Echinocandin heteroresistance causes prophylaxis failure and facilitates breakthrough *Candida parapsilosis* infection

Authors: Bing Zhai [1,2,3,*,#], Chen Liao [4,*], Siddharth Jaggavarapu [5,6,7,8,*], Thierry Rolling [1,2], Sean A. Bergin[9], Mergim Gjonbalaj [2,3], Edwin Miranda [10], N. Esther Babady [2,10], Geraldine Butler [9], Ying Taur [1,2,11], Joao B. Xavier [4,], David S. Weiss [5,6,7,8,#], and Tobias M. Hohl [2,3,11,#]

Institutions: [1] CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2] Infectious Disease Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[3] Immunology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[4] Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[5] Emory Antibiotic Resistance Center, Atlanta, GA, USA.

[6] Department of Microbiology and Immunology, Emory University, Atlanta, GA, USA.

[7] Emory Vaccine Center, Atlanta, GA, USA.

[8] Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA.

[9] School of Biomolecular and Biomedical Science, Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland.

[10] Clinical Microbiology Service, Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[11] Department of Medicine, Weill Cornell Medical College, New York, NY, USA.

[*] : Contributed equally.

[#] : Address correspondence to: hohlt@mskcc.org, david.weiss@emory.edu or to bing.zhai@siat.ac.cn

Abstract

Antimicrobial drug resistance is a major problem in the prevention and treatment of infectious diseases. For patients on antimicrobial prophylaxis during cancer treatment and organ transplantation, breakthrough infections represent a significant and often unexplained cause of morbidity. Here, in high-risk patients on micafungin prophylaxis, we reveal that heteroresistance, the presence of a phenotypically unstable, low frequency subpopulation of resistant cells (~1 in 10,000), causes breakthrough bloodstream infections by *Candida parapsilosis* strains that were misclassified as susceptible by standard antimicrobial susceptibility testing. Genome-matched micafungin heteroresistant strains were isolated from the intestinal reservoir prior to invasive disease. To identify heteroresistance, we harnessed whole genome sequence analysis of 118 clinical *C. parapsilosis* strains and extracted six genetic features into a predictive machine learning model that distinguished heteroresistant strains with high accuracy. These data directly implicate heteroresistance in unexplained prophylaxis failure and highlight a proof-of-principle approach to detect this phenomenon and inform clinical decisions.

Introduction

Antimicrobial resistance leads to over one million annual deaths worldwide and is a severe public health problem [1]. Heteroresistance is a form of antimicrobial resistance in which a minority subpopulation of resistant cells (sometimes as rare as one in a million cells) co-exists with a majority population of susceptible cells [2-5]. Treatment with a given antimicrobial drug kills the susceptible cells but does not kill the rare resistant cells, which survive and replicate rapidly in the presence of the drug. Due to the low frequency of resistant cells, heteroresistance

may be undetected by standard antimicrobial susceptibility testing (AST) which typically relies on liquid-based assays that detect the outgrowth of resistant cells after overnight incubation.

While heteroresistance can mediate treatment failure in murine infection models [6-9], it has been challenging to determine its role in antimicrobial failure in humans. Among other confounding variables, many serious infections are treated with multiple antibiotics, often preventing a clear link between failure and resistance to a single drug. In contrast to therapeutic uses, prophylaxis typically relies on administration of a single antimicrobial drug against a potential pathogen, and thus provides a simplified scenario in which to study the impact of drug-resistant phenotypes on breakthrough infections.

In patients undergoing allogeneic hematopoietic cell transplantation (allo-HCT), randomized prospective clinical trials have established antifungal prophylaxis as the standard of care to prevent fungal infections and reduce all-cause mortality [10-13]. Among the three major classes of antifungal drugs, azoles and echinocandins have been approved by the FDA for prophylaxis [14]. Echinocandin drugs (i.e., micafungin, caspofungin, and anidulafungin) inhibit β-1,3-glucan synthase, which is encoded by the *FKS1* gene and has no mammalian homologs [15]. Since echinocandin drugs exhibit a low toxicity profile and a paucity of drug-drug interactions, they are commonly used for antifungal prophylaxis in high-risk clinical scenarios.

Candida parapsilosis is one of the most common *Candida* species isolated among patients with invasive and superficial candidiasis [16]. In particular, life-threatening *C. parapsilosis* infections and septic shock occur in premature neonates, patients on total parenteral nutrition, immune compromised patients, and in patients following gastrointestinal surgery. *C. parapsilosis* can survive in the gastrointestinal tract as a commensal organism and translocate into the bloodstream in the context of dysbiosis and mucosal injury [17], yet can also be isolated from

environmental reservoirs including the soil, insects, and domesticated animals [18]. Compared to other *Candida* species, *C. parapsilosis* exhibits higher minimal inhibitory concentrations (MICs) to echinocandins due to a natural P660A polymorphism in its drug target Fks1 protein [19], yet classical resistance to echinocandins israrely observed in this organism [20]. Nonetheless, echinocandin prophylaxis or exposure has been linked to an increased frequency of *C. parapsilosis* infections in different medical facilities [21,22], with an unclear cause of drug failure.

Here, we describe heteroresistance to micafungin in *Candida parapsilosis* strains that colonize the intestinal tract prior to breakthrough bloodstream infection in allo-HCT recipients. We observed that micafungin heteroresistance could lead to echinocandin prophylaxis failure, representing a major advance in our understanding of the clinical relevance of heteroresistance. Moreover, we showed that contemporary AST misclassified heteroresistant *C. parapsilosis* strains as susceptible, highlighting an urgent need to develop novel predictive tools to identify heteroresistance and inform clinical decision making. To advance the rapid diagnosis of micafungin heteroresistant phenotypes, we developed a quantitative index system to evaluate the levels of heteroresistance, and built a machine learning model to predict micafungin heteroresistance from six genetic features. We parameterized the 6-feature model with the genome sequences and heteroresistance indices of 118 *C. parapsilosis* strains. This model could accurately identify 90% of micafungin-heteroresistant *C. parapsilosis* bloodstream strains, highlighting the potential future application of this approach to facilitate effective antifungal drug selection and thereby improve patient care.

Results:

Micafungin heteroresistance in clinical *Candida* strains.

To analyze breakthrough fungal bloodstream infections (BSIs) during allo-HCT, we performed a cohort study that included 983 allo-HCT events in 952 patients from January 2016 to December 2020 at Memorial Sloan Kettering Cancer Center (MSKCC) (Extended Data Fig. 1). In the majority of allo-HCTs (n = 966), patients received antifungal prophylaxis with intravenous micafungin (100-150 mg/day) starting seven to ten days prior to allo-HCT for a median of 17 days, prior to switching to triazole-based prophylaxis per institutional practices [23]. There were 30 breakthrough fungal BSIs from transplant day 0 to 30, including 25 breakthrough infections while patients received micafungin prophylaxis and 5 breakthrough infections while patients received triazole prophylaxis (Extended Data Fig. 1). We noted a clear enrichment of *Candida parapsilosis* as the etiologic agent in micafungin breakthrough infections (12 out of 25 cases, Fig. 1a), whereas *C. parapsilosis* breakthrough infections were less frequent in patients on triazole-based prophylaxis (1 out of 5 cases). Other *Candida* species did not show a clear enrichment while allo-HCT patients received either micafungin or triazole prophylaxis (Fig. 1a). In total, we were able to retrieve 15 *C. parapsilosis* bloodstream strains from 13 cases of micafungin breakthrough candidemia for further analysis (Extended Data Fig. 1): 12 strains from 10 patients between transplant day 0 to 30, and 3 strains from 3 micafungin-breakthrough cases later in the transplant process (day > 30).

All 15 *C. parapsilosis* bloodstream strains were classified susceptible to micafungin by broth microdilution (Table 1) and E-test (i.e., Epsilometer test; Fig. 1b) assays. However, these conventional ASTs do not have sufficient sensitivity to determine if there is a very low frequency subpopulation of replicating micafungin resistant cells in these strains, i.e.,

heteroresistance. The population analysis profile (PAP) assay is the gold standard for detecting heteroresistance [4] (Extended Data Fig.2). In this study, we define micafungin heteroresistance as a strain with a resistant cell frequency of at least 0.0001% but lower than 5% at 2x the breakpoint concentration (2 µg/ml, EUCAST [24]) or higher (≥ 4 µg/ml micafungin). Using the PAP assay, we found that 11 out of the 15 strains (e.g., strain 65) harbored a minor subpopulation of micafungin resistant cells that survived at 2x the breakpoint concentration or above (≥ 4 µg/ml, Fig. 1c) [25,26]. The resistant cells rapidly replicated and became dominant during micafungin exposure (Fig. 1d), indicating that they were not persister cells which cannot replicate exponentially in the presence of an antimicrobial drug [27]. However, the frequency of resistant cells reverted to the baseline level during subsequent growth in the absence of micafungin (Fig. 1e), which demonstrates that the resistant cells are phenotypically unstable and are not conventional, stable resistant mutants. In contrast, micafungin-susceptible *C. parapsilosis* strains (e.g., strain 804) did not contain a resistant subpopulation, and could be entirely killed by micafungin exposure (Fig. 1f-h). These findings demonstrate the phenomenon of heteroresistance (HR) to micafungin in *C. parapsilosis* clinical strains. This phenotype extended to other members of the echinocandin drug class: the strains heteroresistant to micafungin were largely heteroresistant to caspofungin and anidulafungin as well (Extended Data Fig. 3).

All bloodstream *C. parapsilosis* strains were isolated on drug-free media and sub-cultured for clinical AST, therefore, the clinical MICs of the strains were consistent with the MICs of the susceptible subpopulation (Table 1). From the PAP assay, we found that 11 heteroresistant strains contained a resistant subpopulation (frequency ranging from 0.01% to 1.7%) with high MICs between 64 to 128 µg/ml, which were 32- to 64- fold higher than those of the susceptible cells. In contrast, the other 4 bloodstream strains were completely susceptible to

micafungin at 2 µg/ml (Table 1). Taken together, 10 out of the 13 BSI patients (76.9%) developed breakthrough *C. parapsilosis* bloodstream infections caused by micafungin heteroresistant strains (Table 1).

Micafungin heteroresistance is associated with breakthrough bloodstream infection.

A recent study from our group demonstrated that 12% of allo-HCT patients experienced intestinal colonization by *C. parapsilosis* complex species prior to donor cell engraftment [23]. *C. parapsilosis* expansion within the intestinal tract represented the critical reservoir for fungal BSI via translocation across the intestinal mucosa into the bloodstream [17]. To examine the micafungin susceptibility of fungi that colonize the intestinal tract during allo-HCT, we recovered 85 fecal *C. parapsilosis* strains from 30 unique patients (Extended Data Fig. 1, Extended Data Fig. 4). We found that 43 strains (50.6%) were classified heteroresistant by PAP. In 16 of 30 patients, we recovered only susceptible *C. parapsilosis* strains, and none of these patients developed a fungal BSI. In contrast, among 14 patients with heteroresistant strains detected from their feces, five patients developed breakthrough *C. parapsilosis* BSI (Fig. 2a). These data reveal a strong correlation ($P = 0.014$; Fisher's Exact Test) between breakthrough BSI and intestinal colonization with micafungin heteroresistant *C. parapsilosis* strains.

We next investigated the genetic similarities between the intestinal and BSI *C. parapsilosis* strains in these five patients with breakthrough infections. Consistent with our previous study [17], the bloodstream strains from a given patient all had extremely high genomic identity (>99.999% population average nucleotide identity, popANI [28]) with strains in the intestinal tract prior to infection (Fig. 2b). Importantly, matched intestinal and bloodstream strains from each patient exhibited the same micafungin heteroresistant phenotype in PAP

assays. In the two patients who were also colonized with susceptible strains in the intestinal tract prior to BSI (Fig. 2b). Patient 2 had sequential colonization with two distinct clusters of susceptible strains and heteroresistant strains, and the intestinal heteroresistant strains led to the ensuing BSI. Patient 3 had co-colonization with susceptible and heteroresistant strains in the intestinal tract, leading to BSI with both strains (Fig. 2b). Thus, *C. parapsilosis* strains that exhibited micafungin heteroresistance and caused BSI were present in the intestinal reservoir prior to translocation into the bloodstream. Taken together, we observed that patients colonized with micafungin heteroresistant strains had a higher probability of developing a breakthrough fungal BSI than those colonized with susceptible strains.

Quantification of micafungin heteroresistance by mathematical modeling.

Given the critical role of micafungin heteroresistance in developing breakthrough infections, the precise and rapid identification of heteroresistant strains is vital to inform clinical decision making. A quantitative measurement system for micafungin heteroresistance that enables accurate evaluation of the phenotype across strains would facilitate the development of a novel diagnostic approach. To develop such a quantitative measurement system, we first assembled a collection of 118 clinical *C. parapsilosis* strains and performed PAP analyses on each. This collection includes 15 *C. parapsilosis* strains from 13 breakthrough BSIs (Table 1), together with 14 additional bloodstream strains (all classified as susceptible by standard AST, Supplementary Table 1) from patients who were hospitalized during the same period (2016-2020) at MSKCC. These 14 strains included two *C. parapsilosis* strains from one allo-HCT patient on triazole prophylaxis and 12 strains from 12 independent non-allo-HCT patients. We also retrieved a total of 89 fecal *C. parapsilosis* strains from 58 representative fecal samples from

the MSKCC fecal fungal strain biobank, for a total of 118 fecal and bloodstream *C. parapsilosis* strains (Extended Data Fig. 4).

The PAP curves of the 118 *C. parapsilosis* strains exhibited a wide range of survival percentages across a concentration gradient of micafungin (Fig. 3a). We developed a heteroresistance index by condensing each PAP curve into a single value based on the curve features (Fig. 3b). Feature extraction was carried out using non-negative matrix factorization (NMF) [29], which decomposes the survival curve of each strain as a weighted sum of two strain-independent signature curves (Fig. 3b). Compared to other dimensionality reduction techniques (e.g., principal component analysis), a unique advantage of NMF is that the non-negativity of feature weights guarantees that any heteroresistance index constructed from these weights is also non-negative. The reconstruction accuracy of NMF had an R^2 of 0.93 (Extended Data Fig. 5a), indicating that the dimensionality reduction preserved most of the information of the original data. The two signature curves represented the extreme gradient responses of susceptible and heteroresistant populations (Fig. 3c), and the two weights correlated with their relative abundances in the population of each strain (Fig. 3d). The weights associated with the two signatures were correlated linearly ($R^2 = 0.99$).

This quasi-linear relationship allowed us to use only the weight associated with the heteroresistant feature after normalization as the heteroresistance index (Extended Data Fig. 5b). We observed a continuous, and not a binary, distribution in the heteroresistance index from 0 to 1 across all 118 *C. parapsilosis* strains (Fig. 3e). To aid in interpretation and susceptibility classification, we drew a line between low and high heteroresistance indices to achieve a binary heteroresistance classification. We determined the threshold at 0.22 (Extended Data Fig. 5c) by maximizing its classification consistency with the clinical metric we introduced above—a strain

was classified as heteroresistant if it harbored a subpopulation of cells that survived at micafungin concentrations above the breakpoint (≥ 4 μg/ml). The cutoff separated all 118 strains into a heteroresistant (HR, n = 64) and a susceptible (S, n = 54) cluster (Fig. 3d, e), with only four mismatches (Fig. 3e, arrows), all of which had a heteroresistance index near the cutoff value. Therefore, the HR indices generated by NMF are highly consistent with the clinical metric (114 out of 118 strains; 96.6%) and laid a solid foundation for the downstream analysis of this phenotype.

Associations of micafungin heteroresistance with genetic variations in *C. parapsilosis* strains.

Since strain-level differences in the responses to antifungal drugs have been attributed to genetic differences from single-nucleotide to chromosome-level variations [2], we performed whole genome sequencing analyses of the 118 strains to identify genetic components associated with micafungin heteroresistance in *C. parapsilosis*. We profiled genetic variations across the 118 strains at different genomic scales by mapping reads against the completed reference genome CDC317 [30].

We first examined chromosome-level variation as aneuploidy has been reported in drug heteroresistance in *Candida parapsilosis* and other human fungal pathogens [31-33]. We identified eight strains (6.8%, seven heteroresistant and one susceptible strains) with aneuploidy: four strains with an extra copy of chromosome VI, three strains with trisomy of partial chromosome III and chromosome IV, and one strain with an extra copy of chromosome V (Extended Data Fig. 6a). In sum, we did not observe a link between aneuploidy and heteroresistance in *C. parapsilosis* strains cultured under baseline conditions.

Across the 118 genomes, we identified 1,689 open reading frames (ORFs) that had an amplification or a deletion in at least one strain (including those affected by aneuploidies, Supplementary Table 2), 1,885 small insertion/deletions (InDels) and 20,177 single nucleotide variants (SNVs, Supplementary Table 3). To test whether these genetic variations correlated with heteroresistance, we encoded them with the following rules: 1) using continuous integers for ORF copy numbers; 2) using two binary dummy variables (heterozygous or not, and homozygous alternative or not) for nonsynonymous SNVs/InDels. Taken together, we identified 1,858 non-redundant features. Using principal component analysis, we showed that both ORF copy number variations (CNVs) and SNVs/InDels distinguished, though not perfectly, heteroresistant and susceptible strains (Extended Data Fig. 6b). We identified 198 features that possessed a significant association with the heteroresistance indices after multi-test correction (Supplementary Table 4).

Construction of a whole-genome SNP-based phylogeny identified two clades enriched with susceptible strains (S-clade) and one clade enriched with heteroresistant strains (HR-clade) (Fig. 4), suggesting a high correlation between heteroresistance and phylogeny. To identify the genetic variants that explain heteroresistance beyond phylogeny, we evaluated the associations of the 198 selected genetic features with heteroresistance indices after controlling for their phylogenetic distances in a multivariate phylogenetic LASSO (Least Absolute Shrinkage and Selection Operator) model. Only one CNV feature, *CPAR2_104610*, had non-zero positive LASSO regression coefficients, indicating that this ORF is correlated with heteroresistance in both phylogeny-dependent and -independent manners. Taken together, micafungin heteroresistance of *C. parapsilosis* is strongly associated with genetic variations and phylogeny, which provides a basis for developing a sequence-based diagnostic approach for this phenotype.

Diagnosis of micafungin heteroresistance using a six-feature machine learning model.

Equipped with the heteroresistance index and the genetic variants profile, we employed a Random Forest (RF) model to predict micafungin heteroresistance from genetic features. To find the minimum set of features required, we ran a two-step feature selection procedure that selectes only six features from the 1,858 non-redundant features, including three CNV features and three SNV features (Fig. 4, bottom; Fig. 5a; Extended Data Fig. 7a). CNV1 and CNV2 are adjacent ORFs on chromosome I, and some amplifications cover both ORFs [34]. However, CNV1 and CNV2 were evaluated independently in the RF models and were both retained by recursive feature elimination algorithm. Using the six features, we then implemented two train-test data split strategies for model development and validation.

First, the 118 strains were split into five folds: iteratively, four folds (80% data) were used to train a model, which was then tested using the remaining fold (20% data). This process was repeated five times and produced five sets of predictions. We showed that the predicted heteroresistance indices correlated with their observed values (R^2: 0.63 ~ 0.84, Fig. 5b) with an overall R^2 value of 0.75. There is a strong phylogenetic signal contributing to the prediction of heteroresistance (R^2 of 0.69 for phylogenetic prediction alone, compared to R^2 of 0.75 for the RF model (Extended Data Fig. 7b). This reflects the observation that many of the heteroresistant isolates fall into one of the three major clades (Fig. 4). However, phylogenetic analysis requires whole genome sequencing and is not suitable for clinical diagnostic purposes.

Second, we predicted heteroresistance indices of the 29 bloodstream strains with available clinical MIC (test set) data by training a RF model based on the PAP results from the 89 fecal strains. Only four strains were misclassified by the RF model (Fig. 5c). To explore the mechanisms underlying the misclassification, we used SHAP (SHapley Additive exPlanations [35])

to interpret the model predictions by decomposing the predicted indices as a sum of contributions from each feature on a strain-by-strain basis. In both false-positive cases (strains 800, 813), the prediction of heteroresistance received strong supports from all three SNVs and the CNV of *CPAR2_805410*, despite most strains with the same features were indeed heteroresistant. The HR indices of the two false-negative predictions (strains 807, 2384) were both close to the HR threshold value (0.22), suggesting the challenges of classification of their drug susceptibility. Further investigations with SHAP of these two cases revealed that they received strong support of being heteroresistant from SNV3 (strain 807) and high copy number of *CPAR2_104600* (strain 2384), but the combination of the other features offset the support in decision-making (Fig. 5c). Although the RF model had four misclassifications, it had an overall superior accuracy (86.2% vs. 31.0%) than the MIC-based laboratory approach (Fig. 5d). More importantly, the RF model correctly identified 18 out of 20 heteroresistant strains while AST detected none (90% vs 0%). These data demonstrate that the model prediction has high sensitivity in detecting micafungin heteroresistance, which is completely missed by standard AST. Taken together, these data show a highly accurate interpretable machine learning model based on genomic features as a proof-of-concept diagnostic tool to facilitate identification of antimicrobial heteroresistance in the clinic.

Discussion

In this study, we show that micafungin heteroresistance promotes breakthrough *C. parapsilosis* infections in allo-HCT patients on micafungin prophylaxis. While both heteroresistant and susceptible *C. parapsilosis* strains could colonize the intestinal tract, heteroresistant strains were highly enriched among those causing bloodstream infections.

Although host immune, anatomic, and microbiota-dependent factors may contribute to fungal pathogenesis and bloodstream translocation, patients colonized with micafungin heteroresistant strains in the intestinal tract had a higher likelihood of developing breakthrough BSI. Importantly, we did not observe any instances of classical, stable resistance to micafungin in this study. Taken together, these data highlight heteroresistance, rather than classical resistance, as potentially being the major cause of breakthrough *C. parapsilosis* infections during echinocandin prophylaxis.

Heterogeneous antifungal resistance includes heteroresistance and tolerance [2]. In azole-tolerant *C. albicans* strains, subpopulations of fungi grow slowly in drug concentrations above the MIC, and the frequency of the resistant cells ranged from 5% to 90% [2,7], a much higher frequency than that observed with heteroresistant *C. parapsilosis* strains (mostly between 0.01% to 0.2%, Table 1). The low frequency of the resistant subpopulation in the *C. parapsilosis* strains in this study was in line with the range observed among azole heteroresistant *Cryptococcus neoformans* (0.01% to 25%) [32], and likely explains why we did not observe visible colonies in the zone of inhibition in E-tests (Fig. 1b).

The PAP assay is the gold standard for detecting resistant subpopulations in a heteroresistant strain, however, there is no agreed upon numerical parameter to quantify the level of heteroresistance. Previous studies quantified antifungal heteroresistance or tolerance by area under the curve (AUC) or fraction of growth (FoG), respectively [7,32,36]. Here, we showed that the heteroresistance indices generated by NMF can be used in reverse to accurately reconstruct PAP curves (Extended Data Fig. 8). This quantification approach is very powerful when combined with our RF model: the RF-predicted heteroresistance indices can be further transformed to the

percentages of resistant subpopulations under a range of micafungin concentrations to inform clinical decision-making.

The *C. parapsilosis* strains in this study had heterozygosity levels ranging from 0.025 to 0.055 heterozygous single nucleotide variants (SNVs) per kb, which was 3-6 fold lower than that of the reference genome CDC317 (0.158 per kb [37]). The complexity and diversity of *C. parapsilosis* genomes are in general lower than that of other *Candida* species, and may facilitate this proof-of-concept approach to identify heteroresistance-linked genetic features. The 118 *C. parapsilosis* strains in this study, although including all five clades on a phylogeny tree of 170 *C. parapsilosis* strains from around the world [34], were all collected from one hospital. Future investigations that include *C. parapsilosis* strains with more diversified genetic backgrounds may help revise the currently selected HR features and establish an even more accurate prediction model.

Compared to phenotype-based AST and PAP assays, our genotype-based approach can bypass the time needed to culture the isolate in the presence of antibiotics which take days to weeks for slow-growing organisms. The clade enriched heteroresistant strains on the *C. parapsilosis* phylogenetic tree indicates phylogeny as a strong predictor of micafungin heteroresistance. The phylogenetic analyses require whole genome sequencing that is both time- and computional power-consuming. In contrast, the input feature values in our six-feature prediction model can be obtained by quantitative PCR that targets specific genomic regions. Due to the increasing prevalence of antimicrobial resistance (AMR) and decreasing sequencing costs, genotype-based advanced machine learning models have the potential to revolutionize AMR diagnostics and support rapid, personalized clinical decisions in therapy and infection control. Since many microbial pathogens, including *C. parapsilosis*, colonize in the GI tract prior to

translocation into the bloodstream, a future research direction could extend the current model to detect heteroresistance from shotgun metagenome data of fecal samples, without the need for strain isolation, to further reduce turnaround time.

Resistance or tolerance to echinocandin drugs has been attributed to chromosomal mutations in the *FKS* or *MSH2* genes in *Candida* species [19,38]. We did not identify mutations in *FKS* or *MSH2* genes that were associated with micafungin heteroresistance among the *C. parapsilosis* strains examined in this study. In fact, studies on azole tolerance have suggested that unlike conventional resistance that often resulted from mutations of the direct drug target, tolerance or heteroresistance may rely on changes in several central stress response pathways [2]. Beyond the development of a heteroresistance diagnostic approach, comparative genomic analysis of data from diversified clinical *C. parapsilosis* strains may also provide clues to the underlying genetic mechanisms of heteroresistance. The three SNV features identified by the RF models all locate at Chromosome 2, indicating that mutations and the chromosomal dynamics of Chromosome 2 might be linked to this phenomenon. We also identified ORFs such as *CPAR2_104610*, which had copy number variations, that were associated with micafungin heteroresistance. The specific contributions to heteroresistance of these genes requires future investigation.

Finally, we examined if heteroresistance to echinocandins exists in another medically relevant *Candida* species, *Candida auris*. *C. auris* are highly drug-resistant pathogenic fungi that have rapidly spread across the globe since their identification in 2009. More than 99% of the *C. auris* isolates in the United States are resistant to fluconazole, 50-60% are resistant to amphotericin B, and less than 10% are resistant to echinocandins [39-41]. Thus, echinocandin drugs are the first-line empiric therapy for *C. auris* infections, as recommended by the CDC [42]. PAP

analysis of *C. auris* strains from the CDC/FDA AR Bank detected some with micafungin heteroresistance, and these strains exhibited unstable resistant subpopulations that grew to be dominant when exposed to micafungin (Extended Data Fig. 9). Interestingly, recent reports describe echinocandin treatment failures in strains that were classified as susceptible, along with emergence of resistance while the patients were being treated with echinocandins [39,40]. It is worthwhile to determine if some of these isolates are heteroresistant to echinocandins, and thus if micafungin heteroresistance may be a cause of treatment failure. Such potential treatment failures, together with the prophylaxis failures that we observed in allo-HCT recipients, would highlight an urgent need for rapid detection of heteroresistance in clinical settings.

In summary, we describe heteroresistance to echinocandin drugs in *Candida* species. This drug resistance phenotype remains undetected by current diagnostic methods and can result in unexplained prophylaxis failures. The development of a diagnostic test based on six genomic features holds the promise of supporting rapid, improved clinical decision making to enhance and individualize prophylaxis and treatment strategies for these life-threatening infections.

Methods:

Inclusion criteria of patients and *C. parapsilosis* strains.

Adult patients that underwent allo-HCT at Memorial Sloan Kettering Cancer Center (MSKCC) between Jan. 1st, 2016 to Dec. 31st, 2020 were examined for breakthrough fungal bloodstream infections. The medical record of non-allo-HCT patients with *Candida parapsilosis* bloodstream infections from the same study period (2016 to 2020) were also reviewed. A total of 29 bloodstream *C. parapsilosis* strains allo-HCT and non-allo-HCT patients (Extended Data Fig. 4) were obtained from clinical micro lab.

Fecal biospecimens from allo-HCT recipients were drawn from a prospective fecal collection biobank, which has been described previously [43]. Fecal samples from a subset of 205 patients (transplant day -10 to 30) had been examined for the presence of viable fungal strains. A total of 30 patients were confirmed with intestinal colonization by *C. parapsilosis* from culture methods. Fecal *C. parapsilosis* strains was isolated and save from each culture positive fecal sample as previously described [17]. Two *C. parapsilosis* strains from each fecal samples (except the day 2 fecal sample in Patient 2, which was isolated only one strain) from 16 fecal samples from these five patients with *C. parapsilosis* BSI were randomly picked for further antifungal susceptibility and genomic analysis. For the 25 patients with intestinal colonization but no fungemia, two *C. parapsilosis* strains from two different *C. parapsilosis* culture positive fecal samples from each patient were randomly selected. Fecal biospecimens from healthy donors [44] were also examined for the *C. parapsilosis* culture positivity. Two fecal samples were identified and two *C. parapsilosis* strains from each sample were picked. Taken together, there were a total of 85 fecal *C. parapsilosis* strains included in this study.

Population analysis profile assay (PAP)

YPD agar (Sigma) plates were made with each antifungal at the indicated concentrations. The stocks of micafungin (Biovision), caspofungin (Biovision) and anidulafungin (Biovision) were made by dissolving each drug in molecular biology grade water (Corning). Overnight cultures of all the isolates were grown in 2 mL of YPD broth (RPI) in 8 mL round-bottom tubes (Corning) at 35°C with 200 rpm shaking. Ten-fold serial dilutions of the overnight cultures were made in PBS in 96-well plates (Corning). 5 µL of each dilution, from undiluted to 10^{-6}, were spotted on the agar plates. Plates were incubated for 48 to 72 hours at 35°C before the CFUs were enumerated. The CFUs from the plates with an antifungal were divided by the CFUs from the no antifungal plate to calculate the % survival.

Time-kill assays

Two mL of YPD broth containing 2 µg/mL or 0.5 µg/mL of micafungin for *C. parapsilosis* and *C. auris*, respectively, were inoculated with 10^6 CFU from an overnight culture. Cultures were incubated at 35°C with shaking at 200 rpm for up to 72 hours. Aliquots of 50 µL were removed at indicated time points and the CFU were counted by plating serial dilutions on plates with and without micafungin at the same concentrations as the respective broth cultures.

Antifungal resistance stability assay

For antifungal treatments, *C. parapsilosis* cultures were sub-cultured (1:100) from a drug-free overnight culture into YPD with 2 µg/mL micafungin and grown at 35°C with 200 rpm shaking for 48 hours. Micafungin-free subcultures were performed by diluting the micafungin culture 1:1,000 into fresh YPD media and the cultures were grown for 24 hours. Cultures were diluted in PBS and plated on YPD-agar plates with and without micafungin to determine the proportion of resistant *C. parapsilosis*.

E-tests

Micafungin E-test strips were obtained from Biomerieux (Marcy-l'Étoile, France). The E-tests were performed according to the manufacturer's instructions. Briefly, 5 colonies from an overnight YPD agar plate were dissolved into sterile PBS (Lonza). The OD was adjusted to 0.5 McFarland standard. The solution was applied to RPMI 1640 plates with MOPS and 2% Glucose (Remel) with a cotton swab. The E-test strip was laid on the plate and the images were captured after 48 hour incubation at 35°C.

Non-negative matrix factorization (NMF)

We assume, for each strain, the observed survival percentages (matrix X) at different micafungin concentration is a linear combination of the typical gradient responses of susceptible and heteroresistant subpopulations (matrix H) weighted by their mixing ratios (matrix W). The linear decomposition was approximated by NMF [29]. Under the constraints that both H and W are non-negative, the objective function minimized by NMF is $0.5||X - WH||_{kl}^2$, where $|| \cdot ||_{kl}$ is the Kullback-Leibler distance. Before applying NMF, the survival percentages were first log-transformed and then increased by 4 units such that the matrix X is non-negative (all values range between 0 and 6). After NMF, we normalized matrix H such that its values at micafungin=0 equal 6 and updated matrix W accordingly to keep their product the same.

Whole genome sequencing and Genetic variants calling

Selected 118 *C. parapsilosis* strains were streaked from -80°C glycerol stocks on YPD agar and cultured at 30°C for 2 days. Single colony of each strain was picked and inoculated into 2ml YPD liquid media and culture at 30°C, 240 rpm for overnight. Genomic DNA extraction and whole genome sequencing were performed as previously described [17].

Reads were mapped against *C. parapsilosis* CDC317 reference genome [30] using BWA-mem with parameters -M -R 'RG' where 'RG' represents read group information generated for each strain for compatibility with further downstream analysis. Duplicated reads were marked using the GATK MarkDuplicates tool as per GATK best practices. SNVs and InDels were identified for each strain using the HaplotypeCaller tool from GATK. The resulted GVCF files were combined using the CombineGVCFs tool and jointly genotyped using the GenotypeGVCFs tool from GATK. The resulted SNVs, InDels, and mixed-type variants were first hard filtered using the following criteria. Filter settings for SNVs include mapping quality (<40), variant quality score (<30), quality by depth (<2), strand odds ratio (>3), Fisher strand bias (>60), mapping quality rank-sum test (<−12.5) and read position rank-sum test (<−8). Filter settings for InDels and mixed-type variants include quality by depth (<2), variant quality score (<30), Fisher strand bias (>200) and read position rank-sum test (<−20). Very low coverage regions with a read depth <10 and a genotype quality <20 were set to missing on a per-sample basis. Monomorphic (no alternative alleles are called for any of the samples) and multiallelic (all sites at which only alternative alleles are called) variants, clusters of three or more SNPs called in a 10-basepair window, and variants in repetitive regions identified by TandemRepeatFinder [45] were removed. In addition, we removed SNVs that are within 10 bp of InDels and allowed only one variant within a cluster of InDels separated by 5 or fewer base pairs to pass. The remaining variants surviving the filtering process were exported in table format using the GATK VariantsToTable tool.

Genetic variants annotation

SNVs and InDels were annotated using the SnpEff tool [46]. The variants with keyword "synonymous_variant" in annotation were removed for the downstream data analysis.

Gene-level copy number variation

Distributions of read coverage at each annotated open reading frame (ORF) and across the genome were calculated using BEDTools coverage and genomecov tools [47] respectively. The ORF-level copy number was computed by multiplying the coverage ratio (the ratio of mean read coverage at each ORF to that across the genome) by the expected ploidy (diploid), and then rounded to the nearest integer.

Phylogenetic tree construction

We followed the approach in our previous study [17]. The filtered VCF file of genetic variants was input into the RRHS tool [48] to generate equal-length haploid FASTA sequences for each strain (parameter -n 1 generates one sequence per individual). The phylogenetic tree was constructed from these sequences using RAxML [49] with parameters -T 20 (number of threads) -f a (rapid bootstrap analysis algorithm) -m GTRGAMMA (model of nucleotide substitution) -p 12345 -x 12345 -# 1000.

Feature extraction

The genetic features used in univariate statistical test, regular LASSO (least absolute shrinkage and selection operator), phylogenetic LASSO, and Random Forest model were extracted from SNVs, InDels, and CNVs profiled above. Each SNV or InDel was represented by two features: a binary variable for whether this variant is heterozygous and a binary variable for homozygous alternative allele. Therefore, all three possible categories (homozygous reference, homozygous alternative, and heterozygous) of an SNV or InDel can be represented by their combinations: 0/0, 0/1 and 1/1. We combined redundant SNV/InDel features that have exactly the same binary data pattern despite these variants may be biologically different (e.g., SNV3 in

Fig. 5a). The CNV feature of an ORF is encoded by its copy number (an integer). The total number of non-redudnent SNV, InDel, and CNV features is 1,858.

Phylogenetic LASSO

The phylogenetic LASSO (least absolute shrinkage and selection operator) model was also developed in our previous study for bacterial genomics [50]. Briefly, the input (matrix I) and output (matrix O) data of LASSO were multiplied by a transformer: $I_{new} \rightarrow (T'T)^{-1}T'I$, $O_{new} \rightarrow (T'T)^{-1}T'O$, where matrix T is obtained from the phylogenetic variance-covariance matrix G by Cholesky decomposition such as $TT' = G$. The phylogenetic variance-covariance matrix was computed using the "vcv.phylo" function of the ape library [51]. Given new input matrices (I_{new} and O_{new}), the LASSO method determines the coefficient vector β that associate genetic features (matrix X) with heteroresistance index (matrix Y) by minimizing $||Y - X\beta||_2^2/2\mathrm{N} + \alpha||\beta||_1$, where α is a constant that multiplies the regularization term, and $||\cdot||_1$ and $||\cdot||_2$ are L1 and L2 norm respectively. We used the function LassoCV implemented in the scikit-learn Python library with CV=5 (5-fold cross validation) and n_alphas=1,000 (number of α along the regularization path).

Random Forest (RF) regression models

RF models were built using the RandomForestRegressor function with 1,000 trees. We developed a two-stage strategy for feature selection and parameter optimization. In the first step, we pre-selected 14 unique features from the 1,858 features obtained above with non-zero coefficients by regular LASSO (this should not be confused with phylogenetic LASSO). Using these 14 features (Extended Data Fig. 7a), RF parameters were optimized from a predetermined set of values: the number of features to consider when looking for the best split in RF (max_features = square root, log2, 16%, 32%, 64%, 81%, 100% of all features), the maximum depth of the tree in RF (max_depth=2,4,8,16,32,64,128), the minimum number of samples

required to split an internal node in RF (min_samples_split = 2, 4, 8, 16), and the minimum number of samples required to be at a leaf node (min_samples_leaf = 1, 2, 4). We split all 118 strains into 5 folds using the StratifiedKFold method based on their heteroresistance labels. For each permutation of the 5-fold cross validation, we used 4 folds of data to build a RF regression model and chose the optimal parameter combination using GridSearchCV that internally performed a nested, within-training-data 5-fold cross validation. By out-joining the optimal values chosen by GridSearchCV using the 5 models, we reduced the total number of possible parameter combinations from 588 to 18.

In the second step, we implemented a recursive feature elimination algorithm, where the weakest feature (lowest Gini importance) was iteratively removed. In each iteration, we built a RF model using all 118 strains and the RF parameter values were again optimized by GridSearchCV—but, this time the possible combinations are limited to the 18 sets we determined in the first step. Since the minimum number of features is 6 (see Fig. SX), the best parameter set chosen by GridSearchCV using the 6-feature RF model was considered as final and used throughout the study. The final set of parameter values are: max_features = 0.16, max_depth = 4, min_samples_split = 2, min_samples_leaf=1.

The RandomForestRegressor, StratifiedKFold, and GridSearchCV are all functions from the scikit-learn Python library. We used R^2 as the evaluation metric for all RF models. When fitting a RF model, the sample weight of each strain was assigned to be 1 divided by the total number of identical strains (including itself). Any two strains were identical if their average nucleotide identity (ANI), which represents the nucleotide-level genomic similarity of all orthologous genes, is at least 99.999%. The pairwise ANI values among all 118 *C. parapsilosis* isolataes were computed using the inStrain package [28].

Statistical analysis

In the univariate analysis, we used Kruskal-Wallis test for SNVs/InDels and Spearman correlation for ORF CNVs. Both functions (kruskal and spearmanr) are implemented in Python from the Scipy's stats library. We determined the P-value threshold after multiple-test correction: the critical P-value cutoff of 0.05 was divided by the number of independent tests (21,350 features). This threshold (2.3×10^{-6}) results in 279 unique significant features and is likely not conservative. We then decreased the threshold to 5.0×10^{-9} and reported the 198 significant features (roughly top 200) at this more stringent significance level in Table S3.

Other softwares

Data from population analysis profile, time-kill, and antifungal susceptibility assays were graphed GraphPad prism version 9 (San Diego). Figures were plotted by Adobe Illustrator 2021. Principal components were computed in Python using the pca function. SHAP values were computed by TreeExplainer [35] for the RF model built on the 89 strains that have not been tested clinically. Box plot, swarm plot and cluster map were generated using the corresponding functions in the Python's Seaborn package. The infinitesimal jackknife variances of the heteroresistance indices predicted by RF models were calculated using the forestci package [52].

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon request. All sequencing data generated in this study will be deposited in the Sequence Read Archive upon acceptance.

Acknowledgements

We thank E. Pamer and K. Wolfe for discussions. This work was supported by the National Key Research and Development Program of China 2021YFA0911300 (B.Z.), National Institutes of Health (NIH) grant nos. R01 AI093808 (T.M.H.), R21 AI105617 (T.M.H.), R21 AI156157 (T.M.H.), U19 AI158080 (D.S.W.), R01s AI141883 and AI148661 (D.S.W), U01 AI124275 (J.B.X.), R01 AI137269 (J.B.X. and Y.T.), an Investigator in the Pathogenesis of Infectious Diseases Award from the Burroughs Wellcome Fund (T.M.H. and D.S.W.), the Ludwig Center for Cancer Immunotherapy (T.M.H.), the Susan and Peter Solomon Divisional Genomics Program (T.M.H.), Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant no. RO-5328/1-2 (T.R.), Science Foundation Ireland grant nos. 19/FFP/6668 (G.B.), 18/CRT/6214 (G.B.), and all authors from Memorial Sloan Kettering Cancer Center were supported by NIH P30 CA008748.

Author contributions

The study was conceived by B.Z., C.L., S.J., D.S.W. and T.M.H.. Clinical data were collected and analyzed by Y.T. and B.Z.. Collection of bloodstream infection strains was done by E.M. and N.E.B.. Fecal sample processing and fecal strain collection were done by B.Z., T.R. and M.G.. Antifungal population analysis profiling, time-killing, and stability assays were performed by S.J.. Whole genome sequencing analyses were done by C.L., S.B. and G.B.. Mathematical modeling construction and analyses were done by C.L. and J.B.X. The manuscript was written by B.Z., C.L., S.J., D.S.W. and T.M.H.. All co-authors reviewed and edited the manuscript.

Tables:

Table 1. Micafungin susceptibility profiles of 15 bloodstream *C. parapsilosis* stains from micafungin breakthrough infections. Clinical MIC data and susceptibility interpretation were based on the YeastOne assay. MIC-S: inhibitory concentrations of the susceptible subpopulation of each strain; MIC-R: inhibitory concentrations of the resistant subpopulation of each strain based on the population analysis profiling (PAP) assay. S, susceptible; HR, heteroresistant. N.D. – not detected;N.A. – not available. Fold difference indicates the fold change between MIC-S and MIC-R in each strain. Resistance frequency indicates the proportion of resistant cells in each strain that survived at the micafungin breakpoint of 2 ug/ml [24]. Strains 65 and 67 were both isolated from patient 1 in Fig. 2b, and strains 485 (HR) and 1191 (S) were isolated from patient 3 in Fig. 2b.

Figure legends:

Fig. 1: Micafungin heteroresistance in bloodstream strains of *C. parapsilosis*. (a) Distribution of fungal species that led to micafungin or triazole breakthrough fungal bloodstream infections in allo-HCT patients at MSKCC (between 2016 to 2020). "Others" included four episodes of breakthrough BSI by *Candida kefyr*, *Trichosporon asahii*, *Capronia munkii*, and *Saccharomyces cerevisiae*. (b) Micafungin E-tests of *C. parapsilosis* strains 65 and 804. (c-h): Micafungin population analysis profile (PAP) of (c) strain 65 and (f) 804. The % survival was calculated compared to growth on drug-free plates. Strains 65 (d) and 804 (g) were treated with micafungin at 2 µg/mL and cultures were plated at indicated time points for enumeration of total (black, circles) and resistant (colored, squares) cells. The frequency of micafungin-resistant subpopulation of strains 65 (e) and 804 (h) was quantified after growth in YPD broth for 24 h (Pre-treatment), passage in YPD broth with 2 µg/mL micafungin (In micafungin) for 48 h and following passage in YPD broth with no micafungin for 24 hours (Subculture). (i) Graphical illustration of the dynamics of the resistant subpopulation in a micafungin heteroresistant strain in the presence (+ mica) or absence of micafungin (-mica). Blue cells are susceptible, orange cells are resistant (unstable), and grey cells are those killed by micafungin.

Fig. 2: Intestinal colonization with micafungin heteroresistant *C. parapsilosis* strains is associated with breakthrough bloodstream infections. a) Distribution of allo-HCT recipients with *C. parapsilosis* intestinal colonization and breakthrough bloodstream infection (BSI). Fisher's Exact test gave a *P* value = 0.014. b) Longitudinal medical data from five allo-HCT recipients with *C. parapsilosis* breakthrough BSI, including antifungal drug administration (first row), the time of fungal BSI identification and the drug susceptibility of bloodstream strain(s)

(second row), the time of fecal strain isolation and the drug susceptibility of fecal strains (third row). PopANI (population average nucleotide identity) between blood (star) and fecal (circle) *C. parapsilosis* strains of each patient were shown in a heatmap to the right of its corresponding longitudinal data panel. The grids on the grey dash line indicate self-comparisons of each *C. parapsilosis* strain. Color bar indicates the scale of popANI, white color represents 99.999% identity, as a common threshold of "identical strains" [53]. Patient 1 and patient 2 in this figure were patient 2 and patient 3, respectively, in a previous study [17].

Fig. 3: Quantification of micafungin heteroresistance in *C. parapsilosis*. a, Population analysis profile (PAP) curves indicating the proportion of cells surviving at the indicated micafungin concentrations for each of the 118 fecal and bloodstream *C. parapsilosis* strains from the MSKCC. **b**, A computational framework for heteroresistance index. We assume that the PAP curve of each strain (matrix X) is a linear combination of signature curves (matrix H) weighted by their mixing proportions (matrix W). The decomposition is approximated by 2-component non-negative matrix factorization (NMF). For each strain, its heteroresistance index (matrix I) is derived from the weights of the signature curves (matrix W). **c**, Signature PAP curves of heteroresistant and susceptible subpopulations. **d**, Scatter plot of the signature curve weights across all strains. **e**, Heteroresistance index of the 118 strains ordered from low to high. The face color of each bar represents the binary classification of heteroresistance using the clinical metric (survival percentage > 0.0001% at 4 ug/ml micafungin), which agrees with the NMF-derived heteroresistance index with only 4 exceptions (indicated by arrows). The separation of S (susceptible) and HR (heteroresistant) clusters in panel d and e are determined by maximizing the consistency of classifying heteroresistance using the two approaches.

Fig. 4: The phylogenetic tree constructed from the 118 *C. parapsilosis* strains.

The heteroresistance (HR) index and the values of six features used to build a predictive model (see Fig. 5 for feature description and model evaluation) are aligned to the phylogenetic tree based on the tree structure. S-clade I and S-clade II: clades enriched with susceptible (S-) strains; HR-clade: clade enriched with heteroresistant strains. Allo-HCT (allogenic stem cell transplantation); BSI (bloodstream infection); HR (heteroresistance); Hom ref (homozygous reference allele); Het (heterozygosity with one reference allele and one alternative allele); Hom alt (homozygous alternative allele).

Fig. 5: Random Forest model outperforms clinical testing in identifying *C. parapsilosis* micafungin heteroresistance. a, Functional annotation of the six genetic features chosen by feature selection. Mutations represented by the SNV features: SNV1 (Chr2, 1574350, G->A); SNV2 (Chr2, 2251778, C->T); SNV3 (combines two loci with the same binarized SNV pattern on Chr2, 1283814, T->G; Chr2, 1283825, A->C). ORF (open reading frame); *S.* (*Saccharomyces*) *cerevisiae*; Hom alt (homozygous alternative allele); Het (heterozygosity with one reference allele and one alternative allele). **b**, Comparison of predicted and measured heteroresistance indices for the 118 *C. parapsilosis* strains. Each dot represents one strain. Notably, these predictions were made from 5-fold cross-validation (folds indicated by colors), where a different model was generated in each permutation by training on four folds (80% data) of strains and then used to predict the heteroresistance indexes of the remaining fold (20% data). The error bar represents the infinitesimal jackknife variance. The R^2 of each model in the 5-fold cross validation and their combined R^2 are indicated. **c**, Decomposed feature contributions (SHAP values) to the predicted (Pred.) heteroresistance indexes for the 29 BSI

isolates with AST data. FN: false negative predictions; FP: false positive predictions. A strain is predicted to be heteroresistant if the predicted index > 0.22. The SHAP bias is a constant and represents the expected heteroresistance value. **d**, Comparison of Random-Forest-model-based diagnosis to antimicrobial susceptibility testing (AST) for the detection of micafungin heteroresistance within a subset of 29 BSI strains. #HR: number of heteroresistant strains; #S: number of susceptible strains.

Extended data:

Extended Data Fig. 1 | Information on allo-HCT patients included in this study. The dashed line boxed areas indicate the number of patients included in respective figures, and the anatomic source (i.e., bloodstream or fecal) of analyzed fungal isolates. Grey-shaded box indicates the bloodstream *C. parapsilosis* strains listed in Table 1.

Extended Data Fig. 2 | The schematic demonstration of antimicrobial heteroresistance by population analysis profiling (PAP) assay. X-axis shows the fold of breakpoint concentration, Y-axis shows the ratio of survival population under certain drug concentrations. Resistant (red), heteroresistant (orange), and susceptible (blue) curves are representative curves for each drug susceptibility classification. Experimental PAP curves that fall in the red/orange/blue shade area suggest the strains are resistant/heteroresistant/susceptible, respectively.

Extended Data Fig. 3 | Caspofungin and anidulafungin susceptibility phenotypes of *C. parapsilosis* strains. Strains 65, 794, 795, 2384 (orange) are heteroresistant to micafungin; strains 804, 811, 2386 (blue) are susceptible to micafungin. Strain 795 is susceptible to anidulafungin, strain 804 is heteroresistant to caspofungin. Except for these two cases, caspofungin and anidulafungin susceptibility in the tested strains were consistent with micafungin susceptibility results.

Extended Data Fig.4 | Information on *C. parapsilosis* strains in this study. Dashed line gated areas represent the same patient or strain groups in the indicated figures. Grey-shaded box indicates the strains that listed in Table 1.

Extended Data Fig. 5 | Construction of the heteroresistant index. a) Reconstruction accuracy. As shown in Fig. 3b, NMF approximates the original survival percentage data (matrix X) by the reconstructed data (matrix product $H \cdot W$). b) Geometric illustration of the computation of heteroresistance index. For each data point in Fig. 3d, it was first projected onto the diagonal (x+y=1) for normalization. The normalized data point was then projected onto the x axis. The distance from the origin to this projected point along the axis of heteroresistance cell signature weight is called the heteroresistance score. c) Determination of the heteroresistance index cutoff that best discriminates heteroresistant from susceptible strains. For a given cutoff value, each *C. parapsilosis* strain from the MSK cohort was classified as heteroresistant or susceptible using two criteria: (1) whether the computed heteroresistance index is above the cutoff value and (2) whether it grows at 4 ug/ml micafungin in the PAP assay. The number of mismatches was computed as the total number of strains with different classifications between the two approaches.

Extended Data Fig. 6 | Association assays between genomic variables and micafungin heteroresistance. a) Chromosome-level and large-scale structural variations. Windowed coverage ratio is defined as the mean number of reads averaged over each 5kb fixed-size window divided by that averaged across the genome. The heteroresistant indices and their classification labels (S/HR) of the 8 strains with additional chromosome copies are indicated in the parenthesis. The chromosome ploidy numbers were confirmed using the GATK DetermineGermlineContigPloidy tool. The comparison of heteroresistance indices between the strains with and without trisomy is visualized using box plot overlaid on top of a swarm plot (each dot represents one strain), with central line representing the median, box limits

representing the first and third quartiles, and whiskers extending to the smallest and largest values or at most to 1.5× the interquartile range, whichever is smaller. Chr: chromosome; S: susceptible; HR: heteroresistant/heteroresistance; ns, non-significant (Kruskal-Wallis test). b) Principal component (PC) analysis of all 118 *C. parapsilosis* strains based on the SNV/InDel-(left) or CNV-(right) features. For SNVs/InDels, we encode homozygous or heterozygous variants as separate features that take binary values (0/1). For CNVs, the actual copy numbers are used.

Extended Data Fig. 7 | Validation of the Random Forest (RF) prediction models. a) Feature selection by recursive feature elimination. We determined that the minimum number of features is 6 because the cross-validated error started to decrease steadily when further features were excluded. b) 5-fold Cross-validation using principal components of phylogenetic variance-covariance matrix as input features. Each element of the variance-covariance matrix represents the shared branch length between a pair of strains. The R^2 value (0.69) is slightly lower than that of the 6-feature RF prediction model (0.75), suggesting the feature-based prediction exceeds the phylogeny-based prediction.

Extended Data Fig. 8 | Reconstruction of PAP curves from heteroresistance index. Comparison of experimental survival curves (solid lines) and those reconstructed from their heteroresistance indexes (dashed lines).

Extended Data Fig. 9 | Micafungin heteroresistance in *Candida auris* clinical strains. *C. auris* strains 386 (a, b, c, d) and 381 (e, f, g, h) were obtained from CDC/FDA AR Bank.

Micafungin susceptibility tests were performed on the two *C. auris* strains, include: Etests of strains 386 (a) and 381 (e); population analysis profile (PAP) of strain 386 (b) and 381 (f), with the % survival calculated by comparison to growth on drug-free plates. Strains 386 (c) and 381 (g) were treated with micafungin at 0.5 µg/mL and cultures were plated at indicated time points for enumeration of total (black, circles) and resistant (colored, squares) cells. The frequency of micafungin-resistant subpopulation of strains 386 (d) and 381 (h) was quantified after growth in YPD broth for 24 h (Pre-treatment), passage in YPD broth with 0.5 µg/mL micafungin (In micafungin) for 48 h and following passage in YPD broth with no micafungin for 24 hours (Subculture).

Supplementary information:

Supplementary Table 1: Patients with *C. parapsilosis* bloodstream infection in this study. 29 *C. parapsilosis* strains were recovered from 26 patients. 14 patients were infected post allo-HCT (patient 1-14). 12 patients were with other primary diseases (patient 14 - 29). The clinical MICs (by YeastOne assay) and the heteroresistant (HR) or susceptible (S) phenotypes by PAP assay were listed.

Supplementary Table 2: List of open reading frames with copy number variations (CNV).

Supplementary Table 3: List of small insertion/deletions (InDels) and single nucleotide variants (SNV).

Supplementary Table 4: List of genetic features that possessed a significant association with the heteroresistance indices after multi-test correction.

Reference:

1 Antimicrobial Resistance, C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629-655, doi:10.1016/S0140-6736(21)02724-0 (2022).

2 Berman, J. & Krysan, D. J. Drug resistance and tolerance in fungi. *Nat Rev Microbiol* **18**, 319-331, doi:10.1038/s41579-019-0322-2 (2020).

3 Andersson, D. I., Nicoloff, H. & Hjort, K. Mechanisms and clinical relevance of bacterial heteroresistance. *Nat Rev Microbiol* **17**, 479-496, doi:10.1038/s41579-019-0218-1 (2019).

4 El-Halfawy, O. M. & Valvano, M. A. Antimicrobial heteroresistance: an emerging field in need of clarity. *Clin Microbiol Rev* **28**, 191-207, doi:10.1128/CMR.00058-14 (2015).

5 Band, V. I. & Weiss, D. S. Heteroresistance: A cause of unexplained antibiotic treatment failure? *PLoS Pathog* **15**, e1007726, doi:10.1371/journal.ppat.1007726 (2019).

6 Sionov, E., Chang, Y. C., Garraffo, H. M. & Kwon-Chung, K. J. Heteroresistance to fluconazole in Cryptococcus neoformans is intrinsic and associated with virulence. *Antimicrob Agents Chemother* **53**, 2804-2815, doi:10.1128/AAC.00295-09 (2009).

7 Ben-Ami, R. *et al.* Heteroresistance to Fluconazole Is a Continuously Distributed Phenotype among Candida glabrata Clinical Strains Associated with In Vivo Persistence. *mBio* **7**, doi:10.1128/mBio.00655-16 (2016).

8 Band, V. I. *et al.* Antibiotic failure mediated by a resistant subpopulation in Enterobacter cloacae. *Nat Microbiol* **1**, 16053, doi:10.1038/nmicrobiol.2016.53 (2016).

9 Xiong, Y. *et al.* Impact of Carbapenem Heteroresistance Among Multidrug-Resistant ESBL/AmpC-Producing Klebsiella pneumoniae Clinical Isolates on Antibiotic Treatment in Experimentally Infected Mice. *Infect Drug Resist* **14**, 5639-5650, doi:10.2147/IDR.S340652 (2021).

10 Fisher, B. T. *et al.* Effect of Caspofungin vs Fluconazole Prophylaxis on Invasive Fungal Disease Among Children and Young Adults With Acute Myeloid Leukemia: A Randomized Clinical Trial. *JAMA* **322**, 1673-1681, doi:10.1001/jama.2019.15702 (2019).

11 Walsh, T. J. *et al.* Caspofungin versus liposomal amphotericin B for empirical antifungal therapy in patients with persistent fever and neutropenia. *N Engl J Med* **351**, 1391-1402, doi:10.1056/NEJMoa040446 (2004).

12 Slavin, M. A. *et al.* Efficacy and safety of fluconazole prophylaxis for fungal infections after marrow transplantation--a prospective, randomized, double-blind study. *J Infect Dis* **171**, 1545-1552, doi:10.1093/infdis/171.6.1545 (1995).

13 Goodman, J. L. *et al.* A controlled trial of fluconazole to prevent fungal infections in patients undergoing bone marrow transplantation. *N Engl J Med* **326**, 845-851, doi:10.1056/NEJM199203263261301 (1992).

14 Epstein, D. J., Seo, S. K., Brown, J. M. & Papanicolaou, G. A. Echinocandin prophylaxis in patients undergoing haematopoietic cell transplantation and other treatments for haematological malignancies. *J Antimicrob Chemother* **73**, i60-i72, doi:10.1093/jac/dkx450 (2018).

15 Douglas, C. M. *et al.* Identification of the FKS1 gene of Candida albicans as the essential target of 1,3-beta-D-glucan synthase inhibitors. *Antimicrob Agents Chemother* **41**, 2471-2479, doi:10.1128/AAC.41.11.2471 (1997).

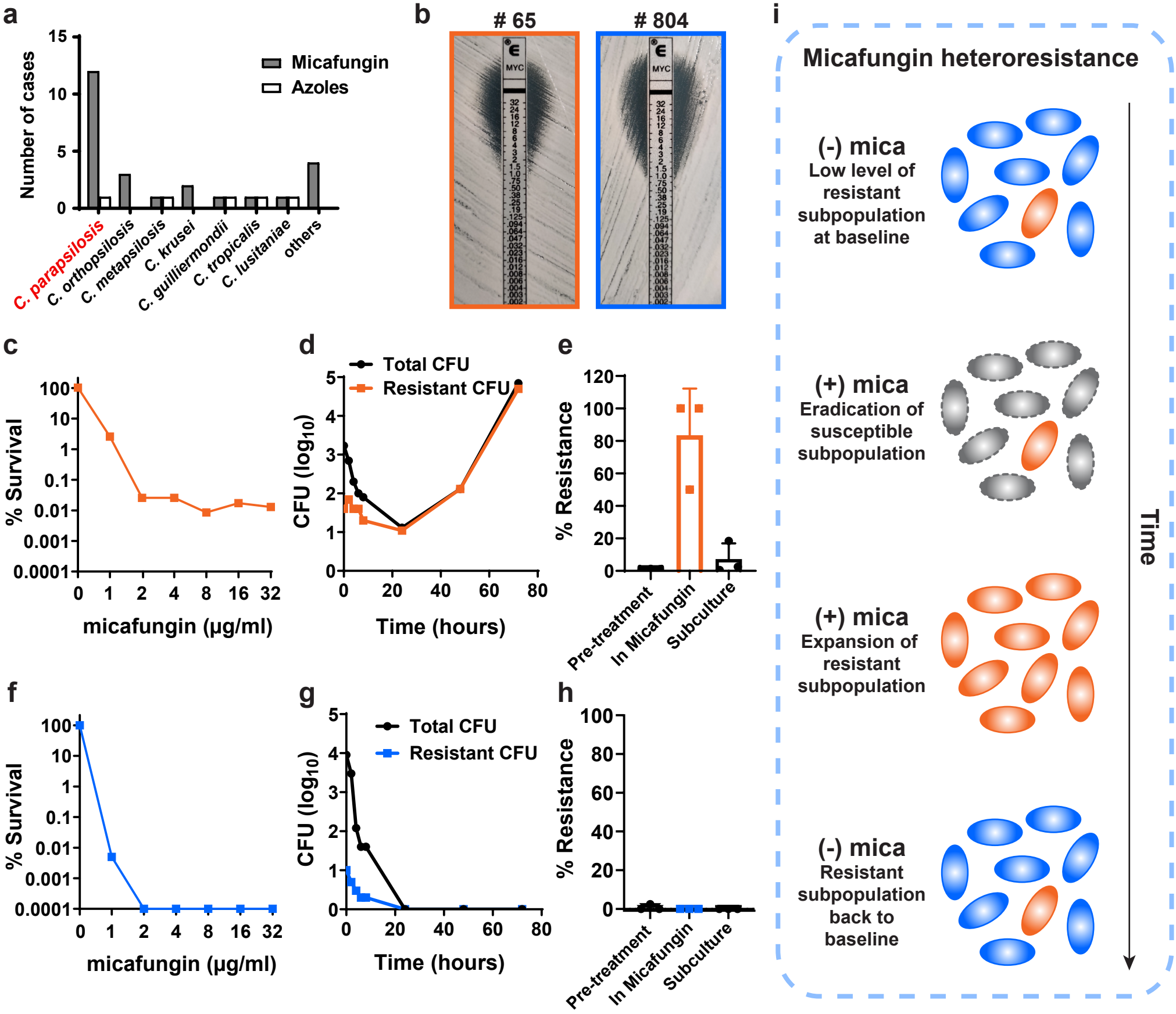
16 Toth, R. *et al.* Candida parapsilosis: from Genes to the Bedside. *Clin Microbiol Rev* **32**, doi:10.1128/CMR.00111-18 (2019).

17 Zhai, B. *et al.* High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis. *Nat Med* **26**, 59-64, doi:10.1038/s41591-019-0709-7 (2020).
18 Trofa, D., Gacser, A. & Nosanchuk, J. D. Candida parapsilosis, an emerging fungal pathogen. *Clin Microbiol Rev* **21**, 606-625, doi:10.1128/CMR.00013-08 (2008).
19 Garcia-Effron, G., Katiyar, S. K., Park, S., Edlind, T. D. & Perlin, D. S. A naturally occurring proline-to-alanine amino acid change in Fks1p in Candida parapsilosis, Candida orthopsilosis, and Candida metapsilosis accounts for reduced echinocandin susceptibility. *Antimicrob Agents Chemother* **52**, 2305-2312, doi:10.1128/AAC.00262-08 (2008).
20 Pfaller, M. A., Moet, G. J., Messer, S. A., Jones, R. N. & Castanheira, M. Geographic variations in species distribution and echinocandin and azole antifungal resistance rates among Candida bloodstream infection isolates: report from the SENTRY Antimicrobial Surveillance Program (2008 to 2009). *J Clin Microbiol* **49**, 396-399, doi:10.1128/JCM.01398-10 (2011).
21 Wang, E. *et al.* The ever-evolving landscape of candidaemia in patients with acute leukaemia: non-susceptibility to caspofungin and multidrug resistance are associated with increased mortality. *J Antimicrob Chemother* **70**, 2362-2368, doi:10.1093/jac/dkv087 (2015).
22 Lortholary, O. *et al.* Recent exposure to caspofungin or fluconazole influences the epidemiology of candidemia: a prospective multicenter study involving 2,441 patients. *Antimicrob Agents Chemother* **55**, 532-538, doi:10.1128/AAC.01128-10 (2011).
23 Rolling, T. *et al.* Haematopoietic cell transplantation outcomes are linked to intestinal mycobiota dynamics and an expansion of Candida parapsilosis complex species. *Nat Microbiol* **6**, 1505-1515, doi:10.1038/s41564-021-00989-7 (2021).
24 Giske, C. G., Turnidge, J., Canton, R., Kahlmeter, G. & Committee, E. S. Update from the European Committee on Antimicrobial Susceptibility Testing (EUCAST). *J Clin Microbiol* **60**, e0027621, doi:10.1128/JCM.00276-21 (2022).
25 Pfeiffer, C. D. *et al.* Breakthrough invasive candidiasis in patients on micafungin. *J Clin Microbiol* **48**, 2373-2380, doi:10.1128/JCM.02390-09 (2010).
26 Pfaller, M. A. *et al.* Clinical breakpoints for the echinocandins and Candida revisited: integration of molecular, clinical, and microbiological data to arrive at species-specific interpretive criteria. *Drug Resist Updat* **14**, 164-176, doi:10.1016/j.drup.2011.01.004 (2011).
27 Lewis, K. Persister cells, dormancy and infectious disease. *Nature Reviews Microbiology* **5**, 48-56, doi:10.1038/nrmicro1557 (2007).
28 Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727-736, doi:10.1038/s41587-020-00797-0 (2021).
29 Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791, doi:10.1038/44565 (1999).
30 Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* **459**, 657-662, doi:10.1038/nature08064 (2009).
31 Sionov, E., Chang, Y. C. & Kwon-Chung, K. J. Azole heteroresistance in Cryptococcus neoformans: emergence of resistant clones with chromosomal disomy in the mouse brain

during fluconazole treatment. *Antimicrob Agents Chemother* **57**, 5127-5130, doi:10.1128/AAC.00694-13 (2013).
32 Stone, N. R. *et al.* Dynamic ploidy changes drive fluconazole resistance in human cryptococcal meningitis. *J Clin Invest* **129**, 999-1014, doi:10.1172/JCI124516 (2019).
33 Yang, F. *et al.* Aneuploidy Underlies Tolerance and Cross-Tolerance to Drugs in Candida parapsilosis. *Microbiol Spectr* **9**, e0050821, doi:10.1128/Spectrum.00508-21 (2021).
34 Bergin, S. A. *et al.* Resistance to miltefosine results from amplification of the *RTA3* floppase or inactivation of flippases in *Candida parapsilosis*. Preprint at https://www.biorxiv.org/content/10.1101/2021.12.16.473093v2.full (2021).
35 Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**, 56-67, doi:10.1038/s42256-019-0138-9 (2020).
36 Rosenberg, A. *et al.* Antifungal tolerance is a subpopulation effect distinct from resistance and is associated with persistent candidemia. *Nat Commun* **9**, 2470, doi:10.1038/s41467-018-04926-x (2018).
37 Pryszcz, L. P., Nemeth, T., Gacser, A. & Gabaldon, T. Unexpected genomic variability in clinical and environmental strains of the pathogenic yeast Candida parapsilosis. *Genome Biol Evol* **5**, 2382-2392, doi:10.1093/gbe/evt185 (2013).
38 Healey, K. R. *et al.* Prevalent mutator genotype identified in fungal pathogen Candida glabrata promotes multi-drug resistance. *Nat Commun* **7**, 11128, doi:10.1038/ncomms11128 (2016).
39 Ostrowsky, B. *et al.* Candida auris Isolates Resistant to Three Classes of Antifungal Medications - New York, 2019. *MMWR Morb Mortal Wkly Rep* **69**, 6-9, doi:10.15585/mmwr.mm6901a2 (2020).
40 Park, J. Y., Bradley, N., Brooks, S., Burney, S. & Wassner, C. Management of Patients with Candida auris Fungemia at Community Hospital, Brooklyn, New York, USA, 2016-2018(1). *Emerg Infect Dis* **25**, 601-602, doi:10.3201/eid2503.180927 (2019).
41 Kilburn, S. *et al.* Antifungal Resistance Trends of Candida auris Clinical Isolates, New York-New Jersey, 2016-2020. *Antimicrob Agents Chemother*, aac0224221, doi:10.1128/aac.02242-21 (2022).
42 CDC. *Recommendations for treatment of Candida auris infections*, <https://www.cdc.gov/fungal/candida-auris/c-auris-treatment.html> (2021).
43 Taur, Y. *et al.* Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. *Sci Transl Med* **10**, doi:10.1126/scitranslmed.aap9489 (2018).
44 Kim, S. G. *et al.* Microbiota-derived lantibiotic restores resistance against vancomycin-resistant Enterococcus. *Nature* **572**, 665-669, doi:10.1038/s41586-019-1501-z (2019).
45 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
46 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
47 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

48 Lischer, H. E., Excoffier, L. & Heckel, G. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of microtus voles. *Mol Biol Evol* **31**, 817-831, doi:10.1093/molbev/mst271 (2014).
49 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
50 Yan, J. *et al.* Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network. *PLoS Comput Biol* **13**, e1005677, doi:10.1371/journal.pcbi.1005677 (2017).
51 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).
52 Polimis, K., Rokem, A. & Hazelton, B. Confidence Intervals for Random Forests in Python. *The Journal of Open Source Software* **2**, doi:10.21105/joss.00124 (2017).
53 West, P. T. *et al.* Genetic and behavioral adaptation of Candida parapsilosis to the microbiome of hospitalized infants revealed by in situ genomics, transcriptomics, and proteomics. *Microbiome* **9**, 142, doi:10.1186/s40168-021-01085-y (2021).

Strain #	Clinical MIC	Clinical designation	PAP result	MIC-S	MIC-R	Fold difference	Resistance frequency
65	1	S	HR	2	128	64	0.01%
67	1	S	HR	2	64	32	0.01%
478	1	S	HR	1	64	64	0.03%
485	1	S	HR	2	128	64	0.03%
794	2	S	HR	1	64	64	1.7%
795	2	S	HR	1	64	64	0.2%
802	1	S	HR	2	128	64	0.1%
804	1	S	S	1	N.D.	N.A.	N.A.
811	1	S	S	1	N.D.	N.A.	N.A.
812	2	S	HR	4	128	32	0.02%
1191	1	S	S	2	N.D.	N.A.	N.A.
2384	2	S	HR	2	128	64	0.01%
2386	1	S	S	1	N.D.	N.A.	N.A.
2387	1	S	HR	2	128	64	0.02%
2390	0.5	S	HR	2	128	64	0.01%


a
Number of cases
Micafungin
Azoles
C. parapsilosis
C. orthopsilosis
C. metapsilosis
C. krusei
C. guilliermondii
C. tropicalis
C. lusitaniae
others
b
65
804
MYC
c
% Survival
micafungin (µg/ml)
d
Total CFU
Resistant CFU
CFU (log10)
Time (hours)
e
% Resistance
Pre-treatment
In Micafungin
Subculture
f
% Survival
micafungin (µg/ml)
g
Total CFU
Resistant CFU
CFU (log10)
Time (hours)
h
% Resistance
Pre-treatment
In Micafungin
Subculture
i
Micafungin heteroresistance
(-) mica
Low level of resistant subpopulation at baseline
(+) mica
Eradication of susceptible subpopulation
(+) mica
Expansion of resistant subpopulation
(-) mica
Resistant subpopulation back to baseline
Time

a

Allo-HCT recipients with C.p. intestinal colonization n = 30

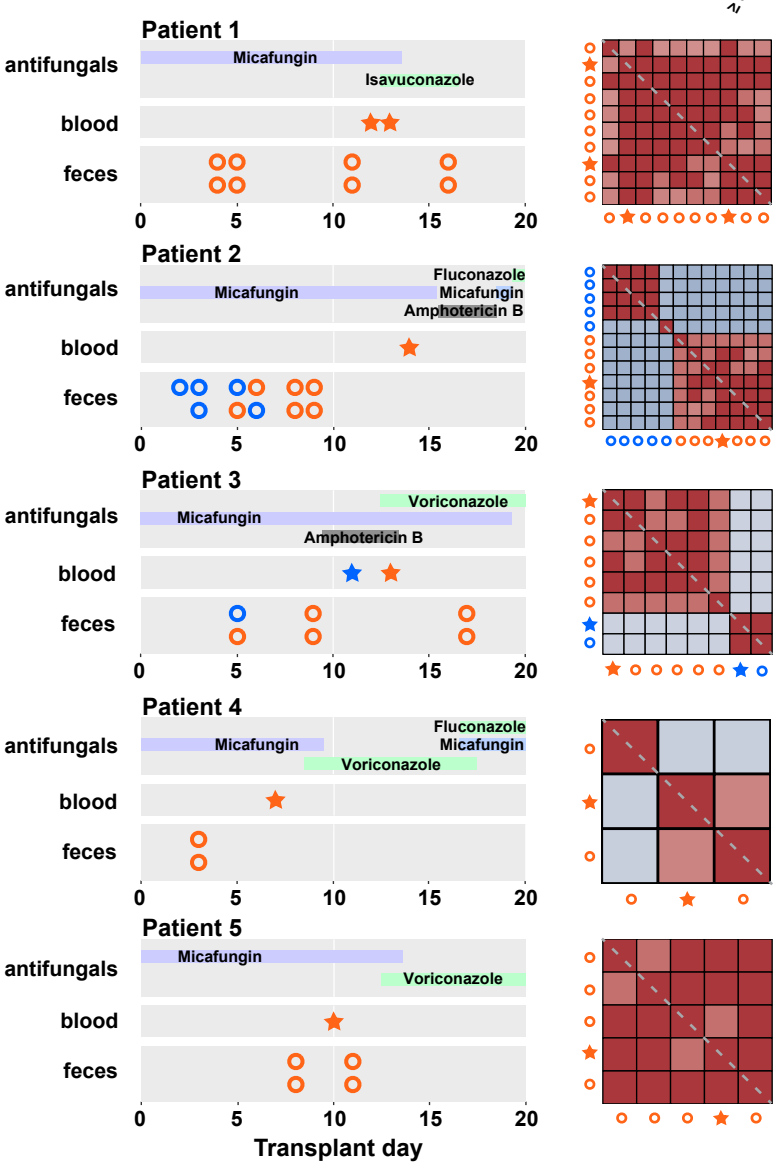
Presence of HR C.p. strains in feces	Breakthrough BSI (−)	Breakthrough BSI (+)
(−)	16	0
(+)	9	5

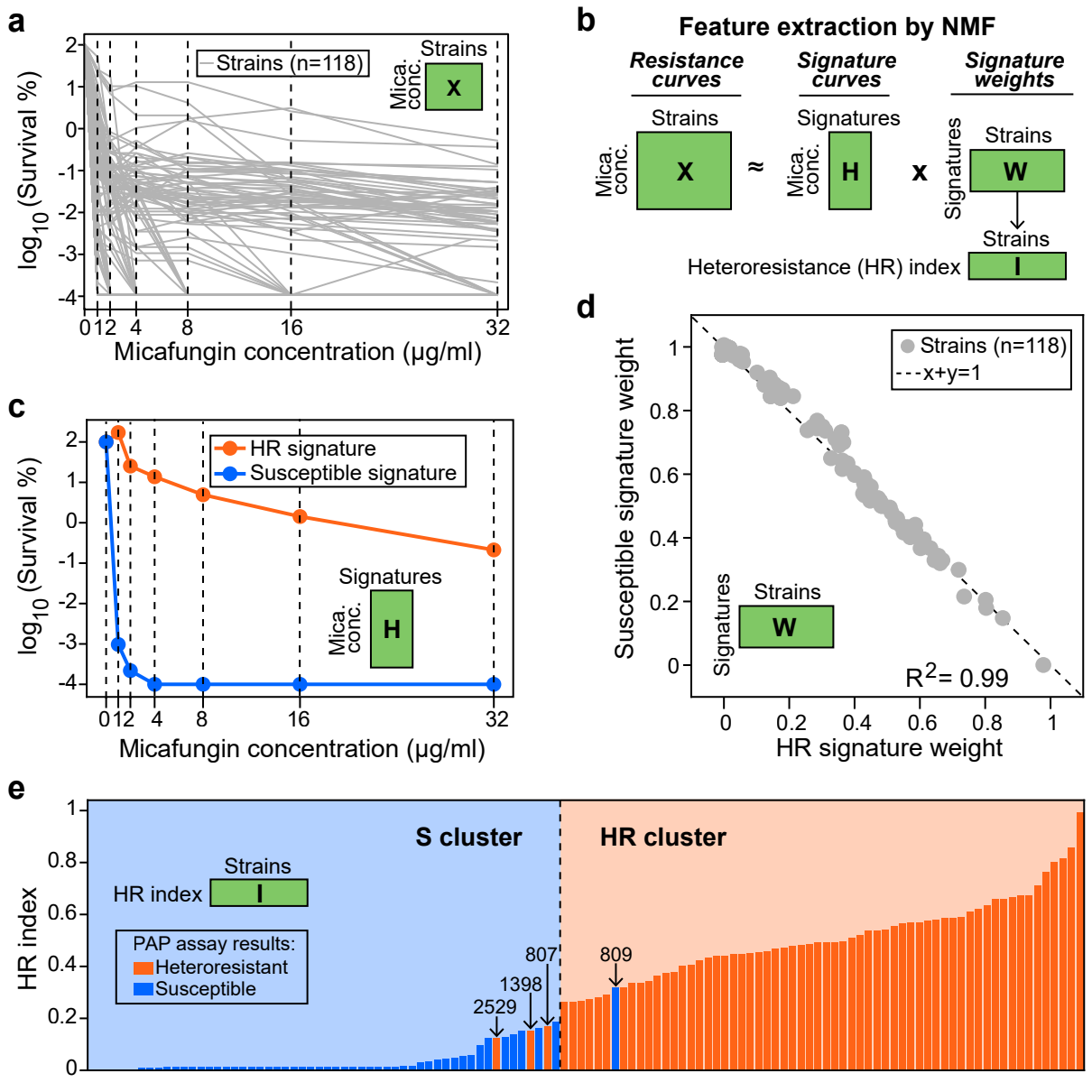
b

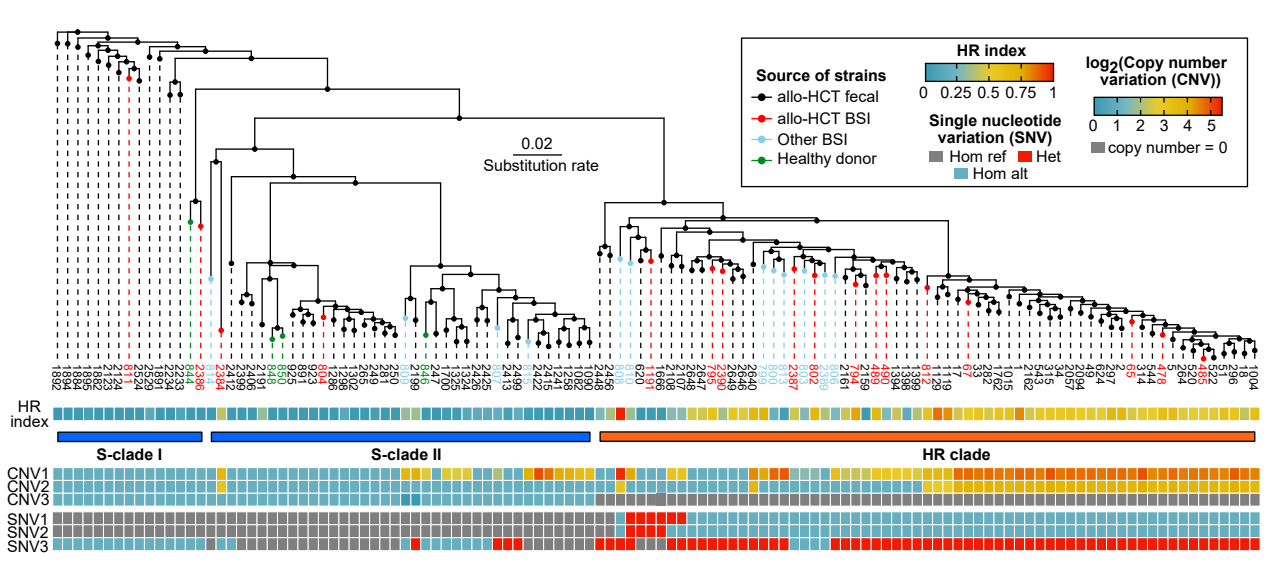
Isolation site: ★ Blood; ○ Feces

Phenotype: Heteroresistant; Susceptible

popANI (%): ≤99, 99.9, 99.99, 99.999, 99.9999, 99.99999, ≥99.999999



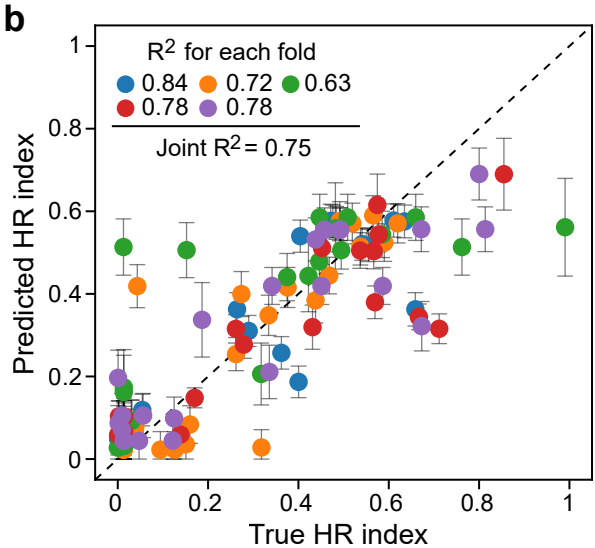

a
Strains (n=118)
Strains
Mica. conc.
X
log10(Survival %)
Micafungin concentration (μg/ml)
b
Feature extraction by NMF
Resistance curves
Signature curves
Signature weights
Strains
Mica. conc.
X
≈
Signatures
Mica. conc.
H
x
Signatures
Strains
W
Strains
Heteroresistance (HR) index
I
c
HR signature
Susceptible signature
Signatures
Mica. conc.
H
log10(Survival %)
Micafungin concentration (μg/ml)
d
Strains (n=118)
x+y=1
Susceptible signature weight
Signatures
Strains
W
R2= 0.99
HR signature weight
e
S cluster
HR cluster
Strains
HR index
I
PAP assay results:
Heteroresistant
Susceptible
807
809
1398
2529
HR index


Source of strains
allo-HCT fecal
allo-HCT BSI
Other BSI
Healthy donor
HR index
0 0.25 0.5 0.75 1
Single nucleotide variation (SNV)
Hom ref
Het
Hom alt
log2(Copy number variation (CNV))
0 1 2 3 4 5
copy number = 0
0.02
Substitution rate
HR index
S-clade I
S-clade II
HR clade
CNV1
CNV2
CNV3
SNV1
SNV2
SNV3

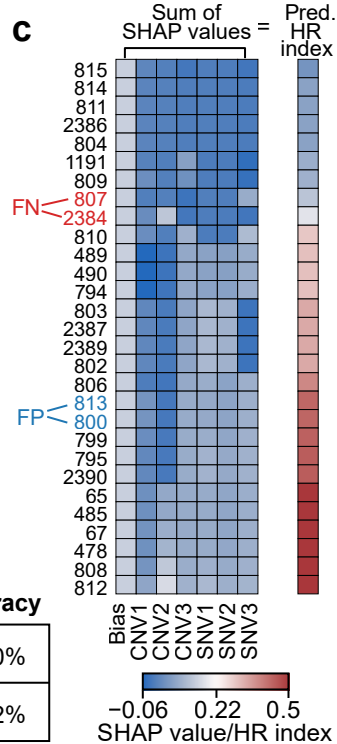
a

Feature	Description	ORF	ORF ortholog in *S. cerevisiae*
CNV1	Amplification	CPAR2_104610	*RTA3*, a lipid translocase
CNV2	Amplification	CPAR2_104600	*MAK16*, 66S pre-ribosomal particles
CNV3	Amplification	CPAR2_805410	*FAA2-3*, predicted acyl-CoA synthetase
SNV1	Hom alt	CPAR2_207400	*CAT8*, DNA-binding transcription factor
SNV2	Hom alt	Not in an ORF	N.A.
SNV3	Het	Not in an ORF	N.A.

b



c



d

	#HR (Detected/Total)	#S (Detected/Total)	Accuracy
AST	0/20 (**0.0%**)	9/9 (100.0%)	31.0%
RF	18/20 (**90.0%**)	7/9 (77.8%)	86.2%