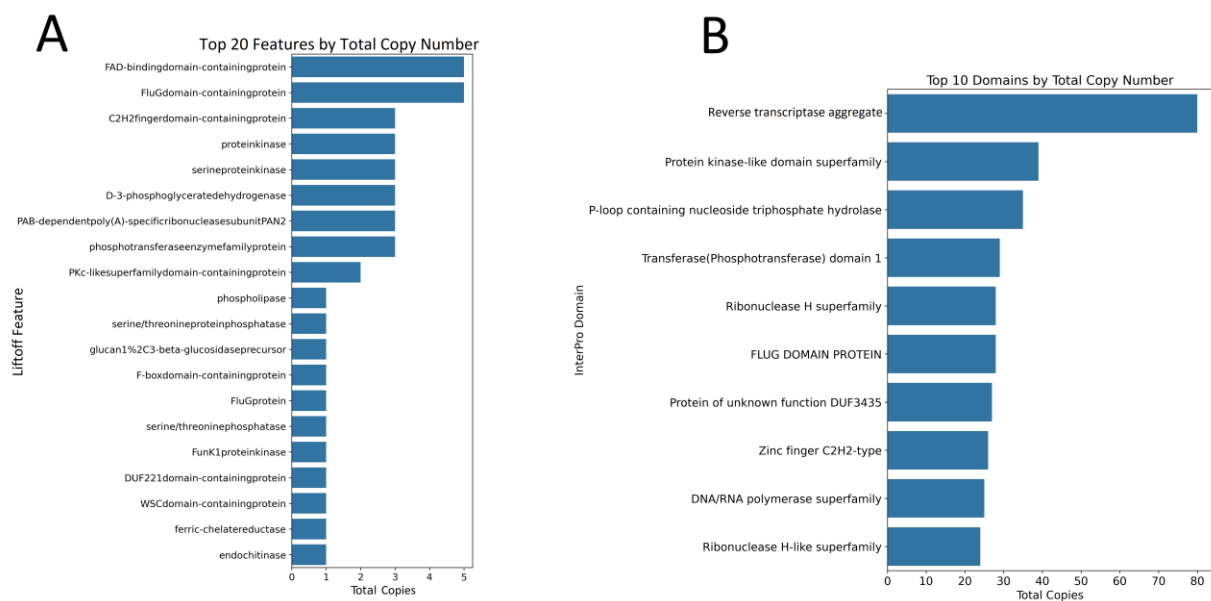
### 3.4.1 Lift-over annotations

There were few features found from the lift-over annotation, only 44 for all 50 elements, with most of the features relatively commonplace and non-specific, such as FAD-binding domains, protein-kinases and phospholipase (Figure 5A). However, a single annotation of ferric-chelate reductase was present inside one of the *Starships*. Some ferric reductases are known to assist yeasts to survive inside hosts such as humans by extracellularly reducing iron ions transported by siderophores, which are excreted by *H. capsulatum* to collect iron from the environment. This allows for intake of the collected iron through the yeast membrane (Martínez-Pastor & Puig 2020, Valdez *et al.* 2022). None of the characteristic virulence factors for *H. capsulatum* were found, although it is possible that several of the InterPro domains identified describe one of these proteins.

### 3.4.2 InterProScan

A total of 195 features were found by InterProScan, with most domains comparable to those annotated by the lift-over annotation. Protein kinases, phosphotransferases and RNAses appear to be common within the elements, along with other ubiquitous protein domains commonly found across life, such as zinc fingers (Figure 5B). However, several features characteristic of *Starships* and other TEs were found as well. For instance, 30 copies of DUF3435 were identified. DUF3435 is considered to be an important conserved domain within all *Starship* YRs and is used by *Starfish* to phylogenetically classify the captains. 6 chromo shadow domains were also found, which are likely involved in assisting the *Starship* to access parts of the genome during transposition (Supplemental Table 7).

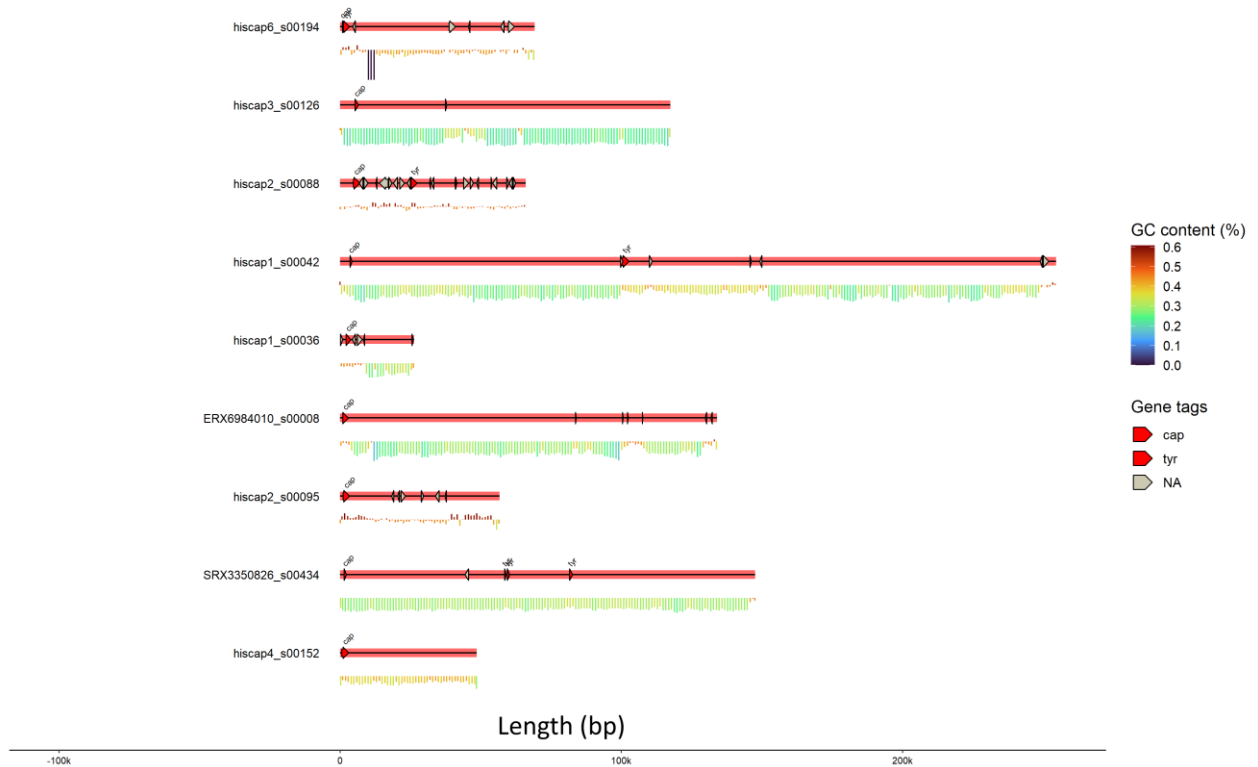
Furthermore, 80 reverse transcriptase domains were identified in aggregate (Figure 5B), although it is likely some of the 80 reverse transcriptases are in fact overlaps, since InterProScan uses several databases to identify functional domains. These reverse transcriptase domains are concentrated on only 17 *Starships* (Supplemental Table 8), with the others appearing to completely lack any reverse transcriptase domains. These domains are rare within active *Starships*, since they transpose through YRs, with reverse transcriptases instead indicating the presence of nested and dead Class 1 transposons or viruses.



**Figure 5: A summary of the most common genetic features annotated within the 50 identified *H. capsulatum* Starships. A) The 20 most common features found by the initial lift-over annotation conducted before Starfish was run, represented in Supplemental Table 6. B) The 10 most common domains found by InterProScan, which searches for functional domains against an existing database. The “Reverse transcriptase aggregate” domain is a merger of all different identified reverse transcriptase domains, this has not been done for other domains such as “protein kinase” and was merely done to illustrate the total number of times a reverse transcriptase-adjacent domain was identified. Values taken from Supplemental Table 7.**

### 3.4.3 Several elements have low GC-content

Several of the identified elements contain large regions where the GC-content is 10-30%, which is below the average of 40-45% for the *H. capsulatum* assemblies they were identified from (Figure 6). These GC-poor regions appear to be more common in elements and element regions lacking annotated genes, this can for instance be seen in hiscap1\_s00042, where the GC content is noticeably higher in the regions that contain annotated genes. This lack of annotation is possibly due to a lack of genes in those regions, since AT-rich regions often lack coding sequences. Additionally, low GC content is common among repetitive elements and could be linked to the abundance of reverse-transcriptase domains. However, there are too few *Starships* to make a strong 95% confidence inference regarding any correlation between the length of an element and its GC-content, or the number of annotations. Hiscap6\_s00182 for instance, is one of the longer elements and displays a GC content rate in the high 40%-region (Supplemental Figure 1).



**Figure 6: A subset of 9 Starships sampled from Supplemental for their abundance of low/high GC content and their length. hiscap4\_s00152 and hiscap2\_s00088 were chosen as examples of Starships with a more neutral GC content. "Gene tags" are assigned from existing Starfish and liftoff annotations. In these plots, "cap" and "tyr" are YRs identified by Starfish, where cap is considered the *Starships* captain.**

## 4 Discussion

### 4.1 Identified Elements

In total, 50 curated and non-duplicate *Starships* were found for *H. capsulatum*, compared to the current number of 6 present in Starbase. The BLASTn results suggest at least 4 of these ships have been identified previously, which is unsurprising considering that the Starbase elements were obtained from a previous analysis of the NCBI assemblies. Starbase also contains an additional 31 *H. capsulatum* elements which have not yet been curated or dereplicated; many of these elements are also likely present among the 50 *Starships* identified in this study. The non curated and dereplicated elements present in Starbase were not used for the BLASTn alignment database since it was determined that their quality and validity was questionable, without false positives, low confidence elements and duplicates having been removed.

The 14 pre-constructed *H. capsulatum* assemblies were overrepresented in element counts compared to the 65 assemblies constructed from SRA reads for this specific study. 122 out of the 200 initially identified elements were found in the pre-constructed assemblies, despite there being considerably fewer assemblies in number (Supplemental Table 2). This discrepancy was further extended after both the manual filtering of false positives and the removal of duplicates, both of which were biased against the assemblies constructed for this study. Since 59 of the SRA assemblies were constructed solely from Illumina short-reads, many of the contigs containing putative *Starships* or empty insertion sites were considerably smaller compared to those in Figure 2. Such smaller contigs were often almost exclusively composed of the putative element, making it difficult to determine if its flanking regions aligned to another contig. Furthermore, as previously mentioned in 3.1, the manual removal of duplicates was also biased towards elements found in the NCBI assemblies due to them being higher up in the alphabet.

The main explanation for the SRA assemblies being underrepresented in *Starship* counts is that most of these assemblies were made using short-reads, not long reads. Large MGEs such as *Starships* are known to be difficult to find without long reads, since they are often too large and repetitive for short-reads to form a contig that completely encompasses the element (Urquhart A *et al.* 2024). This is further confirmed by the fact that 51 of the 200 initially identified elements belonged to the 6 long-read SRA assemblies (Supplemental Table 1, Supplemental Table 2). All except 3 of the elements found within the long-read SRA assemblies were filtered out as duplicates, meaning only 3 remain in the final 50 *Starships*, which suggests that many of the SRA long-reads were used to assemble the 14 pre-existing assemblies downloaded from NCBI due to the abundance of duplicate *Starships*.

## 4.2 Phylogenies

The phylogenomic tree of *H. capsulatum* assemblies in Appendix Figure A2 shows several clear geographic clades for the fungi, and several uncertain clades. The most distinguished clades; “India”, “North America 1” and “North America 2” largely correspond with those established in more extensive phylogenetic studies (Jofre *et al.* 2022). The other 3 known monophyletic clades, Panama, South America and Africa are less defined in the tree due to a lack of information about the geographic origin of most samples. For instance, the metadata SRX21308638 only states the sample was collected in the US, but due to belonging to the same clade as “hismis1” and “hismis2”, it was labelled under the “North America 1” clade (Supplemental Table 5). Most of the 79 samples used lacked concrete geographic metadata, with most samples belonging to the United States simply because the sequencing was done in a US laboratory or hospital. However, despite these issues, the assembly tree does strongly suggest that all assemblies used for *Starship* identification belong to *H. capsulatum* Appendix Figure A1, Appendix Figure A2.

### 4.3 Annotations

Figure 5 along with Supplemental Tables 6 and 7 show few complete or discernible genes within the identified elements. Both features that are present many times, such as protein kinase domains, or only once, such as armadillo-type folds, are very common domains that provide limited information on the protein they belong to. None of the annotated features, from either lift-over or InterProScan can be directly linked to proteins known to be involved in *H. capsulatum* infectivity, such as Hsp60 or CBP1 (Valdez *et al.* 2022). While it is possible that genes encoding these proteins are present as cargo within several of the identified elements, the limited specificity of annotations makes it hard to determine. One of the many identified protein kinase domains could for instance be present in one of the proteins involved in survival within a human host.

A major limitation to this study was the lack of transcriptomic data. As mentioned in 2.1, this was done because most of the SRA reads lacked corresponding RNA reads, making it easier to simply not include such information, since only a limited number of assemblies would contain a corresponding transcriptome. Furthermore, the reliance on lift-over annotation could have limited the number of annotations produced, making it more difficult to identify features and domains of interest. A complete fungal annotation pipeline, such as Funannotate (Palmer & Stajich 2020), which also includes InterProScan as part of its pipeline, could have produced superior annotations. However, it must also be noted that many genes simply lack existing database entries, so even a more thorough annotation process would likely have missed many cargo genes of potential interest.

The identified *H. capsulatum* *Starships* display a consistency in being “messy”, displaying many nested YRs within the elements (Supplemental Figure 1), large regions with low GC-content (Figure 6) and a large amount of reverse transcriptase domains (Figure 5). Together, these results suggest a relative abundance of TEs having become buried within the larger *Starship* elements compared with *Starships* from other fungi. Regions with low or high GC content rates are common signs of repetitive regions. Long repetitive regions indicate in turn the presence of smaller TEs and a lack of conservation in the region, which could suggest that the *Starships* containing such features provide few contributions to the fitness of the host fungus. However, there was no discernible pattern between the elements which according to InterProScan contained reverse transcriptase domains, and the elements which displayed large regions with low GC content. For instance, hiscap3\_0126 (Figure 6) has no identified reverse-transcriptase domains, but has a low GC content in a majority of the element. It is therefore equally possible that there is no direct link between the GC content and the presence of nested TEs within the identified *Starships*.

Some of the identified elements contain many (2-3) nested YRs which are not active captains, but were conserved enough for *Starfish annotate* to consider (Supplementary Figure 1). These nested YRs could be the result of past *Starships* which have since become inert after

transposing inside an existing element, such as what appears to be the case for the Enterprise *Starship* hiscap2\_s00111 (Figure 4). Furthermore, *Starfish* was able to identify these genes as *Starship* YRs and assign them to one of the *Starship* families. Alternatively, these YRs could have been captured by the *Starships* during transposition. If the YRs are from inert *Starships*, a correlation can potentially be found between the abundance of GC-poor regions, inert captains and retrotransposons. RIP is present in many fungi, where it induces GC to AT point mutations toward repetitive regions, this is believed to be a defense mechanism against TEs and other selfish genetic elements. RIPs would further explain the abundance of AT-rich regions in the *Starships*, as the presence of retrotransposons and *Starships* would induce the activation of RIPs, and therefore the abundance of repetitive AT regions. Finally, it is possible that some of the nested *Starships* are present because RIP has created several AT-rich regions, since it is hypothesized that captains prefer AT-rich regions when inserting (Urquhart *et al.* 2023). Such a scenario could potentially create a feedback-loop, where the insertion of *Starships* induces RIP, which in turn makes it more likely for more *Starships* to insert, further inducing RIP. However, since there is limited knowledge on how *Starships* transpose, such a scenario could easily be incorrect.

Finally, the effects on natural selection caused by elements with low GC content and an abundance of retrotransposons must be considered. The abundance of such elements within *H. capsulatum* is an outlier, compared to other fungal species investigated by the Vogan lab. *Starships*, like other MGEs, are generally considered to be deleterious to the host organism, due to their selfish replication and ability to transpose. The trade-off between the beneficial cargo of a *Starship*, and the deleterious nature of TEs, has been an important consideration in the past (Urquhart *et al.* 2023). This is especially true regarding *Starships* that appear to lack beneficial cargo genes, seemingly containing only an abundance of selfish TEs within, yet are still carried by *H. capsulatum*. One explanation is that these *Starships* which are seemingly deleterious do carry advantageous cargo, which has not been annotated properly yet, that provides enough of a fitness advantage for the fungal lineages carrying them to survive. Alternatively, the deleterious effects of such elements within *H. capsulatum* could be minimal, making their existence have a nearly neutral effect on fitness, instead of deleterious. Whether that is connected to *H. capsulatum* in isolation, or if a similar abundance of *Starships* with low GC-content and many retrotransposons also exists in other fungi remains to be seen.

## 5 Future Perspective and Conclusion

### 5.1 Future Perspective

Future studies on *H. capsulatum* could benefit from transcriptomic data, as mentioned in 4.3, along with more extensive *de novo* annotations using annotation pipelines such as Funannotate. The search for past HGT events was constrained towards only comparing against known *Starships* due to time constraints, however, it would be a simple task to run a BLAST search against a database containing, for instance, all publicly available Pezizomycotina genomes in the future.

Concerning the apparent “messiness” of the *H. capsulatum* *Starships*, an investigation on the GC-content and abundance of retrotransposons within *Starships* of other fungi could be valuable. Such an investigation could provide statistical significance to the claims made by this study, that *Starships* with low GC-content and many retrotransposons are overrepresented within *H. capsulatum*. If this is correct, the reasons for why such elements are overrepresented could potentially provide valuable information on *Starships* and *H. capsulatum*.

### 5.2 Conclusion

This thesis aimed to find new *Starships* within the human pathogen *Histoplasma capsulatum*, place them within the existing *Starship* tree of life, find signs of past horizontal gene transfer to other species and identify cargo genes that could potentially contribute toward survival within a human host. In total, 50 elements were identified from 79 *H. capsulatum* assemblies, with at least four elements having been identified previously. These elements were successfully placed within the existing *Starship* tree of life and were mostly grouped within the same monophyletic *Starship* families as previous elements. No evidence was found to support past HGT events to other fungal species, however, this search was limited to comparing against existing high confidence *Starships* from other species. Finally, no cargo genes connected to survival and infection of a human host were found from the annotations. However, the annotations showed a large number of *Starships* containing reverse transcriptase domains, which is uncharacteristic for *Starships*. These annotations, together with an abundance of low GC content in many *Starships*, suggests that *H. capsulatum* *Starships* are messy in relation to known *Starships*. Investigating why so many *H. capsulatum* elements contain nested retrotransposons and what implications they have for the fitness of the fungus could help us understand both *Starships* and the *Histoplasma* fungus better.

## 6 Ethical Aspects and Conflicts of Interest

This thesis focuses on *Histoplasma capsulatum*, a known human pathogen whose infection can prove fatal. Therefore, many of the genetic samples used for this study can be traced to humans infected with the fungus, either from directly sequencing the fungal samples, or storing the fungus as a culture. These samples were downloaded from public databases, such as the Sequence-read archive, and did not disclose any sensitive private information regarding the origin of the samples. Although it was not investigated for this thesis, it is hoped that all genetic samples taken from humans were uploaded to public databases with the consent of the donor.

## 7 Acknowledgements

I would like to thank my supervisor, Aaron Vogan, for providing me the opportunity to work with this project within Sysbio for six months, and for his guidance on the project, its aims and the results obtained. I am also deeply grateful to Adrian Forsythe for his expert assistance during the day-to-day bioinformatics work which constituted most of this project, his help and knowledge of the *Starfish* pipeline and the feedback he provided during the writing of this thesis was invaluable. Finally, I would like to thank Fabien Burki for his questions and feedback during the project, my examiner Siv Andersson and course administrator Lena Henriksson for their assistance.

## References

- Andersson JO, Doolittle WF, Nesbø CL. 2001. Are There Bugs in Our Genome? *Science* 292: 1848–1850.
- Andrews S. 2010. FastQC A Quality Control tool for High Throughput Sequence Data.
- Arkhipova IR, Yushenova IA. 2019. Giant Transposons in Eukaryotes: Is Bigger Better? *Genome Biology and Evolution* 11: 906–918.
- Ayad LAK, Pissis SP, Polychronopoulos D. 2018. CNEFinder: finding conserved non-coding elements in genomes. *Bioinformatics (Oxford, England)* 34: i743–i747.
- Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A. 2025. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* 53: D444–D456.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
- Cambareri EB, Jensen BC, Schabtach E, Selker EU. 1989. Repeat-induced G-C to A-T Mutations in *Neurospora*. *Science* 244: 1571–1575.
- Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* 38: 5825–5829.
- Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* 65: 997–1008.
- Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Computational Biology* 7: e1002195.
- Edgar RC. 2022. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* 13: 6968.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30: 1575–1584.

Ewels P, Magnusson M, Lundin S, Källner M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32: 3047–3048.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* 38: 276–278.

Fouché S, Oggenfuss U, Chanclud E, Croll D. 2022. A devil’s bargain with transposable elements in plant pathogens. *Trends in Genetics* 38: 222–230.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* 3: 722–732.

Gluck-Thaler E, Forsythe A, Puerner C, Stajich JE, Croll D, Cramer RA, Vogan AA. 2024. Giant transposons promote strain heterogeneity in a major fungal pathogen. *bioRxiv* 2024.06.28.601215.

Gluck-Thaler E, Ralston T, Konkel Z, Ocampos CG, Ganeshan VD, Dorrance AE, Niblack TL, Wood CW, Slot JC, Lopez-Nicora HD, Vogan AA. 2022. Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. *Molecular Biology and Evolution* 39: msac109.

Gluck-Thaler E, Vogan AA. 2024. Systematic identification of cargo-mobilizing genetic elements reveals new dimensions of eukaryotic diversity. *Nucleic Acids Research* 52: 5496.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.

Hackl T, Ankenbrand M, Adrichem B van, Wilkins D, Haslinger K. 2024. gggenomes: effective and versatile visualizations for comparative genomics. doi 10.48550/arXiv.2411.13556.

Hall JPJ, Harrison E, Baltrus DA. 2021. Introduction: the secret lives of microbial mobile genetic elements. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377: 20200460.

Harshil Patel, Maxime U Garcia, Adam Talbot, Sateesh\_Per, Moritz E. Beber, Esha Joshi, Daisy Wenyan Han, nf-core bot, Edmund Miller, James A. Fellows Yates, nicolae06, Dave Carlson, Robert Syme, Phil Ewels, Jose Espinosa-Carrasco, Maxime Borry, Alain Domissy, MCMandR, Oliver Ziff, jahdoos, Kevin Menden, sirclockalot. 2024. nf-core/fetchngs: nf-core/fetchngs v1.12.0 - Titanium Platypus. doi 10.5281/ZENODO.10728509.

Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics (Oxford, England)* 32: 1323–1330.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35: 518–522.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* 47: D309–D314.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431: 569–573.

Jofre GI, Singh A, Mavengere H, Sundar G, D’Agostino E, Chowdhary A, Matute DR. 2022. An Indian lineage of *Histoplasma* with strong signatures of differentiation and selection. *Fungal Genetics and Biology* 158: 103654.

Johnson CM, Grossman AD. 2015. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual review of genetics* 49: 577–601.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37: 540–546.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* 19: 1639–1645.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD,

Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen H-C, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang S-P, Yeh R-F, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Patrinos A, Morgan MJ, International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research C for GR, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Department of Genome Analysis I of MB, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, Multimegabase Sequencing Center TI for SB, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Cold Spring Harbor Laboratory LAHGC, GBF—German Research Centre for Biotechnology:, \*Genome Analysis Group (listed in alphabetical order also includes individuals listed under other headings):, Scientific management: National Human Genome Research Institute UNI of H, Stanford Human Genome Center:, University of Washington Genome Center:, Department of Molecular Biology KUS of M, University of Texas Southwestern Medical Center at Dallas:, Office of Science UD of E, The Wellcome Trust: 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8: 48.

Makałowski W, Gotea V, Pande A, Makałowska I. 2019. Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. I: Anisimova M (red.). *Evolutionary Genomics: Statistical and Computational Methods*, s. 177–207. Springer, New York, NY.

Manni M, Berkeley MR, Seppy M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* 38: 4647–4654.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* 14: e1005944.

Martínez-Pastor MT, Puig S. 2020. Adaptation to iron deficiency in human pathogenic fungi. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1867: 118797.

McGowan J. 2024. BUSCO phylogenomics.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37: 1530–1534.

Palmer JM, Stajich J. 2020. Funannotate v1.8.1: Eukaryotic genome annotation. doi 10.5281/zenodo.4054262.

Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. 2019. Large-scale sequence comparisons with sourmash. *F1000Research* 8: 1006.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* 70: e102.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K,

Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddelloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326: 1112–1115.

Sepúlveda VE, Márquez R, Turissini DA, Goldman WE, Matute DR. 2017. Genome Sequences Reveal Cryptic Speciation in the Human Pathogen *Histoplasma capsulatum*. *mBio* 8: 10.1128/mbio.01339-17.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* 37: 1639–1643.

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. 2015. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology* 64: 778–791.

Urquhart A, Vogan AA, Gluck-Thaler E. 2024. Starships: a new frontier for fungal biology. *Trends in Genetics* 40: 1060–1073.

Urquhart A, Chong NF, Yang Y, Idnurm A. 2022. A large transposable element mediates metal resistance in the fungus *Paecilomyces variotii*. *Current Biology* 32: 937-950.e5.

Urquhart A, O'Donnell S, Gluck-Thaler E, Vogan AA. 2025. A natural mechanism of eukaryotic horizontal gene transfer. 2025.02.28.640899.

Urquhart A, Vogan AA, Gardiner DM, Idnurm A. 2023. Starships are active eukaryotic transposable elements mobilized by a new family of tyrosine recombinases. *Proceedings of the National Academy of Sciences* 120: e2214521120.

Valdez AF, Miranda ,Daniel Zamith, Guimarães ,Allan Jefferson, Nimrichter ,Leonardo, and Nosanchuk JD. 2022. Pathogenicity & virulence of *Histoplasma capsulatum* - A multifaceted organism adapted to intracellular environments. *Virulence* 13: 2137987.

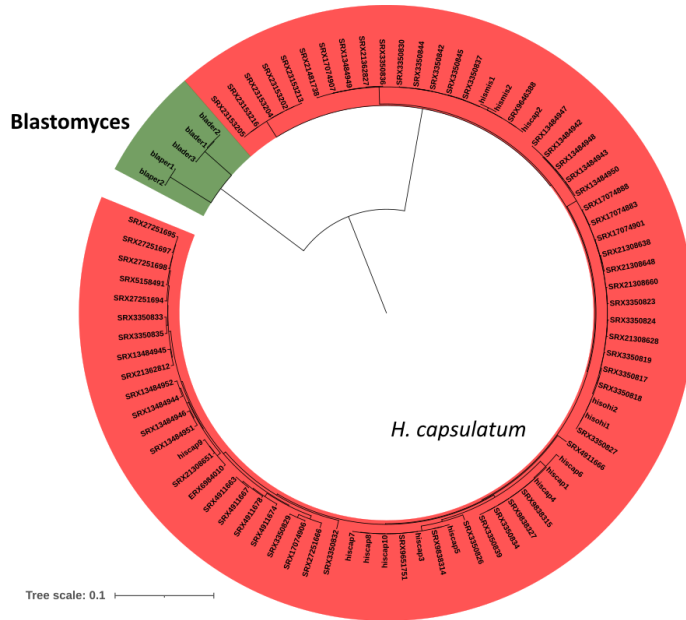
Vogan AA, Ament-Velásquez SL, Bastiaans E, Wallerman O, Saupe SJ, Suh A, Johannesson H. 2021. The Enterprise, a massive transposon carrying Spok meiotic drive genes. *Genome Research* 31: 789.

Vogan AA, Ament-Velásquez SL, Granger-Farbos A, Svedberg J, Bastiaans E, Debets AJ, Coustou V, Yvanne H, Clavé C, Saupe SJ, Johannesson H. 2019. Combinations of Spok genes create multiple meiotic drivers in *Podospora*. *eLife* 8: e46454.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* 9: e112963.

World Health Organization. 2022. WHO Fungal Priority Pathogens List to Guide Research, Development and Public Health Action, 1st ed. World Health Organization, Geneva.

## Appendix A – Figures



**Figure A1:** A maximum likelihood phylogenomic tree of conserved BUSCO features shared by the 79 *H. capsulatum* assemblies used to search for *Starships* (red), and 5 *Blastomyces* assemblies (green). The tree was rooted at the *Blastomyces* branch.

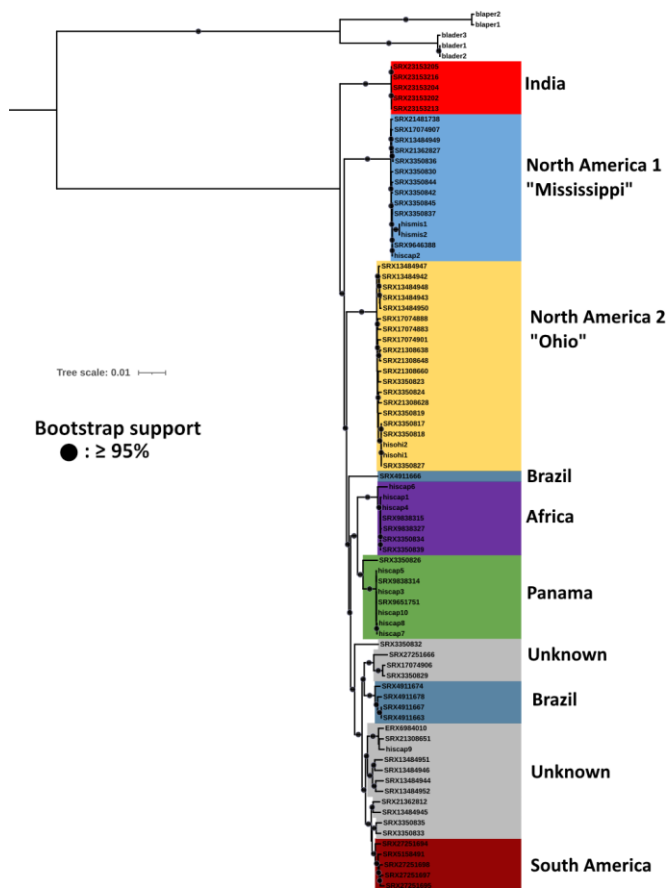


Figure A2: A maximum likelihood phylogenomic tree of conserved BUSCO features shared by the 79 *H. capsulatum* assemblies used to search for *Starships*. The tree is the same as Appendix Figure A1 but displayed to showcase the putative geographic clades of the 79 assemblies. Geographic location of *H. capsulatum* assemblies can be seen in Supplemental Table 5. Most assemblies in Supplemental Table have uncertain geographic origins, but were anchored to the tree with the help of phylogenetically related assemblies in the table which did have more certain origins.

## Appendix B – Supplemental Material

Supplemental tables and figures can be found in the attached .zip file.