UNIVERSITAT DE BARCELONA

Comparative genomics of recent adaptation in *Candida* pathogens

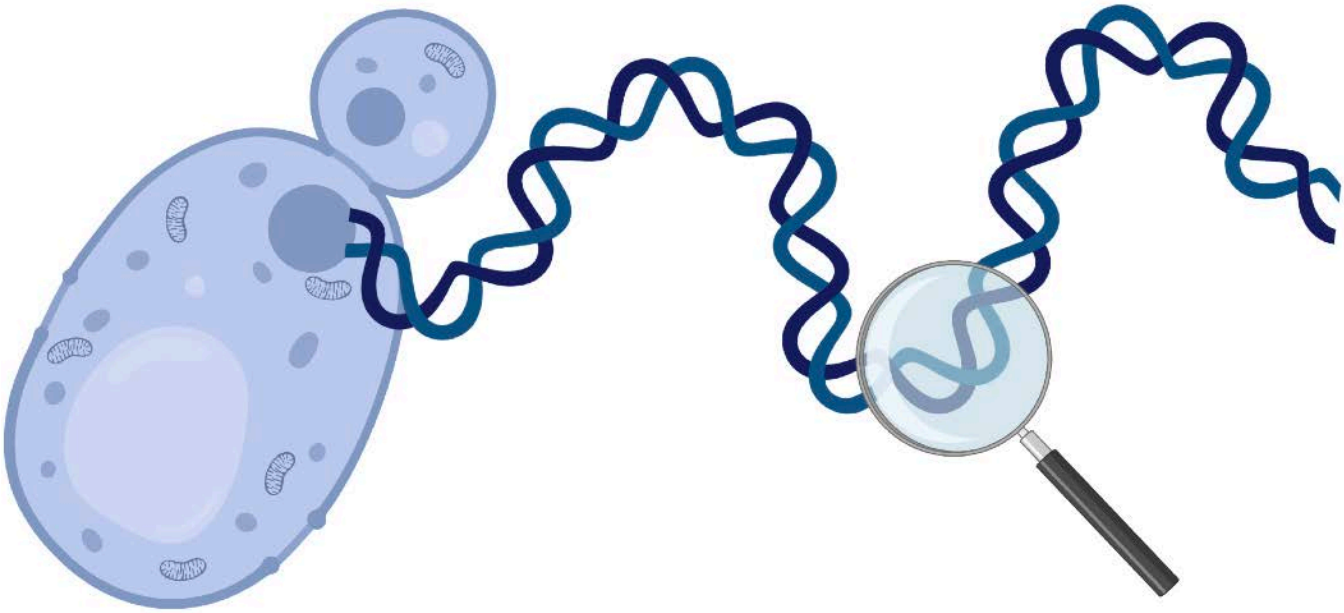
Miquel Àngel Schikora Tamarit

Comparative genomics of recent adaptation in *Candida* pathogens

PhD candidate: Miquel Àngel Schikora Tamarit

Supervisor: Toni Gabaldón Estevan

Barcelona, 2023



Comparative genomics of recent adaptation in *Candida* pathogens

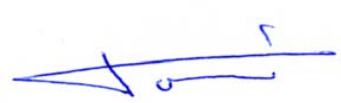
Dissertation presented by **Miquel Àngel Schikora Tamarit**,

opting for the qualification of PhD by the University of Barcelona (UB)

This doctoral thesis was performed under the direction of Dr. Juan Antonio Gabaldón Estevan,

ICREA Professor affiliated at the Barcelona Supercomputing Center

and the Institute for Research in Biomedicine (IRB Barcelona)



Juan Antonio Gabaldón Estevan	Josep Lluís Gelpí Buchaca	Miquel Àngel Schikora Tamarit
PhD thesis director	*PhD thesis tutor*	*PhD candidate*

PhD Thesis

Barcelona Supercomputing Center

Institute for Research in Biomedicine (IRB Barcelona)

Doctoral program in Biomedicine

Barcelona, 2023

UNIVERSITAT DE BARCELONA

BSC

IRB BARCELONA

ACKNOWLEDGEMENTS

Agraïments

Aquest projecte ha estat un viatge llarg i emocionant, i ha estat possible gràcies totes les persones que m'heu fet meu costat durant aquests anys. De manera directa o indirecta, considero que aquest projecte també és mèrit vostre. No tinc espai suficient per a agrair tot el que voldria, però intentaré mencionar algunes coses que per a mi han estat significatives.

En primer lloc, al Toni, per donar-me aquesta oportunitat, per tot el suport, i per deixar-me explorar tantes idees, per més esbojarrades que semblessin inicialment. També t'agraeixo que hagis creat un grup genial, on tots ens sentim persones, a banda de científics.

A tots els companys actuals i passats del Comparative Genomics lab que heu estat al meu costat. Moltes gràcies per contribuir a fer del nostre grup un espai científic i humà immillorable. D'una banda, per tota l'ajuda científica, idees i aprenentatges que m'heu aportat. A la Marina, per acompanyar-me tan bé durant els meus primers passos en el món de la genòmica comparativa i els fongs. A la Cinta, per a ajudar-me a entrar en el camp de la genòmica de poblacions i el variant calling. Al Guifré, per fer-me veure com d'interessant és anar més enllà de l'antropocentrisme hegemònic. A la Valentina, per les masterclasses sobre comunicació científica que són les teves presentacions i consells. Al Moisès, per les nostres converses sobre estadística i sobre el particular funcionament del món acadèmic. A l'Olfat, per transmetre'm la teva passió i curiositat sobre la recerca bioinformàtica translacional. Al Diego, per totes les nostres converses sobre pipelines i containers, i per tots els consells de com fer perSVade millor. A Miguel, por transmitirme su pasión por los hongos y la investigación básica. A Uciel, por todas las conversaciones sobre los *pros* y *cons* del futuro académico. A Juan Carlos, por hacerme ver que las ideas científicas requieren muchas iteraciones para ser geniales. To Veronica, for all our discussions about variant calling. To Vlad, for showing me that my lines of research have a future in the lab. To Hrant, for being a role model about how bioinformaticians in the field of *Candida* pathogens should be. To Matteo, for your help and advice about how to write genomic pipelines, which allowed me to overcome my "code pride" with perSVade. Finally, to Ewa, for being an amazing companion in all our shared projects, and for showing me that being a scientist is about questioning everything.

D'altra banda, per tots els dinars, cafès, sopars, concerts, karaokes, escape rooms virtuals, retreats, excursions, congressos, i també pels moments de penes compartides. Em venen moltíssims moments especials envoltats de tots vosaltres. Especialment amb la Marina, Guifré, Valentina, Olfat, Toni, Diego, Juan Carlos, Hrant, Matteo, Vlad, Uciel, Ewa, Ester, Manu, Martina, Giacomo, Jesse, Saioa i Moisès.

I also thank all the former members of the Single Cell Behavior Lab. To Lucas, for guiding me in my first steps in the world of science, and for introducing me to the fascinating world of yeast. Moltes gràcies també al Carlos, l'Alsu, la Lorena i la Júlia per crear un ambient genial, i per introduir-me en el món de la biologia molecular i la bioinformàtica.

Més enllà de l'àmbit científic, m'agradaria agrair el suport de totes les persones que heu estat al meu costat durant aquests anys. Als membres de la Simfònica Tekhné per crear un espai de desconnexió i música cada dimarts, que m'ajuda moltíssim a funcionar. Als amics del cor Noctes, especialment al Guillem, Marc, Jaume, Helena, Irene, Maria i Àlex, per tirar endavant aquest projecte espectacular, i també per compartir amb mi tants moments especials i de música preciosa. Als companys de the Rayberis, especialment a l'Ada, per tota la música, viatges, concerts, festes i sopars compartits, i per haver-me ajudat a superar moltes barreres mentals amb la trompeta. Moltes gràcies als Pachucos, especialment al Jordi i al Juan Carlos, per tot el suport, i per tots els *Benedictus* interpretats a llocs inapropiats. També als companys del màster, especialment a la Winona, l'Aida i la Marta, pels nostres sopars, festes, excursions i divagacions sobre ciència. Finalment, als amics biòlegs humans, Marta, Carla, Mariona, Mariona, Adri i Josep, per tots els sopars, converses, viatges, batalles de neu, festes, snorkels, escape rooms i excursions. Em doneu la vida. Marta, Carla, Mariona i Josep, estic molt content de que haguem pogut compartir aquest viatge tortuós del doctorat, i que arribem junts al final.

Moltes gràcies a tots els membres de la família, ja que m'heu ajudat a ser el científic i persona que sóc avui. An meinen Vater, für die Unterstützung, Bergtouren, gemeinsame Musik und für unsere Gespräche über Philosophie und Evolution. Al meu padrí Jordi, per introduir-me en el món de la recerca biomèdica des de ben aviat. Al Joaquim, per totes les converses, excursions, dinars en família, i per les nostres converses sobre fongs patògens i els *Cordyceps*. Al Frederic, per totes les converses profundes, la música compartida, concerts i els moments amb amics i en família, i per les nostres discussions sobre el *vibrato*. A la meva mare, per ser el pal de paller de la nostra família malgrat les dificultats, per cuidar-me sempre, per tot el suport i per fer-me sentir molt valorat. També per tots els concerts, música compartida, els moments en família, les classes de cant i per sempre mostrar un interès per la meva recerca. Finalment, a la Marina, per ser una companya de viatge fantàstica, que mai m'hauria pogut imaginar que trobaria. D'una banda, has estat la meva *sidekick* de la tesi, i t'agraeixo tot el suport científic, els *pomodoros*, els consells sobre presentacions, text i figures, i per escoltar les meves divagacions. D'altra banda, moltes gràcies per escoltar-me sempre, per ser molt observadora, pel suport en els moments difícils, pels viatges, la música compartida, per entrar a formar part de la meva família, pels cants a la natura, pel submarinisme, per les nostres divagacions, i per tots els moments petits i grans que espero que seguim compartint.

ABSTRACT

Abstract in english

Fungal infections pose a serious health threat, affecting >1,000 million people and causing ~1.5 million deaths each year. The problem is growing due to insufficient diagnostic and therapeutic options, increased number of susceptible patients, expansion of pathogens partly linked to climate change and the rise of antifungal drug resistance. Among other fungal pathogens, *Candida* species are a major cause of severe hospital-acquired infections, with high mortality in immunocompromised patients. Various *Candida* pathogens constitute a public health issue, which require further efforts to develop new drugs, optimize currently available treatments and improve diagnostics. Given the high dynamism of *Candida* genomes, a promising strategy to improve current therapies and diagnostics is to understand the evolutionary mechanisms of adaptation to antifungal drugs and to the human host. Previous work using *in vitro* evolution, population genomics, selection inferences and Genome Wide Association Studies (GWAS) have partially clarified such recent adaptation, but various open questions remain. In the three research articles that conform this PhD thesis we addressed some of these gaps from the perspective of comparative genomics.

First, we addressed methodological issues regarding the analysis of *Candida* genomes. Studying recent adaptation in these pathogens requires adequate bioinformatic tools for variant calling, filtering and functional annotation. Among other reasons, current methods are suboptimal due to limited accuracy to identify structural variants from short read sequencing data. In addition, there is a need for easy-to-use, reproducible variant calling pipelines. To address these gaps we developed the “personalized Structural Variation detection” pipeline (perSVade), a framework to call, filter and annotate several variant types, including structural variants, directly from reads. PerSVade enables accurate identification of structural variants in any species of interest, such as *Candida* pathogens. In addition, our tool automatically predicts the structural variant calling accuracy on simulated genomes, which informs about the reliability of the calling process. Furthermore, perSVade can be used to analyze single nucleotide polymorphisms and copy number-variants, so that it facilitates multi-variant, reproducible genomic studies. This tool will likely boost variant analyses in *Candida* pathogens and beyond.

Second, we addressed open questions about recent adaptation in *Candida*, using perSVade for variant identification. On the one hand, we investigated the evolutionary mechanisms of drug resistance in *Candida glabrata*. For this, we used a large-scale *in vitro* evolution experiment to study adaptation to two commonly-used antifungals: fluconazole and anidulafungin. Our results show rapid adaptation to one or both drugs, with moderate fitness costs and through few mutations in a narrow set of genes. In addition, we characterize a novel role of *ERG3* mutations in cross-resistance towards fluconazole in

anidulafungin-adapted strains. These findings illuminate the mutational paths leading to drug resistance and cross-resistance in *Candida* pathogens. On the other hand, we reanalyzed ~2,000 public genomes and phenotypes to understand the signs of recent selection and drug resistance in six major *Candida* species: *C. auris, C. glabrata, C. albicans, C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis*. We found hundreds of genes under recent selection, suggesting that clinical adaptation is diverse and complex. These involve species-specific but also convergently affected processes, such as cell adhesion, which could underlie conserved adaptive mechanisms. In addition, using GWAS we predicted known drivers of antifungal resistance alongside potentially novel players. Furthermore, our analyses reveal an important role of generally-overlooked structural variants, and suggest an unexpected involvement of (para)sexual recombination in the spread of resistance. Taken together, our findings provide novel insights on how *Candida* pathogens adapt to human-related environments and suggest candidate genes that deserve future attention. In summary, the results of this thesis improve our knowledge about the mechanisms of recent adaptation in *Candida* pathogens, which may enable improved therapeutic and diagnostic applications.

Sinopsi en català

Les infeccions fúngiques representen una greu amenaça per a la salut, afectant a més de 1.000 milions de persones i causant aproximadament 1,5 milions de morts cada any. El problema està augmentant a causa d'unes opcions terapèutiques i diagnòstiques insuficients, l'increment del nombre de pacients susceptibles, l'expansió dels patògens parcialment vinculada al canvi climàtic i l'augment de la resistència als fàrmacs antifúngics. D'entre diversos fongs patògens, els llevats del gènere *Candida* són una causa important d'infeccions nosocomials, amb una alta mortalitat en pacients immunodeprimits. Diverses espècies de *Candida* constitueixen un problema de salut pública, cosa que requereix més esforços per a desenvolupar nous medicaments, optimitzar els tractaments disponibles i millorar els diagnòstics. Tenint en compte el dinamisme genòmic d'aquests patògens, una estratègia prometedora per millorar les teràpies i diagnòstics actuals és comprendre els mecanismes evolutius d'adaptació als fàrmacs antifúngics i a l'hoste humà. Treballs anteriors utilitzant l'evolució *in vitro*, la genòmica de poblacions, les inferències de selecció i els estudis d'associació de genoma complet (GWAS, per les sigles en anglès) han aclarit parcialment aquesta adaptació recent, però encara hi ha diverses preguntes obertes. En els tres articles que conformen aquesta tesi doctoral, hem abordat algunes d'aquestes preguntes des de la perspectiva de la genòmica comparativa.

En primer lloc, hem abordat qüestions metodològiques relatives a l'anàlisi dels genomes de les espècies *Candida*. L'estudi de l'adaptació recent en aquests patògens requereix eines bioinformàtiques adequades per a la detecció, filtratge i anotació funcional de variants genètiques. Entre altres raons, els mètodes actuals són subòptims a causa de la limitada precisió per identificar variants estructurals a partir de dades de seqüenciació amb lectures curtes. A més, hi ha una necessitat d'eines computacionals per a la detecció de variants que siguin senzilles d'utilitzar i reproduibles. Per abordar aquestes mancances, hem desenvolupat el mètode bioinformàtic "personalized Structural Variation detection" (perSVade), una eina que permet la detecció, filtratge i anotació de diversos tipus de variants, incloent-hi les variants estructurals, directament des de les lectures. PerSVade permet la identificació precisa de les variants estructurals en qualsevol espècie d'interès, com ara els patògens *Candida*. A més, la nostra eina prediu automàticament la precisió de la detecció d'aquestes variants en genomes simulats, la qual cosa informa sobre la fiabilitat del procés. Finalment, perSVade es pot utilitzar per analitzar altres tipus de variants, com els polimorfismes de nucleòtid únic o els canvis en el nombre de còpies, facilitant així estudis genòmics integrals i reproduibles. Aquesta eina probablement impulsarà les anàlisis genòmiques en els patògens *Candida* i també en altres espècies.

En segon lloc, hem abordat algunes de les preguntes obertes sobre l'adaptació recent en els llevats *Candida*, utilitzant perSVade per a la identificació de variants. D'una banda, hem investigat els mecanismes evolutius de resistència als fàrmacs antifúngics en *Candida glabrata*. Per a això, hem utilitzat un experiment

d'evolució *in vitro* a gran escala per estudiar l'adaptació a dos antifúngics comuns: el fluconazol i l'anidulafungina. Els nostres resultats mostren una adaptació ràpida a un o ambdós fàrmacs, amb un cost per al creixement moderat i a través de poques mutacions en un nombre reduït de gens. A més, hem caracteritzat un paper nou de les mutacions en *ERG3* en la resistència creuada al fluconazol en soques adaptades a anidulafungina. Aquests descobriments aclareixen els processos mutacionals que condueixen a la resistència als fàrmacs i a la resistència creuada en els patògens *Candida*. D'altra banda, hem re-analitzat aproximadament 2.000 genomes i fenotips disponibles en repositoris públics per a comprendre els senyals genòmics de selecció recent i de resistència a fàrmacs antifúngics, en sis espècies rellevants de *Candida*: *C. auris*, *C. glabrata*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* i *C. orthopsilosis*. Hem trobat centenars de gens sota selecció recent, suggerint que l'adaptació clínica és diversa i complexa. Aquests gens estan relacionats amb funcions específiques de cada espècie, però també trobem processos alterats de manera similar en diferents patògens, com per exemple l'adhesió cel·lular, cosa que indica fenòmens d'adaptació conservats. A part, utilitzant GWAS hem predit mecanismes esperats de resistència a antifúngics i també possibles nous factors. A més, les nostres anàlisis revelen un paper important de les variants estructurals, generalment poc estudiades, i suggereixen una implicació inesperada de la recombinació (para)sexual en la propagació de la resistència. En conjunt, els nostres descobriments proporcionen noves perspectives sobre com els patògens *Candida* s'adapten als entorns humans, i suggereixen gens candidats que mereixen investigacions futures. En resum, els resultats d'aquesta tesi milloren el nostre coneixement sobre els mecanismes d'adaptació recent en els patògens *Candida*, cosa que pot permetre el disseny de noves teràpies i diagnòstics.

TABLE OF CONTENTS

TABLE OF CONTENTS

1. GENERAL INTRODUCTION

1. GENERAL INTRODUCTION

1.1. Fungal pathogens pose a pressing threat

Fungi include a diverse set of organisms that play many important roles across ecosystems. In this chapter we introduce the relevance of fungi as a growing threat for food security, biodiversity and human health. In addition, we describe why *Candida* fungal pathogens are a public health issue, which deserves broader attention.

1.1.1. The fungal kingdom is highly diverse

The kingdom ”Fungi” includes >100,000 described and ~2-4 million estimated species (*1*). The morphologies of these organisms range from single-celled yeasts to multicellular mushrooms and filamentous molds (*2*, *3*). Fungi are key players of ecosystems due to various functions in organic matter digestion, influencing the soil and atmospheric availability of carbon, oxygen, phosphorus and nitrogen. In addition, fungi establish symbiotic and parasitic relationships with bacteria, plants and animals, which are relevant for biodiversity dynamics (*4*). This ability to infect animals sometimes constitutes a threat for human health (*5*), which is the main motivation for this PhD project. Consistent with their broad importance, fungi have been found across diverse environments, including the stratosphere (*6*), antarctic glaciers (*7*) or the gut of insects and mammals (*8*, *9*). All in all, fungi are ubiquitous, morphologically diverse and key ecosystemic players.

Consistent with these broad functions, the evolutionary divergence of fungi is also high. Current work has defined eight phyla (or groups) in fungi: Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Zoopagomycota, Mucoromycota, Glomeromycota, Basidiomycota and Ascomycota (*10*, *11*) (**Figure 1A**). Chytridiomycota includes free-living saprobes and parasites belonging to three classes: Chytridiomycetes, Monoblepharidomycetes and Hyaloraphidiomycetes (*12*, *13*). Neocallimastigomycota is a group of non-parasitic anaerobes with flagella, formed by a single family with a debated phylogenetic position (Neocallimastigomycota may actually be within Chytridiomycota) (*10*, *14*). Blastocladiomycota includes several zoosporic fungi with various morphologies and ecological capacities (*15*). Zoopagomycota is an early-diverging clade of non-flagellated fungi with common animal association, composed by Zoopagomycotina, Entomophthoromycotina and Kickxellomycotina (*10*, *16*). Mucoromycota is a large group of saprobes, plant parasites and ectomycorrhizal species, composed by the Mortierellomycotina and Mucoromycotina subphyla (*16*). Glomeromycota is a group of obligate plant symbionts forming mycorrhizae (*17*). Basidiomycota is a species-rich phylum comprising an array of lifestyles and morphologies, including multicellular mushrooms, divided into Pucciniomycotina, Ustilagomycotina, and Agaricomycotina. A

common aspect of them is that they have basidia, a specialized cell type related to sporulation (*10*, *18*). Finally, Ascomycota is the largest phylum comprising more than half of the described species, and contains three main classes: Taphrinomycotina, Saccharomycotina and Pezizomycotina. Fungi within this phylum range from simple yeasts to fungi with highly complex macroscopic fruiting bodies (*19*). In summary, fungi is a kingdom with a high genetic diversity, likely above other traditional kingdoms, as illustrated by the fact that the sequence divergence within Saccharomycotina (the budding-yeasts subphylum of Ascomycota) is comparable to the divergence found within plants and animals (*20*). In the following paragraphs we provide an overview about why some of these fungal species are a threat for food security, biodiversity and human health.

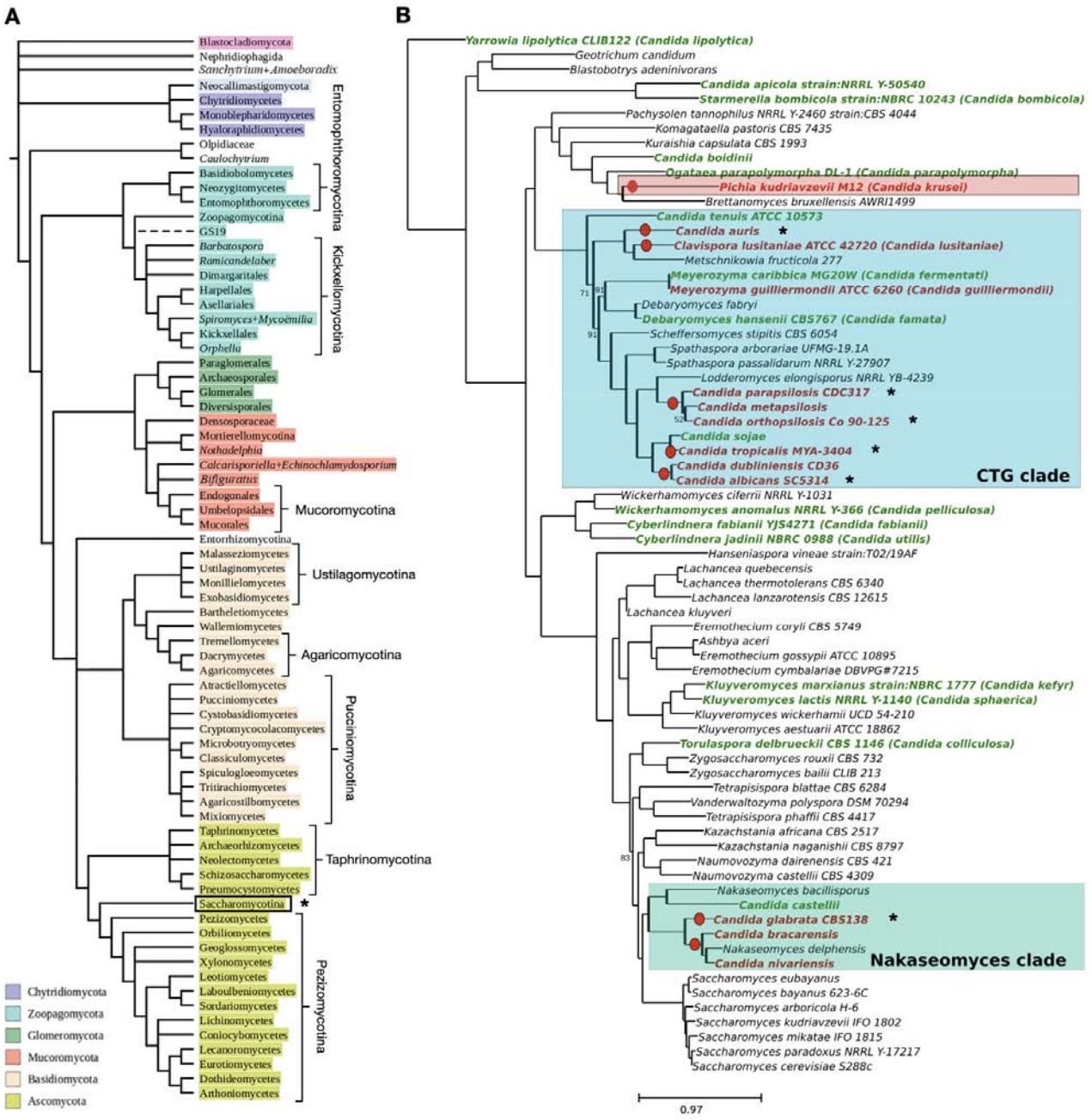


Figure 1. Diversity in fungi and Saccharomycotina. (A) Fungal tree of life, adapted from (*10*), with the main phyla outlined in colors. The leaves of the tree contain relevant fungal groups. Groups with no color and dashed branches have unclear taxonomic affiliation (*incertae sedis*), according to (*10*). The square brackets indicate the main clades within each phylum, mentioned in section 1.1.1, that are composed by different subgroups. The asterisk indicates the position of Saccharomycotina, studied in this PhD thesis. **(B)** Phylogenetic tree of Saccharomycotina species with a focus on major *Candida* pathogens, adapted from (*21*). Pathogenic / non-pathogenic *Candida* species are colored in red / green, respectively. The asterisks indicate the species studied in this PhD thesis.

1.1.2. Fungi are a threat for food security and biodiversity

Some fungal species are a major threat to food security because they can cause widespread damage to crops, leading to significant reductions in yield and quality (*22*, *23*). Current estimates indicate that fungal plant pathogens spoil around 30% of crop production, which can result in food shortages, increased prices, and decreased nutritional value of available food (*23*, *24*). This damage can occur at various stages of plant growth, and the causal agents are spread across fungal phyla. Several species belonging to the Pezizomycotina group in Ascomycota are major plant pathogens. Among them, *Magnaporthe oryzae*a is a filamentous fungus causing rice blast disease. *Botrytis cinerea* is a necrotrophic organism targeting mostly wine grapes. *Fusarium* species affect various cereals *Blumeria graminis* generates powdery mildew of grasses in many cereals such as wheat and barley. *Mycosphaerella graminicola* underlies the Stritici blotch (STB) disease in wheat, often in temperate regions. Finally, *Colletotrichum* species cause anthracnose spots and blights in various aerial plant parts, and also cause postharvest rots. Similarly, some Basidiomycota are important plant pathogens. For example, *Puccinia* species (from the Pucciniomycotina group) destroy different wheat organs, including stems and leaves. In addition, *Ustilago maydis* (Ustilagomycotina) generates corn smut. These pathogens are particularly concerning in developing countries in the tropic where crops are grown in conditions that are prone to fungal attacks (i.e. higher humidity and temperature) (*25*). The number of countries affected by such tropical crop diseases is growing, likely driven by higher temperatures derived from climate change (*26*). In summary, all these organisms are common fungal pathogens underlying massive crop losses, which could be addressed with new fungicides, better crop management practices and improving current crop varieties (*27*, *28*).

Beyond direct plant pathogenesis, fungi are a threat towards food security due to the production of mycotoxins, which cause serious health problems in humans when consumed. Mycotoxins are secondary metabolites produced by some fungal species, which can persist in food like cereals, dried fruits, spices and nuts, even after processing and cooking (*29*). The most important such compounds are produced by *Aspergillus*, *Penicillium* and *Fusarium* molds, all from the Pezizomycotina group in Ascomycota. Among them, *Aspergillus* species produce aflatoxins and ochratoxin A. Similarly, *Penicillium* species generate ochratoxin A and patulin. Finally, *Fusarium* species produce fumonisins, type A and type B trichothecenes (*30*). In humans, these compounds can generate cancer (aflatoxins, ochratoxin A, patulin, fumonisins), renal failure (ochratoxin A), teratogenesis (fumonisins), neurologic alterations, gastrointestinal damage and/or immune suppression (trichothecenes) (*31–35*). Moreover, mycotoxins can also contaminate animal feed, causing health problems in livestock and reducing the quality and quantity of animal-based food (*36*). These damaging effects of mycotoxins could be worsened by the fact that some mycotoxigenic fungi may benefit from climate change factors (increased temperatures, elevated CO_2 and drought stress) (*30*, *36*). In addition, the impact of mycotoxins is higher in developing countries that have insufficient food safety policies (*37*). In

summary, mycotoxins compound the dangers of fungi for food security through various health alterations, which require improving food safety and quality control measures.

In addition to food security issues, some pathogenic and opportunistic fungi can cause severe diseases in wild animals and plants, which drive significant biodiversity loss (*38*). For example *Batrachochytrium dendrobatidis* and *Batrachochytrium salamandrivorans* (Chytridiomycetes within the Chytridiomycota phylum) cause chytridiomycosis in amphibians, which has generated the decline or extinction of >500 species (*39*, *40*). To put things in scale, this constitutes the greatest biodiversity loss generated by infections. This is partly due to human activities, since the international trade of amphibians increases the outbreaks of such chytrid pathogens (*41*). Similarly, *Pseudogymnoascus destructans* (from the Pezizomycotina group in Ascomycota) causes white-nose syndrome in bats, which has reduced by 90% several bat populations across North America (*40*, *42*). In addition to these well-studied diseases, various fungi cause devastating disease in snakes, lizards, dolphins, birds and lizards (*40*). However, this threat towards wild animals is mostly overlooked, and some studies claimed the need of improved control strategies, such as bat vaccination, raising public awareness, sterilization of fungal reservoir species, antifungal treatment of tadpoles, probiotic therapy and/or biological control of fungi with mycoviruses (*40*, *43*). As with animals, some fungi have driven the decline in populations of wild plants. For example *Cryphonectria parasitica* (from the Pezizomycotina group in Ascomycota) causes chestnut blight, which has devastated chestnut trees across Europe and America (*44*). In addition to killing individual organisms, this decline of tree species due to fungal diseases can lead to changes in carbon or nitrogen cycles, leading to cascading effects that threaten the stability of the ecosystem (*45*). In conclusion, the pathogenic capabilities of fungi on animals and plants are significant contributors to biodiversity loss, which deserves further attention.

In summary, fungal pathogens are an overlooked issue driving food scarcity and biodiversity loss.

1.1.3. Fungi are a threat for human health

As with other animals, fungal pathogens can cause serious diseases in humans, which constitutes a growing public health issue. They affect >1,000 million people and cause an array of diseases ranging from mild skin alterations to invasive disseminated infections. These severe fungal infections cause ~1.5 million deaths each year worldwide, which is comparable to the death toll attributed to tuberculosis (*46*, *47*). This problem has been increasing over the last decades mostly due to the growing population of susceptible patients. On the one hand, this can be attributed to tradeoffs in the recent advances of medicine. For instance, the widespread accessibility to immune-disrupting therapies (including chemotherapy or immunosuppressants after organ transplants) has generated an increased population of immunocompromised patients highly prone to such infections (*46*, *48*, *49*). Similarly, the increased survival of neonates has generated more fungal diseases, as these patients are highly susceptible to infections (*50*). Another reason is the extensive usage of antibiotics, which can dysregulate the microbiome promoting fungal overgrowth (*51*). On the other hand, several emerging diseases or conditions (often associated with aging) increase the susceptibility towards fungal infections, which compounds the problem. Among others, these include non-tuberculous mycobacterial infections, HIV/AIDS, kidney and liver disease, respiratory infections caused by viruses such as influenza or SARS-CoV2 and chronic obstructive pulmonary disease (COPD) (*47*, *52*–*54*).

This issue is aggravated by common therapeutic failure and inefficient diagnostic tools. There are only four families of antifungal drugs suitable for systemic infections: pyrimidines (5-flucytosine), echinocandins (caspofungin, anidulafungin and micafungin), azoles (such as fluconazole, itraconazole, voriconazole or posaconazole) and polyenes (amphotericin B). 5-flucytosine is converted into 5-fluorouracil in the fungal cell, which is metabolized into compounds that impair RNA and DNA synthesis, generating fungistatic or fungicidal effects depending on the species (*55*, *56*). Echinocandins are fungicidal drugs that target the 1,3-ß-D-glucan synthase, essential for fungal cell wall homeostasis (*57*). Azoles are fungistatic compounds that inhibit the lanosterol 14α-demethylase enzyme, essential for ergosterol biosynthesis and membrane integrity (*58*). Finally, polyenes are fungicidal drugs that directly bind to ergosterol and block its function (*59*). These compounds have frequent toxicity in humans (affecting liver and kidneys), high cost, restricted therapeutic range and rising drug resistance (*48*, *60*–*63*). For example, 5-flucytosine is mostly considered an adjuvant because it is only effective in combination with other drugs, due to frequent resistance when used as monotherapy (*55*). In addition, the effective drugs are often unavailable in many low and middle-income countries, which can contribute to increased mortality of fungal infections (*47*, *64*, *65*). Furthermore, investment in antifungal drug development is low due difficult identification of therapeutic targets and the fact that patients often have severe comorbidities (which alter the assessment of effectiveness and make clinical studies particularly expensive) (*63*, *66*). Regarding diagnostics, common methods are inefficient

because they are slow (as they need culturing the pathogen) and often lack satisfactory specificity and sensitivity to detect the infecting species or drug susceptibility profiles (*67*, *68*).

Beyond the problems with increasing susceptible populations, ineffective therapies and inaccurate diagnostics, climate change likely worsens the effects of fungal pathogens on humans. Global warming may select for environmental fungi adapted to higher temperatures, sometimes close to those in the human body. This is an issue because the high body temperature in humans is a barrier for most fungal infections, and thus the increase of such thermotolerant fungi may promote the emergence of new pathogens (*40*, *69*, *70*). In addition, climate change may expand the geographical distribution of human fungal pathogens (as with plant pathogens) and their vectors, resulting in higher prevalence of these infections (*71*, *72*). Finally, increased climatological extreme events, such as floods or hurricanes, may promote more frequent aerosolization of these pathogens and/or higher implantation via traumatic wounds (*71*). Such factors likely explain the recent emergence and expansion of some human fungal pathogens, such as *Candida auris* (*71*, *73*). In summary, human fungal pathogens are a rising, but overlooked, threat for public health, requiring urgent improvements in therapies and diagnostics. This constitutes the main justification for this PhD project. In the following paragraphs we provide an overview about the most important human fungal pathogens.

There are >600 fungal species that are potential human pathogens (*74*), but a few of them stand out as the most important threats. The World Health Organization (WHO) has recently elaborated a list of the most important fungal pathogens, taking into account fatality rates, morbidity, prevalence, geographic distribution, transmissibility dynamics, long term effects, prevalence of antifungal drug resistance, effectiveness of diagnostics and treatment options. Among them, *Cryptococcus neoformans*, *Aspergillus fumigatus,* some *Candida* species (C. *auris*, C. *glabrata*, C. *albicans*, C. *tropicalis* and C. *parapsilosis*), *Histoplasma* species, eumycetoma-causative agents, Mucorales and *Fusarium* species have been classified as high or critical priority targets for research and public health action (*47*). In the following paragraphs we describe these major human fungal pathogens and their most impactful health consequences.

Several of these important pathogens belong to the Pezizomycotina group in Ascomycota. Among them, *Histoplasma* species are environmental molds found all over the world that cause histoplasmosis, affecting the lungs and the central nervous system. This disease can affect healthy individuals, but it has particularly high mortality (21-53%) among HIV immunocompromised patients. In addition, it can cause severe outbreaks. Treatment options are available (based on azoles and/or polyenes), as drug resistance remains moderate (although rarely monitored) (*47*, *75*). Conversely, *Aspergillus fumigatus* is a mold globally distributed in the environment. In humans, it can cause aspergillosis, which ranges from mild allergic reactions to severe invasive infections in immunocompromised patients. The fungus is usually inhaled into

the lungs but can spread to other parts of the body, such as the central nervous system. Azole resistance is on the rise, often due to the agricultural use of this antifungal drug, which generates resistant strains that have high mortality (47%-100% depending on the study) (*3*, *47*). In addition, *Fusarium* species are a broad group of molds causing fusariosis in humans, an invasive disease of the lungs, eyes, central nervous system and internal organs. These species are mostly environmental and globally distributed, and they affect mostly immunocompromised patients. Such invasive fusariosis has death rates ranging from 43% to 67%, mostly due to high intrinsic resistance towards available antifungal drugs (*47*, *76*). On another line, eumycetoma is a deep tissue infection that results in low mortality, but serious disability. It is caused by various Pezizomycotina fungi (*Madurella* species, *Falciformispora senegalensis*, *Curvularia lunata*, *Scedosporium* species, *Zopfia rosatii*, *Acremonium* and *Fusarium* species) that can enter the body through skin breaks. The exact global incidence is uncertain, but it is particularly common among young, male farmers in tropical, low-income countries. Antifungal treatment is available to treat eumycetoma and resistance is not a major concern, but amputation of the infected area is often necessary (in ~39% of cases) (*47*, *77*).

Also within Ascomycota, yeasts from the paraphyletic *Candida* genus (in the Saccharomycotina group) are important human fungal pathogens. These species are often human commensals, but they can become opportunistic pathogens upon immunocompromised states or antibiotic usage (*51*, *78*). The most common effects of *Candida* infections are mild mucosal alterations, such as vulvovaginal candidiasis, which affects 75% of women at least once during their lives (*47*, *48*, *79*). Furthermore, *Candida* pathogens can cause severe invasive infections involving the blood (candidemia), heart, central nervous system, bones, eyes or other internal organs, often in immunocompromised individuals (*47*, *48*). Such invasive infections have a mortality between 20-60% (*47*), and candidemia is responsible for >400,000 life-threatening infections worldwide each year, mostly in developed countries (*48*, *80*). There are >30 *Candida* species that can be pathogenic, distributed all over the Saccharomycotina tree (*51*, *81*) (**Figure 1B**), but the most relevant in terms of public health are *C. auris*, *C. glabrata*, *C. albicans*, *C. tropicalis* and *C. parapsilosis* (*47*). Antifungal resistance towards one or several major drug classes is rising in some of these species (such as *C. auris* and *C. glabrata*), which hampers clinical management of these infections (*60*, *61*, *82*). In addition, a common therapeutic limitation for *Candida* infections is that effective antifungal drugs (mostly echinocandins) are often unavailable (especially in low and middle income countries) (*47*, *83*, *84*). Since *Candida* species are the main object of study of this project, the next section (1.1.4) describes more in depth the clinical relevance of the most important species.

Finally, there are many relevant non-Ascomycota human pathogens, including *Cryptococcus neoformans* and Mucorales species. *Cryptococcus neoformans* (from the Agaricomycotina group in Basidiomycota) is an environmental opportunistic pathogen causing invasive cryptococcosis in immunocompromised patients

(mostly HIV+ individuals). Inhalation of spores from the environment is the most common source of infection, and serious complications include spread to the brain and blood. The mortality is high (41%-61%), likely due to azole resistance and the lack of necessary antifungal drugs in many countries (*47*, *85*). In addition, various globally-distributed Mucorales species (within the Mucoromycotina group of Mucoromycota), such as *Mucor*, *Rhizopus* or *Lichthiemia* species, cause a broad range of infections referred to as mucormycosis. The disease is spread through spore inhalation or skin breaks, and affects eyes, gastrointestinal tract and central nervous system. It affects mostly immunocompromised patients, but also in individuals with poorly managed diabetes mellitus and those who have sustained skin or soft tissue injuries. Invasive mucormycosis has a mortality ranging from 23% to 80% in adults and 72% in children. This is partly due to intrinsic resistance towards some azoles and echinocandins, and occasional resistance to polyenes in these species (*47*, *86*).

Various studies have suggested that the lack of effective treatments and diagnostic tools for most of these human fungal pathogens requires further research to improve public health actions. Among these, *in vitro* and *in vivo* studies may be key to find synergistic effects between drugs, which could optimize treatment regimes (*87*). In addition, global incidence and distribution of antifungal resistance is unknown for most of these pathogens, which requires better surveillance studies (*47*, *88*). Similarly, it is necessary to explore the effectiveness of preventive measures such as vaccination or prophylactic therapies, particularly for patients undergoing chemotherapy or organ transplants (*47*, *89*). In addition, the use of genomic tools such as Next Generation Sequencing or proteomics may be useful to improve diagnostics (*68*). Furthermore, we need a better understanding about the concrete risk factors in patients to optimize clinical management of these infections (*90*, *91*). Finally, a better understanding of the evolutionary (genomic) events driving clinical adaptation (to the host and/or to antifungal drugs) could be useful to optimize treatments and even develop new drugs (*92–94*), which constitutes a major driver of this PhD project.

All in all, fungal pathogens are an overlooked public health issue, deserving further attention. In the next section we describe in depth the clinical relevance of the subset of human pathogens studied in the project: *Candida* species.

1.1.4. *Candida* species are major human pathogens

As introduced in the previous section, yeasts from the *Candida* group are widespread human opportunistic pathogens, which present many therapeutic and diagnostic challenges. In this section we describe the clinical relevance of major *Candida* pathogens investigated in this project. These include five species considered among the high/critical priority targets of the WHO (*C. auris*, *C. glabrata*, *C. albicans*, *C. tropicalis* and *C. parapsilosis*) (*47*) and *C. orthopsilosis*, an emerging hybrid yeast pathogen (*95*). These *Candida* organisms grow mostly as budding yeasts, and they can form true hyphae and/or pseudohyphae upon various conditions (which is a relevant virulence mechanism) (*96*–*98*). At the genome level, some species are diploid (*C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis*) and others are mostly haploid (*C. auris* and *C. glabrata*) (*99*). From a phylogenetic perspective they are highly diverse, distributed in a paraphyletic manner across the Saccharomycotina tree (*21*) (**Figure 1B**). This paraphyly is due to the fact that the original *Candida* genus included yeasts with ability to form hyphae or pseudohyphae, but lacking experimentally-observed sexual cycles and spore formation (*100*). Most of these taxa (*C. auris*, *C. albicans*, *C. tropicalis, C. parapsilosis* and *C. orthopsilosis*) belong to the 'CTG clade' (formerly known as the '*Candida* clade'), which has this name because its species reassigned the leucine CTG codon to encode serine (*101*). *C. albicans* and *C. tropicalis* belong to close sister lineages within the CTG clade (*21*). Similarly, *C. parapsilosis* and *C. orthopsilosis* are close taxa belonging to the same species complex: the *C. parapsilosis* species complex (*102*). Conversely, *C. auris*'s lineage branches early in the CTG clade. Finally, *C. glabrata* belongs to the distant Nakaseomyces clade (some studies name this species as *Nakaseomyces glabrata* (*103*)), which is closer to the *Saccharomyces cerevisiae* (brewer's yeast) than to the other *Candida* species (*21*, *49*). To put this in perspective, the sequence divergence between *C. albicans* and *C. glabrata* is above the divergence between humans and fishes (*20*). In summary, there is a high diversity across *Candida* species, which justifies the need to handle them differently in the clinics (*47*, *104*). The following paragraphs describe in more depth their clinical impact.

Candida albicans is a common commensal yeast, found in mouth, throat, vagina, gut and skin. It can become an opportunistic pathogen and cause mild disease such as oropharyngeal, esophageal, cutaneous and vulvovaginal candidiasis. In immunocompromised patients, however, this species may produce severe invasive infections of blood, eyes, bones, central nervous system, heart and internal organs (*47*, *105*). Such invasive infections result in hospital stays between two weeks and four months, and the associated mortality is between 20-50%, even with the usage of effective antifungals (*47*). *C. albicans* is the *Candida* species with the highest prevalence, and this has been so for many years. However, the rise in prevalence of other *Candida* species has driven a decrease in the relative prevalence of *C. albicans* (as compared to these other species) (*46*). Antifungal therapy usually includes echinocandins followed by azoles (if necessary), and resistance towards them is still uncommon (*47*, *106*, *107*). However, azole resistance is rising in many low

and middle income regions, which is concerning due to common unavailability of echinocandins in the hospitals of such countries (as mentioned in the previous section) (*83*). The accuracy of diagnostic tools is high, except in some disease forms with low positivity rates in blood cultures (i.e. abdominal candidiasis) (*47*, *108*). The most clinically-relevant knowledge gaps for this species include unknown incidence of invasive forms across the last 5 years, and the lack of studies about complications and sequelae. Accordingly, amongst other important fungal pathogens, the WHO ranked *C. albicans* as priority rank 2/19 of public health importance and rank 13/19 for Research & Development (R&D) (*47*).

Candida tropicalis (phylogenetically close to *C. albicans*) is another common human commensal, which can produce opportunistic invasive infections of blood, central nervous system, bones, heart, bones and internal organs. Such invasive infections affect immunocompromised patients, often including neonates (*109*, *110*). *C. tropicalis'* infections result in long hospital stays and have mortality rates of 55-60% in adults and 26-40% in pediatric patients (*47*). Although the incidence is understudied, the trends over the last few years show an increase in this species (*46*, *110*). The diagnosis of this species is often inaccurate (i.e. some colorimetric methods confuse *C. glabrata* and *C. tropicalis*), and very variable across hospitals (*47*, *111*). Azole resistance is common (ranging from 20-80%, depending on the study), so echinocandins are the first line of therapy (*47*, *112*, *113*). From a clinical perspective, there is a need for further studies about global incidence, morbidity, risk factors and length of hospitalization in this species. In addition, evaluation of synergies between different antifungal compounds could be key to improve current treatments. These knowledge gaps explain why WHO ranked *C. tropicalis* as priority 10/19 of public health importance and rank 11/19 for R&D among the most important fungal pathogens (*47*).

The *Candida parapsilosis* species complex includes two emerging pathogens investigated in this project: *C. parapsilosis (sensu stricto)* and *C. orthopsilosis* (*102*). *C. parapsilosis* is a commensal yeast that can cause invasive infections (including in blood, bones, heart, central nervous system, eyes and internal organs) in immunocompromised patients. Common susceptible patients are those undergoing bone marrow transplants or receiving immunosuppressive therapies for cancer and/or organ transplants. In addition, as with *C. tropicalis*, neonates are increasingly at risk to suffer *C. parapsilosis* infections (*47*, *114*). The mortality of such invasive infections ranges between 20-45%, even with active antifungals available (*47*). As with *C. tropicalis*, the exact annual incidence rates are unknown, but the prevalence seems rising, and *C. parapsilosis* has become the second cause of candidemia (after *C. albicans*) in various regions, surpassing *C. glabrata* (*46*, *47*, *115*). Azole resistance is still moderate (~10% in some regions), and resistance to echinocandins and polyenes is still very rare (*47*, *113*). However, *C. parapsilosis* has an intrinsic reduced susceptibility towards echinocandins as compared to other *Candida* species. This means that, although these drugs are technically effective for most *C. parapsilosis* isolates (they have some susceptibility), echinocandins are less efficient in clearing these infections (*116*, *117*). Despite this, echinocandins are the

primary therapeutic choice for this species, often followed by azoles (*47*, *106*). More concerningly, this species has a high propensity to form biofilms (in prostheses and implants), which result in drug resistance (*118*, *119*). The most clinically-relevant knowledge gap for this species is the lack of systematic surveillance about prevalence, mortality and clinical complications. Accordingly, amongst other important fungal pathogens, the WHO ranked *C. parapsilosis* as priority rank 13/19 of public health importance and rank 6/19 for R&D (shared with *C. glabrata*) (*47*).

Candida orthopsilosis is another member of the *C. parapsilosis* species complex. The distinction between different species within this complex (including *C. parapsilosis sensu stricto*, *C. orthopsilosis* and *C. metapsilosis*) is fairly recent, and studies evaluating the specific impact of *C. orthopsilosis* in the clinics are scarce (*102*, *120*). For instance, in contrast to the taxa described above, the role of *C. orthopsilosis* as a commensal member of the healthy human microbiome is unclear and understudied (*102*). However, this species can cause invasive infections like candidemia, mostly affecting immunocompromised individuals (*121*, *122*). The prevalence of this pathogen is poorly understood, with studies suggesting that it accounts for 1-30% of all infections caused by members of the *C. parapsilosis* species complex (*121*, *122*). In addition, the yearly incidence of this pathogen is increasing, so that it is considered an emerging pathogen (*123*). Regarding antifungal resistance, various studies suggested that, as in *C. parapsilosis sensu stricto*, this species has intrinsic reduced susceptibility to echinocandins, which may result in therapeutic failure (*122*, *124*). From a clinical perspective, we consider that the most relevant knowledge gaps include the lack of studies on effective therapies, mortality and global incidence rates.

Still within the CTG clade, *Candida auris* is a recently emerged pathogen that colonizes the skin and can cause invasive infections of blood, eyes, central nervous system, eyes, heart, bones and internal organs. It generates nosocomial infections amongst the immunocompromised, including patients undergoing oncological therapies, organ transplant and mechanical ventilation. This pathogen generates clinical outbreaks which are difficult to manage, requiring specific strategies to prevent transmission (*47*, *125*). For instance, proper disinfection of surfaces containing this pathogen is not trivial, which enhances its spread in hospitals (*126*). Invasive *C. auris* infections generate hospital stays (in median) of 46-68 days in adults and 70-140 in pediatric patients, which is longer than for other *Candida* species. In addition, their mortality rate is between 29-53% (*47*). The global incidence is understudied, but there has been a steady rise in *C. auris* infections since its first identification in 2009 (*82*, *127*). In addition, various countries reported increased cases associated with the COVID-19 pandemic, likely due to the rise in immunocompromised patients using mechanical ventilation (*128*). From the diagnostic perspective, *C. auris* is commonly misdiagnosed due to the need of specific lab protocols, often not applied (*129*). Echinocandins are the first line of treatment, sometimes followed by other drug classes (mostly azoles) (*47*, *84*). Concerningly, resistance to azoles (mostly fluconazole) ranges from 87-100%, while the resistance levels to other drug classes are moderate

(8-35% for amphotericin B (polyene) and 0-8% for echinocandins) (*47*, *82*). In addition, *C. auris* has intrinsic lower echinocandin susceptibility, which may hamper treatment (*125*). Furthermore, in contrast to most other *Candida* species, resistance to several drug classes in the same isolate is frequent, and there are various pan-resistant isolates (resistant to all antifungal drug classes) (*47*, *130*, *131*). From a clinical perspective, there is an urgent need for studies about global incidence and possible preventive strategies. In addition, evaluation of synergies between different antifungal drugs could be key to improve available treatments. This explains why the WHO ranked *C. auris* as priority rank 4/19 of public health importance and rank 8/19 for R&D among the most important fungal pathogens (*47*).

Finally, *Candida glabrata* (belonging to the Nakaseomyces clade) is a global cause of candidiasis. From the ecological perspective, it is likely an environmental yeast which becomes an opportunistic pathogen in immunocompromised individuals (*132*). *C. glabrata* can cause invasive infections involving blood, central nervous system, heart, bones, eyes and internal organs. These severe infections generate hospital stays of 2-8 weeks, with a mortality of 20-50% (*47*, *133*). The epidemiological data over the last decade suggests that the prevalence of *C. glabrata* infections is increasing, and that this species is the second most-frequent cause of candidiasis worldwide (following *C. albicans*) (*113*, *134*). Specifically, a recent survey reported that 46.7% *Candida* isolates are *C. albicans*, while 18.7% are *C. glabrata* (*113*). This species has intrinsically lower susceptibility to azoles and develops frequent azole resistance, so that the common treatment is based on echinocandins. In addition, in recent years echinocandin resistance has been rising, generating multi-drug resistant strains that are difficult to manage (*106*, *135*, *136*). The most clinically-relevant knowledge gaps for this species are the lack of preventive strategies, absence of surveillance in developing countries, understudied risk factors and insufficient data about mortality and complications. Accordingly, amongst other important fungal pathogens, the WHO ranked *C. glabrata* as priority rank 6/19 of public health importance and rank 6/19 for R&D (shared with *C. parapsilosis*) (*47*).

All in all, various *Candida* species are a growing threat for human health, which require more funding to develop new drugs and/or optimize currently available treatments.

1.2. Comparative genomics: a tool to understand mechanisms of recent adaptation in *Candida*

To optimize treatments and diagnostics for *Candida* infections, a promising approach is to improve our understanding about the evolutionary mechanisms underlying recent adaptation to the host and to antifungal drugs (*92*, *93*, *137*). This involves the study of genetic changes (within each species) that underlie variation in clinically-relevant phenotypes such as antifungal drug susceptibility, host cell adhesion, invasion processes, biofilm formation, and host immune escape features (*118*, *138–141*).

In addition to drug susceptibility, these other phenotypes are essential for host adaptation and virulence. Adhesion to host cells and tissues (i.e. macrophages, endothelial or epithelial cells) is a key process for initial colonization, persistence and further establishment of the infection. *Candida* pathogens achieve this through specialized cell wall adhesion proteins, such as epithelial adhesins (*EPA*) in *C. glabrata* or agglutinin-like sequence (*ALS*) proteins in *C. albicans* (*142*, *143*). Following colonization and attachment, multiple invasion mechanisms allow *Candida* pathogens to penetrate host tissues and spread the infection. For instance, the most common mechanism in *C. albicans* is the morphological switch from a yeast commensal form to a hyphal (filamentous) invasive form. Such hyphae express secreted hydrolytic enzymes and toxic peptides (such as candidalysin), which degrade tissue barriers, promote nutrient acquisition and enable spread into the bloodstream (*143–146*). Conversely, *C. glabrata* cannot form hyphae (only pseudohyphae, a result of incomplete budding), and its invasion mechanisms rely on a stealthier (rather understudied) approach: induction of endocytosis by host cells (*146*, *147*). In addition, *Candida* pathogens form biofilms, which are structured communities of yeasts or hyphae (depending on the species) and the extracellular matrix (ECM), held together through cell-cell and cell-ECM adhesion. Biofilm formation is a key process because such communities have increased adhesion to host tissues, persistence on implanted medical devices, and antifungal resistance (*118*, *119*, *148*). Finally, the capacity to escape the host immune response, either by avoidance of recognition or by surviving immune attacks, is an essential virulence factor of *Candida* pathogens. For instance, *C. albicans* can escape the innate immune response by remodeling the cell wall and masking superficial β-glucan, a Pathogen-Associated Molecular Pattern that triggers immune defense (*147*, *149*). Another example is the persistence of *C. glabrata* within macrophages (through induced endocytosis), which has been proposed as a mechanism to reduce immune inflammatory responses and improve survivability of the fungus (*147*, *150*).

A good understanding about the evolutionary processes shaping these phenotypes may clarify the molecular mechanisms underlying virulence and drug resistance, which remain elusive (particularly in non-*albicans Candida* species). This could be useful to improve current diagnostics and therapies. For

example, novel drugs targeting such virulence mechanisms (i.e. biofilm or hyphae formation) may yield lower human toxicity and less frequent drug resistance than current antifungals (which target cell growth) due to several reasons (*151*). First, since fungi and humans are relatively evolutionary close (*152*), genes involved in housekeeping functions (i.e. cell growth) are more likely to be shared than fungal-specific virulence genes. This likely explains high human toxicity of polyenes, azoles and flucytosine (*151*). Second, antifungals that target cell growth impose higher selective pressures that result in drug resistance, as compared to drugs that target virulence genes that do not compromise fungal viability (*151*). In summary, understanding the virulence mechanisms of *Candida* species can be key to define therapeutic targets with improved clinical outcomes. Similarly, finding drug resistance mutations can aid the development of faster, nucleic acid-based tools to diagnose susceptibility profiles and choose optimal therapies (*68*). Furthermore, targeting such resistance mechanisms may be a relevant therapeutic strategy, as shown by the efficacy of combining azoles and drug efflux inhibitors in *Candida* species (*153*, *154*).

This topic is of particular importance in *Candida* pathogens because they have highly dynamic genomes within species (*95*, *155–158*) and even within clonal populations in a host (*104*, *159*). On the one hand, this diversity is generated by vertical evolution since the common ancestor of all the strains within a species. *Candida* species reproduce mostly asexually (clonally) (*160*), accumulating *de novo* genetic variants such as small variants (Single Nucleotide Polymorphisms (SNPs) and small insertions or deletions (INDELs)), complex structural variants (duplications, deletions, insertions and other rearrangements) and chromosomal aneuploidies (*156*, *161–164*).

On the other hand, reticulate evolution (by sexual reproduction and hybridization) has played an important role in the intraspecific diversification of *Candida*. For instance, although clonal reproduction is considered the primary propagation mode, recent studies suggested that *Candida* species can sexually mate, in contrast to the classical assumption of exclusive asexual reproduction (*132*, *160*). Mecanistichally, this is possible due to convoluted cycles, such as the parasexual cycle in *C. albicans* and *C. tropicalis*, which generates recombinant offspring from the mating of parental cells with complementary mating types (*165*, *166*). Accordingly, genomic surveys have found evidence for meiotic recombination between genetically-divergent strains in *C. glabrata* (*156*), *C. auris* (*167*, *168*), *C. albicans* (*158*), *C. orthopsilosis* (*102*), *C. parapsilosis* (*169*) and *C. tropicalis* (*170*). These findings suggest that (para)sexual reproduction has influenced the intraspecific diversification of *Candida* species. On another line, genomic studies have suggested that all or some isolates of *C. albicans*, *C. orthopsilosis* and *C. tropicalis* have a hybrid origin (*95*, *102*, *157*, *171*). Such isolates belong to hybrid lineages that resulted from the mating of two parental divergent lineages, which increased the genomic dynamism of post-hybridization diversification. For example, the ancestral highly-heterozygous hybrid likely underwent sequential loss of heterozygosity (LOH) events in particular

sections of the genome, resulting in gene conversion or loss of certain regions (*172*). All these evolutionary processes explain why *Candida* species have highly variable genomes.

Such genomic changes have been linked to rapid adaptation in changing environments, likely underlying the emergence of antifungal resistance (*136*, *173*, *174*), increased host cell adhesion (*175*), host immune escape (*175*), loss of virulence (*138*), changes in biofilm formation (*176*) and the emergence of new pathogenic species (*172*). Thus, the comparative analysis of full genomes has provided promising novel insights into these evolutionary mechanisms. Compared with traditional molecular methods, current high throughput genome sequencing approaches provide a more comprehensive picture of genetic changes and do not require prior knowledge on potentially relevant loci. This has revolutionized the way in which *Candida* pathogens can be studied.

In this chapter, we review major genomic approaches that have been instrumental in studying recent (within-species) adaptation to drugs and to the host in *Candida* pathogens. Given the (rather wide) scope of this work, we focus only on studies about the techniques and species related to the project. For instance, we use a narrow definition of genomics, restricted to the study of genome sequences, and not the broader meaning including transcriptomics or epigenomics. Similarly, despite the importance of genomic changes across species (more ancestral) (*49*, *177*), we focus on studies of (more recent) intraspecific variation as these are related to this PhD project. In addition, we do not review evolution-agnostic screening approaches, such as transposon-mediated random mutagenesis, although these have been useful to pinpoint mechanisms of drug resistance and host adaptation in *Candida* (*178*). Finally, we do not systematically review all studies of recent adaptation, but rather provide an overview, with illustrative examples that contextualize the objectives and results presented in this PhD thesis. The first four following sections (1.2.1 - 1.2.4) are focused on particular genomic approaches (see **Figure 2** and **Figure 3**), and the last section (1.2.5) lists the specific literature gaps addressed in this project.

1.2.1. Comparing to learn: population genomics

Population genomics refers to the comparative study of natural genomic variation within and across populations of a given species. Once a reference genome of a species is available, population genomics can be performed by sequencing additional isolates to identify genetic variants. Such variants can inform about the genetic structure of a species, identify sub-clades, reconstruct the history of populations, and identify genomic regions under selection and/or underlying specific phenotypes. Population genomic techniques have been used to correlate genomic and phenotypic variation across isolates of *Candida* pathogen species (**Figure 2**) (*95*, *155–159*). Most population genomics studies reconstruct a phylogeny showing the relationships between the isolates, which helps to understand the evolutionary process underlying the emergence of a given phenotype of interest. The populations of most *Candida* pathogens include clearly separated clades, and some of these clades have particular drug resistance or virulence properties (*156*, *158*). For example, some clades in *C. auris* and *C. albicans* lost the ancestral azole drug resistance (*155*) and virulence capabilities (*158*), respectively. Similarly, different clades in *C. tropicalis* have variable fluconazole and flucytosine resistance (*179*, *180*). Conversely, there is not a clear association between clade identity and antifungal resistance patterns in *C. glabrata* (*156*, *181*). These examples illustrate why studying the relationship between phylogenetic clades and phenotypes provides insights into evolutionary mechanisms of adaptation.

Beyond describing when and where phenotypes appeared, population genomics techniques have been used to infer underlying evolutionary mechanisms. A common approach in *Candida* is to analyze variants correlated with a given phenotype, but without rigorously testing the statistical significance of the genotype-phenotype associations. Despite this lack of statistical analysis, such approaches have provided useful insights, and they may be the only option to analyze small strain collections (which is common in *Candida*) where statistical association testing would be underpowered (*181–183*). In addition, such studies are insightful because they generate hypotheses that can be experimentally validated. On the one hand, various studies analyzed variants in genes underlying known mechanisms of adaptation. For example, a recent study in *C. auris* found that *ERG11* (encoding the lanosterol 14α-demethylase, target of azoles) point mutations and copy-number variants are associated with azole resistance, while *FKS1* (encoding the 1,3-ß-D-glucan synthase, echinocandin target) mutations are related to echinocandin resistance (*155*). Another study in *C. glabrata* found that *FKS1* mutations are related to echinocandin resistance, while *PDR1* and *CDR1* variants are associated with azole resistance, likely because these mutations promote increased azole efflux (*181*). In addition, a recent work in *C. albicans* analyzed known drug resistance genes in a collection of isolates to define variants conferring resistance towards azoles (in the *ERG11*, *TAC1*, *MRR1* and *UPC2* genes) and echinocandins (in the *FKS1* gene) (*184*).

On the other hand, various studies compared the genomes of genetically-close strains with phenotypic differences. The genes affected by variants among such strains were useful to draw hypotheses about evolutionary mechanisms underlying the phenotype. For instance, a population genomics study proposed that loss-of-function mutations in the *SFL1* and *ZCF294* transition factors (regulators of hyphal formation, necessary for systemic infection) are correlated to loss of virulence in clade 13 of *C. albicans* (*158*). Another example is provided by an analysis of clonal serial isolates in *C. glabrata*, which predicted that non-synonymous mutations in *SIR4* (involved in telomeric silencing and regulation of biofilm formation and cell adhesion) could yield reduced cell adhesion and thus more efficient systemic dissemination (*182*, *185*). Conversely, a study of natural variation in *C. tropicalis* suggested that strains with a frameshifting mutation in the *BAT22* gene (encoding an amino-acid aminotransferase) cannot grow on valine or isoleucine as nitrogen sources, resulting in reduced virulence (*157*, *183*). These examples illustrate how population genomics can be used to find both novel and previously-expected mechanisms of resistance and host adaptation.

However, such approaches are limited because they i) mostly investigate a few genes, ii) are biased towards evolutionary mechanisms expected *a priori* and iii) mostly lack rigorous statistical association testing. To address this and perform unbiased exploratory analyses, Genome-Wide Association Studies (GWAS) are a promising, more comprehensive, alternative. These consist in systematically screening naturally-occurring variants to find those that are statistically associated with the phenotypes of interest (called GWAS hits). The genes and pathways affected by such variants inform about underlying evolutionary mechanisms, allowing not only the confirmation of previous hypotheses but also the generation of new ones (*186–188*).

Methodologically, the challenge in GWAS is that the detection of causal variants (those that drive the phenotype) is confounded mainly by two factors: multiple testing and population stratification. Multiple testing is problematic because there are typically thousands of tested variants, resulting in false positives by chance alone (*189*). To tackle this, a common practice is to correct the association p values, using methods like bonferroni, False Discovery Rate (FDR) or permutation-based corrections like the maxT method (*189–192*). In addition, population stratification refers to the presence of genetically-distinct subpopulations, leading to spurious GWAS associations driven by this stratification rather than true causal genotype-phenotype relationships (*189*). This is a common issue in most organisms, but particularly in those that predominantly reproduce in an asexual way (*193*, *194*). To address the effects of population stratification, various GWAS methods have been used in microorganisms (bacteria and fungi), falling into two categories depending on the source of the association signal: either allele counts or phylogenetic convergence (*186*, *193*, *194*). Each of these methods has their own strengths and pitfalls to detect true causal variants, reviewed in the paragraphs below.

Allele counting methods (invented for human GWAS) identify variants (alleles) significantly over-represented in strains with a given phenotype (i.e. drug resistance) relative to control strains (i.e. isolates susceptible to the studied drug), while controlling the effect of population stratification (*193*). Common approaches rely on generalized linear regression, where stratification is controlled by either i) principal components as fixed covariates (*194*, *195*) and/or ii) a genetic relatedness matrix as a random effect in a linear mixed model (LMM) (*103*, *194*, *196*).

Despite their power, such methods are only useful to detect causal variants if a set of requirements are fulfilled. First, the studied collection should have a balanced set of strains with and without the phenotype of interest, ensuring that population stratification is not a main driver of such phenotypic diversity (*189*, *197*). Second, the size of the collection should be large enough, typically with >100 isolates for traits with simple genetic architectures where a few variants drive the phenotype (*187*, *197*) (i.e. drug resistance in *Candida* (*93*, *198*)). Third, sexual recombination should be frequent in the analyzed population. This ensures that, once a causal variant appears in the population, it segregates randomly with various other (passenger) variants through recombination, generating diverse phenotypes. In the population, this results in small blocks of variants under linkage disequilibrium (LD), so that genomic regions with GWAS hits are narrow enough to pinpoint relevant evolutionary mechanisms (*197*, *199*). For instance, in populations with exclusive clonal (asexual) reproduction, once a variant appears in the population it is in LD with all other variants, difficulting the distinction between causal and passenger mutations (*194*, *197*). All in all, allele counting methods are powerful tools to do GWAS in collections with balanced phenotypes, large sample sizes and sexual reproduction.

Conversely, convergence-based (or homoplasy-counting) methods (invented for bacterial GWAS) rely on finding variants that appear convergently with the phenotype multiple independent times along a phylogenetic tree of strains (*186*, *193*). For instance, the PhyC method uses ancestral state reconstruction (ASR) to find variants that emerged on tree nodes with the phenotype more often than expected by chance, relative to nodes with no phenotype (*192*, *200*). Similarly, the synchronous method from hogwash (equivalent to the simultaneous score approach from treeWAS (*201*)) uses ASR to find variants whose transition (appearance / loss) is correlated with the transition in the phenotype (*192*). Convergence-based methods intrinsically control the effects of population stratification because they take into account the phylogenetic relationships between strains (*186*). Even in clonal populations with strong stratification partially underlying the phenotype, these methods can detect causal variants as long as some degree of convergence exists (*192*, *193*).

As with allele counting methods, convergence-based approaches are more suitable for specific types of populations. First, they only work well in clonally-reproducing species, with recombination involving only

small genomic regions (i.e. through plasmid exchange and integration in bacteria). This is necessary to ensure that a phylogenetic tree (essential for such analyses) is a good representation of the evolutionary history of the population (*192*, *193*). Second, in small collections convergence-based approaches outperform allele counting methods because the number of convergence events required for significance is lower (as compared to allele counts) (*192*, *193*). In fact, their use is actually limited to small sample sizes due to large computational requirements (*202*). All in all, convergence-based methods are suitable for collections with asexual reproduction and small sample sizes.

A significant limitation of all these GWAS methods is that they are less suitable for phenotypes with an expectation of allelic heterogeneity, which happens when multiple (causal) variants in a genomic region (i.e. a gene) can underlie the phenotype. If this is the case, the association signal for individual variants is low, particularly for recently-appeared (rare) variants, reducing the power of GWAS methods testing single variants (*188*, *203*). For instance, in human diseases driven by haploinsufficiency, any Loss of Function (LoF) mutation (i.e. frameshift, stop codon or alteration of splice sites) in the causal gene can yield the disease phenotype (*188*, *204*). Similarly, and relevant to this PhD thesis, drug resistance towards azoles or echinocandins in *Candida* pathogens may be caused by different variants of the same gene (i.e. *PDR1* for azoles or *FKS1* for echinocandins) (*173*, *181*). An established way to address this is to first group variants with equivalent functional effects (i.e. altering a gene or pathway), and then test whether there is an association between any variant in a given group and the phenotype. This results in one test per group of variants, and it has been shown to improve the detection of relevant associations in both allele counting (*188*, *203*, *205*) and convergence-based methods (*192*, *206*). All in all, while allelic heterogeneity is an issue for GWAS, variant grouping techniques can be a solution.

Recent efforts have used allele counting GWAS techniques to understand recent adaptation in *Candida*. For instance, LMM-based GWAS was used to find variants associated with azole resistance in *C. glabrata*, which revealed two regulatory SNPs in *CST6*. This gene encodes a transcription factor that regulates the *EPA6* adhesin, and the regulatory variants may modulate drug resistance due to the changes in adhesion and biofilm formation (*207*). A similar study found various non-synonymous SNPs in adhesin genes associated with azole, echinocandin and flucytosine resistance in *C. glabrata*, supporting the role of adhesion and biofilm formation in the development of resistance (*103*). In addition, a recent work in *C. orthopsilosis* used cosine similarity metrics to pinpoint 19 gene-disrupting SNPs associated with caffeine and azole susceptibility (*208*). These examples illustrate the potential of GWAS to understand recent adaptation in *Candida*.

However, the usage of allele counting methods in these studies was likely underpowered due to i) small collection sizes (<50 strains), ii) the fact that *Candida* species have predominant asexual reproduction and

iii) the lack of grouped association testing. These limitations likely explain why the cited GWAS studies did not find significant variants in the known drivers of drug resistance (i.e. *FKS1/2*, *ERG11* and/or *PDR1* genes, mentioned above) (*103*, *207*). Convergence-based methods are a promising alternative as they are actually suited for asexual populations with small sample sizes. Although studies using such methods in *Candida* are missing (to the best of our knowledge), convergence-based GWAS has been used to study recent adaptation in other fungal pathogens. For instance, treeWAS (*201*) was used to pinpoint drivers of azole resistance in *Aspergillus fumigatus* (*209*). In addition, a study in *C. albicans* used CAPRIB (*210*), a tool detecting convergence between amino acid substitutions and changes in a given phenotype, to find variants associated with responsiveness to farnesol, which regulates yeast-to-hyphae switching (*211*). This last study suggests that convergence GWAS approaches could be suitable for *Candida* species. Furthermore, given the allelic heterogeneity of clinically-relevant phenotypes in *Candida* (*173*, *181*), methods considering variant grouping could be essential to fully understand mechanisms of adaptation. In summary, GWAS approaches, particularly those based on convergence that also consider grouping of variants, are a promising, yet underused, tool to study recent adaptation in *Candida* pathogens.

On another line, the detection of genomic signatures of positive selection has been instrumental to understand recent adaptation in *Candida* pathogens from genome sequences alone. Such traces of selection may be particularly useful to study non-measurable phenotypes, such as naturally-occurring drug resistance, immune escape and/or pathogen transmissibility (i.e. within patients) (*212*–*214*). To find these signatures, there are two commonly-used approaches, based on either selective sweep detection or the ratio between non-synonymous and synonymous substitution rates (*212*). Below we provide an overview of them, with a focus on their usage in *Candida*.

On the one hand, sweep-detection methods rely on the fact that beneficial alleles (under positive selection) rapidly spread in the population. In sexually-reproducing species, this leads to a drop in nucleotide diversity around the selected allele due to meiotic recombination, a phenomenon known as ‘selective sweep’. This loss in diversity occurs within the extent of LD in the species, as neutral (passenger) alleles are carried along with the advantageous allele due to genomic proximity. By scanning these local fluctuations in diversity, sweep-detection methods pinpoint regions containing alleles under positive selection (*212*, *215*). These approaches have been used on (sexual) fungal plant pathogens (*212*, *216*), but they have limited power to detect selection in clonally-propagating species like bacteria or *Candida* pathogens. In such species, once an adaptive allele appears it is linked to all other genomic variants, which does not result in local drops in diversity that would be detected by sweep-detection methods (*217*, *218*). This likely explains the absence of studies to detect positive selection through sweep detection in *Candida* species.

On the other hand, various studies inferred the signatures of positive selection in protein-coding genes from the ratio between the rate of non-synonymous and synonymous substitutions. The underlying assumption is that coding sequences under positive selection accumulate an excess of (potentially adaptive) non-synonymous variants, as compared to the accumulated (near neutral) synonymous variants. For instance, dN/dS ratios are popular to identify selection due to their straightforward application and intuitive interpretability. For any coding sequence in a given period of time, dN refers to the number of accumulated non-synonymous substitutions per non-synonymous site, while dS is the number of synonymous substitutions per synonymous site. A dN/dS ratio of 1 indicates equivalent rates of non-synonymous and synonymous substitutions, suggesting neutral evolution of the gene. Conversely, dN/dS > 1 indicate that non-synonymous substitutions are occurring at a higher rate than synonymous substitutions, suggesting positive selection (*219–221*). Although the capacity of dN/dS to measure selection within a population could be limited (*212*, *222*), it has been broadly used to understand within-species adaptation in some organisms (*223*, *224*), including *Candida* pathogens. For instance, a recent study of *C. glabrata* clinical isolates found signs of positive selection in adhesins, ribosomal proteins and mitochondrial structural genes (*225*). Similarly, dN/dS metrics were used in *C. albicans* to validate the (required) house-keeping nature of genes used for multilocus sequence typing (a technique for clade identification based on a handful of conserved genes) (*226*).

A conceptually similar measure is the ratio between non-synonymous and synonymous nucleotide diversity (π_N/π_S), a suitable measurement for population genomic data. For any coding sequence in a given strain containing SNPs relative to the reference genome, π_N is the number of non-synonymous SNPs per non-synonymous site, and π_S is the equivalent measure for synonymous SNPs. As with dN/dS, a high π_N/π_S is a typical hallmark of positive selection (*227–231*). A few studies used such π_N/π_S metrics to understand recent adaptation in *Candida* pathogens. For instance, a study in *C. glabrata* found traces of positive selection (high π_N/π_S) in *ESC1* (regulator of sub-telomeric silencing), which may have impacted mating-type switching and the dynamics of sub-telomeric genes (*156*). In summary, these examples illustrate the potential of dN/dS and π_N/π_S metrics to measure adaptation in *Candida* pathogens.

All in all, population genomic techniques, including GWAS and detection of selection, are powerful tools to understand how naturally-occurring genomic variation underlies recent adaptation in *Candida* pathogens.

1.2.2. Seeing it live: directed evolution

While population genomics can be powerful, it has two main limitations for understanding the mechanisms of drug resistance and host adaptation. First, the study of natural variation is mostly useful for phenotypes that emerged independently several times in a population. Thus, population genomics may be underpowered to study adaptation when available data on genotypes and phenotypes is limited. For example, such studies may be unsuited to study drug resistance of a recently developed compound (i.e. beauvericin (*232*)), where there may be insufficient natural phenotypic variation. In addition, the virulence mechanisms of a new emergent fungal pathogen (i.e. *C. orthopsilosis* (*95*)) may be difficult to study with population genomic techniques due to the limited number of strains. Finally, such approaches may be sub-optimal to study genomic determinants of a phenotype that emerged only once in the population (i.e. the virulence loss at the common ancestor of the isolates in clade 13 from *C. albicans* (*158*)). Second, large divergence between isolates with different phenotypes complicates the (key) distinction between causal and passenger mutations (*187*), a relevant issue given the highly diverse and stratified nature of *Candida* populations (*155*, *156*, *158*).

Directed (artificial) evolution (either *in vitro* (*104*, *233*) or *in vivo* (*138*)) of such clinically-relevant phenotypes followed by sequencing (usually whole-genome sequencing) of the adapted strains offer a promising solution to overcome some of these problems. In directed evolution experiments the conditions are controlled, and the phenotypes under study are 'forced' to appear in otherwise wild type strains by using selective regimes. This means that phenotype-causing mutations are expected to be fixed in the selected populations and, if the process is repeated, they are expected to appear recurrently. This usually simplifies the detection of such mutations as compared with population genomics studies (**Figure 2**). For instance, such approaches have been used to understand the *in vitro* evolution of azole resistance in a few strains of *C. glabrata*, which revealed the importance of mutations in *PDR1*, *CgHxt4/6/7* hexose transporters and/or upregulation of adhesins (*92*, *234*). In addition, similar experiments were performed in *C. auris* (*93*, *139*), *C. parapsilosis* (*235*) and *C. albicans* (*236*). The mechanisms of host adaptation have also been partially studied through directed *in vivo* evolution experiments. For instance, a study of *C. albicans* evolved avirulent strains (starting from a virulent parental) in murine models and found that changes in *EFG1* and *FLO8*, related to hyphal growth, were related to the loss of virulence (*138*). This exemplifies how evolution in the host can yield strains with lower virulence (*237*). Another study found that passing *C. albicans* through murine models (in vivo) results in highly diverse populations, which revealed genomic changes underlying host adaptation (*238*). In summary, using directed evolution coupled with genome sequencing has allowed the exploration of drug resistance and host adaptation in controlled settings.

However, such methods may not entirely recapitulate the natural evolutionary process. For example, natural evolution of drug resistance and host adaptation involves synergistic action of multiple selective forces that may be absent from some artificial settings settings. In addition, directed evolution experiments rely on detecting events of parallel evolution (i.e. a gene with new mutations in multiple independently-evolved lineages. This is a limitation because parallel evolution can be either i) a relevant sign of positive selection (underlying the studied adaptation) or ii) a confounding effect of heterogeneous mutation rates across genes (*239*). We thus consider that complementary approaches (such as population genomics) are also key to obtain the complete picture. The following sections describe some of the genetic alterations underlying these phenotypes.

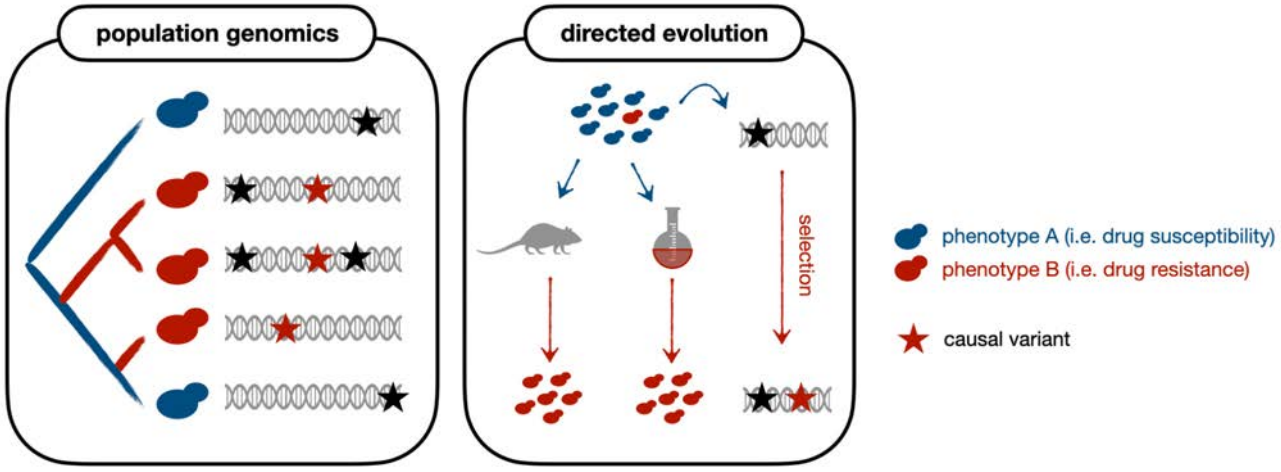


Figure 2. Comparative genomic techniques that have been used to understand recent adaptation in *Candida*. Population genomics (left) is the comparative study of genomic variation within and across populations of a given species. This technique has been used to correlate genomic and phenotypic variation across strains of *Candida* species. The example shows a gene with several variants (in red) that underlie the emergence of the phenotype of interest (i.e. drug resistance) in some strains. Note that, since there is some divergence between strains, it is not trivial to distinguish causal (red) from passenger (black) variants. Directed evolution (right) experiments consist in using selective regimes (either *in vivo* (left) or *in vitro* (right)) to 'force' the appearance of the phenotypes under study. The selected strains are sequenced to identify variants underlying the phenotype. This approach simplifies the detection of causal variants as compared to population genomics studies because the evolutionary conditions are more controlled. This example represents a gene that acquired a single causal variant (in red) driving the phenotype of interest during artificial evolution.

1.2.3. Small variants bringing large change

Most research linking genomic variants to drug resistance or host adaptation in *Candida* is focused on finding small variants (SNPs and INDELs) between isolates of a given pathogenic species. These strains may be different isolates from a population (*155–157*), pairs of parent-daughter lineages from a directed evolution experiment (*93*, *138*) or serial isolates from a given patient (*159*, *182*). To identify such small variants, most studies use custom pipelines that include mapping of sequencing reads to a reference genome and a variant calling step, which is performed with algorithms like HaplotypeCaller from the Genome Analysis Toolkit (GATK) (*240*) or freebayes (*241*). This is usually followed by variant annotation (using tools like the Ensembl Variant Effect Predictor (VEP) (*242*) or SNPeff (*243*)) to prioritize candidate mutations. Some automatic pipelines that call and interpret small variants directly from the reads have been developed to simplify this task, such as the YMAP online tool (*244*) that has been specifically developed for fungi. The following paragraphs describe examples of how small variants underlie recent adaptation (see also **Figure 3**). Since this PhD thesis has a focus on antifungal drug adaptation (see Results sections 3.3 and 3.4), we provide an overview about established resistance mechanisms towards azoles, echinocandins and polyenes.

Azole resistance in *Candida* species involves a plethora of evolutionary mechanisms related to small variants. First, point mutations in the drug target enzyme (lanosterol 14α-demethylase, encoded by *ERG11*) decrease the drug binding affinity, generating resistance. This has been shown in *C. albicans*, *C. auris*, *C. tropicalis* and *C. parapsilosis* (*155*, *245–247*). Second, small Gain-of-Function (GoF) variants in the *UPC2* (*UPC2a* in *C. glabrata*) transcription factor (TF) drive overexpression of *ERG11*, leading to sufficient enzyme activity even in the presence of the drug. This resistance mechanism exists in *C. albicans*, *C. tropicalis*, *C. parapsilosis*, *C. glabrata* (*245*, *248*) and maybe *C. auris* (*249*). Third, LoF mutations in *ERG3* (encoding the C5 sterol desaturase) may change sterol profiles in a way that is adaptive in the presence of azoles, since these drive production of toxic intermediates through this same enzyme (*250*). This has been proposed in *C. albicans*, *C. parapsilosis* and *C. auris* (*139*, *235*, *245*). Fourth, GoF mutations in the *PDR1* (*C. glabrata*), *TAC1b* (*C. auris*) and *TAC1* (*C. albicans*, *C. tropicalis* and *C. parapsilosis*) TFs drive the overexpression of azole efflux pumps (*CDR1*, *CDR2* and *SNQ2*), generating resistance (*93*, *139*, *182*, *245*, *247*, *251*, *252*). Fifth, similar GoF mutations in the TFs *MRR1* (*C. albicans*, *C. parapsilosis* and *C. tropicalis*) and *MRR1a* (*C. auris*) promote the overexpression of *MDR1* azole efflux pump (*245*, *246*, *253*, *254*). Sixth, LOH affecting the *ERG11*, *TAC1* and *MRR1* genes may promote the acquisition of azole-resistance alleles, which has been proposed in *C. albicans* (*245*, *246*). In short, small variants generate azole resistance by altering either the drug target, the produced sterol intermediates and/or drug efflux.

Conversely, known echinocandin resistance mechanisms are less varied, but also predominantly involve small variants. The main mechanism involves amino acid substitutions in 'hot spot' regions of its target: the 1,3-ß-D-glucan synthase, essential for cell wall homeostasis. This enzyme is encoded by the genes *FKS1* (in *C. auris*, *C. albicans*, *C. tropicalis, C. parapsilosis* and *C. orthopsilosis*) or *FKS1* and *FKS2* (two redundant paralogs in *C. glabrata*). As with *ERG11* mutations, such amino acid changes likely impair the binding of the drug, generating resistance. This mechanism has been shown in *C. glabrata*, *C. albicans*, *C. auris*, *C. tropicalis*, *C. orthopsilosis* and *C. parapsilosis* (*155*, *245*, *247*, *255*). In addition, LoF variants in *ERG3* may underlie echinocandin resistance in *C. parapsilosis* (*235*, *256*) and *C. albicans* (*247*), suggesting a (still poorly understood) link between ergosterol biosynthesis and cell wall homeostasis. In short, small variants generate echinocandin resistance by altering either the drug target and/or ergosterol biosynthesis.

Finally, currently-described polyene resistance mechanisms, although less understood, also involve small variants. For instance, point mutations affecting enzymes of the ergosterol biosynthesis pathway (*ERG2, ERG3, ERG5, ERG6* and *ERG11*) may confer resistance because they result in reduced levels of ergosterol in the cell membrane. This has been proposed for *C. albicans*, *C. glabrata* and *C. auris* (*177*, *245*, *246*, *257*). In addition, mutations in the *FLO8* TF of *C. auris* may drive resistance due to changes in biofilm formation and adhesion, a common drug-resistance mechanism (*139*, *258*). Furthermore, given the common alteration of ergosterol biosynthesis, cross-resistance between polyenes and azoles (a phenomenon by which adaptation to a given compound generates resistance towards another drug) is common (*245*). All in all, polyene resistance, while still poorly understood, has been linked to small variants modifying ergosterol and biofilm function.

Beyond drug resistance, a few studies linked small variants to other clinical phenotypes variable across strains of a given species, underlying processes of recent host adaptation. For example, in *C. glabrata* GoF changes in *PDR1* can yield increased host adherence (through *EPA1* adhesin overexpression) (*251*) and reduced macrophage uptake (resulting in immune escape) (*175*). These mutations also caused azole resistance, which exemplifies how a single variant can drive both phenotypes. Another example in *C. glabrata* is the proposed effect of non-synonymous mutations in *SIR4*, which may yield reduced cell adhesion and more efficient systemic dissemination (*182*). Conversely, LoF small variants in various transcriptional regulators of yeast-to-hyphae transitions (*FLO8*, *SFL1* and *ZCF294*) have been linked to reduced virulence and increase commensal behavior in *C. albicans* (*138*, *158*). These examples show that reduced virulence through altered morphogenesis could be a common recent adaptive mechanism. All in all, small variants have significant, yet understudied, effects on host adaptation.

1.2.4. Do not neglect structural variants

Beyond small variants, complex Structural Variants (SVs) (i.e. chromosomal aneuploidies, duplications, deletions, inversions, translocations, insertions and other rearrangements) have been shown to modulate differences in drug resistance and other clinically-relevant phenotypes across strains of *Candida* pathogens. Most current studies analyzed the role of a subset of SVs, copy-number variants (CNVs), identified from changes in genomic read depth (or coverage). These can be classified into aneuploidies (CNVs spanning whole chromosomes) and small CNVs (affecting narrower genomic regions).

On the one hand, whole-chromosome aneuploidies (losses or gains) have been proposed to drive adaptation, likely due to changes in the copy number and/or expression of certain genes within aneuploid chromosomes (*245*, *259*, *260*). Such large changes may enable fast adaptation under certain stressors, but they are regarded as suboptimal due to the high fitness cost of altering so many genes. Thus, aneuploidies likely represent a transient initial step towards stress adaptation, being reverted when i) the stress disappears or ii) further adaptive point mutations, carrying a smaller fitness cost, are acquired (*261*). For instance, chromosomal duplications have been linked to azole resistance in *C. glabrata* (*262*, *263*), *C. auris* (*93*) and *C. albicans* (*264*), likely due to overexpression of genes encoding drug target enzymes (which lower the impact of the drug) and/or efflux pumps (reducing the intracellular drug concentration). In addition, aneuploidies may generate multidrug resistance for compounds with different mechanisms of action. For example, a study in *C. albicans* found that chromosome 2 trisomy promotes cross-resistance to hydroxyurea and caspofungin, while chromosome 5 monosomy generates resistance to azoles and echinocandins (*264*). Similarly, most aneuploidies in *C. albicans* yielded condition-specific fitness benefits (*265*), suggesting they are major drivers of adaptive evolution. This has also implications for the emergence of virulence and host adaptation. For example, it has been proposed that aneuploidies in *C. glabrata* drive increased levels of secreted aspartyl proteases and phospholipase B, necessary for virulence and survival within macrophages (*263*). In addition, chromosome 7 trisomy in *C. albicans* may underlie more efficient gastrointestinal colonization and systemic infection due to increased *NRG1* expression leading to reduced filamentation (*266*). These examples illustrate the role of aneuploidy for recent, rapid adaptation.

On the other hand, smaller CNVs (i.e. duplications generating overexpression of *ERG11*, *TAC1* and/or *CDR1* genes) have been linked to azole resistance in *C. glabrata* (*207*, *267*), *C. albicans* (*268*) and *C. auris* (*155*). Such CNVs have also been linked to changes in virulence and host adaptation, including CNVs in adhesins in *C. glabrata* (*207*) or CNVs in cell wall and stress-response genes in *C. auris*, potentially responsible for dealing with environmental fluctuations (*162*). In addition, amplifications of the *ALS4* adhesin genes have been linked to increased adhesion and biofilm formation in *C. auris* (*269*). Another example was found in *C. glabrata*, where increased adhesion and biofilm formation in three strains was associated with deletions of

AWP13, a GPI-anchored adhesin, perhaps resulting in a rewiring of cell-cell and/or cell-ECM adhesion (*156*). This example additionally illustrates how gene loss can drive important phenotypes in these pathogens. In summary, read-depth based CNV calling revealed that SVs play a fundamental role in *Candida* species.

However, current CNV-focused studies of SVs in *Candida* have several limitations. First, such CNV calling techniques have limited resolution (often ignoring very small CNVs) and lack precision in defining breakpoints positions (*93*, *270*). Second, read depth can be noisy and biased by factors like GC content, read mapping errors, mappability or the distance to the telomere (*244*, *271*, *272*). Note that distance to the telomere is relevant because, in some samples, loci that are closer to the telomere have higher coverage, creating a "smiley pattern" (*244*, *273*). Although these can be partly addressed (as in (*244*)), such biases reduce the accuracy of CNV calling. Third, CNVs are only a subset of all SVs, and the role of more complex SVs (like inversions, translocations or insertions) has been mostly overlooked. This is likely due to technically difficult SV detection from commonly-used short reads, which may be solved by either i) relying on long reads (*274*) or ii) implementing recent methods for accurate short read-based SV calling (*275–277*). This is a significant knowledge gap because there is some (limited) evidence that such variants exist and may underlie adaptation in the populations of *Candida* pathogens (*156*, *161*, *236*, *278*). These results indicate that further research should consider the role of complex SVs in drug resistance and host adaptations of *Candida* pathogens.

In summary, structural variants underlie common mechanisms of adaptation in *Candida* pathogens, although the specific contribution of complex rearrangements is poorly understood.

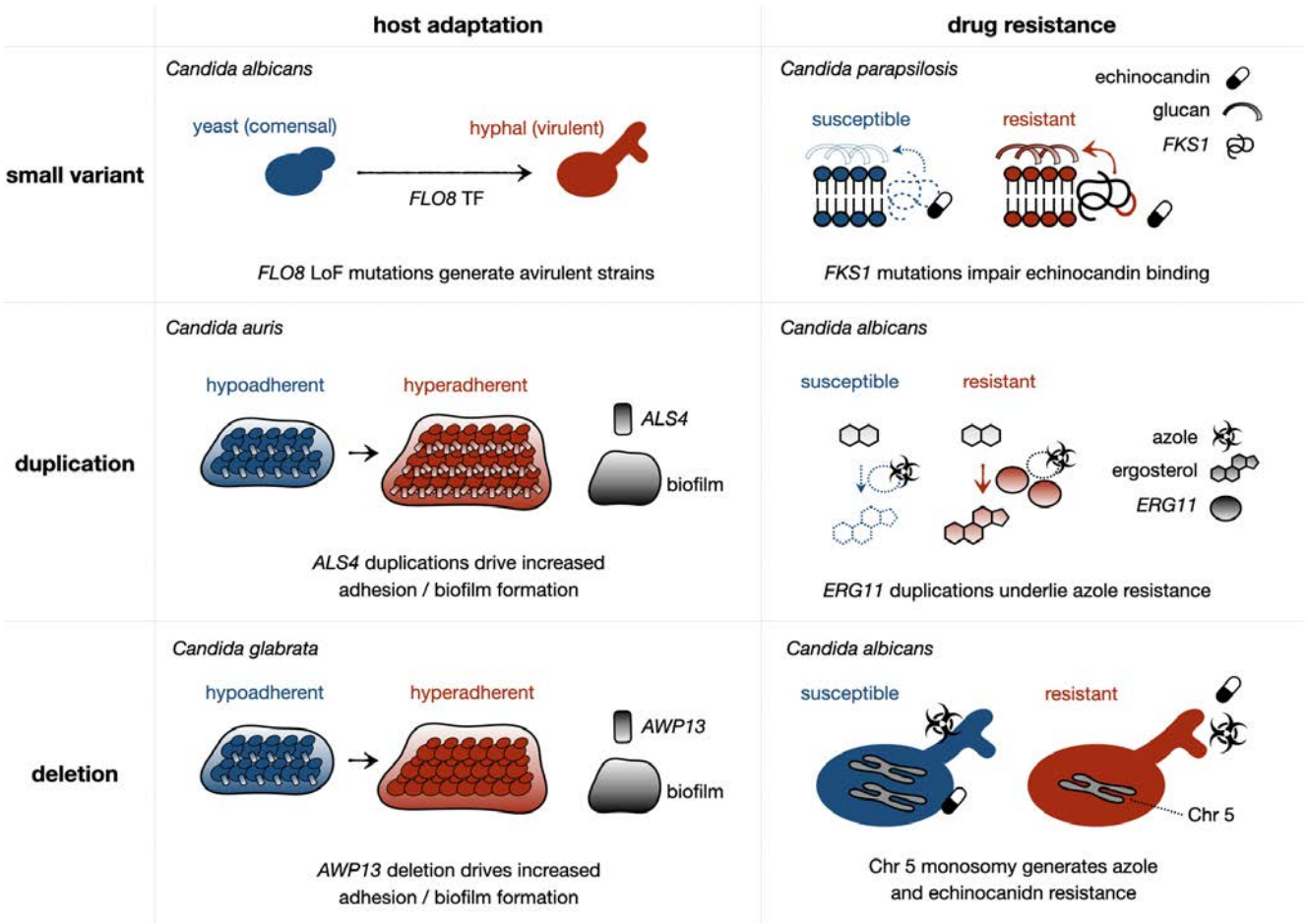


Figure 3. Several types of genomic variants can modulate host adaptation and drug resistance. We show illustrative examples of how small variants (top), duplications (middle) and deletions (bottom) can alter host adaptation (left) or drug resistance (right) across isolates of the same species. This figure is a graphical support to the sections 'Small variants bringing large change' (1.2.3) and 'Do not neglect structural variants' (1.2.4) of the Introduction.

1.2.5. Knowledge and methodological gaps addressed in this PhD thesis

Despite the evidence provided above, there are still many gaps in our understanding about the evolutionary processes underlying recent adaptation to antifungal drugs and to the host in *Candida*. In the paragraphs below we review the gaps that we addressed in this PhD project, related to both methodological pitfalls and missing knowledge about recent adaptation. First, we addressed methodological hurdles related to the bioinformatics analysis of genomic variants in *Candida* pathogens, including i) the general overlooking of SVs and ii) the lack of reproducible computational pipelines. Second, we improved our limited understanding about i) the mechanisms of *in vitro*-evolved drug resistance in *C. glabrata*, ii) the mechanisms of clinical resistance in *C. glabrata*, *C. auris* and *C. albicans* and iii) the genomic signatures of recent, clinically-relevant selection in major *Candida* species. Below we provide a detailed overview of these gaps, as well as their relationship to the objectives and results of this project.

As reviewed above, current studies in *Candida* mostly focused on the role of small variants (SNPs and INDELs) and/or coverage-based CNVs in recent adaptation (*139*, *156*, *157*). Given the limitations of coverage-based CNV calling and the general overlooking of more complex SVs (translocations, insertions or inversions), the contribution of SVs to recent adaptation remains largely unexplored. Most studies had this limitation because they used short-read sequencing, which is not suitable for accurate SV calling in the absence of benchmarking studies for the species of interest. Existing short-read-based SV callers have been mostly tested on humans and some model organisms, and it remains unclear how applicable they are in other species. More specifically, the choice of parameters to filter such variants, essential for accurate SV calling, remains a non-trivial task in the absence of previous benchmarking analyses. This is particularly relevant for SV calling because some of these algorithms can have low recall, making parameter choices a key step of the analysis. Such tools may be inaccurate because i) they rely on indirect evidence like discordant read pairs or split reads and ii) SVs often appear around repetitive elements that bias short read mapping (*277*, *279*).

Although long-read sequencing technologies are more suited for SV calling (*274*), an accurate method for short-read-based SV calling in any species of interest (i.e. *Candida*) could be necessary for various reasons. First, this would allow calling all variant types from a single genomic library, reducing prices and complexity of genomic experiments and analyses. This is rather complicated with long reads because these are less suited for small variant calling (*274*, *280–282*). Second, such an algorithm would allow the re-analysis of available short read-based genomes to study SVs. This is particularly relevant in *Candida* pathogens because there are thousands of short read-based genomes available (*155*, *158*), which could be used to clarify the role of SVs in recent adaptation. In summary, the role of SVs has been mostly overlooked in *Candida*

pathogens, and there is a need for improved methods to call and filter such complex variants. To address this we developed a method for accurate SV calling across eukaryotes (see section 3.2).

Another methodological hurdle is that, despite the availability of open and useful software (*240*, *244*, *283*), there are no easy-to-use established pipelines for variant calling, filtering and functional analysis in *Candida* pathogens. Different studies used unique approaches, generating problems to compare and reproduce the generated results (*95*, *155*, *158*). In addition, typical pipelines rely on the integration of multiple tools, each dedicated to very specific analysis steps. For instance, a typical (simplified) pipeline to study genetic variants in *Candida* may involve i) trimming and quality control of the raw reads, ii) read mapping, iii) calling of small variants, iv) coverage inference to find CNVs, v) variant filtering and vi) functional variant annotation. Each of these steps may in turn involve several specific software tools, so that variant calling pipelines require long development time and highly specialized knowledge. Such pipelines are usually not reproducible due to either complex software dependencies or unavailability of the underlying code (*95*, *155*, *158*). These technical issues are a significant barrier for genomics research in *Candida* pathogens.

To solve these gaps, a pertinent solution could be a computational pipeline to call, filter and annotate variants directly from short reads, with straightforward usability for non-experts. This pipeline should have easy and reproducible installation, ideally using containerization tools like Docker or Singularity (*284*, *285*) to address the issue of complex dependencies. In addition, such a tool should be flexible to deal with diverse species (each with unique genomic features like ploidy or repeat content) and variable experimental designs (i.e. with varying read depths). Some efforts have been made in this direction, such as the online YMAP pipeline (*244*) or MutantHuntWGS (*286*), but there is still a need for improved tools. All in all, current bioinformatic tools for variant analysis in *Candida* pathogens are suboptimal. To address this, we developed a flexible, reproducible pipeline to identify and interpret several variant types in the organism of interest (see section 3.2).

Regarding missing knowledge, there are open questions about the evolutionary mechanisms by which *C. glabrata* adapts (rather frequently) to azoles and echinocandins (*245*). Such knowledge gaps are particularly relevant in this species given its large evolutionary divergence to other *Candida* (see **Figure 1B**), likely having distinct resistance mechanisms (*49*, *104*). First, most of the evidence about resistance mechanisms comes from the exploration of genes previously expected to contribute to resistance (*181*, *247*), so that the genome-wide underpinnings of resistance remain largely unexplored. Second, while some studies explored such genome-wide mechanisms based on natural variation (*92*, *234*), they had limited power due to small sample sizes and lack of rigorous association testing. This is particularly relevant for the lesser-studied echinocandins. Third, the evolutionary mechanisms of adaptation to serial and/or combined exposure of these two drug classes remain unexplored. This is relevant to better understand the suitability of such serial

or combined administration of these drugs in the clinics (*47*, *106*). Fourth, the fitness tradeoffs associated with acquisition of resistance, relevant to understanding how drug adaptation could impair viability and pathogenesis (*287*), deserve further attention. Fifth, despite the importance of CNVs to this adaptation (*207*, *263*), the contribution of complex SVs remains largely unexplored. In summary, the mechanisms of resistance towards major antifungals remain obscure *C. glabrata*. To infer them, we used a large-scale *in vitro* evolution approach coupled with sequencing and phenotyping of adapted clones (see section 3.3).

More broadly, the mechanisms of resistance in *Candida* clinical isolates are understudied due to similar reasons, such as i) the exclusive focus on known driver genes, ii) underpowered analyses and iii) overlooking of SVs. Despite the suitability of artificial techniques like *in vitro* evolution (*92*, *139*, *288*), understanding how resistance evolves in clinical isolates is essential to fully understand the process in its natural niche. Given the availability of thousands of genomes with associated phenotypic data (*155*, *156*, *289*), GWAS approaches integrating data from multiple studies can be a way to address these gaps. Such a meta-analysis would be particularly relevant for species like *C. glabrata*, *C. auris* and *C. albicans*, with hundreds of sequenced clinical isolates (*289*). However, the few existing GWAS works in *Candida* used methods with limited power, based on allele counting and without variant collapsing (unsuited for such mostly-asexual organisms) (*103*, *207*, *208*). In addition, each of the current studies focused on a single species, with particular analytical methods, so that the similarities in resistance mechanisms across taxa remain elusive. Understanding such similarities could be key to enable species-specific treatments and diagnostics. To address this we gathered public datasets and performed a convergence-based GWAS about the resistance towards echinocandins, polyenes and azoles in *C. glabrata*, *C. auris* and *C. albicans* (see section 3.4).

Finally, the genomic signatures of positive selection, which may reveal mechanisms of recent adaptation, remain largely unexplored in *Candida* pathogens. Some studies addressed this using dN/dS and π_N/π_S metrics (*156*, *225*, *226*), but there are many knowledge gaps, some of them similar to open questions mentioned above. First, current studies mostly focused on specific groups of genes (i.e. to study mating type loci (*156*) or the housekeeping nature of MLST genes (*226*)), so that the genome-wide signatures of selection remain obscure. Second, most of our knowledge about selection comes only from a few species (i.e. *C. glabrata* and *C. albicans*). Third, given that dN/dS (or similar metrics) only refer to SNPs, the contribution of INDELs and SVs to selective processes remains less studied. Technically, this is challenging to address since there is no consensus theory defining neutral INDELs and SVs, a key requirement to find signs of selection (*219*, *221*, *290*). Despite this, analyzing the genes with an excess of such variants can hint to selective processes, as illustrated by the fact that adhesin genes have recurrent CNVs in *C. glabrata*, which could imply adaptive processes involving adhesion and biofilm formation (*156*, *161*, *207*). Fourth, as with clinical drug resistance, there is a lack of multi-species studies that may hint to the similarities in recent adaptation across different *Candida* pathogens. Fifth, given that these pathogens alternate between host

and environment (*114*, *132*, *291*), typical calculations of dN/dS and π_N/π_S (which average the effect of all variants) may reflect ancient selection unrelated to clinical adaptation. For instance, a recent study in *Mycobacterium tuberculosis* (bacteria) showed that signatures of selection in a given gene may change drastically over time (*214*). Thus, to fully understand recent selection it may be necessary to analyze only recently-appeared variants. To address these gaps we analyzed the genome-wide signs of recent selection in public genomes of *C. glabrata*, *C. auris*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis* (see section 3.4).

In summary, in this project we tried to address a wide range of methodological and knowledge gaps.

2. OBJECTIVES

2. OBJECTIVES

Our goal was to improve our understanding about evolutionary mechanisms driving recent, clinically-relevant adaptation in *Candida* pathogens (to the host and to antifungal drugs), from the perspective of comparative genomics.

To address this general objective we set the following specific objectives:

1. To develop a computational method that simplifies the detection and analysis of several variant types (with a particular focus on complex structural variants) in non-model organisms such as *Candida* pathogens.

2. To identify the mechanisms and tradeoffs of azole and echinocandin resistance in *C. glabrata*, by using *in vitro* evolution.

3. To understand the mechanisms of naturally-occuring clinical drug resistance towards azoles, echinocandins and polyenes in *C. glabrata*, *C. auris* and *C. albicans*.

4. To infer the genomic signatures of recent selection, potentially underlying adaptation, in clinical isolates of six major *Candida* pathogens: *C. glabrata*, *C. auris*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis*.

5. To understand the contribution of complex structural variants to recent adaptation, in relation to the specific objectives described in the points 2-4 above.

3. RESULTS

3.1

Report of the PhD thesis director

3.1. Report of the PhD thesis director

Dr. Juan Antonio Gabaldón Estevan, ICREA Professor affiliated at the Barcelona Supercomputing Center and the Institute for Research in Biomedicine (IRB Barcelona), as director of the PhD thesis of Miquel Àngel Schikora Tamarit, reports that, in the development of his PhD project "Comparative genomics of recent adaptation in *Candida* pathogens" the candidate Miquel Àngel Schikora Tamarit has actively participated in two research articles published in peer-reviewed journals, and one additional study currently under peer review:

1. PerSVade: personalized structural variant detection in any species of interest
Miquel Àngel Schikora-Tamarit and Toni Gabaldón
Published in Genome Biology
Impact factor: 18.01 (2021), Quartile: Q1 (2022)
2022 Aug 16;23(1):175. doi: 10.1186/s13059-022-02737-4.

This article constitutes the first chapter of the results of this PhD thesis, found in section 3.2. In this work, Miquel Àngel Schikora wrote the code, performed the bioinformatic analyses and prepared all the figures. In addition, he contributed to the conception of the study, the interpretation of results and writing of the manuscript.

2. Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen *Candida glabrata*
Ewa Ksiezopolska*, Miquel Àngel Schikora-Tamarit*, Reinhard Beyer, Juan Carlos Nunez-Rodriguez, Christoph Schüller and Toni Gabaldón
* equal contribution
Published in Current Biology
Impact factor: 10.9 (2022), Quartile: Q1 (2022)
2021 Dec 6;31(23):5314-5326.e10. doi: 10.1016/j.cub.2021.09.084. Epub 2021 Oct 25.

This article constitutes the second chapter of the results of this PhD thesis, found in section 3.3. In this work, both Miquel Àngel Schikora and Ewa Ksiezopolska share the first co-authorship. Miquel Àngel Schikora performed all the bioinformatic analyses. He also generated most of the statistical results and figures. In addition, he contributed to the writing of the manuscript and the interpretation of the results. Note that Ewa Ksiezopolska, who performed the experiments and some of the statistical analyses, used it for her

thesis entitled “Genomic changes driving the acquisition of multidrug resistance in *Candida glabrata*”, by the Universitat Pompeu Fabra.

3. Genome-wide signatures of recent selection and drug resistance across *Candida* opportunistic pathogens

Miquel Àngel Schikora-Tamarit and Toni Gabaldón

This article constitutes the third chapter of the results of this PhD thesis, found in section 3.4. It is currently under review in Nature Microbiology. In this work, Miquel Àngel Schikora performed the bioinformatic analyses and prepared all the figures. In addition, he contributed to the conception of the study, the interpretation of results and writing of the manuscript.

The above-mentioned work was performed as part of this Phd dissertation: “Comparative genomics of recent adaptation in *Candida* pathogens”.

Dr. Juan Antonio Gabaldón Estevan

Thesis director

3.2

PerSVade: personalized structural variant detection in any species of interest

Schikora-Tamarit and Gabaldón
Genome Biology (2022) 23:175
https://doi.org/10.1186/s13059-022-02737-4

Genome Biology

SOFTWARE Open Access

PerSVade: personalized structural variant detection in any species of interest

Miquel Àngel Schikora-Tamarit[1,2] and Toni Gabaldón[1,2,3,4*]

*Correspondence:
toni.gabaldon@bsc.es

[1] Barcelona Supercomputing Centre (BSC-CNS), Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain
[2] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
[3] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[4] Centro Investigación Biomédica En Red de Enfermedades Infecciosas, Barcelona, Spain

Abstract

Structural variants (SVs) underlie genomic variation but are often overlooked due to difficult detection from short reads. Most algorithms have been tested on humans, and it remains unclear how applicable they are in other organisms. To solve this, we develop perSVade (personalized structural variation detection), a sample-tailored pipeline that provides optimally called SVs and their inferred accuracy, as well as small and copy number variants. PerSVade increases SV calling accuracy on a benchmark of six eukaryotes. We find no universal set of optimal parameters, underscoring the need for sample-specific parameter optimization. PerSVade will facilitate SV detection and study across diverse organisms.

Keywords: Structural variants, Variant calling, Short reads, Parameter optimization

Background

Structural variants (SVs) are large changes (typically >50 bp) in the DNA between individuals that alter genome size (duplications and deletions) or generate rearrangements (inversions, translocations, and interspersed insertions) [1, 2]. In eukaryotes, SVs can drive clinically relevant phenotypes including cancer [3–5], neurological diseases [6, 7], or antifungal drug resistance [8, 9]. In addition, SVs may generate significant intraspecific genetic variation across many taxa like humans [10–12], songbirds [13], or rice plants [14]. Despite their role on human health and natural diversity, most genomic studies overlook SVs due to technical difficulties in calling SVs from short reads [15]. This means that the role of SVs remains largely unexplored across eukaryotes.

Inferring SVs from short reads is challenging because it relies mostly on indirect evidence coming from de novo assembly alignment, changes in read depth, or the presence of discordantly paired / split reads in read mapping analysis [16–21]. Long-read-based SV calling may avoid some of these limitations, but short read-based SV calling remains a cost-effective strategy to find SVs in large cohorts [14, 15, 22]. Recent benchmarking studies compared the performance of different tools in human genomes and found that SV calling accuracy is highly dependent on the methods and filtering strategy used [15,

23, 24]. Such studies are useful to define "best practices" (optimal methods and filtering strategies) for SV calling in human samples. However, few studies have investigated the accuracy of these tools on non-human genomes. It is unclear whether the human-derived "best practices" for SV calling can be reliably used in other species. We hypothesize that this may not be the case for genomes with different contents of repetitive or transposable elements, which constrain the short read-based SV calling accuracy [24]. In summary, current tools for short-read-based SV calling are often unprepared for non-human genomes, which hinders the study of SVs in most organisms.

To overcome this limitation, we developed the *personalized structural variation detection* pipeline, or perSVade (pronounced "persuade"), which is designed to adapt a state-of-the-art SV calling pipeline to any sample/individual of any genome/species of interest. PerSVade detects breakpoints (two joined regions that exist in the sample of interest and not in the reference genome) from short paired-end reads and summarizes them into complex SVs (deletions, inversions, tandem duplications, translocations, and interspersed insertions). The pipeline provides automated benchmarking and parameter selection for these methods in any genome or sequencing run, which is useful for species without such recommended parameters. PerSVade provides an automated report of the SV calling accuracy on these simulations, which serves to estimate the confidence of the results on real samples. Beyond SV detection, perSVade can be used to find small variants (single-nucleotide polymorphisms (SNPs) and insertions/deletions (IN/DELs)) and read depth-based copy number variation (CNV), all implemented within a flexible and modular framework.

The following sections describe perSVade and its SV calling performance on various datasets of both simulated and real genomes with SVs.

Results

PerSVade: a pipeline to call and interpret structural variants in your species of interest

PerSVade identifies SVs from a paired-end WGS dataset and a reference genome as sole inputs. It identifies breakpoints from the aligned reads with *gridss* [21] and summarizes them into actual SVs (insertions, translocations, deletions, inversion, and tandem duplications) with *clove* [25]. We followed the recent recommendation of using a single, high-performing algorithm for breakpoint calling instead of using multiple software [24]. We chose *gridss* because of its high accuracy in several benchmarking studies [23, 24]. In addition, our pipeline generates a functional annotation of the variants, which is useful to evaluate the altered genomic regions and aid downstream analyses. In summary, perSVade is a pipeline to find and interpret SVs from most eukaryotic sequencing datasets (Fig. 1).

A key feature of perSVade is the parameter optimization step (implemented in the "optimize_parameters" module and shown in Additional file 1: Figure S1). There are no specific recommendations for filtering the outputs of *gridss* and *clove* in most species, and it is unclear whether the parameters validated on model organisms are universal. Similarly, the performance of these algorithms on different sequencing formats (i.e., varying read lengths, coverage, or insert size) is not easy to predict. To solve this automatically, perSVade "optimize_parameters" generates simulated genomes (based on the reference genome and input dataset) with SVs and chooses the most accurate

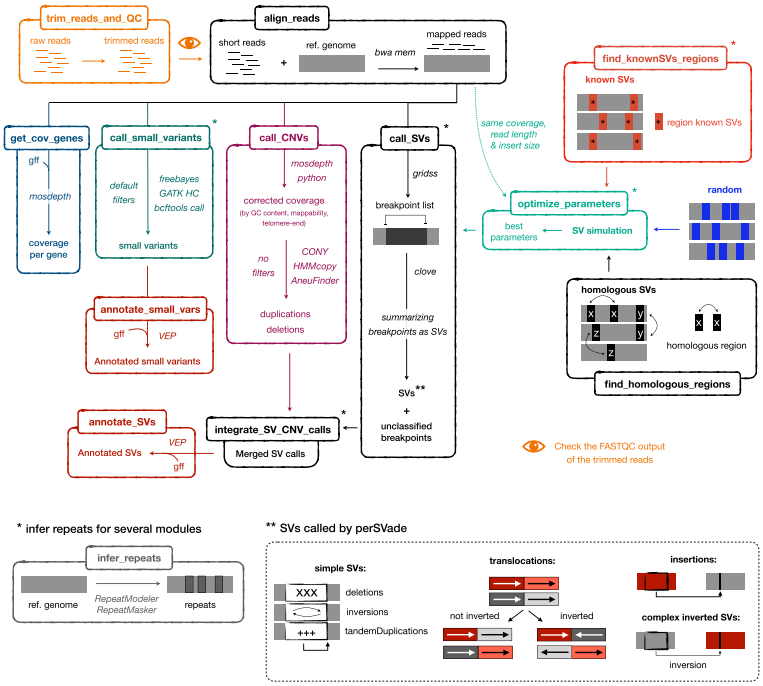


Fig. 1 Schematic representation of the modular workflow of PerSVade. This figure shows the modules of perSVade (each represented in a different box and executable with a single command), which may be combined following the drawn arrows. The italic text describes the algorithms used at each step. The pipeline identifies either structural variants (SVs) (module "call_SVs"), coverage-derived copy number variants (CNVs) (module "call_CNVs"), small variants (module "call_small_variants"), and/or changes in the coverage per gene (module "get_cov_genes") from aligned short paired-end reads (obtained with the module "align_reads"). The different types of SVs output by "call_SVs" are drawn at the bottom for clarity. In addition, the module "trim_treads_and_QC" can be used to trim the reads and perform quality control with *FASTQC* before read alignment. On another note, several modules ("call_SVs," "find_knownSVs_regions," "integrate_SV_CNV_calls," "optimize_parameters," and "call_small_variants") use an annotation of genomic repeats that can be obtained with the module "infer_repeats" (bottom left). The most novel aspect of perSVade is the automatic parameter optimization for SV calling adapted to the input (implemented in the module "optimize_parameters"). This is achieved through simulations of SVs on the reference genome, which can be randomly placed ("random"), around regions with previously known SVs ("known") or on regions with pairwise homology ("homologous"). The modules "find_knownSVs_regions" and "find_homologous_regions" can be used to infer these "known" and "homologous" regions, respectively. In addition, the variants found with "call_SVs" and "call_CNVs" can be combined with the module "integrate_SV_CNV_calls." Finally, the modules "annotate_SVs" and "annotate_small_vars" can be used to obtain a functional annotation of the variants. See "Methods" for more details. In addition, note that Additional file 1: Figure S1 includes a more detailed representation of how "optimize_parameters" works

filters (with the highest harmonic mean between precision and recall (*F*-value)) for these simulations. To account for different mechanisms of SV formation, the simulations can be either (1) randomly placed across the genome ("random" simulations), (2) around regions with previously known SVs ("known" simulations), or (3) around regions with homologous sequences ("homologous" simulations). We consider that "known" and "homologous" simulations are more realistic than the "random" ones. See "Methods" for further details. Regardless of the simulation type, the optimized filters can be used for the SV calling on real data, potentially yielding the highest possible performance. The accuracy of the optimized filters on different simulations is reported as a tabular file, which is useful to define the expected calling accuracy. We hypothesize that this

accuracy may vary across species and/or sequencing formats, and perSVade can infer it on any input sample. All in all, perSVade automatically finds the best filters and reports the expected calling accuracy for each input sample.

We validated the usability of perSVade by running it on available sequences for six phylogenetically diverse eukaryotes with different genome sizes (*Candida glabrata* (12 Mb), *Candida albicans* (14 Mb), *Cryptococcus neoformans* (19 Mb), *Arabidopsis thaliana* (120 Mb), *Drosophila melanogaster* (144 Mb), and *Homo sapiens* (3163 Mb)), with three WGS runs per species (yielding datasets with 6.75×10^6–1.59×10^9 reads, see "Methods"). We ran the pipeline using parameter optimization with "random," "known," or "homologous" simulations. In addition, we ran perSVade with default parameters as a baseline, useful to evaluate the impact of parameter optimization (the core and most novel feature of perSVade) on calling accuracy and resource consumption. We found that the computational burden (running time and memory used) was highly variable among datasets and correlated with genome and dataset sizes. As expected, parameter optimization increased resource consumption in all cases. This burden was particularly high for the human datasets, which may hinder the usage of perSVade on such large genomes if computational resources are limited (Additional file 1: Figure S2). However, we consider that such choices should be left to the user based on these results, since the increased accuracy due to parameter optimization may outweigh resource costs. Taken together, our analysis indicates that perSVade can be used for SV calling in a wide range of eukaryotes and sequencing datasets.

PerSVade's parameter optimization improves calling accuracy in simulated datasets

In order to clarify the impact of parameter optimization on calling accuracy, we measured the performance of perSVade's SV calling on these samples and simulations. We found that the *F*-value after parameter optimization on "random" and "known" simulations was high (between 0.75 and 1.0) in most samples and SV types (with one exception in *Drosophila melanogaster* that yielded an *F*-value ~ 0.5). The *F*-value on "homologous" simulations was often lower (depending on the species), suggesting that SVs happening on regions with pairwise homology may be more difficult to resolve. As expected, the accuracy on "random" SVs was higher than on more realistic simulations ("known" and "homologous"), suggesting that it may overestimate real data accuracy. In general, the *F*-value was higher than the "default" setting in most species (except in *C. neoformans*), and the improvement was dramatic in some SV types (i.e., the *F*-value went from <0.1 to >0.95 in *C. glabrata*'s deletions or insertions) (Fig. 2). In addition, we found that parameter optimization increases recall rather than precision, which is >0.95 in most simulations and SV types (Additional file 1: Figure S3). We also found that using a single set of (global) parameters optimized for all SV types in a given sample yields an accuracy that is as high as using a set of parameters specifically for each SV type (Additional file 1: Figure S4). This validates our approach of running SV calling once (with a single set of parameters) for each sample. Taken together, our results suggest that parameter optimization yields maximum performance by improving the recall of SVs as compared to default parameters.

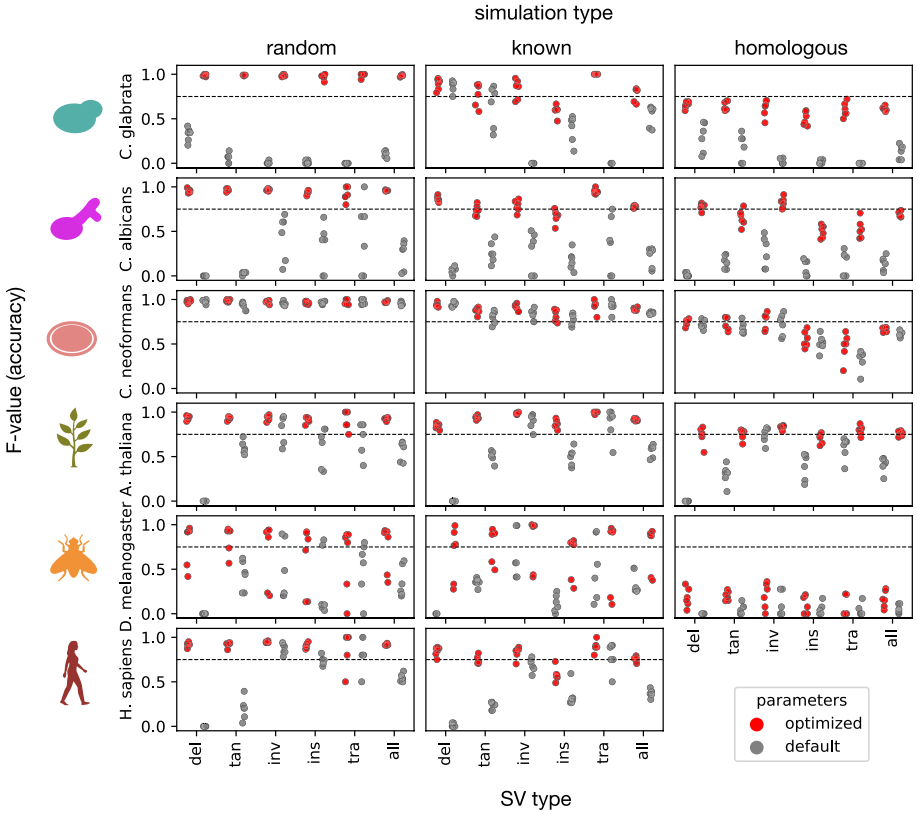


Fig. 2 PerSVade's parameter optimization improves the SV calling accuracy on simulations. We ran perSVade's SV calling on three samples / species for six eukaryotes (see "Methods") using either "random,""known," or "homologous" simulations. These plots show the *F*-value of either default (gray) or optimized (red) parameters (for each sample and simulation type) on these simulations. The *x* axis represents the type of SV (deletions (del), tandem duplications (tan), inversions (inv), insertions (ins), translocations (tra), and the average of all SVs (all)). Note that Additional file 1: Figure S3 shows the corresponding precision and recall, from which the *F*-value is calculated

We next explored whether different runs of perSVade (i.e., in different species or simulation types) yield similar parameters, which may clarify how necessary this optimization is. We hypothesized that each sample and simulation type combination may require specific parameters that would not necessarily work for other samples. To test this, we first compared the chosen parameters across different runs, which appeared to be sample-specific (Additional file 1: Figure S5A). This suggests that there is not a universal recipe (i.e., filtering parameters) for SV calling with perSVade. However, another (null) hypothesis could be that different parameter sets have similar outcomes, without changing the SV calling accuracy. This question was highly important to us. If perSVade's optimization converges to equivalent parameter sets in different samples we would not need the optimization on every sample (i.e., we could re-define one of these parameters as default). In order to sort this out, we evaluated how different parameter sets (either "default" ones or those that are defined as "optimum" for a given sample) work on simulated genomes related to other samples. The results of this analysis are shown in Fig. 3 and Additional file 1: Figure S6. As hypothesized, not all the parameter sets yield accurate results on all samples, with large differences between species (Fig. 3A). However, we found that parameters optimized for one

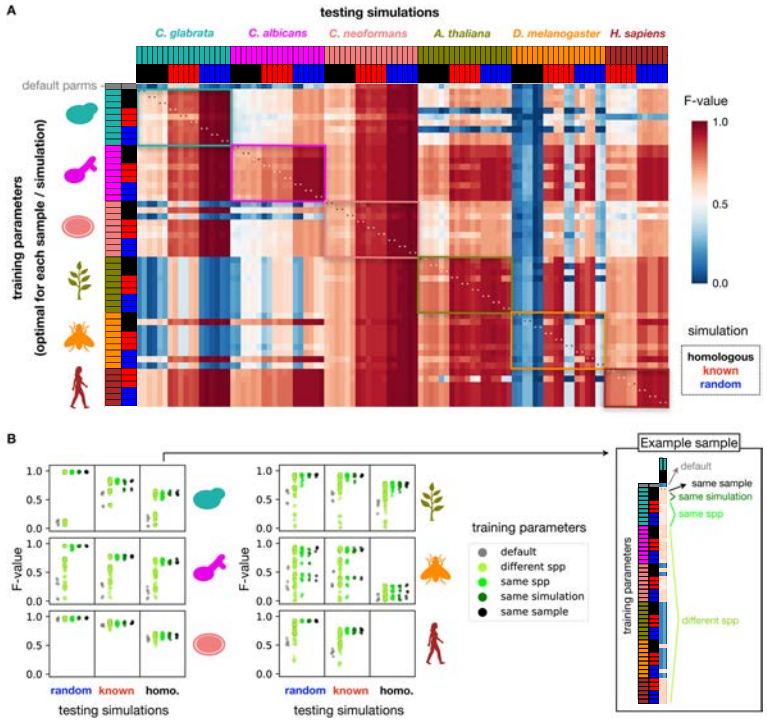


Fig. 3 There is no universal recipe for SV calling across all species. **A** In order to assess whether perSVade's parameter optimization is necessary for a given combination of sample and simulation (mentioned in Fig. 2), we measured the SV calling accuracy of each optimized parameter set on the other combinations. Each row indicates a different "training" parameter set optimized for each sample and simulation type in all species. In addition, the first row refers to the default parameters. Each column represents a simulation from a given sample / simulation type to be "tested." The heatmap shows the *F*-value of each parameter set on each tested simulation (hereafter referred to as "testing instance"). Note that the species are ordered alike in rows and columns. In addition, note that each sample (from a given species and simulation type) yielded one set of training parameters and two simulated genomes tested here, which explains why there are two columns for each row. The colored boxes indicate testing instances where the training and testing species are equal. The asterisks refer to instances where both the sample and type of simulation are equal in the training and testing (equivalent to the "optimized" parameters from Fig. 2). Note that Additional file 1: Figure S6 shows the corresponding precision and recall, from which the *F*-value is calculated. **B** We summarized the data shown in **A** to compare how similar types of training parameters performed on each species (in the rows) and type of simulations (in the columns). Each point corresponds to a testing instance, matching one cell from the heatmap in **A**. The "default" and "same sample" reflect testing instances where the training parameters were either un-optimized or optimized specifically for each sample, respectively. The "different spp" group includes instances where the training parameters were from different species. The "same spp" group shows testing instances with both training parameters and tested simulations from a different sample of the same species. The "same simulation" reflects instances with the same training and testing sample, but different simulation types. For clarity, the right box shows how the training parameters are grouped for a set of "homologous" simulations based on one example *C. glabrata* sample (which corresponds to the first two columns in **A**)

sample are mostly accurate on samples of the same species, regardless of the simulation type (Fig. 3B). Of note, the parameters yielded by "random" simulations were accurate on "homologous" and "real" simulations (Fig. 3). This indicates that running perSVade on "random" simulations (the cheapest setting in terms of resources) yields accurate parameters for more realistic simulations and possibly real SVs. On another line, we found that the different parameters changed mostly the SV calling recall, and not the precision (Additional file 1: Figure S6).

To understand why certain parameter choices impact SV calling accuracy, we tested how changing each parameter in isolation (keeping all others with default values) affects accuracy in these different species and simulations (Additional file 1: Figure S5). We first used these data to assess whether the change of single parameters drives the optimization process. We measured, for each parameter and sample, the ratio between the single-parameter-change *F*-value (where only one parameter has the optimal value) and the maximum *F*-value (obtained with the set of parameters where all parameters are optimized). We find that 78.05% of these parameter-sample instances have an *F*-value ratio below 0.75 (Additional file 1: Figure S5B), suggesting that the optimal accuracy is mostly reached by a complex interplay between different (at least 2) parameters, rather than being driven by a single-parameter change.

This analysis also serves to evaluate the impact of different parameters on SV calling accuracy. For example, we find that the set of vcf "FILTER" tags defining accepted breakpoints (*wrong_FILTERtags* parameter) drastically affects the accuracy in *C. glabrata*, in a way that requiring de novo assembly support for breakpoints (default behavior) is too conservative. A similar (but smaller) effect is observed in *C. albicans*, but not in the other species, which could be due to unique genomic features and/or technical properties in the *Candida* samples driving worse assembly performance. In addition, the coverage thresholds that define tandem duplications and deletions (*min_rel_coverage_dup* and *max_rel_coverage_del*, respectively) determine accuracy in a way that is dependent on ploidy, likely because diploid species (*C. albicans*, *D. melanogaster*, and *H. sapiens*) require a less conservative threshold to accept heterozygous variants. Importantly, these three parameters (*wrong_FILTERtags*, *min_rel_coverage_dup* and *max_rel_coverage_del*) explain why default parameters are suboptimal in most cases, as the default values can be too conservative in different species. On another line, the minimum number of supporting reads per variant (*min_Nfragments*) changes accuracy, with sample-specific effects (see *D. melanogaster* and *A. thaliana*), which we attribute to varying coverage, read lengths, or sequencing quality. Finally, filtering out variants that overlap any repetitive elements (*filter_overlappingRepeats*) generally reduces accuracy for realistic simulations ("homologous" and "known"), likely due to the fact that real variants could appear around such repeats. Conversely, there are other parameters that have minimal effects on accuracy (*dif_between_insert_and_del*, *filter_noReadPairs*, *max_to_be_considered_small_event*, *maximum_length_inexactHomology*, *maximum_microhomology*, *maximum_strand_bias*, *min_QUAL*, *min_af*, *min_length_inversions*, *range_filt_DEL_breakpoints*) (Additional file 1: Figure S5B). However, these parameters can have an impact in some samples and, since perSVade only considers parameter values that can change the filtering in each sample (see "Methods"), we consider that they should not be removed from the "optimize_parameters" module.

Our analysis also showed that the need for parameter optimization is different for each species. An illustrative example is the dramatic difference between *C. neoformans* and *C. glabrata* (Fig. 3A), which provides further insights on the role of various parameters. We found that the parameter choice is irrelevant in *C. neoformans*, while *C. glabrata* samples required specific optimization (Fig. 3A). We consider that this is unlikely driven by intrinsic genomic differences between the two species, as both have small (<20Mb) haploid genomes with low content of simple repeats (0.98% in *C. glabrata* and 0.80% in

C. neoformans) or low-complexity regions (0.16% in *C. glabrata* and 0.21% in *C. neoformans*). We hypothesized that *C. glabrata* samples have an excessively high coverage (>300×, while *C. neoformans* samples have a 30×–40× coverage (Additional file 1: Table S1)) which may constrain SV calling accuracy and require optimized parameters. To test this, we measured the accuracy of different parameter sets on the *C. glabrata* simulations with randomly downsampled coverages. As hypothesized, we find that most parameters are accurate on the *C. glabrata* with 30× coverage, while simulations with lower (10×) and higher (100×–500×) coverage require specific parameters (Additional file 1: Figure S7). These results suggest that 30×–40× could be the optimal coverage for perSVade, which is reasonable given that *gridss* was developed for human datasets with similar coverage. However, there are still differences between the *C. neoformans* and the downsampled (30×) *C. glabrata* samples. For example, there are two parameter sets optimized for the high-coverage *C. glabrata* samples (both requiring at least 30 supporting reads per SV) which are accurate on all *C. neoformans* simulations (Fig. 3A), but not on the *C. glabrata* 30× (Additional file 1: Figure S7). This suggests that there are different genomic features between these species (i.e., content of simple repeats) constraining the accuracy. These findings indicate that both technical variation (i.e., changes in coverage) and different genomic features underlie the observed differences in SV calling accuracy between species. Importantly, this also illustrates that perSVade can adapt to each sample and yield optimal results.

In summary, our results suggest that parameter optimization is necessary for maximum performance in each species and dataset and that there is a complex interplay between parameters.

PerSVade's parameter optimization improves the calling accuracy in datasets with defined sets of real SVs

The performance of SV calling on simulations may not be equivalent on real data, as SVs often appear around repetitive or low-complexity regions which hamper their detection [24, 26–28]. It is thus possible that we overestimated the real accuracy in our simulations. We partially addressed this with our analysis based on "realistic" simulations ("known" and "homologous"), where the inferred accuracy was lower (Fig. 2) and potentially closer to the real one. To further validate the usage of perSVade for real SV calling, we tested it on datasets with known SVs, which were available for the human samples tested above (i.e., Fig. 3). We ran perSVade (using different simulation types) on the same three datasets, which had previously defined deletions and inversions (see "Methods" for details).

We used these data to assess the accuracy of perSVade on real datasets, using different sets of parameters (optimal for each simulation and sample from the six species tested above, shown in Fig. 3). As expected, we found a lower *F*-value on real datasets (Fig. 4) as compared to the simulated genomes (Figs. 2 and 3), with high precision and lower recall (Fig. 4B). In addition, parameter optimization improved the *F*-value modulating both precision and recall (Fig. 4B). However, the other results described in the simulations' analysis (related to the performance of the pipeline and the universality of the parameters) are qualitatively equivalent in these real datasets (Fig. 4). Taken together,

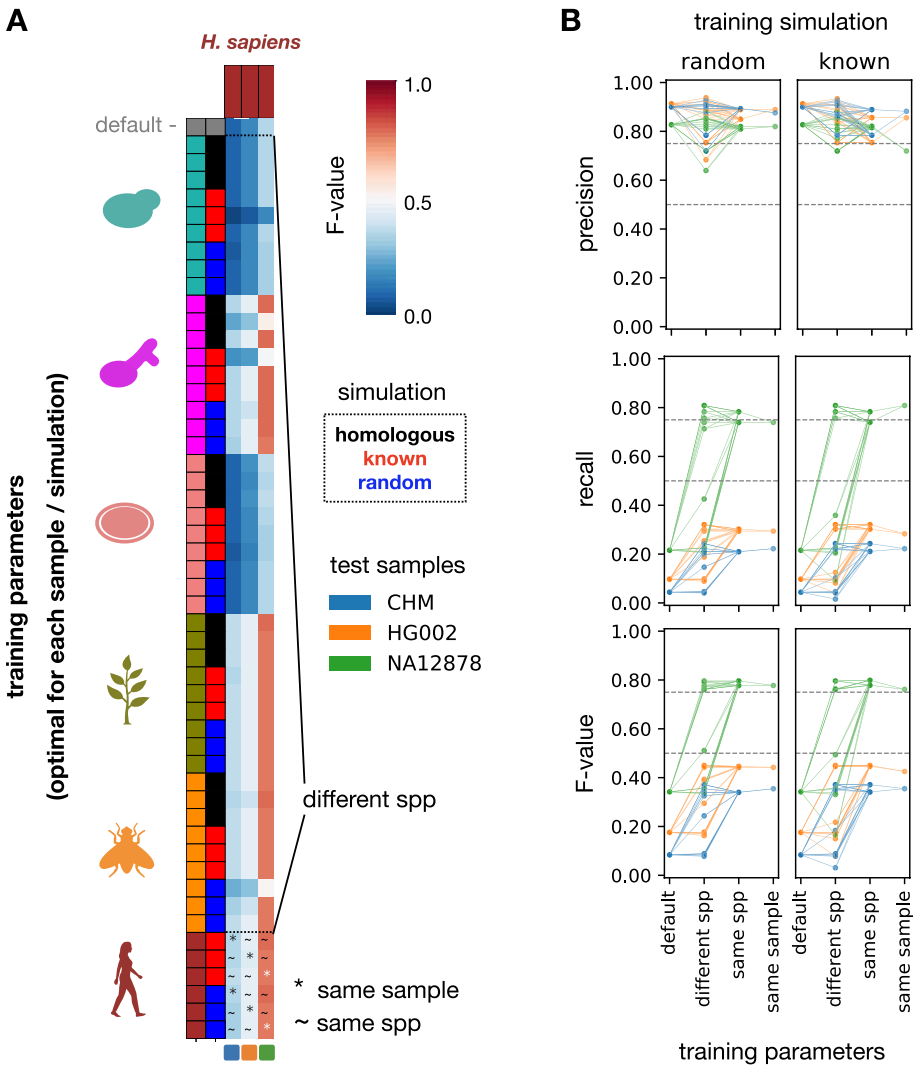


Fig. 4 PerSVade's parameter optimization improves the SV calling accuracy on datasets with known real SVs. **A** To test perSVade's performance on real SVs, we measured how the parameters optimized for several simulations in different species (see Fig. 3) work on three human samples (CHM, HG002, and NA12878) with defined sets of real SVs. Each row indicates one of these different "training" parameters optimized for each sample and simulation type. In addition, the first row refers to the default parameters. Each column represents a sample with defined real SVs to be "tested." The heatmap shows the *F*-value of each parameter set on each tested real sample (hereafter referred to as "testing instance"). In addition, we divide the testing instances into different groups ("default," "different spp," "same spp," and "same sample"), which are relevant to understand the **B** panel. The "different spp" group refers to instances where the training and testing species were different. The "~" (same spp) refers to instances where the training and testing samples were different, but from the same species. Finally, the "*" (same sample) refers to instances where the training and testing samples were the same. **B** We summarized the data shown in **A** to compare how similar types of training parameters performed on each testing sample (each represented by a different color). Each row corresponds to a different accuracy measure. Each point corresponds to a testing instance (matching one cell from the heatmap in **A** in the bottom "*F*-value" plots). The "default" and "same sample" reflect testing instances where the training parameters were either un-optimized or optimized specifically for each sample, respectively. The "different spp" group includes instances where the training parameters were from a different, non-human, species. The "same spp" group shows testing instances with both training parameters and tested simulations from different samples of the same species. In addition, each column represents testing instances where the training parameters were based on "random" or "known" simulations, respectively. Note that the different groups of "training parameters" are equivalent to those shown in **A**

our analysis indicates that perSVade improves SV calling in real datasets (similarly to simulated genomes).

Discussion

Despite large variation of genomic features across taxa, SV detection approaches in non-model organisms tend to rely on tools and parameters developed for other species (generally human). We hypothesized that this "one size fits all" approach is suboptimal, and likely biased towards certain species. To test this idea and overcome the problem, we developed perSVade, a flexible pipeline that automatizes the calling and filtering of structural variants (SV) across eukaryotes. PerSVade is a modular method to automatically adapt a state-of-the-art SV calling pipeline to any sample/species of interest. PerSVade uses simulations to choose the optimal filters for each sample and report the calling accuracy, which can inform about the reliability of the results. This will allow users to be aware of the accuracy in their datasets (i.e., perSVade may be inaccurate in some datasets due to low coverage, short read lengths, or excessive repeats in the genome) and make informed choices.

We validated the broad usability of perSVade by testing it on simulations and real datasets for a wide range of eukaryotes (with genomes of 12–3000 Mb and datasets including 10^7–10^9 reads). We found that there is a significant computational burden related to parameter optimization, which may hinder its usage on large genomes. This means that perSVade may be particularly cost-effective for small genomes (i.e., <200 Mb). However, the chosen settings will likely depend on the available resources, and some users may consider that the resources spent (see Additional file 1: Figure S2) are worth it given that parameter optimization yields improved accuracy (see below).

This testing also revealed that, as we hypothesized, parameter optimization improves the calling accuracy on both simulations and datasets with real, previously defined SVs. We found that the optimization mostly improves the recall rather than precision (which is generally high regardless of the used parameters). However, there are some exceptions (mostly in the testing on real SVs), suggesting that optimization can be necessary for reaching both high recall and precision in some samples. In addition, perSVade's optimization yielded unique parameter sets for each sample, which were often inaccurate on other datasets. This means that there is no universal set of parameters that work well for all samples, which justifies the need for parameter optimization and a tool like perSVade to automate such a task. Conversely, we found some trends that can be useful to skip parameter optimization in some cases. For instance, parameter sets were often accurate across datasets of the same species (which could be due to differences in coverage and/or intrinsic genomic features). In addition, parameters resulting from "random" simulations performed well in more realistic ("known" and "homologous") simulations as well as in real SV datasets of the same species, indicating that they can be used for maximum performance. Based on these findings, we propose the following recommendations for a cost-effective usage of perSVade:

– For SV calling on many datasets of one species with similar properties (similar coverage, read length, and insert size), run perSVade using "random" simulations on one sample, and use the optimized parameters for the other samples (skip-

ping optimization). The reported calling accuracy may be overestimated since the simulations are not realistic, but the chosen parameters are expected to be optimal. This strategy may be particularly suited to large genomes, where users would avoid the computational burden of optimizing parameters for each sample.

- For approximating the real SV calling accuracy, run perSVade on realistic simulations ("homologous" or "known"), which may report an accuracy that is closer to the real one.
- For SV calling on large datasets, consider the following options to speed up the process. First, rationally design the parameters (based on parameters optimized for similar samples (see first point) and/or the benchmarking shown in this work) instead of inferring them with the "optimize_parameters" module for every new sample. Second, skip marking duplicates in read alignment, which can be very costly. Third, limit the simulations to a subset of chromosomes in the "optimize_parameters" module. Fourth, randomly downsample your reads (i.e., to 30×), which may improve both performance and accuracy (see Additional file 1: Figure S7).

We note that perSVade is not a fundamentally new algorithm for SV detection but rather a pipeline implementing existing algorithms. This is why we did not compare it with other such methods (like *manta* [20] or *delly* [29]). The novelty of our pipeline lies in the automatic parameter selection feature, which is unique (to the best of our knowledge) for short read-based SV calling. We thus centered our testing on the accuracy of different parameters on SV calling. In fact, some recent approaches specifically developed for human genomes [22, 30] may outcompete perSVade in human samples. However, such methods rely on previously defined sets of known SVs, which are not available in most taxa. We thus consider that our pipeline will be mostly useful in species without such specific methods available. For example, perSVade was used in a recent study to find SVs associated with antifungal drug resistance in the non-model yeast *Candida glabrata* [9], which successfully validated all (8/8) the predicted variants using PCR.

An open question is whether a similar parameter optimization strategy can be applied to SV calling pipelines based on other algorithms. Several studies have shown that the filtering strategies (considering parameters like read coverage, variant quality and vcf "FILTER" tags) largely affect the calling accuracy in various algorithms [23, 24]. This suggests that most SV callers could be boosted with a parameter optimization strategy such as the one described here for *gridss* and *clove*. However, due to high heterogeneity in SV callers, each algorithm may require a custom pipeline to deal with caller-specific parameters, outputs, and SV types.

Finally, perSVade also includes modules for CNV identification and SNP/INDEL calling, as a way to automate the finding of other broadly used genomic variants. In addition, it includes variant annotation features to ease the functional interpretation of these variants for downstream analyses. In summary, perSVade is a Swiss-knife-like framework to identify many types of variants with a few bash commands. We consider that this tool will be useful to understand the role SVs in different phenotypes and organisms, particularly those with no specific recommendations.

Conclusions

- perSVade can identify SVs from short reads with sample-optimized parameters using a few bash commands.
- perSVade can predict the SV calling accuracy on simulated genomes, which informs about the reliability of the calling process in an automatic manner.
- perSVade's parameter optimization improves the SV calling accuracy on simulated variants for five eukaryotic organisms, and on a reference dataset of validated human variants.
- We found no universal set of "optimal" parameters, which underscores the need for species-specific parameter optimization for SV calling.

Methods

PerSVade pipeline

PerSVade has several modules that can be executed independently (each with a single command) and/or combined to obtain different types of variant calls and functional annotations. The following sections describe how each of these modules work, and Fig. 1 shows how they can be combined.

Module "trim_reads_and_QC"

This module runs *trimmomatic* [31] (v0.38) with default parameters for the input reads followed by *fastqc* [32] (v0.11.9) on the trimmed reads. These trimmed reads may be used for downstream analysis after checking that they are reliable according to the output of *fastqc*.

Module "align_reads"

This module runs *bwa mem* [33] (v0.7.17) to align the short reads, generating a sorted .bam file (using *samtools* [34] (v1.9) with marked duplicates (through *GATK MarkDuplicatesSpark* [35] (v4.1.2.0)), that is the core input of several downstream modules ("call_SVs," "optimize_parameters," "call_CNVs," "call_small_variants," and "get_cov_genes"). If *--skip_marking_duplicates* is specified, this module skips the marking of duplicate reads (default behavior), which may be useful to speed up the process in large datasets.

Module "call_SVs"

This module uses *gridss* [21] to infer a list of breakpoints (two regions of the genome—two breakends—that are joined in the sample of interest and not in the reference genome) from discordant read pairs, split reads, and de novo assembly signatures. The breakpoints are summarized into SVs with *clove* [25] (v 0.7). Importantly, this module (and others) runs *clove* without the default coverage filter to classify deletion-like (DEL-like) and tandem duplication-like (TAN-like) breakpoints into actual deletions and tandem duplications. Instead, perSVade "call_SVs" calculates the relative coverage of the regions spanned by such breakpoints (using *mosdepth* [36]). This information is used to define the final set of deletions (DEL-like breakpoints with a coverage below a "max_rel_coverage_to_consider_del" threshold) and tandem duplications (TAN-like

breakpoints with a coverage above a "min_rel_coverage_to_consider_dup" threshold). This setting allows a separate thresholding for the classification of DEL and TAN-like breakpoints, which is a novel feature of perSVade as compared to the current implementation of *clove*. Note that this module requires as an input a set of parameters to filter the *gridss* and *clove* outputs. These parameters may be inferred using the module "optimize_parameters" (described below) or rationally designed based on the benchmarking shown here (which can be useful to speed up the process). In addition, the user can use a set of default parameters, described in the section "Filters used by perSVade" below. Note that these default parameters are inspired by previous filtering strategies from [21, 37, 38].

The final output of this module is a set of files with the called variants (one file for each variant type), which belong to these types:

- Simple SVs: deletions, inversions, and tandem duplications (duplication of a region which gets inserted next to the affected region). This module outputs one .tab file for each of these SV types.
- Translocations: whole-arm balanced translocations between two chromosomes, which can be inverted or not. There is one .tab file for translocations.
- Insertions: a region of the genome is copied or cut and inserted into another region. Note that these are not de novo insertions (i.e., of DNA not present in the reference), which are actually not called in this module. There is one .tab file for insertions.
- Unclassified SVs: One .tab file reports all the variants that are called by *clove* and cannot be assigned to any of the above SV types. These include *clove*'s unclassified breakpoints (which could be part of unresolved/unknown complex variants) and complex inverted SVs (which are non-standard SVs). These types of SVs are not included in the simulations performed by "optimized parameters" (see below), so that their accuracy is unknown. This is why we group them together into a single file.

These separate files have a tabular format, where each variant is represented in a single line. In addition, the module "integrate_SV_CNV_calls" (see below) generates a single .vcf file with all the variants together, represented in a way that is focused on how each SV affects particular regions of the genome (useful for functional annotation). PerSVade's github wiki [39] includes further information on the output formats.

On another line, note that *gridss* does report de novo insertions, but the usage of short reads limits the calling to small events, which may miss many real de novo insertions. This is why we decided to not consider such variants as a trustful output in this module or the "optimize_parameters." However, "call_SVs" saves the raw *gridss* output, and the unfiltered small de novo insertions can be obtained (although these should be treated with caution). In addition, note that these *de novo* insertions are different from non-template insertions happening around the breakends of actual SVs. Non-template insertions are likely the product of DNA repair after the rearrangement, and they are considered in the "integrate_SV_CNV_calls" (see below).

Module "optimize_parameters"

To find optimal parameters for running "call_SVs" in a given input dataset, this module generates two template (haploid) simulated genomes (the number can be customized

with *--nsimulations*) with up to 50 SVs of each of five types (insertions, translocations, deletions, inversion, and tandem duplications) with *RSVsim* [40] (v1.28) and custom python (v3.6) scripts (which use *biopython* [41] (v1.73)). By default, this template genome contains all chromosomes in the reference, but this can be customized with the *--simulation_chromosomes* argument to only simulate a subset of chromosomes and speed up the process. For each template genome, the module simulates reads with *wgsim* [42] (v1.0) and *seqtk* [43] (v1.3) with a read length, insert size, and coverage matching that of the input dataset. Note that the read simulation is performed according to a user-defined zygosity and ploidy (through the mandatory argument "*--simulation_ploidies*") to resemble various organisms. For example, if "*--simulation_ploidies diploid_hetero*" is specified, this module simulates reads with heterozygous SVs by merging reads from both the reference genome and the simulated genome with SVs in a 1:1 manner. Importantly, *--simulation_ploidies* can have multiple values, so that for each template simulated genome and simulation ploidy this module generates unique simulated reads with the specified ploidy and zygosity. For example, if "*--nsimulations 2 --simulation_ploidies diploid_hetero,diploid_homo*" is set, this module generates four simulated reads. First it generates two template genomes, and for each of them it simulates reads with either heterozygous or homozygous SVs. Note that "*--simulation_ploidies*" can include any combination of "haploid," "diploid_homo," "diploid_hetero," and "ref:<nref>_var:<nvar>" (where <nref> / <nvar> are the number of reference / alternative chromosomal copies, respectively). For example, setting "*--simulation_ploidies* ref:3_var:1" simulates reads assuming a tetraploid genome, where three chromosomes are like the reference and one has the SVs. This flexibility in setting ploidies / zygosity allows adapting this module to polyploid genomes or complex samples (i.e., pools of different samples of a population).

For each set of simulated reads (from one template genome with a specific ploidy and zygosity), perSVade "optimize_parameters" then tries several combinations (>278,000,000,000 by default, although this can be user-defined with the argument *--range_filtering_benchmark*) of parameters to run *gridss* and *clove* and filter their outputs. The detailed explanation about the used filters can be found in the section "Filters used by perSVade" below**.** To reduce the number of parameter combinations to be optimized, the pipeline discards parameter values that do not change breakpoint filtering as compared to an unconservative set of parameters. This means that the set of parameters to be optimized are limited to those that can be relevant, and these could be different in any run. One of these possible filters includes removing SVs that overlap repetitive elements, which may be inferred with the module "infer_repeats" (see below). This module selects the combination of filters that yield the highest *F*-value (the harmonic mean between precision and recall) for each SV type in each template simulated genome and ploidy/zygosity (see the section "Comparing sets of SVs to calculate precision and recall" below for more information on how accuracy is calculated). These filters are optimized for each simulation, and thus may not be accurate on independent sets of SVs (due to overfitting). In order to reduce this effect, perSVade "optimize_parameters" selects a final set of "best parameters" that work well for all simulations, ploidies/zygosities, and SV types. This set of best parameters may be used in the "call_SVs" module. The accuracy (*F*-value, precision, recall) of these parameters on each simulation and SV type is reported in a tabular file, which serves to evaluate the expected calling accuracy. Note that we default the number of template simulated genomes

to two in order to have a meaningful evaluation of overfitting (which likely requires more than one template genome). In addition, note that setting several simulation ploidies can be useful to select parameters that work well for different ploidies/zygosities.

All plots are generated using *python* (v3.6) and the libraries *seaborn* [44] (v0.9.0) and *matplotlib* [45] (v3.3.0). In addition, the *python* packages *scipy* [46] (v1.4.1), *scikit-learn* [47] (v0.21.3), *psutil* [48] (v5.7.2), and *pandas* [49] (v0.24.2) are used for scripting and various statistical calculations. On another line, *pigz* [50] (v2.4) and *gztool* [51] (v0.11.5) are used for fast compression steps. Finally, perSVade "optimize_parameters" uses *picard* [52] (v2.18.26) to construct a sequence dictionary for the provided reference genome.

By default, the simulated events are placed randomly across the genome. However, real SVs often appear around repetitive elements or regions of the genome with high similarity (e.g., transposable elements insertions) [24, 26–28]. This means that random simulations may not be realistic, potentially leading to overestimated calling accuracy and a parameter selection inaccurate for real SVs [24]. To circumvent this, perSVade "optimize_parameters" can generate more realistic simulations occurring around some user-defined regions (i.e., with previously known SVs or homologous regions) provided with the *--regions_SVsimulations* argument. Importantly, perSVade provides an automatic way to infer such regions through the modules "find_knownSVs_regions" and "find_homologous_regions" (described below). Beyond setting custom regions, users may want to tune the number of simulated SVs (through the *--nvars* argument) to be realistic in the samples/species of interest. In addition, note that the variant size is proportional to genome length, which ensures that long genomes have larger sections under SV.

Finally, note that Additional file 1: Figure S1 includes a detailed graphical representation which can be useful to understand how this module works.

Module "find_knownSVs_regions"

This module finds regions with known SVs using a provided list of sequencing datasets (with the option *--close_shortReads_table*) from species close to the reference genome. These datasets are processed with perSVade's modules "trim_reads_and_QC," "align_reads," and "call_SVs" (using default parameters) to find SVs. This module then outputs a .bedpe file with the ±100bp regions around the breakends from these SVs. This .bedpe file can be input to the module "optimize_parameters" through the *--regions_SVsimulations* argument in order to perform "known" realistic simulations.

Module "find_homologous_regions"

This module infers homologous regions by defining genomic windows (from the reference genome) of 500 bp as a query for a *blastn* [53] (v2.10.0+) against the same reference genome. Hits with an e-value $<10^{-5}$ that cover >50% of the query regions are defined as pairs of homologous regions, which are written as a .bedpe file. This .bedpe file can be input to the module "optimize_parameters" through the *--regions_SVsimulations* argument in order to perform "homologous" realistic simulations.

Module "call_CNVs"

Copy number variants (CNVs) are a type of SVs in which the genomic content varies (deletions or duplications). The "call_SVs" module (see previous section) identifies some

CNVs (insertions, tandem duplications, deletions, and complex inverted SVs) but it can miss others (i.e., whole-chromosome duplications or regions with unknown types of rearrangements yielding CNVs [8, 54]). PerSVade uses this "call_CNVs" module to call CNVs from read-depth alterations. For example, regions with 0× or 2× read depth as compared to the mean of the genome can be called deletions or duplications, respectively. A straightforward implementation of this concept to find CNVs is challenging because many genomic features drive variability in read depth independently of CNV [55, 56]. In order to solve this, perSVade "call_CNVs" calculates the relative coverage for windows of the genome (using *bedtools* [57] (v2.29.0) and *mosdepth* [36] (v0.2.6)) and corrects the effect of the GC content, mappability (calculated with *genmap* [58] (v1.3.0)), and distance to the telomere (using *cylowess* [59] for nonparametric regression as in [56]). Note that *cylowess* uses the library *cython* [60] (v0.29.21). This corrected coverage is used by *CONY* [61] (v1.0), *AneuFinder* [62] (v1.18.0), and/or *HMMcopy* [63] (v1.32.0) to call CNVs across the genome. Note that we modified the R code of *CONY* to be compatible with the input corrected coverage. The corrected code (used in the pipeline) is available in "scripts/CONY_package_debugged.R" from [39]. PerSVade "call_CNVs" generates consensus CNV calls from the (up to) three programs taking always the most conservative copy number for each bin of the genome. For example, if the used programs disagree on the copy number of a region the closest to 1 will be taken as the best estimate. Note that the parameters obtained in the module "optimize_parameters" cannot be used for this module, since the SV and CNV calling methods are fundamentally different.

Module "integrate_SV_CNV_calls"

This module generates a vcf file showing how SVs (called by the modules "call_SVs" and/or "call_CNVs") alter specific genomic regions. We designed this vcf to be compatible with the Ensembl Variant Effect Predictor [64] (VEP) tool for functional annotation, which can interpret tandem duplication (TDUP) duplication (DUP), deletion (DEL), and breakend-like (BND) events. This requires the decomposition of each variant into such TDUP, DUP, DEL, and BND events (one event in each row of the vcf). For example, each inversion is decomposed into two BND events (two rows in the vcf), one for each end of the inversion. The rationale behind this is that, in terms of functional annotation for inversions, we are interested in genomic features that are around the ends of the inversion, where the rearrangement happens. Each SV can thus be split across multiple rows when it affects more than one region of the genome. All rows related to the same SV are identified by the field variantID in INFO. On top of this, each row has a unique identifier indicated by the field ID. Some SVs generate non-template inserted sequences around the breakends (likely the product of DNA repair after a rearrangement), and each of these is represented in a single row. Note that each of the rows may indicate a region under CNV (with the SVTYPE in INFO as DEL, DUP, or TDUP), a region with some breakend (with the SVTYPE in INFO as BND) or a region with a non-template insertion (with the SVTYPE in INFO as insertionBND) around the breakend. Such non-template insertions are included here because they may modulate the impact of SVs on genomic features, and thus they are relevant for functional annotation. Note that this module also removes redundant calls between the CNVs identified with "call_SVs" (tandem

duplications, deletions and insertions) and those derived from "call_CNVs." To remove redundancy, this module skips any CNV called by "call_CNVs" that overlaps reciprocally (by at least an 80% of the region) a CNV called by "call_SVs" using *bedmap* from the *bedops* tool [65] (v2.4.39). See the FAQ "What is in SV_and_CNV_variant_calling.vcf?" from [39] for more information about the format of this .vcf file.

Module "annotate_SVs"

This module runs the Ensembl Variant Effect Predictor [64] (v100.2) on the vcf output of the module "integrate_SV_CNV_calls" to get the functional annotation of each SV. This requires a .gff file from the user.

Module "call_small_variants"

This module performs small variant (SNPs and small IN/DELs) calling with either *freebayes* [66] (v1.3.1), *GATK HaplotypeCaller* [67] (v4.1.2.0), and/or *bcftools call* [68] (v1.9) and integrates the results into .tab and .vcf files. The section "Calling of small variants" below provides further information on how this calling is performed.

Module "annotate_small_vars"

This module runs the Ensembl Variant Effect Predictor [64] (v100.2) on the vcf output of the module "call_small_variants" to obtain the functional annotation of each variant. This requires a .gff file from the user.

Module "get_cov_genes"

This module runs *mosdepth* [36] (v0.2.6) to obtain the coverage for each gene, which requires a .gff file from the user.

Module "infer_repeats"

This module annotates repetitive elements in a genome, which can be used for the modules "call_SVs," "find_knownSVs_regions," "integrate_SV_CNV_calls," "optimize_parameters," and "call_small_variants." These repeats are inferred with RepeatModeler [69] (v2.0.1) and RepeatMasker [70] (v4.0.9). The user can input these repeats to several modules (with --repeats_file), which will have the following effects:

- If repeats are provided, "optimize_parameters" will assess whether removing SV calls overlapping repeats increases the overall accuracy. If so, the resulting optimized parameters will include a "filter_overlappingRepeats : True." If you use these optimized parameters in "call_SVs," any breakpoint overlapping repeats will be removed.
- If repeats are provided, "call_SVs" may filter out SVs that overlap repeats if the SV filtering parameters include a "filter_overlappingRepeats : True."
- If repeats are provided, "find_known_SVs" will pass them to the "call_SVs" module.
- If repeats are provided, "integrate_SV_CNV_calls" will add a field in the INFO which indicates whether the SVs overlap repeats.
- If repeats are provided, "call_small_variants" will add a field in the tabular variant calling file which indicates whether the SVs overlap repeats.

Alternatively, the user can specify "--repeats_file skip" to avoid the consideration of repeats in all these modules.

Testing SV calling with perSVade on simulated structural variants

To test perSVade's performance on different species, we ran it on paired-end WGS datasets for six eukaryotes (*Candida glabrata*, *Candida albicans*, *Cryptococcus neoformans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Homo sapiens*). To obtain a high number of SVs, we gathered three samples for each species with enough genetic divergence to the reference genome. For this, we first used an automatic pipeline to find these samples running the custom script "scripts/perSVade.py" from [39] with the options *--close_shortReads_table auto --n_close_samples 3 --nruns_per_sample 1 --target_taxID* <*species_taxID*>. This used *entrez-direct* [71] (v13.3), *SRA Tools* [72] (v2.10.9), and *ete3* [73] (v3.1.2) to query the SRA database [74] and find three WGS datasets of close taxIDs (to each <*species_taxID*> according to the NCBI taxonomy species tree [75]) with a coverage >30× and >40% of mapped reads to the reference genome. We could find three such datasets for *C. albicans*, *C. neoformans*, *A. thaliana*, and *D. melanogaster*, which included samples from the same species or genera as the target species, with >65% of the reads mapping to the reference genome. We randomly downsampled the *A. thaliana* and *D. melanogaster* runs to 30× coverage (using *samtools* [34] (v1.9)) for faster computation (using the option *--max_coverage_sra_reads 30*). For *C. glabrata*, we used datasets generated in our lab from three divergent strains (BG2, CST34, and M12, from [9]). All these datasets are listed in Additional file 1: Table S1. Finally, we tested perSVade on three *H. sapiens* datasets previously used for benchmarking SV callers [23, 24]. These included NA12878 (a Genome in a Bottle (GIAB) cell line related to the Ceph family [76, 77]), HG002 (another GIAB project with reads available at [78]), and CHM1/CHM13 (two haploid cell lines sequenced independently [79], for which we merged the raw reads to generate synthetic diploid data). Note that we chose testing datasets with various read lengths and coverages (see Additional file 1: Table S1) to evaluate how perSVade works on realistic diverse scenarios. The reference genomes were taken from the Candida Genome Database [80] (version s02-m07-r35 for *C. glabrata* and "haplotype A" from version A22-s07-m01-r110 for *C. albicans*), GenBank [81] (accession GCA_000149245.3 for *C. neoformans*, GCA_000001735.2 for *A. thaliana* and GCA_000001215.4 for *D. melanogaster*), and UCSC [82] (the latest version of genome hg38 at 06/04/2021 for *H. sapiens*, keeping only chromosomes 1-22, X,Y and the mitochondrial DNA). In addition, we performed quality control of the reads with fastqc [32] (v0.11.9) and trimming with *trimmomatic* [31] (v0.38).

We ran the SV calling pipeline of perSVade (using the modules "align_reads," "call_SVs," and "integrate_SV_CNV_calls") on all these datasets using either "default" or optimized parameters (based on "random," "known," or "homologous" simulations using the modules "optimize_parameters," "find_homologous_regions," and "find_knownSVs_regions"). Note that the default parameters were designed as a baseline to understand the need for parameter optimization. We thus pre-defined these parameters based on standard author recommendations (from previous filtering strategies designed by the *gridss* authors [21, 37, 38]). By comparing the results of such parameters (designed

based on previous usage) and the optimized ones, we could assess the gain in SV calling accuracy associated with parameter optimization. In addition, note that we used the module "infer_repeats" to find repetitive elements in each genome. These were provided to "optimize_parameters" to assess whether filtering out repeats improved SV calling accuracy. In addition, we simulated diploid heterozygous SVs for the diploid genomes (*C. albicans*, *A. thaliana*, *D. melanogaster*, and *H. sapiens*) and haploid SVs for the haploid genomes (*C. glabrata*, *C. neoformans*). Note that we decided to only simulate heterozygous variants in the diploid genomes to create the most challenging scenario for SV calling (since homozygous variants are expected to be easier to find due to higher coverage), as previously done [24]. In addition, the output of the "infer_repeats" module was used to calculate the fraction of the genome with simple repeats or low-complexity regions of *C. glabrata* and *C. neoformans.* We used computational nodes in an LSF cluster (https://www.ibm.com/support/pages/what-lsf-cluster) with 16 cores and either 32 Gb (for *C. glabrata*, *C. albicans*, *C. neoformans*), 64 Gb (for *A. thaliana* and *D. melanogaster*), and 96 Gb (for *H. sapiens*) of RAM for the testing. We first ran the read alignment step (module "align_reads") for all samples, and then used the resulting .bam files as inputs for the other perSVade modules. We calculated the resource consumption (running time and maximum RAM used) for each of these perSVade runs, thus ignoring the resources related to read alignment. Of note, perSVade was run with different parameters for the human datasets to avoid excessive resource consumption and match our computational infrastructure. First, we skipped the marking of duplicate reads on the .bam files (default behavior) with perSVade's *--skip_marking_duplicates* option on the module "align_reads." Second, we ran the simulations on a subset of the genome (only chromosomes 2, 7, 9, X, Y and mitochondrial), by using the *--simulation_chromosomes* argument of the "optimize_parameters" module. Third, we skipped the "homologous" simulations in human samples because we could not finish the inference of pairs of homologous regions (see previous section) due to excessive memory consumption. By running this inference on a few chromosomes, we realized that there are millions of such regions, generating excessively large files. Note that this strategy may be used in general to speed up parameter optimization.

Finally, we tested the accuracy of all the optimized parameters (for each sample / simulation) on the other samples / simulations using the script "testing/get_accuracy_parameters_on_sorted_bam.py" from [39]. In addition, to test the impact of changing each parameter in isolation, we generated sets of parameters where only one parameter is changed to a non-default value. We then used this same script ("testing/get_accuracy_parameters_on_sorted_bam.py" from [39]) to measure the accuracy of each parameter set on each sample / simulation. On another line, to assess whether the high coverage of *C. glabrata* samples (>300×, see Additional file 1: Table S1) constrained SV calling, we measured the accuracy of each parameter set (optimized for each species / simulations) on the *C. glabrata* simulations with varying coverage. For each simulation (based on a sample and a type of simulation (homologous / known / uniform)), we subsampled randomly the reads to get a coverage of 10×, 30×, 50×, 100×, 200×, or 300× using *samtools* [34] and *mosdepth* [36] on the aligned simulated reads. We then used our custom script "testing/get_accuracy_parameters_on_sorted_bam.py" from [39] to test the SV calling accuracy on each downsampled simulation. The section "Comparing sets of

SVs to calculate precision and recall" below provides further information on how accuracy is calculated.

Testing perSVade on real SVs

To validate the usage of perSVade on real data, we focused on public datasets with available short reads and independently defined sets of known SVs. We could find such SVs in the human samples (also used in the testing mentioned above), for which SV callsets of deletions or inversions exist (as done in [24]). We defined as "true SVs" the deletions of NA12878 (defined in [77], available at [83]), the high-confidence deletions of HG002 (available at [84]) and the union of all deletions and inversions found in either CHM1 or CHM13 lines (defined by [79], available at [85]).

We then tested the accuracy of the "training" parameters optimized for each sample and simulation of the six eukaryotes mentioned above (in the section "Testing SV calling with perSVade on simulated structural variants") on these human samples using our custom script "testing/get_accuracy_parameters_on_sorted_bam.py" from [39]. In addition, we removed SVs overlapping simple repeats or low-complexity regions (as inferred by the module "infer_repeats") from this analysis. Note that each of these "true SV" datasets were defined on different reference genomes: the NA12878 and HG002 callsets were based on hg19 and the CHM1/CHM13 was relative to hg38. This means that we could not directly use the optimized training parameters from the human samples from the previous section, since they were all based on hg38. We thus ran perSVade's SV calling and parameter optimization modules on NA12878 and HG002 using the hg19 reference, and used the resulting optimum parameters as "training" for these two samples. For this, we obtained the latest version of hg19 and hg38 genomes at 06/04/2021 from UCSC [82], keeping only chromosomes 1-22, X,Y, and the mitochondrial DNA.

Filters used by perSVade

These are the filters used in the module "call_SVs," whose values may vary across parameter optimization in perSVade (note that most of the *gridss* filters were inspired by the filtering strategy used to generate the somatic call set from [21, 37] and the original *gridss* paper [37, 38])):

- min_Nfragments: Minimum number of reads supporting a breakend in *gridss* to be accepted (default is 5).
- min_af: Minimum variant allele frequency of a breakend in *gridss* to be accepted (default is 0.25).
- min_af_EitherSmallOrLargeEvent: Minimum variant allele frequency (VAF) of a breakend in *gridss* to be accepted, regardless of how VAF is calculated (default is 0.25). Note that VAF is calculated differently depending on if the breakend spans a region longer than the insert size or not (see https://github.com/PapenfussLab/gridss/issues/234#issuecomment-521489484). We regularly (i.e., for the min_af filter) calculate a VAF assuming that the breakend is a small event (vaf_small) or a large event (vaf_large). If the length of the breakpoint is above a certain threshold, related to the insert size ("median insert size + median absolute deviation of the insert size"), perSVade

sets VAF to be "vaf_large" and vice versa. Note that our distinction between small and large breakends could be error prone in some cases, and min_af_EitherSmallOrLargeEvent allows the filtering based on VAF independently of the size of the event. If min_af_EitherSmallOrLargeEvent is above 0, breakends that have both "vaf_small" and "vaf_large" below the set min_af_EitherSmallOrLargeEvent will be discarded.

- min_QUAL: Minimum quality (QUAL field of the vcf file) of a breakend in *gridss* to be accepted (default is 0).
- max_to_be_considered_small_event: Maximum length of a breakpoint in *gridss* to be considered a small event (default is 1000). Events shorter than this value are considered as "small events," which are treated particularly by other filtering steps.
- min_length_inversions: Minimum length of inversion-like breakends in *gridss* to be accepted (default is 40).
- maximum_lenght_inexactHomology: Maximum length of the inexact homology region around a breakend in *gridss* to be accepted (default is 50). This filter is not applied to "small events," as defined by "max_to_be_considered_small_event."
- maximum_microhomology: Maximum length of the exact homology (microhomology) region around a breakend in *gridss* to be accepted (default is 50).
- maximum_strand_bias: Maximum strand bias of a breakend in *gridss* to be accepted (default is 0.99). This filter is only applied to "small events," as defined by "max_to_be_considered_small_event."
- filter_noReadPairs: Discards *gridss* breakends without discordant read pair support (default is false). This filter is not applied to "small events," as defined by "max_to_be_considered_small_event."
- filter_noSplitReads: Discards *gridss* breakends without split-read evidence (default is false). This filter is only applied to "small events," as defined by "max_to_be_considered_small_event."
- filter_overlappingRepeats: Discards *gridss* breakends overlapping repetitive elements (default is false). This will only have an effect if you provide a repeats file as inferred by the module "infer_repeats."
- filter_polyGC: Discards *gridss* breakends with long inserted G or C sequences (>15bp) (default is true).
- wrong_FILTERtags: A set of values in the FILTER field of the *gridss* vcf which flag discarded breakends (default is ["NO_ASSEMBLY"]).
- range_filt_DEL_breakpoints: A range of lengths in which DEL-like breakends (as defined by *gridss*) are discarded if the breakend has a region with inexact homology above 5bp (default is [0, 1]). For example, if set to [500, 1000], DEL-like breakends whose length is between 500 and 1000bp with an inexact homology sequence >5 bp would be discarded.
- dif_between_insert_and_del: The margin given for comparing the length of the inserted sequence (len_seq) on a *gridss* DEL-like breakend and the length of the actual event (len_event) (default is 5). If len_seq > (len_event - dif_between_insert_and_del), the breakend is filtered out. This filter is only applied to "small events," as defined by "max_to_be_considered_small_event."
- max_rel_coverage_to_consider_del: The maximum relative coverage that a region spanning a DEL-like breakpoint (as defined by *clove*) can have to be classified as an actual deletion (default is 0.1). Note that the default is a conservative filter adapted to haploid genomes or homozygous variants.

– min_rel_coverage_to_consider_dup: The minimum relative coverage that a region spanning a TAN-like breakpoint (as defined by *clove*) can have to be classified as an actual tandem duplication (default is 1.8). Note that the default is a conservative filter adapted to haploid genomes or homozygous variants.

Note that all the breakpoints that have at least one breakend that does not pass the filters are discarded by perSVade.

Calling of small variants

PerSVade's small variant calling pipeline (module "call_small_variants") uses three alternative methods (GATK Haplotype Caller (HC) [67] (v4.1.2), freebayes (FB) [66] (v1.3.1), and / or bcftools (BT) [68] (v1.9)) to call and filter single-nucleotide polymorphisms (SNP) and small insertions/deletions (IN/DEL) in haploid or diploid configuration (specified with the *-p* option). The input is the bam file generated by the "align_reads" module. This module defines as high-confidence (PASS) variants those that are in positions with a read depth above the value provided with *--min_coverage*, with extra filters for HC and FB. For HC, it keeps as PASS variants those where (1) there are <4 additional variants within 20 bases; (2) the mapping quality is >40; (3) the confidence based on depth is >2; (4) the phred-scaled *p*-value is <60; (5) the MQRankSum is >−12.5, and (6) the ReadPosRankSum is >−8. For FB, perSVade "call_small_variants" keeps as PASS variants those where (1) quality is > 1 or alternate allele observation count is > 10, (2) strand balance probability of the alternate is > 0, (3) number of observations in the reverse strand is > 0, and (4) number of reads placed to the right/left of the allele are > 1. Then, bcftools (v1.10) and custom python code are used to normalize and merge the variants called by each software into a consensus variant set, which includes only those variants called with high-confidence by *N* or more algorithms This results in one .vcf file with the high-confidence variants for each *N*. Note that this .vcf file only keeps variants for which the fraction of reads covering the alternative allele is above the value provided with *--min_AF* (which may be 0.9 for haploids or 0.25 for diploids). For diploid calls, it defines the genotype with the strongest support (the one called by most programs). In addition, the quality of each variant is calculated from the mean of the three algorithms. Beyond the filtered variant calls, this module writes a tabular file with all the raw variants with various metadata columns (i.e., the programs that called the variant), which can be used to apply a custom filtering of the variants.

Comparing sets of SVs to calculate precision and recall

To measure accuracy in different sets of "called SVs" (in perSVade's simulations and also the testing of the pipeline (related to Figs. 2, 3, and 4 and Additional file 1: Figures S1, S3, S4, S5, S6, S7)), we compared them against the corresponding sets of "known SVs" and calculated the following estimates:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$F\text{-value} = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$$

where true positives (TP) are those in the "called SVs" that match at least one variant from the "known SVs," false positives (FP) are those in the "called SVs" that do not match any from the "known SVs," and false negatives (FN) are those in the "known SVs" that are not matched by any variant from the "called SVs." We define that two SVs are "matching" using a different approach for each type of SV:

- Inversions, tandem duplications, and deletions: both SVs are in the same chromosome, their altered regions are overlapping by 75% and their breakends are <50bp apart.
- Insertions: both SVs have the same origin and destination chromosomes and are both either cut-and-paste or copy-and-paste. In addition, the regions of the origin chromosome are overlapping by 75% and the breakends are <50bp apart. Finally, the starts of the destination chromosomes (insertion sites) in both SVs are <50bp apart.
- Translocations: both SVs have the same origin and destination chromosomes and are both either inverted or not. In addition, the breakpoint positions in both SVs are <50bp apart.

In addition, we calculated "integrated" precision and recall measures (related to Figs. 3 and 4 and Additional file 1: Figure S6) merging all the variants together into single sets of "called SVs" and "known SVs." We used custom python (v3.6) code and *bedmap* from the *bedops* tool [65] (v2.4.39) to calculate all these overlaps. See the section "PerSVade pipeline" above for further information on the meaning of each type of SV.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02737-4.

Additional file 1: Figure S1. Detailed workflow of the 'optimize_parameters' module. Figure S2. PerSVade's parameter optimization requires extra resources. Figure S3. PerSVade's parameter optimization improves the recall of SVs. Figure S4. Global vs SV type-specific parameter optimization. Figure S5. Each sample yields a different set of optimum parameters. Figure S6. PerSVade's parameters optimization mostly changes the recall of SVs in simulations. Figure S7. Coverage constrains SV calling accuracy in *C. glabrata* simulations. Table S1. Datasets used for the testing in simulations in *C. glabrata*, *C. albicans*, *C. neoformans*, *A. thaliana* and *D. melanogaster*.

Additional file 2. Review history.

Acknowledgements
The authors thank Cinta Pegueroles and Marina Lleal for the useful discussions key in the building of perSVade. In addition, we want to thank Hrant Hovhannisyan, Valentina del Olmo, Diego Fuentes, Anna Vlasova, Maria Artigues, Matteo Schiavinato, and Marina Marcet for beta-testing the pipeline and providing us with useful feedback.

Review history
The review history is available as Additional file 2.

Peer review information
Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team

Authors' contributions
MAST wrote the code and performed all bioinformatic analysis. MAST and TG conceived the study, interpreted the results, and wrote the manuscript. TG supervised the project and provided resources. Both authors read and approved the final manuscript.

Authors' information
Twitter handles: @MikiSchikora (Miquel Àngel Schikora-Tamarit); @Toni_Gabaldon (Toni Gabaldón)

Funding
TG group acknowledges support from the Spanish Ministry of Science and Innovation for grant PGC2018-099921-B-I00, cofounded by European Regional Development Fund (ERDF); from the Catalan Research Agency (AGAUR) SGR423; from the European Union's Horizon 2020 research and innovation program (ERC-2016-724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742) and from the Instituto de Salud Carlos III (INB Grant PT17/0009/0023 - ISCIII-SGEFI/ERDF). MAST received a Predoctoral Fellowship from "Caixa" Foundation (LCF/BQ/DR19/11740023).

Availability of data and materials
PerSVade is available at https://github.com/Gabaldonlab/perSVade [39] and can be installed using either conda environments or through a docker image containing the pipeline, available at https://hub.docker.com/r/mikischikora/persvade. The github repository is released under an open source GNU General Public License (GPL). In addition, the code can be accessed in Zenodo through the DOI 10.5281/zenodo.6866529 [86]. The github repository contains detailed examples on how to install and run perSVade using conda, docker, or singularity. We have tested perSVade on several Linux and Mac architectures, and the docker image may be run in any machine in a reproducible way. All the results shown in this paper were generated using the script https://github.com/Gabaldonlab/perSVade/blob/master/scripts/perSVade.py from version 1.0, which is a wrapper to execute several modules with a single command. Since perSVade is an actively used (and maintained) pipeline, we have created a few new releases since version 1.0, which include an improved documentation, more unit tests, and the implementation of an efficient debugging of inputs. Note that these changes do not affect the functionality of the modules as implemented in version 1.0. Hence, we recommend the usage of the latest version (version 1.02.7 at the time of publication), which is the one with the best documentation and usability. In addition, note that this one-liner wrapper is not recommended for broad usage. All the data used for testing perSVade was obtained from the SRA database or public ftp servers, and is listed in Additional file 1: Table S1 and "Methods." All the code necessary to reproduce the results and plots shown in this paper is in https://github.com/Gabaldonlab/perSVade/tree/master/testing.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 29 November 2021 Accepted: 22 July 2022
Published online: 16 August 2022

References

1. Baker M. Structural variation: the genome's hidden architecture. Nat Methods. 2012;9:133–7.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97.
3. Garsed DW, Marshall OJ, Corbin VDA, Hsu A, Di Stefano L, Schröder J, et al. The architecture and evolution of cancer neochromosomes. Cancer Cell. 2014;26:653–67.
4. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011;144. https://doi.org/10.1016/j.cell.2010.11.055.
5. Wang W-J, Li L-Y, Cui J-W. Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. Epigenetics Chromatin. 2020;13:1–17.
6. Ibáñez P, Lesage S, Janin S, Lohmann E, Durif F, Destée A, et al. Alpha-synuclein gene rearrangements in dominantly inherited parkinsonism: frequency, phenotype, and mechanisms. Arch Neurol. 2009;66. https://doi.org/10.1001/archneurol.2008.555.
7. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Autism Consortium, Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. 2008;358:667–75.
8. Todd RT, Selmecki A. Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. Elife. 2020;9. https://doi.org/10.7554/eLife.58349.
9. Ksiezopolska E, Schikora-Tamarit MÀ, Beyer R, Nunez-Rodriguez JC, Schüller C, Gabaldón T. Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen Candida glabrata. Curr Biol. 2021. https://doi.org/10.1016/j.cub.2021.09.084.
10. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
11. Dennis MY, Eichler EE. Human adaptation and evolution by segmental duplication. Curr Opin Genet Dev. 2016;41:44–52.

12. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61. https://doi.org/10.1146/annurev-med-100708-204735.
13. Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, et al. Discovery and population genomics of structural variation in a songbird genus. Nat Commun. 2020;11:1–11.
14. Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, et al. Structural variants in 3000 rice genomes. Genome Res. 2019;29:870–80.
15. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:1–14.
16. Bartenhagen C, Dugas M. Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. Brief Bioinform. 2016;17:51–62.
17. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinformatics. 2014;45:15.6.1–11.
18. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics. 2010;26:1895–6.
19. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
20. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
21. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. 2021;22:1–25.
22. Valls-Margarit J, Galván-Femenía I, Matías-Sánchez D, Blay N, Puiggròs M, Carreras A, et al. GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. bioRxiv. 2021:2021.07.20.453041.
23. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20:117.
24. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10:1–11.
25. Schröder J, Wirawan A, Schmidt B, Papenfuss AT. CLOVE: classification of genomic fusions into structural variation events. BMC Bioinformatics. 2017;18:346.
26. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. 1000 Genomes Project, mapping copy number variation by population-scale genome sequencing. Nature. 2011;470:59–65.
27. Pang AW, Migita O, Macdonald JR, Feuk L, Scherer SW. Mechanisms of formation of structural variation in a fully sequenced human genome. Hum Mutat. 2013;34. https://doi.org/10.1002/humu.22240.
28. Todd SLS, Treangen J. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13:36-46.
29. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28. https://doi.org/10.1093/bioinformatics/bts378.
30. Liu Y, Huang Y, Wang G, Wang Y. A deep learning approach for filtering structural variants in short read sequencing data. Brief Bioinform. 2020;22. https://doi.org/10.1093/bib/bbaa370.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
32. Babraham Bioinformatics - FastQC A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc.
33. Manual Reference Pages for bwa. http://bio-bwa.sourceforge.net/bwa.shtml.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078.
35. GATK MarkDuplicatesSpark. https://gatk.broadinstitute.org/hc/en-us/articles/360036358972-MarkDuplicatesSpark.
36. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34:867–8.
37. Cameron DL, Baber J, Shale C, Papenfuss AT, Valle-Inclan JE, Besselink N, et al. GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. 2019:781013.
38. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27:2050-60.
39. Schikora-Tamarit MÀ, Gabaldón T. perSVade. Github. 2022. https://github.com/Gabaldonlab/perSVade.
40. Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics. 2013;29:1679–81.
41. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
42. Wgsim. https://github.com/lh3/wgsim.
43. Seqtk. https://docs.csc.fi/apps/seqtk/.
44. Seaborn 0.11.2 documentation. https://seaborn.pydata.org/.
45. Matplotlib: visualization with Python. https://matplotlib.org/.
46. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.
47. API design for machine learning software: experiences from the scikit-learn project. https://arxiv.org/abs/1309.0238.
48. Psutil: Cross-platform lib for process and system monitoring in Python. https://github.com/giampaolo/psutil.
49. Pandas. https://pandas.pydata.org/.
50. Pigz: Parallel gzip. https://zlib.net/pigz/.
51. Gztool. https://github.com/circulosmeos/gztool.

52. Picard. Available at http://broadinstitute.github.io/picard/.
53. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:1–9.
54. Newman S, Hermetz KE, Weckselblatt B, Katharine Rudd M. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. Am J Hum Genet. 2015;96:208.
55. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.
56. Abbey DA, Funt J, Lurie-Weinberger MN, Thompson DA, Regev A, Myers CL, et al. YMAP: a pipeline for visualization of copy number variation and loss of heterozygosity in eukaryotic pathogens. Genome Med. 2014;6:1–16.
57. IMH, Quinlan AR. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841.
58. Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: ultra-fast computation of genome mappability. Bioinformatics. 2020;36:3687–92.
59. Cylowess. https://github.com/livingsocial/cylowess.
60. Cython: The Best of Both Worlds. https://ieeexplore.ieee.org/document/5582062.
61. Wei Y-C, Huang G-H. CONY: A Bayesian procedure for detecting copy number variations from sequencing read depths. Sci Rep. 2020;10:1–14.
62. Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DCJ, de Jong TV, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 2016;17:1–15.
63. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, et al. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics. 2006;22. https://doi.org/10.1093/bioinformatics/btl238.
64. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
65. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919–20.
66. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. 2012. (Available at http://arxiv.org/abs/1207.3907).
67. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
68. The official development repository for BCFtools. https://github.com/samtools/bcftools.
69. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
70. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2004;Chapter 4. https://doi.org/10.1002/0471250953.bi0410s05.
71. Entrez Direct: E-utilities on the Unix Command Line. https://www.ncbi.nlm.nih.gov/books/NBK179288/.
72. SRA Tools. https://github.com/ncbi/sra-tools.
73. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635.
74. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;39:D19.
75. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. a comprehensive update on curation, resources and tools. Database. 2020;2020. https://doi.org/10.1093/database/baaa062.
76. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 2017;27. https://doi.org/10.1101/gr.210500.116.
77. Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. BMC Genomics. 2016;17:1–16.
78. HG002 sequencing data. ftp://ftptrace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.60X.1.bam.
79. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27. https://doi.org/10.1101/gr.214007.116.
80. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45:D592.
81. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;48:D84–6.
82. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
83. NA12878 deletions. ftp://ftptrace.ncbi.nlm.nih.gov/giab/ftp/technical/svclassify_Manuscript/Supplementary_Information/Personalis_1000_Genomes_deduplicated_deletions.bed.
84. HG002 structural variants. ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz.
85. CHM1 and CHM13 structural variants. http://eichlerlab.gs.washington.edu/publications/Huddleston2016/structural_variants/.
86. Schikora-Tamarit MÀ, Gabaldón T. perSVade v1.02.7. Zenodo. 2022. https://zenodo.org/record/6866529.

Publisher's Note

ADDITIONAL FIGURES

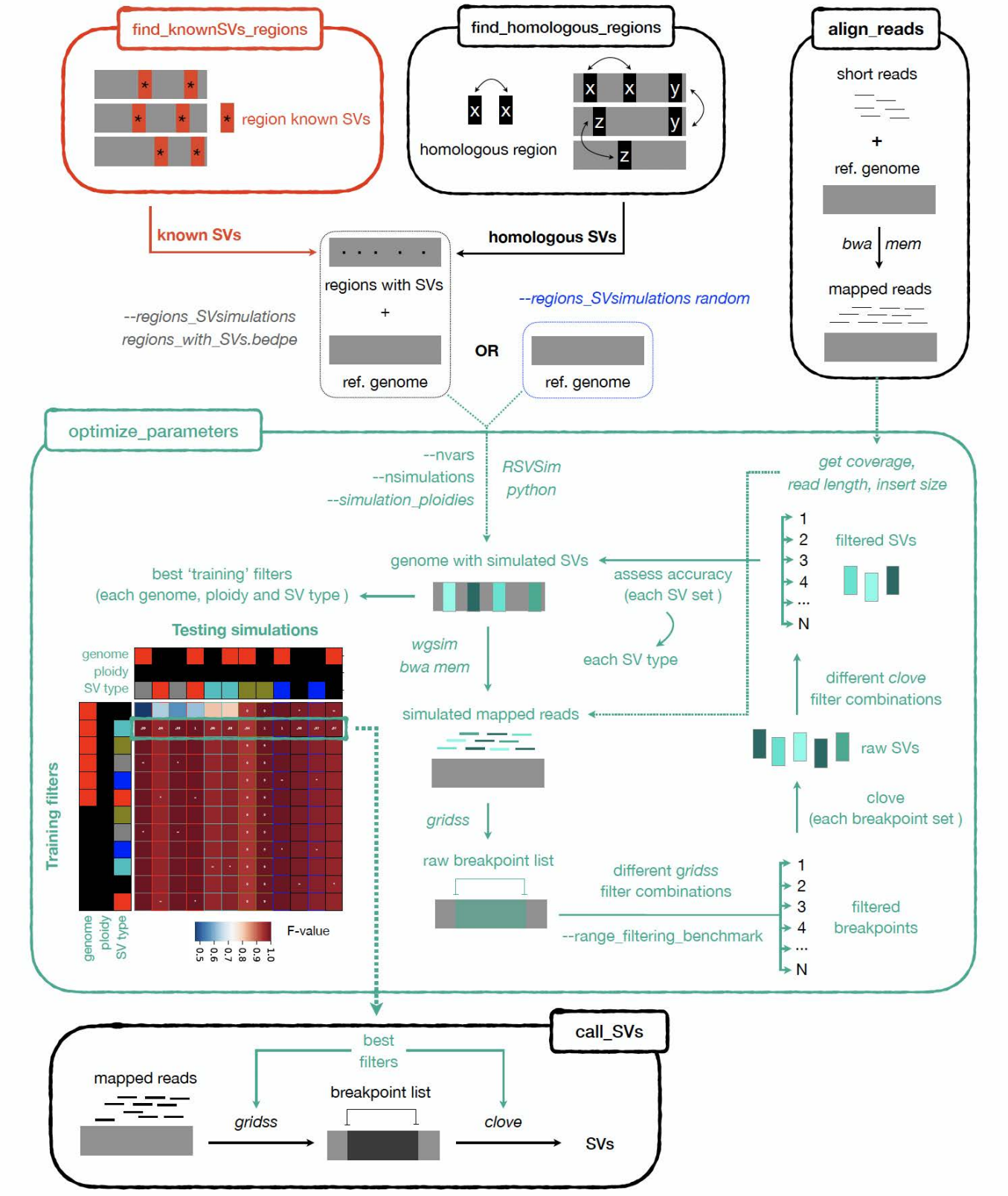


Figure S1. Detailed workflow of the 'optimize_parameters' module. The module 'optimize_parameters' is the core, most novel function of perSVade. It requires the argument --regions_SVsimulations, which specifies the regions of the genome for simulations of SVs. These can be either around some specific

regions (with the argument --regions_SVsimulations <regions>.bedpe) or randomly placed across the genome (with the argument --regions_SVsimulations random). Note that perSVade has modules to infer either regions with previously known SVs (through 'find_knownSVs_regions') or regions with pairwise homology (through 'find_homologous_regions'). The SV simulations around such regions may be more realistic than random simulations, which is why they may be considered. The module 'optimize_parameters' finds a set of optimum parameters through simulations around these regions. By default, it generates two simulated genomes (tunable with --nsimulations) with 50 SVs of each type (tunable through --nvars) based on the reference genome and the provided regions. There are two simulated genomes for each of the desired ploidies/zygosities, tunable through --simulation_plodies. For example, we set '--simulation_plodies haploid' for haploid organisms and '--simulation_plodies diploid_hetero' for diploids (which means that the simulated genomes will have only heterozygous variants) in the testing of perSVade on several organisms (see **Methods**). For each simulated genome perSVade 'optimize_parameters' simulates reads with equal insert size, coverage and read length as the input mapped reads (provided with the argument -sbam). Then it aligns the reads and runs gridss to obtain a list of 'raw breakpoints'. This module then tries several combinations of filters on them (by default >278,000,000, which is tunable through --range_filtering_benchmark) to generate many 'filtered breakpoints'. Each of these is processed with clove to generate a set of 'raw SVs'. PerSVade 'optimize_parameters' next tries several combinations of filters on each of them to get a set of filtered SVs. These are compared against the true set of SVs (inserted in the simulated genome) to calculate the accuracy (F-value) of each combination of *gridss* and clove filters on each simulated genome, ploidy and SV type. These filters are optimized for each simulation, and thus may not be accurate on independent sets of SVs (due to overfitting). In order to reduce this effect, this module tests how each of these 'best' filters perform on all simulations, ploidies and SV types (not only in those that yielded the given filters as optimum). The heatmap shows the F-value for an example sample (BG2 based on random simulations from *Candida glabrata,* see **Methods**), where the filters in the second row are accurate on all simulations (indicating that there is no overfitting on them) and thus they are chosen as the final set of best parameters. Note that the filters in the first row are only accurate on some simulations, suggesting that they are overfitted and thus they are not chosen as good filters. At the end, this module writes the accuracy of these best parameters into a .tab file, which will allow the user to understand how much the results are to be trusted. In addition, these optimized filters (or parameters) are written into a .json file that may be used for calling SVs with 'call_SVs'.

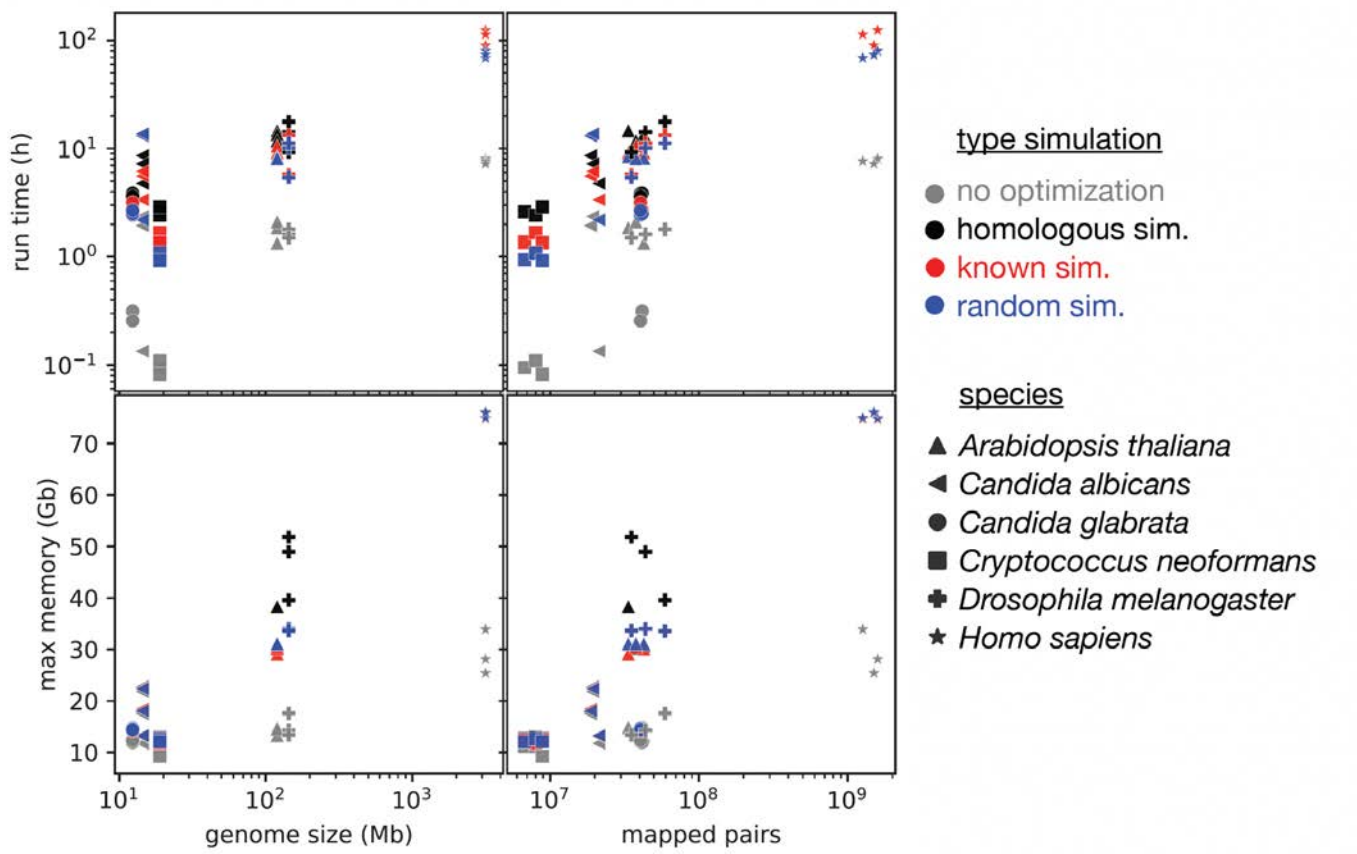


Figure S2. PerSVade's parameter optimization requires extra resources. We tested perSVade's SV calling modules ('optimize_parameters', 'call_SVs' and 'integrate_SV_CNV_calls') on six eukaryotes (three samples per species) using either no parameter optimization (gray) or different types of simulations (black, red, blue) for the 'optimize_parameters' module in a machine with 16 cores. Shown are the running time and maximum RAM used ignoring the resources related to read alignment (which was performed independently). Thus, each point reflects the resources used by 'optimize_parameters' (except in the gray points), 'call_SVs' and 'integrate_SV_CNV_calls'. Of note, perSVade was run with a different setting for the human datasets to avoid excessive resource consumption. First, we skipped the marking of duplicate reads on the .bam files. Second, we ran the simulations on a subset of the genome (only chromosomes 2, 7, 9, X, Y and mitochondrial). Third, we skipped the 'homologous' simulations in human samples due to excessive memory consumption. The x axes reflect the reference genome size (left) and the number of mapped read pairs (right), which are correlated with resource consumption. This data may be useful to allocate computational resources for running perSVade.

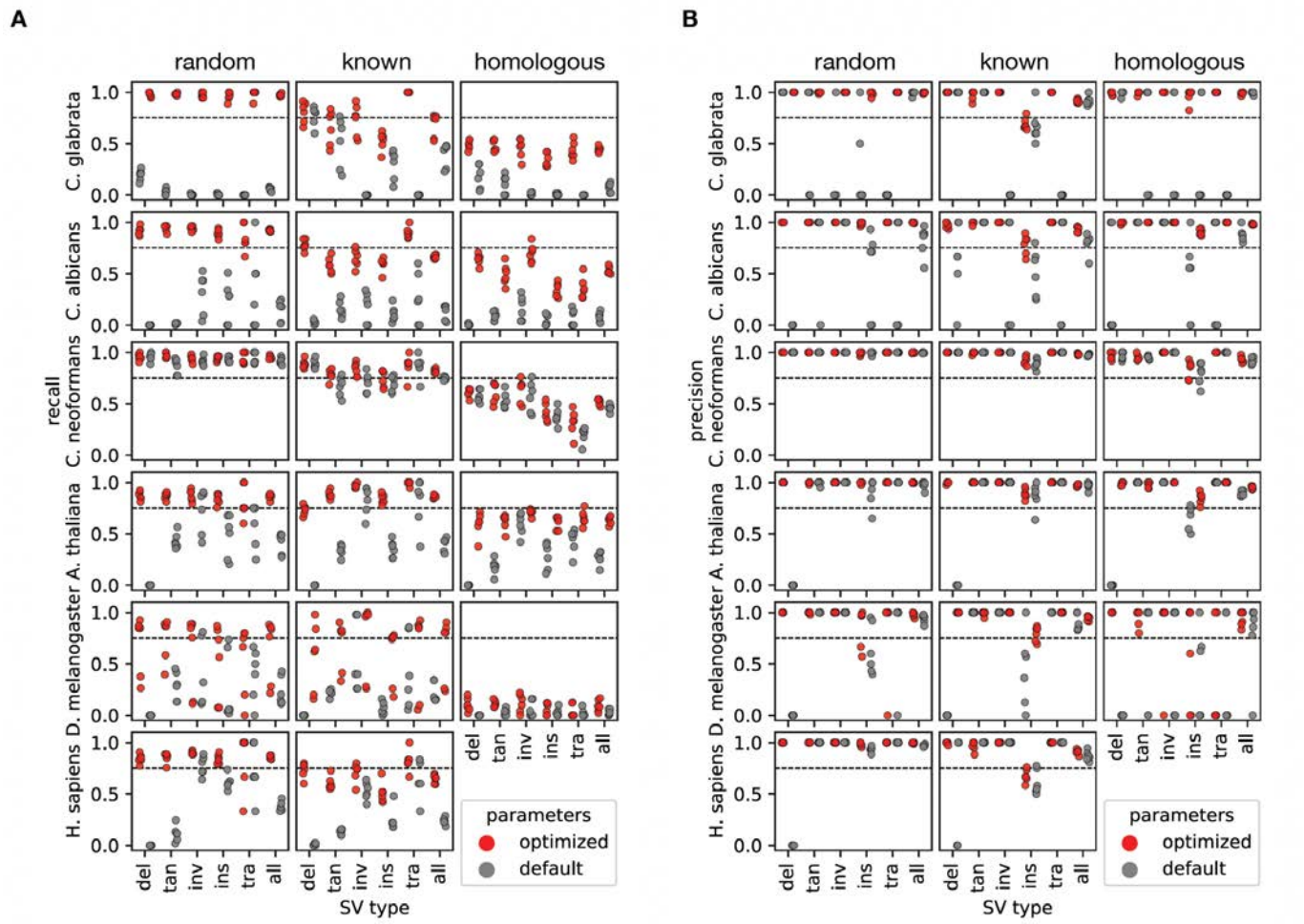


Figure S3. PerSVade's parameter optimization improves the recall of SVs. We ran perSVade's SV calling modules on three samples per species for six eukaryotes (see **Methods**) using either 'random', 'known' or 'homologous' simulations. These plots show the recall (left) and precision (right) of either default or optimized parameters (for each sample and simulation type) on these simulations. The x axis represents the type of SV (deletions (del), tandem duplications (tan), inversions (inv), insertions (ins), translocations (tra) and the average of all SVs (all)).

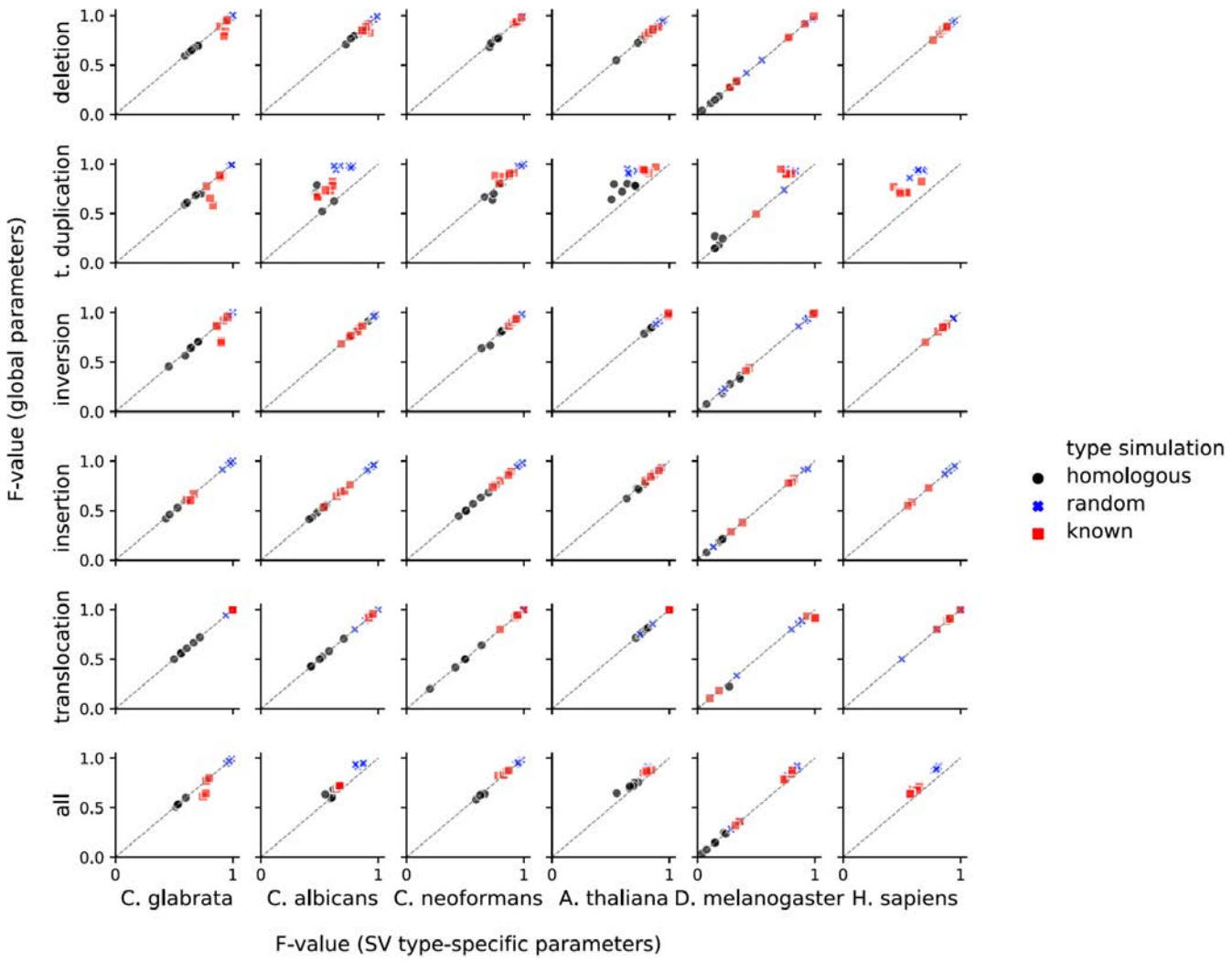


Figure S4. Global vs SV type-specific parameter optimization. To understand whether using separate optimal parameters for each SV type is necessary to achieve maximal accuracy we analyzed the intermediate files from the parameter optimization on different species and simulation types (see **Methods**, **Figure 2,3**). Note that perSVade's parameter_optimization module returns a set of parameters that are optimized for all (not only one) simulated genomes and SV types (global parameters). Understanding whether these global parameters are as accurate as those specifically obtained for each simulated genome and SV type is relevant to validate that a single set of parameters can be used for calling all SV types. To find these global parameters, the module first calculates the best parameters for each simulated genome and SV type (SV type-specific parameters) and then tests the accuracy of these parameters on each simulated genome and SV type (see **Additional file 1: Figure S1**). We used these accuracy measurements to understand whether using SV-type specific parameters is more accurate than using global parameters. Each point in these scatter plots represents the SV calling accuracy for a simulated genome considering a given SV type (rows), species (columns) and type of simulations (colors). The x axis represents the calling accuracy (F-value) when using SV type-specific parameters, while the y axis represents the accuracy when using global parameters. Note that we only considered the cases where the global parameters were obtained from a different simulated genome / SV type than the SV-type specific parameters. In addition, note that global parameters perform slightly better than the SV type-specific ones for tandem duplications in some samples of diploid species. This is due to the behavior of the coverage threshold that defines true tandem duplications

(*min_rel_coverage_to_consider_dup* parameter, see **Methods**), where the rationally-designed value set in global parameters (from non-tandem duplication SV types) is more accurate than any of the values tried in the SV type-specific optimization.

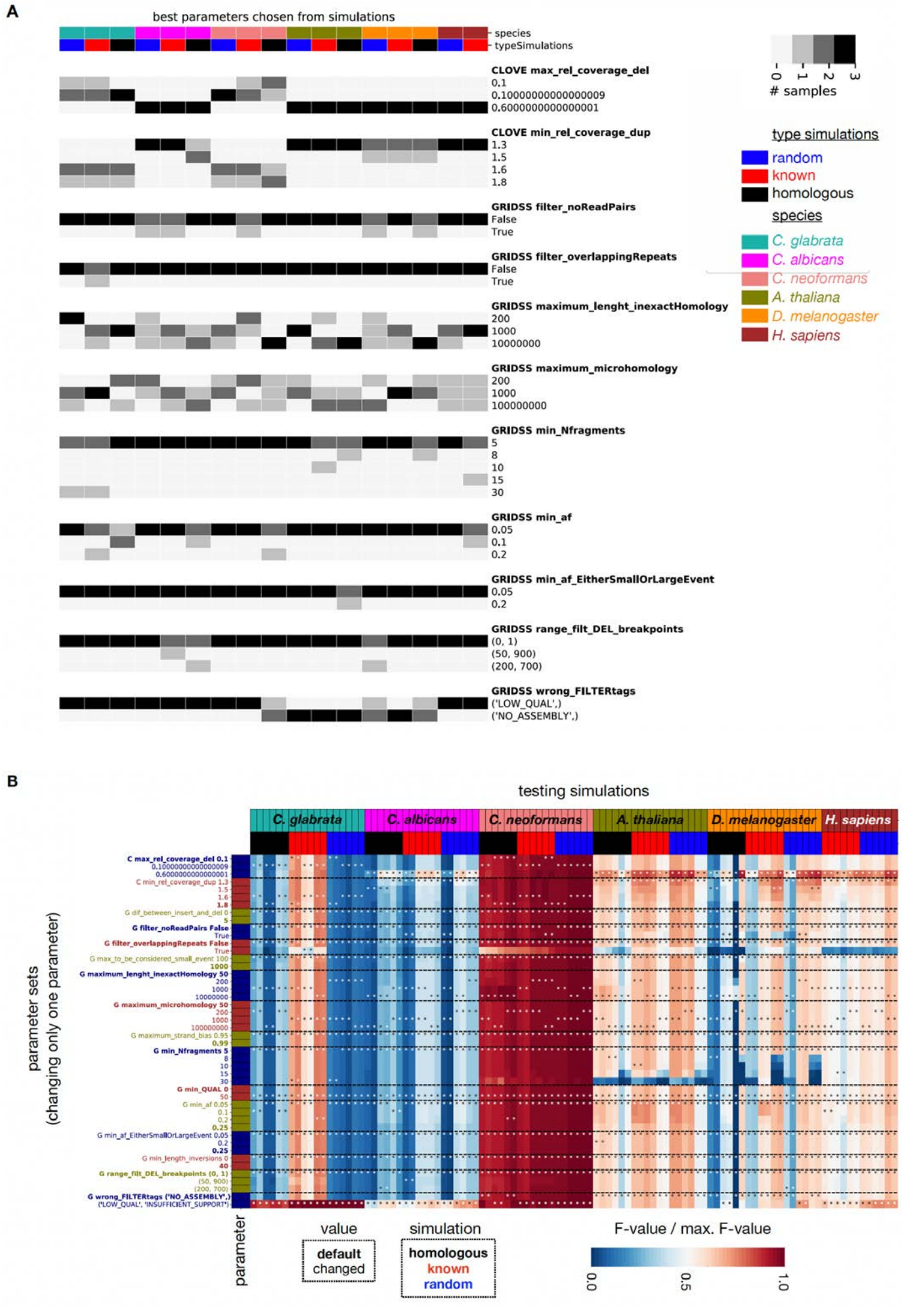

A
best parameters chosen from simulations
species
typeSimulations
samples
type simulations
random
known
homologous
species
C. glabrata
C. albicans
C. neoformans
A. thaliana
D. melanogaster
H. sapiens
CLOVE max_rel_coverage_del
0.1
0.10000000000000009
0.6000000000000001
CLOVE min_rel_coverage_dup
1.3
1.5
1.6
1.8
GRIDSS filter_noReadPairs
False
True
GRIDSS filter_overlappingRepeats
False
True
GRIDSS maximum_lenght_inexactHomology
200
1000
10000000
GRIDSS maximum_microhomology
200
1000
100000000
GRIDSS min_Nfragments
5
8
10
15
30
GRIDSS min_af
0.05
0.1
0.2
GRIDSS min_af_EitherSmallOrLargeEvent
0.05
0.2
GRIDSS range_filt_DEL_breakpoints
(0, 1)
(50, 900)
(200, 700)
GRIDSS wrong_FILTERtags
('LOW_QUAL',)
('NO_ASSEMBLY',)
B
testing simulations
C. glabrata
C. albicans
C. neoformans
A. thaliana
D. melanogaster
H. sapiens
parameter sets
(changing only one parameter)
parameter
value
default
changed
simulation
homologous
known
random
F-value / max. F-value
0.0
0.5
1.0

Figure S5. Each sample yields a different set of optimum parameters. (A) We ran perSVade's 'optimize_parameters' module on six eukaryotes (three samples per species) and a parameter optimization based on either 'random', 'homologous' or 'known' simulations (see **Methods**). Shown are the chosen values for each parameter (only those that changed across samples) in each optimization procedure. Each group of rows refers to the values chosen for one type of parameter (see **Methods**) used for *gridss* or *clove*. The color indicates how many samples (from zero to three) yielded a specific value for each parameter type. For example, the threshold to discard breakpoints (called by *gridss*) based on allele frequency was set to 0.05 in most samples (see "GRIDSS min_af"). However, the optimization for 'known' SVs in *C. glabrata* yielded a threshold of 0.2 in one sample (out of three) (see the second column). **(B)** To understand the effect of changing each parameter, we measured the SV calling accuracy of different parameter sets, each with only one tuned parameter while all other parameters were kept to default values. We tested all the values obtained in an optimal set (those from A) and the default values. This explains why there are some extra parameters in this panel (*dif_between_insert_and_del*, *max_to_be_considered_small_event, min_QUAL, min_length_inversions* and *min_af_EitherSmallOrLargeEvent*). Each row corresponds to a different parameter set. Each column represents a simulation from a given sample / simulation type to be "tested". The heatmap shows the F-value of each parameter set on each tested simulation, relative to the maximum F-value derived from combinatorial parameter optimization (see **Figure 3A**). Each cell is hereafter referred to as 'testing instance'. The asterisks refer to instances where the changing parameter value is the one picked in the optimization of each sample / simulation type. In addition, the row colors indicate which parameter is tuned in each set. The column color indicates the parameter that is changing (each with a different color), and the fontweight of the label indicates whether the parameter set has default values (labels in bold) or has some parameter value changed. This means that all the rows with default values (bold labels) correspond to the same parameter set. We repeat them for each parameter to aid visual comparisons.

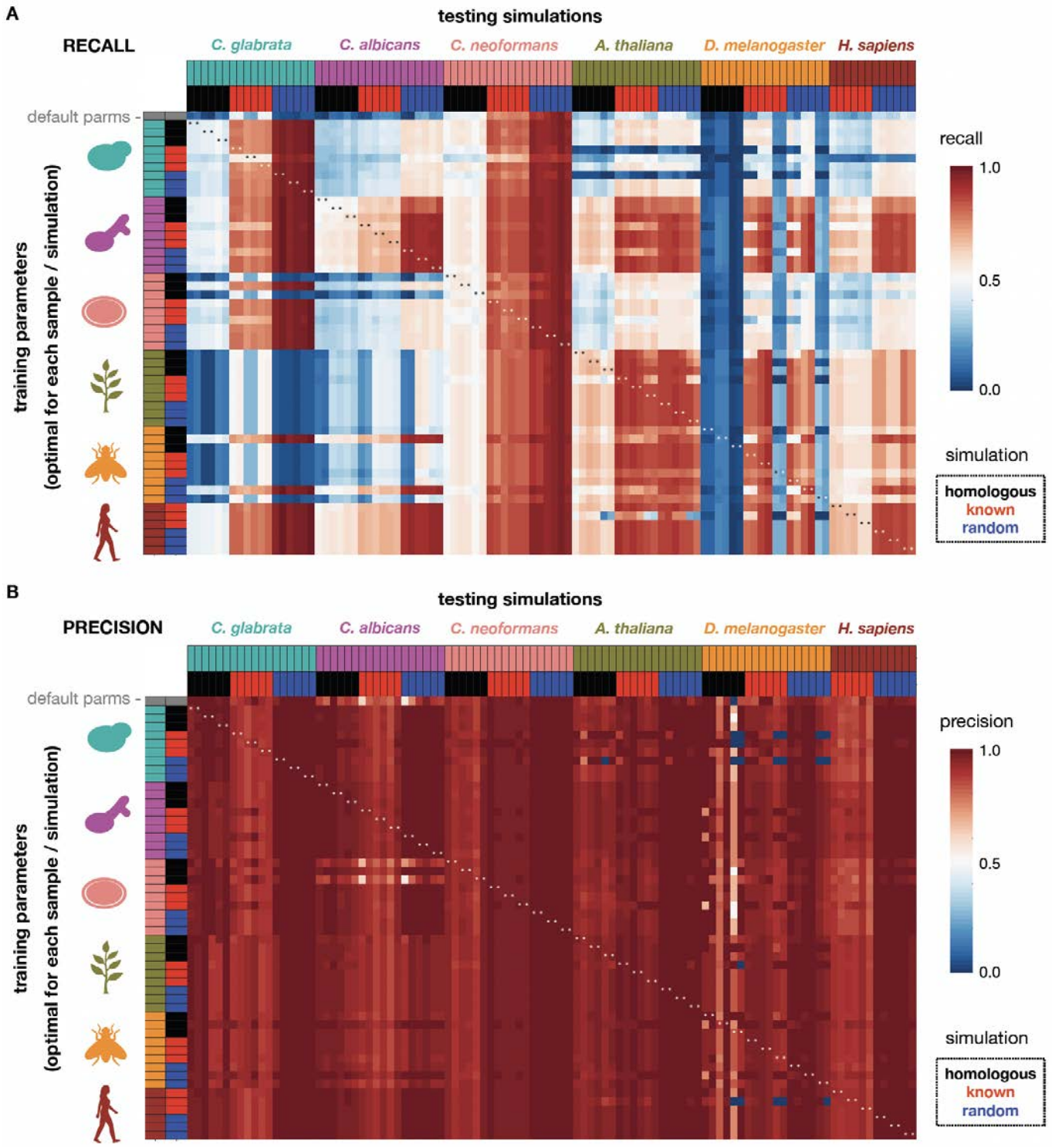


Figure S6. PerSVade's parameters optimization mostly changes the recall of SVs in simulations. To assess whether perSVade's parameter optimization is necessary for all samples / simulations (mentioned in **Figure 2 and Additional file 1: Figure S3**) we measured the SV calling accuracy of each parameter set on the other samples / simulations. Each row indicates a different "training" parameter set optimized for each sample and simulation type in all tested species. In addition, the first row refers to the default parameters. Each column represents a simulation from a given sample / simulation type to be tested. The heatmap shows either the recall (A) or the precision (B) of each parameter set on each tested simulation. Note that the species are ordered alike in rows and columns. In addition, note that each sample (from a given species and simulation type) yielded one set of "training" parameters and two simulated genomes tested here, which explains why there are two columns for each row. The asterisks

refer to testing instances where both the sample and type of simulation are equal in the training and testing (equivalent to the 'optimized' parameters from **Figure 2 and Additional file 1: Figure S3**).

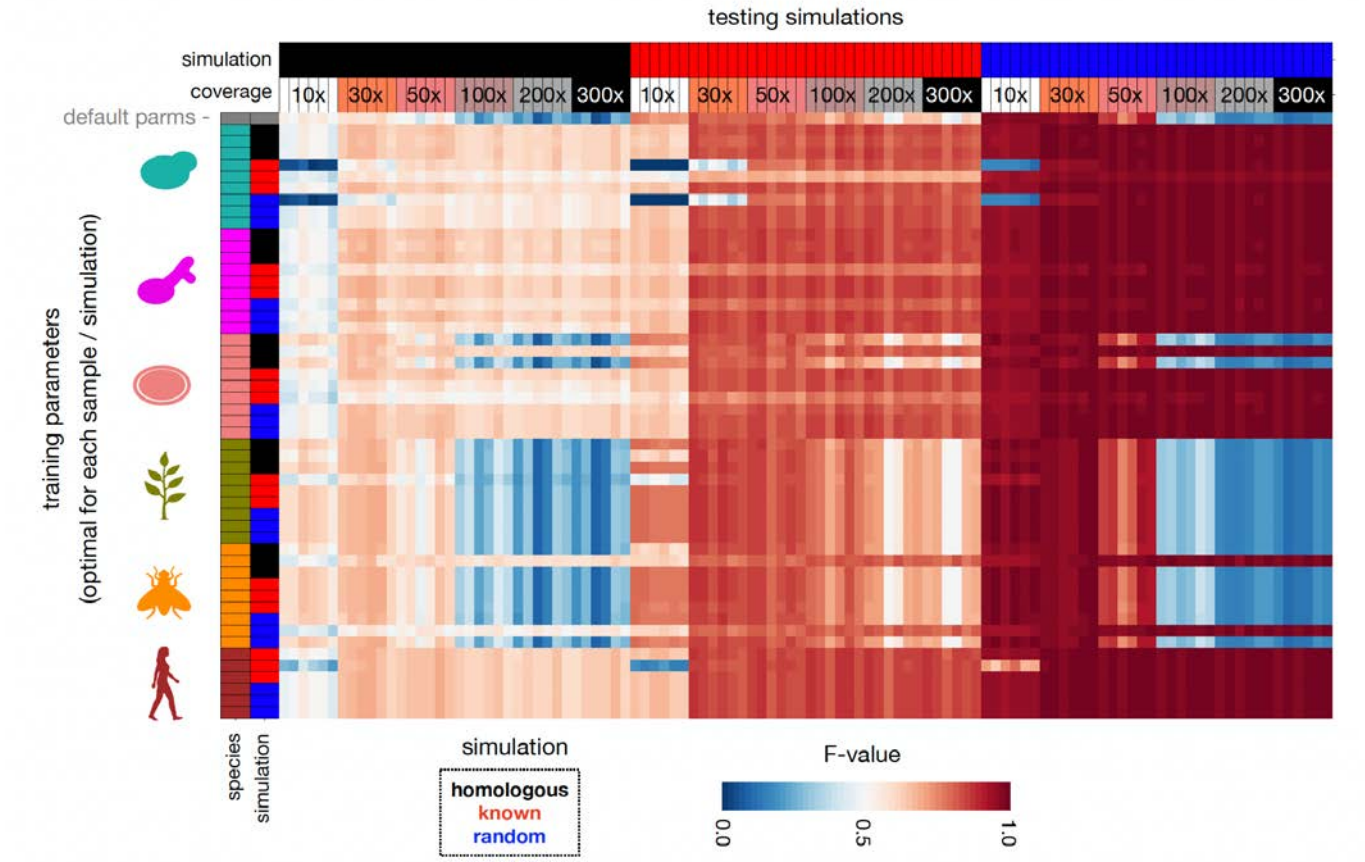


Figure S7. Coverage constrains SV calling accuracy in *C. glabrata* simulations. To assess whether the high coverage of *C. glabrata* samples (>300x, see **Table S1**) constrained SV calling, we measured the accuracy of each parameter set (optimized for each species / simulations (see **Figure 3A**)) on the *C. glabrata* simulations with varying coverage. For each simulation (based on a sample and a type of simulation (homologous / known / uniform)), we subsampled randomly the reads to get a coverage of 10x, 30x, 50x, 100x, 200x or 300x. . Each row indicates a different "training" parameter set optimized for each sample and simulation type in all tested species (the same parameters as in **Figure 3A**). In addition, the first row refers to the default parameters. Each column represents a simulation from a given sample / simulation type / coverage to be tested. The heatmap shows the F-value of each parameter set on each tested simulation. Note that the species symbols correspond to *C. glabrata*, *C. albicans*, *C. neoformans*, *A. thaliana, D. melanogaster* and *H. sapiens*, respectively. Note that the cells related to 300x coverage are similar to the original simulations (**Figure 3A**).

ADDITIONAL TABLES

target_species	target_taxID	sample_taxID	sample_species	SRA_run	% reads map.	cov.
C. glabrata	N/A	N/A	*C. glabrata BG2*	SRR15498429	N/A	319
C. glabrata	N/A	N/A	*C. glabrata* CST34	SRR15498440	N/A	342
C. glabrata	N/A	N/A	*C. glabrata M12*	SRR15498481	N/A	337
C. albicans	5476	1182531	*C. albicans 3153*	SRR641729	94.58	128
C. albicans	5476	1182537	*C. albicans A123*	SRR538772	88.77	89
C. albicans	5476	1182540	*C. albicans A203*	SRR538786	95.91	92
C. neoformans	5207	1423894	*C. neoformans Bt35*	SRR1063293	99.93	30
C. neoformans	5207	1423915	*C. neoformans RSA-MW-1281*	SRR1063017	99.86	37
C. neoformans	5207	1423916	*C. neoformans RSA-MW-5465*	SRR1063214	99.94	40
A. thaliana	3702	38785	*A. arenosa*	SRR4128971	76.76	22
A. thaliana	3702	378006	*A. arenosa x A. thaliana*	ERR5032500	89.59	25
A. thaliana	3702	2608267	*A. arenosa x A. lyrata*	ERR3514861	65.98	21
D. melanogaster	7227	7238	*D. sechellia*	SRR5860659	89.70	37
D. melanogaster	7227	7240	*D. simulans*	ERR1597900	84.44	45
D. melanogaster	7227	7243	*D. teissieri*	SRR13202235	86.61	6

Table S1. Datasets used for the testing in simulations in C. glabrata, *C. albicans, C. neoformans, A. thaliana* and *D. melanogaster*. We chose these datasets automatically from the SRA database for *C. albicans, C. neoformans, A. thaliana* and *D. melanogaster*. In order to have enough SV calls we selected mildly divergent samples (as compared to the reference genome) with a NCBI taxonomy taxon ID (indicated by each sample_taxID) different from the ID of species of interest (target_taxID). However, we only kept samples with most reads mapped (specified in the column '% reads map.') in order to discard datasets from highly divergent taxa. Note that it was not possible to find such samples for *C. glabrata* at the time of this study. We thus used three datasets for *C. glabrata* strains from our lab (see **Methods**). 'N/A' indicates that the column (i.e. taxID or % of mapped reads) was not taken into consideration for selecting these samples. See **Methods** for more information. Note that the 'cov.' column indicates the read depth of each of these samples.

3.3

Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen *Candida glabrata*

Article

Current Biology

Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen *Candida glabrata*

Highlights

- The ability for fast acquisition of drug resistance is widespread in *Candida glabrata*
- Resistance-conferring mutations are very diverse but affect a small number of genes
- Cross-resistance to fluconazole is common in strains adapted to anidulafungin
- *ERG3* mutations often drive fluconazole resistance in anidulafungin-adapted strains

Authors

Ewa Ksiezopolska,
Miquel Àngel Schikora-Tamarit,
Reinhard Beyer,
Juan Carlos Nunez-Rodriguez,
Christoph Schüller, Toni Gabaldón

Correspondence

toni.gabaldon@bsc.es

In brief

Ksiezopolska et al. trace mutational paths leading to drug resistance in the fungal pathogen *Candida glabrata* and uncover new resistance-related genes. Importantly, they find that mutations in the *ERG3* gene underpin the common appearance of cross-resistance to fluconazole in strains adapted only to anidulafungin.

Ksiezopolska et al., 2021, Current Biology *31*, 5314–5326
December 6, 2021
https://doi.org/10.1016/j.cub.2021.09.084

CellPress

CellPress
OPEN ACCESS

Current Biology

Article

Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen *Candida glabrata*

Ewa Ksiezopolska,[1,2,5] Miquel Àngel Schikora-Tamarit,[1,2,5] Reinhard Beyer,[3] Juan Carlos Nunez-Rodriguez,[1,2] Christoph Schüller,[3] and Toni Gabaldón[1,2,4,6,7,*]

[1]Barcelona Supercomputing Centre (BSC-CNS), Life Sciences Department, Jordi Girona 29, 08034 Barcelona, Spain
[2]Institute for Research in Biomedicine (IRB Barcelona), Mechanisms of Disease Program, The Barcelona Institute of Science and Technology, Baldiri Reixac 10, 08028 Barcelona, Spain
[3]Institute of Microbial Genetics and Core Facility Bioactive Substances: Screening and Analysis, University of Natural Resources and Life Sciences, Vienna (BOKU), Konrad Lorenz Strasse 24, 3430 Tulln an der Donau, Austria
[4]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain
[5]These authors contributed equally
[6]Twitter: @gabaldonlab
[7]Lead contact
*Correspondence: toni.gabaldon@bsc.es
https://doi.org/10.1016/j.cub.2021.09.084

SUMMARY

Fungal infections are a growing medical concern, in part due to increased resistance to one or multiple antifungal drugs. However, the evolutionary processes underpinning the acquisition of antifungal drug resistance are poorly understood. Here, we used experimental microevolution to study the adaptation of the yeast pathogen *Candida glabrata* to fluconazole and anidulafungin, two widely used antifungal drugs with different modes of action. Our results show widespread ability of rapid adaptation to one or both drugs. Resistance, including multidrug resistance, is often acquired at moderate fitness costs and mediated by mutations in a limited set of genes that are recurrently and specifically mutated in strains adapted to each of the drugs. Importantly, we uncover a dual role of *ERG3* mutations in resistance to anidulafungin and cross-resistance to fluconazole in a subset of anidulafungin-adapted strains. Our results shed light on the mutational paths leading to resistance and cross-resistance to antifungal drugs.

INTRODUCTION

Each year, fungal infections affect >1 billion people worldwide and cause 1.5 million deaths.[1] Current challenges to overcome this trend include the lack of fast and accurate diagnoses and the rise of antifungal drug resistance.[2] Acquisition of antifungal resistance is particularly worrying, given the limited number of available compounds. The widespread use of antifungal agents to counteract the high clinical, agricultural, and economic burden caused by various fungal pathogens, coupled with the high ability of fungi to adapt to selective pressures, have resulted in an alarming increase in the rates at which fungal species or isolates resistant to one or multiple drugs are identified.[3,4] As a result, we are witnessing a global epidemiological change represented by the increased incidence of previously uncommon species with a greater ability to adapt to drugs, the increased failure of therapies due to adaptation of the infecting clone, and the common appearance and rapid spread of deadly outbreaks caused by resistant lineages. These trends affect all major human fungal pathogenic genera, including *Candida*, *Aspergillus*, *Cryptococcus*, and *Pneumocystis*. Despite the pressing challenge that the emergence of antifungal resistance represents for human health and food security, we have a limited understanding of the evolutionary processes leading to drug adaptation in fungi.[5] Although we know common resistance-conferring mutations and major resistance mechanisms operating in many fungal pathogens, these represent the culmination of an adaptation process. This evolutionary process remains understudied because most of our knowledge derives from already-adapted clones, and from the exploration of a usually limited set of known target genes. In this regard, the use of an *in vitro* evolution approach coupled to whole-genome sequencing represents a promising research avenue.[5]

Candida species are among the main causes of hospital-acquired fungal infections.[1] *C. albicans* is the most common cause of candidiasis, but the relative incidence of non-albicans *Candida* species is on the increase,[6] with *C. glabrata* often being the second most prevalent cause of infection.[6] *C. glabrata* belongs to the Nakaseomyces clade and is phylogenetically closer to *Saccharomyces cerevisiae* than to most other *Candida* pathogens,[7] which may imply different routes for drug adaptation as compared to other *Candida* species. Antifungal resistance in *C. glabrata* is particularly problematic, as this yeast shows a remarkable ability to adapt to both azoles and echinocandins, thus leading to multidrug resistance (MDR).[8–11] Most antifungals commonly used against *Candida* are azoles (e.g., fluconazole [flz]), fungistatic drugs that inhibit a lanosterol demethylase

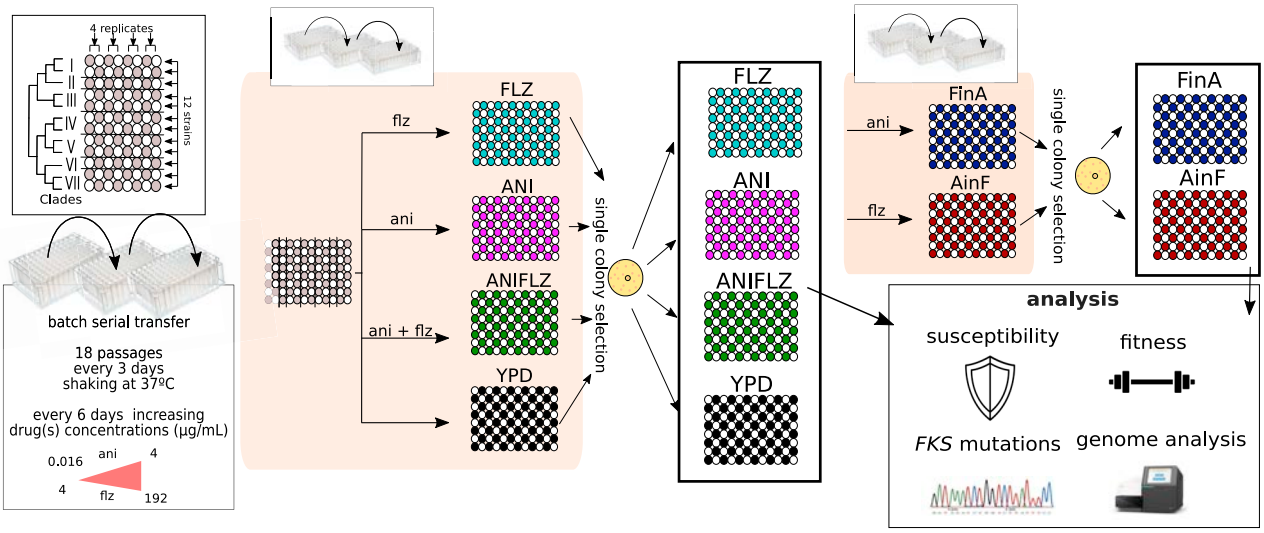


Figure 1. Schematic representation of the *in vitro* evolution experiment

A total of 48 populations, quadruplicates of each of the 12 strains, were grown with increasing concentrations of flz (FLZ samples), ani (ANI), both drugs in combination (ANIFLZ), and no drug (YPD). Subsequently, ANI samples were grown in flz (AinF), whereas FLZ samples were grown in ani (FinA). The experiment involved batch serial transfer of the samples every 3 days, in which every second passage involved an increase in drug concentrations up to 4 and 196 μg/mL ani and flz, respectively (Table S4; STAR Methods). After the final passage, an aliquot was plated for single colony isolation and storage.

encoded by *ERG11*[12], and echinocandins (e.g., anidulafungin [ani]), which inhibit 1,3-β-D-glucan synthase encoded by *FKS* genes[13] and are fungicidal to *Candida*. Most prevalent mechanisms conferring protection against azoles in yeasts involve alterations in the target enzyme or overexpression of drug efflux pumps.[14] Known mechanisms of azole resistance in *C. glabrata* almost exclusively consist of gain-of-function mutations in *PDR1,* which encodes a transcriptional regulator of drug efflux pumps,[15] whereas echinocandin resistance has been linked to non-synonymous variations in two conserved hotspot (HS; i.e., frequently mutated) regions of *FKS* genes.[16] Antifungal drug resistance, tolerance, and adaptation are related to the ability of a cell to respond to stress.[17] Under stress, genome maintenance and repair mechanisms are altered, which may lead to the appearance of resistance phenotypes.[5,18] Rapid adaptation to varying conditions, including exposure to drugs, has been attributed to a remarkable genomic plasticity in *Candida*. In *C. glabrata*, a large degree of genomic and phenotypic variation has been described between and within genetically diverse clades[19,20] and even within clonal populations infecting a patient.[21,22] Previous studies on *in vitro*-acquired drug resistance in *C. glabrata* have evaluated the fitness costs of echinocandin resistance[23] or used transcriptomics to unveil the mechanisms contributing to azole resistance,[24] but the genome-wide genetic alterations involved during this process remain elusive. In addition, the genetic underpinnings of MDR in this pathogen are poorly understood.

RESULTS

C. glabrata has a widespread ability to acquire drug and MDR

Here, we set out to explore the evolutionary adaptation of *C. glabrata* to azoles and echinocandins using an *in vitro* evolution approach coupled to phenotyping and targeted gene and whole-genome sequencing (Figure 1; STAR Methods). To this end, 12 strains representing the 7 previously described *C. glabrata* clades[20] were subjected to increasing concentrations of antifungal drug(s) in the following regimes: fluconazole (FLZ samples; note the use of uppercase letters for samples/conditions as opposed to lowercase letters for the drug), anidulafungin (ANI), and both drugs in combination (ANIFLZ). In addition, to gain insight into mechanisms of cross-resistance, adaptation to serial exposure to both drugs was studied by growing isolates from the final steps of the ANI samples under the flz regime (AinF) and, conversely, final FLZ isolates under ani (FinA). Finally, control populations of all of the strains were grown for the same time without any drug (YPD). The experiment comprised a total of 288 independently evolved populations. When exposed to a single drug or to the two drugs in a sequential manner, all of the populations survived the entire experiment. However, when simultaneously exposed to both drugs, 21 populations (43.75%) died, including all replicates of each of 2 strains from clade I (CST109) and clade III (M12). Nevertheless, populations from other strains from these clades survived, indicating that low adaptation potential is strain- and not clade-specific. We analyzed available parental sequences of the two strains[20] unable to adapt to ANIFLZ and found that they shared eight genes (the *S. cerevisiae* orthologs of *SWI6*, *CDC3*, *LAP2*, *MAD1*, *MNN4*, *RSN1*, and *SQS1* and the gene CAGL0C05313 g) with alterations that were not present in the parentals of the surviving strains within the same clades (Table S1).

We determined susceptibility using the minimum inhibitory concentration (MIC) and the relative area under the curve (rAUC) measurements (Figures 2A, 2B, S1A, and S1B; STAR Methods). All of the surviving strains acquired stable resistance to the exposed agent(s); that is, the resistance phenotype was kept for several generations in standard growth conditions after

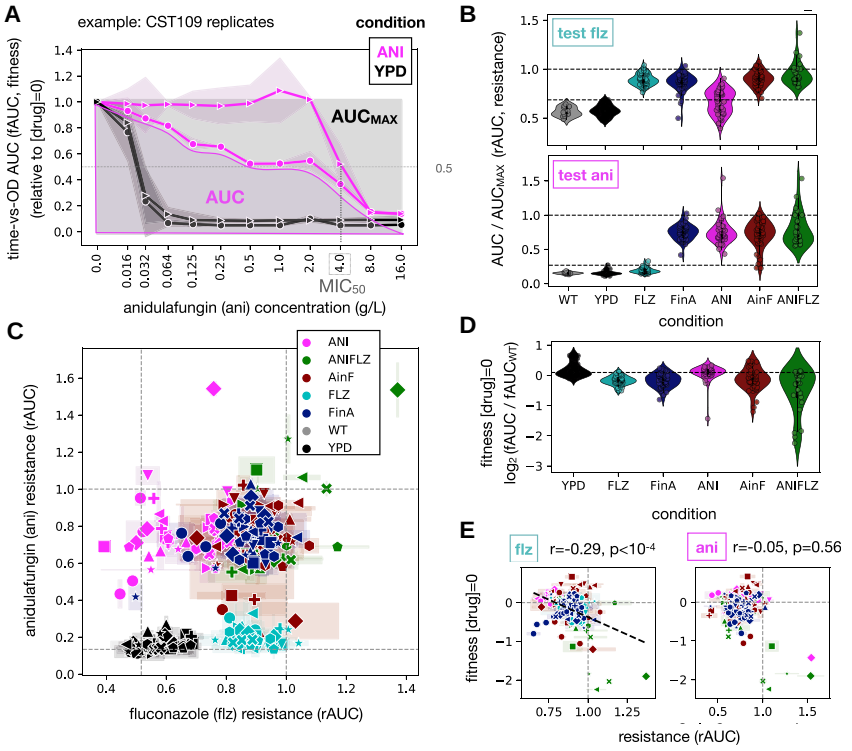


Figure 2. Fitness and drug resistance

(A) We measured relative fitness (the ratio between fitness in each drug concentration versus the no-drug condition [control]) in a time course experiment at several concentrations of flz and ani. Fitness was measured as the area under the time-versus-optical density (OD) curve (fAUC). The graph depicts an illustrative example of two independently evolved replicates of the CST109 strain in the ANI and YPD evolution experiments. The shaded areas represent the median absolute deviation across technical replicates. As a proxy for drug resistance, we defined rAUC as the AUC of these data (normalized by the maximum AUC, in which fitness is maintained across all the range of concentrations [AUC_{MAX}]).50% of growth inhibition, as compared to the no-drug control, is marked as MIC_{50}.

(B) rAUC for flz (top) and ani (bottom) across all samples in our experiments. Each point corresponds to an independently evolved biological replicate. Note that some samples have an rAUC above 1.0, where fitness did not drop upon increasing drug concentration (suggesting high resistance). In addition, Figure S6 includes information about the drug resistance levels among samples with different mutations.

(C) The relationship between ani and flz resistance across all samples. Dashed lines indicate median rAUCs levels for each drug in the YPD samples and $rAUC_{MAX}$ (1.0). Each point corresponds to a biological replicate, and the error bars reflect the median absolute deviation across technical replicates. Each marker corresponds to a different strain.

(D) Fitness in the absence of drug (measured as the $\log_2$ fold change in fAUC (see [A] between each sample and the median fAUC in the WT of the matching strain). Note that Figure S6 includes information about relative fitness levels among samples with different mutations.

(E) Fitness in the absence of drugs is slightly correlated with the levels of flz, but not ani, resistance (rAUC). Spearman's correlation coefficient (r) and p value are shown for flz (left) and ani (right) resistance. The correlation for flz resistance was maintained when considering only samples with mild fitness defects (fitness >-1, $r = -0.22$, $p = 0.0029$). Only resistant samples, defined as those with a $\log_2$ fold increase above 1 as compared to the WT (Figure S1D), were included in this analysis. The individual fitness and susceptibility measurements for each sample can be found in Data S1.

the removal of the selective agent (Figures 2B and S1C; Data S1), indicating that the phenotype is genetically encoded. Unexpectedly, we observed increased resistance to flz in a large subset of ANI samples (21/47, MIC > 256), thus showing that adaptation to ani can frequently induce cross-resistance to flz. The reverse process, acquisition of resistance to ani in FLZ samples, was not observed (Figure S1C). Increased resistance to both drugs (MDR) was often achieved, including all surviving ANIFLZ samples, a majority of AinF (91.6%) and FinA (97.9%) samples, and, due to the mentioned cross-resistance, in 44.7% of ANI samples (Figure 2C; Data S1). In serial drug-exposure experiments, previously acquired resistance was rarely lost during exposure to the second drug (1 FinA and 4 AinF samples), indicating that the phenotype is stable. To assess cross-resistance to other antifungal drugs, we tested the growth of a selected panel of evolved strains on other antifungal drugs (Figure S2D). Similar results were observed for the two tested echinocandins (ani and caspofungin), while the two tested azoles presented more disparate patterns, with few strains growing better on voriconazole (vrz) than on flz (discussed below). None of the tested strains presented improved growth on flucytosine (5-FC, pyrimidine analog) or amphotericin B (ampB, polyene) when compared to wild-type (WT) strains, although a few strains presented higher susceptibility to ampB. We evaluated the fitness costs of acquired resistance using AUC values of growth curves in the absence of the drug as a proxy for fitness (fAUC) relative to the fitness of the unevolved (WT) strain (Figure 2D; Data S1). All flz-exposed samples showed a tendency to reduce fitness ($p < 10^{-5}$, Kolmogorov-Smirnov test), while the mean fitness of ANI samples remained unaltered ($p > 0.05$). Consistently, a small but significant negative correlation between resistance (rAUC) and fitness levels for flz, but not for ani, was detected (Figure 2E). Nevertheless, many of the flz-exposed samples retained fitness levels within 2 standard deviations of the mean of YPD-exposed strains (56% of ANIFLZ, 77% AinF, 81% FLZ, and 68% FinA), and only a few samples (2.9%, 5/8 of them ANIFLZ) had severely reduced fitness levels below 50% of the corresponding WT strain. These results indicate that resistance, including MDR, is often achieved at mild fitness costs. Finally, we evaluated the repeatability of the fitness and susceptibility outcomes in the parallel evolution experiments for replicates and strains subjected to similar conditions. We did so by comparing the distribution of pairwise differences between samples with respect to assayed fitness and susceptibility levels. Our analysis (Figure S1E) indicates that repeatability may be unique to each phenotype and condition, where AinF and ANIFLZ samples have particularly higher phenotypic variability. In addition, we found that variability was similar among evolved samples of the same or different strains (Figure S1E), suggesting that different strains reached similar phenotypes. Interestingly, we found some exceptions

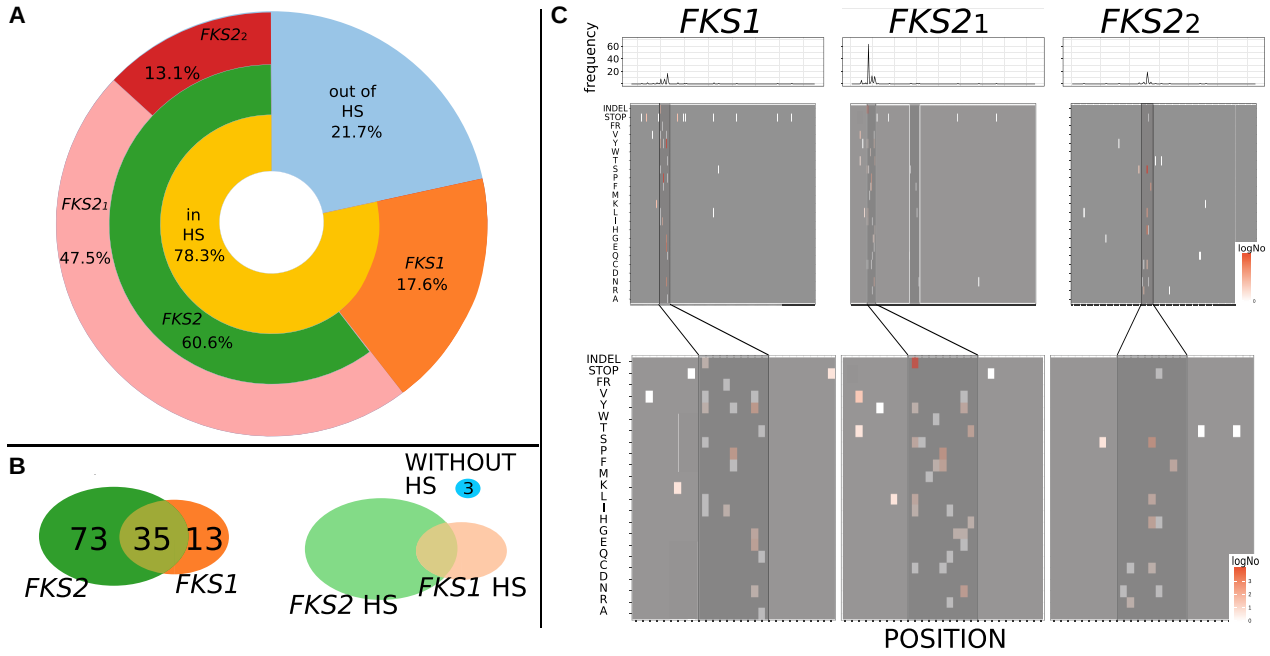


Figure 3. Mutational analysis of *FKS* regions

(A) Distribution of the mutations in studied regions of *FKS*. A non-negligible presence of mutations outside HSs can be observed. Note that Table S5 includes the oligos used for the sequencing. In addition, Data S2 includes the precise mutations.

(B) Distributions of samples according to the presence of mutations in particular *FKS* gene and distribution of samples according to the presence of mutations in *FKS* HSs.

(C) Mutational signatures per sequenced regions: *FKS1* and *FKS2*_1 and *FKS2*_2. Mutated positions are shown as highlighted boxes at the corresponding amino acid in the mutation, over a gray background. Color scale, from white to red, indicates the observed number of mutations (log scale). Darker gray boxes indicate HSs and the white-framed box in *FKS2*_1 marked positions for other possible mutational HSs. The bottom part of the graph represents an enlargement in HSs and mutations in their close proximity.

(including the fitness and flz resistance in YPD-evolved and the fitness of FinA-evolved samples) in which the evolved phenotypes are more consistent among samples of the same strain (Figure S1E).

The *FKS* mutational spectrum in resistant strains expands beyond HS regions

We used a target sequencing approach to screen 121 ani-adapted strains for mutations in the typically surveyed HS of *FKS* genes[25] (Data S2; STAR Methods). In addition, we selected 77 representative (considering clades, susceptibility levels, and *FKS* mutations) samples for whole-genome sequencing and called small variants (SVs), copy-number variations (CNVs), and genomic rearrangements (GRs) appearing *de novo* in each of the evolved samples (Data S3; STAR Methods). All 121 ani-evolved strains presented newly acquired non-synonymous (ns) mutations in the targeted *FKS* regions (Data S2), which indicates that *FKS* mutations may be necessary for ani adaptation. Mutations preferentially occurred in *FKS2* over *FKS1* and in HS1 over HS2 (Figure 3), suggesting a more prevalent role of these loci. Notably, 22% of *FKS* mutations were outside the HS regions. Three resistant strains carried only such non-HS *FKS* mutations (*FKS1*-R1422L and *FKS1*-F708S; *FKS1*-W681L and *FKS2*-K265*; and FKS2-A651T; Data S2), and whole-genome sequencing of these strains revealed no additional mutations outside *FKS* genes that could explain their resistance phenotype (see below). These observations suggest that some of these non-HS *FKS* mutations contribute to resistance and emphasize the importance of studying *FKS* genes beyond HS regions. In addition, we tested whether the distance of non-HS mutations to the actual HS is related to the level of ani resistance in samples harboring only non-HS mutations. We could not find any such significant correlation (Spearman rho = −0.14, $p < 0.11$), suggesting that non-HS and HS mutations confer similar levels of resistance. Overall, the most frequently mutated site in ani-adapted samples was *FKS2*-F659 (63 samples, 52.1%; Data S2), with the most prevalent alteration being F659del (52 samples, 43%), which was the only *FKS* mutation in 26 samples (21.5%). This finding suggests that, as compared to replacements, amino acid deletions may more efficiently prevent the binding of the drug, and reinforces the need to consider this type of mutation. Finally, 26 samples exposed to ani (19.8%) carried a truncation in one of the *FKS* genes (2 of them with a GR breaking the coding region (Figure S3; STAR Methods) in combination with a ns mutation in the other paralog, indicating that this specific combination may facilitate adaptation.

Mutational landscapes in resistant strains reveal a high diversity of genetic alterations affecting a restricted set of recurrently mutated genes

The analysis of genome-wide mutational patterns revealed no newly acquired SVs in YPD samples, while the drug-evolved

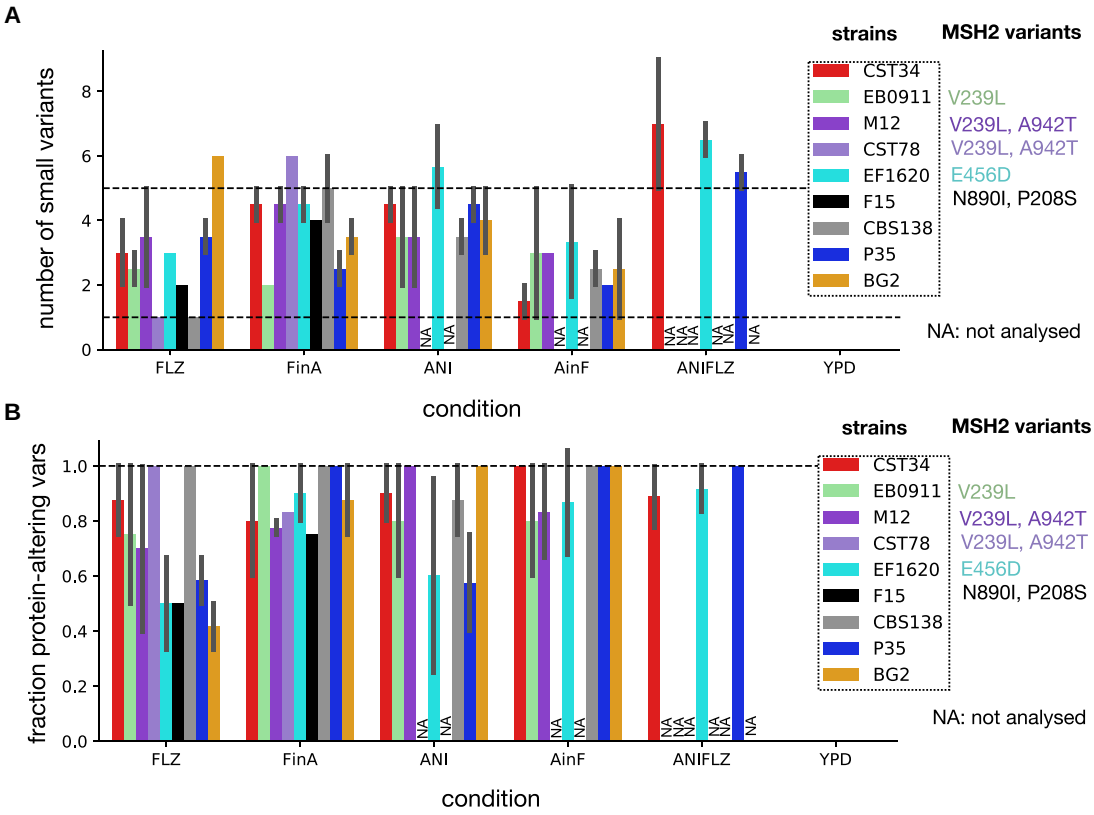


Figure 4. The number of small variants (synonymous and non-synonymous) that appear during the experiment

(A) To select only newly acquired mutations in each drug-evolved sample, we subtracted from called variants those also called in the corresponding WT, YPD, and the parental drug condition (ANI for AinF, and FLZ for FinA), while the corresponding variants called in WT, ANI, AinF, FinA, and FLZ samples were subtracted from those found in the YPD sample. The dashed lines, from bottom to top, correspond to 1 and 5 mutations, respectively. We also represent the presence of $\geq$1 ns variants in the *MSH2* gene in the WT strain. The bars represent the mean number of mutations across biological replicates and the error bars represent the standard deviation.

(B) The same as in (A), but showing the fraction of protein-altering mutations.

strains accumulated a small number (<10) of variants (Figures 4A and 4B). This indicates that susceptible strains are a few mutational steps away from acquiring resistance. Strains carrying distinct *MSH2* variants (Figures 4A and 4B) did not accumulate a different number of mutations, thereby supporting the notion that these represent natural, functional variants rather than hypermutator mutations.[20] As expected,[26,27] we found that aneuploidies were common in experiments involving exposure to flz, but they were not detected in cells exposed only to ani (Figure 5A). Total or partial aneuploidies in chromosome E (ChrE), encompassing *ERG11*, were the most common, appearing in 11/16 FLZ, 4/15 AinF, and 2/6 ANIFLZ samples. Most (10/11) FLZ samples with the ChrE aneuploidy retained it upon further exposure to ani (FinA). One strain presented a partial ChrE aneuploidy resulting from unbalanced translocation with ChrJ (Figure S3D; STAR Methods), suggesting that GRs can drive drug resistance. Importantly, we detected no heterozygous variant in any of the duplicated chromosomes, indicating they have not accumulated new mutations since their duplication, and, therefore, that aneuploidies were adaptive per se and not because they allowed faster evolution of duplicated genes. To investigate whether aneuploidies conferring flz resistance were rendering strains avirulent, we used an *in vivo Galleria mellonella* model (STAR Methods) to assess the virulence of a WT strain and two of its descendant FLZ strains, one of which presented chromosomal duplications in ChrE and ChrI. Our results (Figure 5B) show that all of the descendant strains remained virulent, suggesting that flz resistance or the presence of aneuploidies are compatible with virulence.

To identify mutations likely associated with the resistance trait, we selected genes that were mutated at least twice independently in our experiment. This search identified nine genes (*ERG11*, *PDR1*, *CDR1*, *CNE1*, *EPA13*, *FKS1*, *FKS2*, *ERG3*, *ERG4*; Figure 6). Importantly, all of the resistant strains carried mutations or duplications in at least one of these genes, and the subset of mutated genes largely separated samples by treatment. This strong association of acquired mutations, treatment, and phenotypes indicates that a limited set of genes is central for the acquisition of resistance. The most common altered gene under exposure to flz was *PDR1*, which was in many instances (14/37 strains) accompanied by alterations in *ERG11* (Figure 6; Data S3). Although less common, five resistant strains contained no *PDR1*-related mutations or aneuploidies (Figure 6), indicating that alternative mechanisms confer resistance on their own.

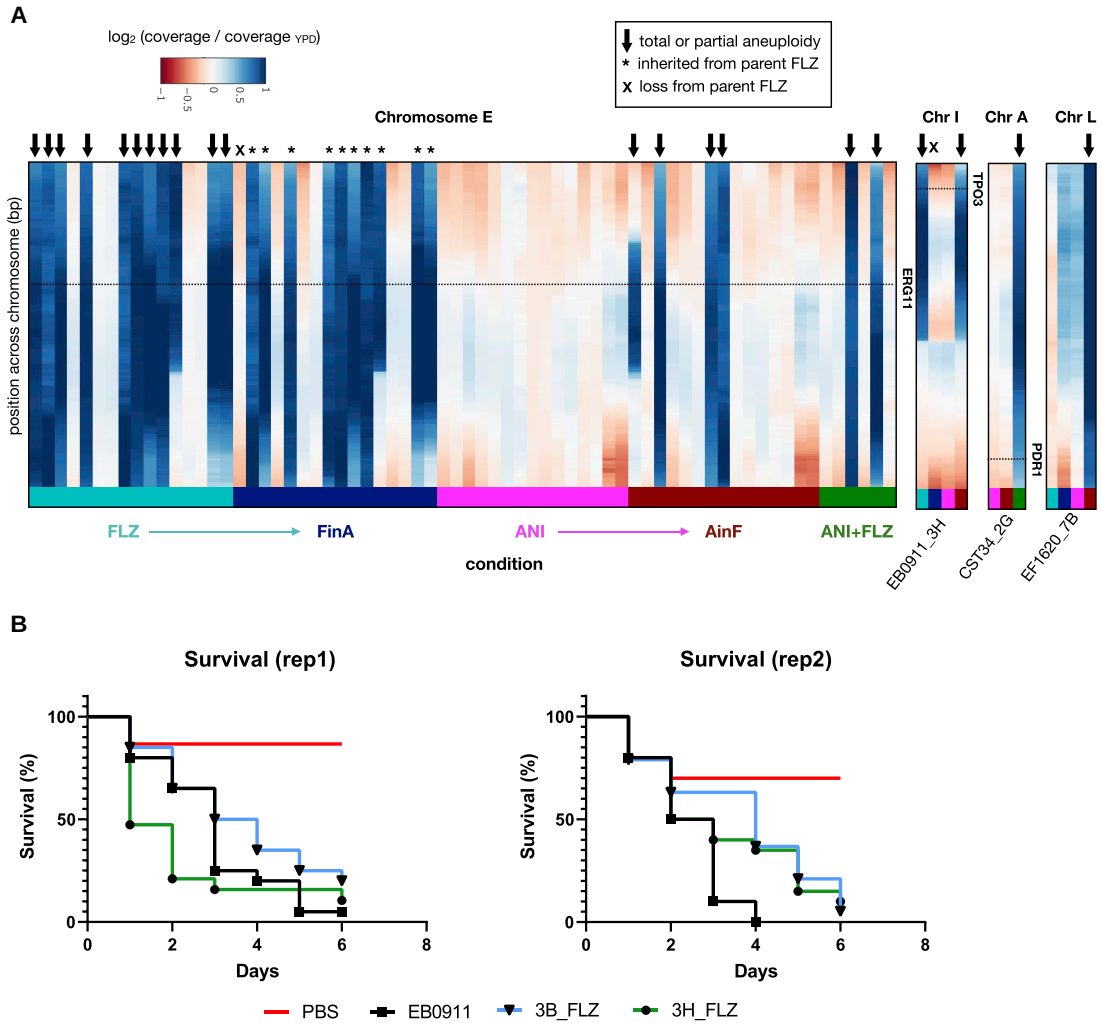


Figure 5. The role of aneuploidies in drug resistance

(A) We calculated the median relative coverage per gene for all samples analyzed in this work. This parameter appeared to be correlated with the distance to the telomere (STAR Methods), so that the $\log_2$ ratio to the YPD (of the corresponding strain) was used as a proxy for the gene copy number. Shown is the rolling-median of this value for windows of 50 genes and chromosomes where large duplications were observed (chromosomes E, I, A, and L). Data for chromosomes I, A and L are shown only for those strains in which aneuploidies are observed. Each column corresponds to a sample (ordered as in Figure 6), and the "*" and "X" correspond to FinA samples in which the parent (FLZ) aneuploidy was maintained or lost, respectively. *ERG11*, *PDR1*, and *TPO3* are genes that we speculate could be driving the selective advantage of the aneuploidy (see Results). All of the values were cut off at 1.0 (2× coverage as compared to the YPD) for clarity. (B) Survival of *Galleria mellonella* larvae during 6 days after inoculation of EB0911 (WT strain) and 2 flz resistant progenies: 3B_FLZ (without aneuploidies) and 3H_FLZ (presenting both ChrE and ChrI duplications).

These strains harbored mutations in *ERG3* (3 AinF strains, discussed below) and *ERG11* (2 strains). Importantly, *ERG11* mutations and aneuploidies in ChrE, bearing this gene, were strongly anticorrelated, with a single ANIFLZ sample carrying both alterations. In this case, the mutation was present in the two alleles, suggesting that the mutation preceded the chromosomal duplication. Among *ERG11* mutations, K152 was the most altered amino acid (12/16 samples), followed by *ERG11*-Y141 (2/16 samples). Although common in other *Candida* species, these mutations have not been commonly reported in *C. glabrata*.[5] Structural analysis revealed that both altered residues were close to the azole binding pocket (Figure S4).

We next assessed whether the catalog of mutations found in our *in vitro* analysis was representative of what can be found in clinical strains. To this end, we compared this catalog with variants found in 393 *C. glabrata* clinical isolates with genomes publicly available at Candidamine (https://candidamine.org/). Our results (Figure S5; STAR Methods) show that the overlap of specific mutations is very low. This low overlap is, however, expected from the actual large diversity of the identified mutations in our experiments (Figure S5B; Data S2 and S3) and is similarly low for mutations identified in actual clinical surveys (e.g., SENTRY[6]). These results suggest that although the set of genes recurrently mutated during the acquisition of resistance is rather

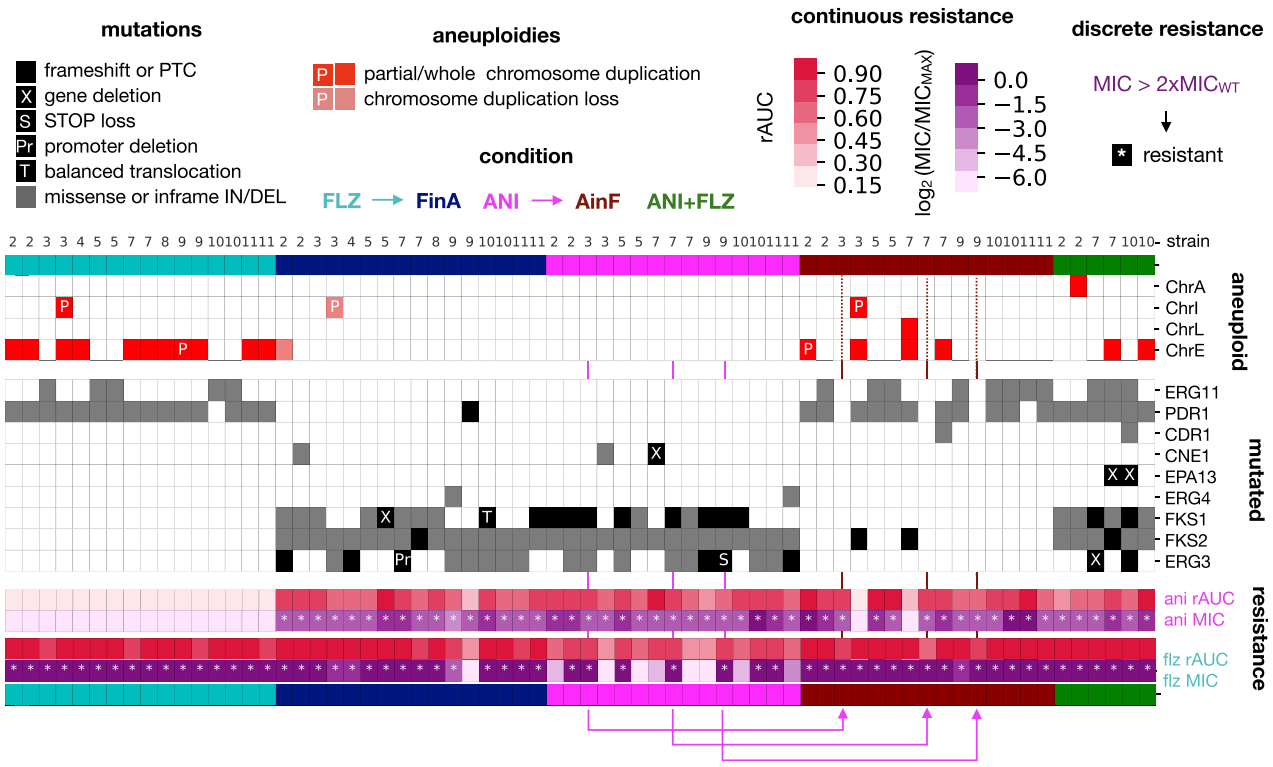


Figure 6. Aneuploidies and recurrently mutated genes

Each drug is associated with a particular set of mutated genes and aneuploidies. Columns represent the evolved samples, each strain indicated by a number: 2, CST34; 3, EB091; 4, CST78; 5, M12; 6, EF1237; 7, EF1620; 8, F15; 9, CBS138; 10, P35; 11, BG2. Replicates of the same strain appear in the same order as in the experimental plate. Colors indicate the experimental condition. Blocks show, from top to bottom, chromosomal alterations, mutated genes, and susceptibility data. Whole and partial (P) chromosomal duplications appearing newly in each condition are marked as red, while losses are marked as light salmon boxes. Protein-altering mutations (gray boxes) and losses (black boxes) of genes appearing in at least 2 drug-evolved samples are shown. Note that we found a balanced translocation in *FKS1* (T) and a deletion in the *ERG3* promoter region (Pr) (Figure S3; Results; STAR Methods). PTC stands for premature termination codon. Pink arrows indicate the parent-daughter relationships for 3 AinF samples that did not present any new alteration in recurrent genes. Note that Figures S3 and S4 and Data S3 provide additional information about these mutations and genomic rearrangements. In addition, Figure S6 shows the association between these mutations and fitness or drug-resistance levels.

limited (nine genes), the number of specific mutations (i.e., which residue is mutated and what type of mutation occurs) is large and highly diverse and only partially covered by our experiment or clinical surveys.

Decreased fitness of some resistance-conferring mutations could hamper their detection in the clinics, as clinical isolates are not obtained in selective conditions. To explore possible fitness trade-offs of specific mutations, we evaluated whether strains harboring each type of mutation had a particular fitness or susceptibility level. Consistent with the fitness results presented above, most of the mutations had no significant effect on fitness in the absence of the drug (Figure S6). However, we found that strains harboring *CNE1* truncations or ChrL and ChrA duplications presented lower fitness, indicating that some resistance mechanisms may generate decreased growth (Figure S6). On another note, we found that most strains had similar flz and ani susceptibility levels independently of the mutation type (i.e., we found no differences in flz resistance among strains with *ERG11* mutations or ChrE duplications) (Figure S6). Finally, we investigated whether there was a correlation between the number of different genes with acquired mutations and fitness/susceptibility levels in any of the evolution conditions. We found no significant Spearman correlation ($p < 0.05$) after removing a single outlier AinF sample with a particularly high number of new mutations and low ani resistance. These results indicate the lack of a general correlation between the numbers of acquired mutations and these evolved phenotypes. Our data suggest that different evolutionary paths drive similar levels of drug resistance and fitness in a strain-independent and mutation-independent manner.

Crosstalk between echinocandin and fluconazole resistance

In the experiments of sequential exposure to the two drugs, all of the samples successfully adapted, in turn, to the two challenges. When adapting to the new drug, most samples (90 of 95) retained the previously acquired resistance, resulting in MDR (Figures 7A, 7B, S1C, and S1D). However, three sequenced samples lost the previously acquired resistance upon the change in selective conditions (according to MIC, see Figures S1C and S1D). These included a FinA sample and two AinF samples. This FinA sample acquired a premature termination codon in *PDR1*, which may

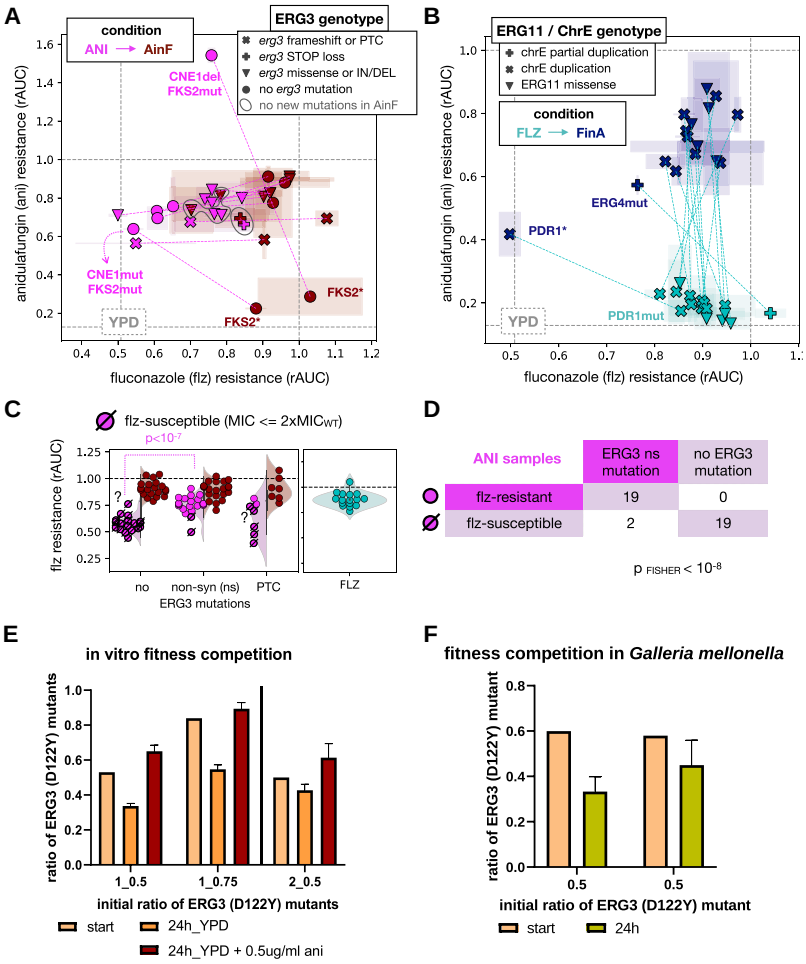


Figure 7. *ERG3* mutations and multidrug resistance

(A) Biplot showing the relationship between resistance (rAUC) toward ani and flz for a series of ANI/AinF related samples. The gray dashed lines indicate the rAUC = 1.0 (where fitness is maintained across the range of concentrations; Figure 1A) and the median rAUC across YPD samples for each of the 2 drugs. Each sample is represented by a symbol, with the color indicating the sample type: ANI (pink) and AinF (red) samples. The pink dashed lines indicate parent-daughter relationships (ANI-AinF) between the samples. The symbols represent different types of *ERG3* mutations, and the gray circles outline 3 samples that did not acquire any new mutation in the recurrent genes in AinF. The 2 ANI samples with alterations in *CNE1*, which lost ani resistance due to truncations in *FKS2*(*) in AinF samples, are marked. One of the ANI samples showed high ani resistance (above 1.0, meaning the fitness was higher in ani than in no drug), but also showed low basal fitness, which means that the high resistance value may be not representative. Error bars reflect the median absolute deviation across technical replicates.

(B) Relationship between rAUC of ani and flz in FLZ (light blue) and FinA (dark blue) samples. The green dashed lines indicate parent-daughter relationship (FLZ-FinA). The gray dashed lines indicate the rAUC = 1.0 (where fitness is maintained across all the range of concentrations; Figure 1A) and the median rAUC across YPD samples for each of the 2 drugs. No acquisition of ani resistance was observed in FLZ samples but only as a result of ani (FinA). The symbols represent the presence of *ERG11* missense mutations or chromosome E aneuploidies. Two FinA samples showed a drop in flz resistance levels. One of them carried a *PDR1* premature termination codon (*), which resulted in susceptibility according to our MIC-based thresholding (STAR Methods) and reduced flz resistance below the median rAUC value of YPD samples. The other sample carried *ERG4* mutation that resulted in a reduction but not a total loss of flz resistance. Error bars reflect the median absolute deviation across technical replicates.

(C) Non-synonymous (including missense and STOP loss) *ERG3* mutations are associated with higher flz resistance (rAUC) in ANI samples. The p value corresponds to a Kolmogorov-Smirnov test. The corresponding AinF and FLZ samples are also shown for comparison of flz-resistance levels. The dashed symbols represent samples that were found to be flz susceptible according to our MIC-based thresholding (STAR Methods). Note that 2 samples (marked with "?") were found as susceptible but have rAUC values in the range of resistant samples. This mismatch is clarified in Figures S7C and S7D. In addition, see Tables S2 and S3 for further information on the *ERG3* mutations found in each sample.

(D) The presence of *ERG3* non-synonymous mutations is correlated with discrete flz resistance in ANI samples. The number of ANI samples in each category and the p value of a Fisher test are shown.

(E and F) Growth competition between ani-resistant strains with and without *ERG3* mutation (note that Table S5 includes the oligos used for sequencing). The y axis presents the calculated ratio of a sample with mutated *ERG3* gene and the x axis ratios aimed at the beginning of the experiments. The error bars represent the standard deviation across technical replicates. (E) *In vitro* fitness competition of 2 pairs of strains: 1-CRISPR transformant *ERG3* (D122Y) versus CRISPR transformant ERG3(WT) with *NAT1* and 2-CRISPR transformant *ERG3* (D122Y) with *NAT1* versus 3H_ANI (*ERG3* WT). The competition was conducted over a 24-h period and in YPD and YPD supplemented with 0.5 μg/mL ani in triplicates. (F) Two independent competition experiments *in vivo*. The fungal burden was obtained from 3 separate larvae for each of the initial mix of populations.

revert the flz resistance conferred by previous mutations in this gene. In the two AinF samples that lost resistance to ani, we found frameshift mutations in *FKS2* downstream of the ani resistance-conferring mutations inherited from the parental ANI samples (Figure 7A). Interestingly, both of the ANI parents carried only one *FKS2* mutation and alterations in *CNE1* ortholog, encoding an endoplasmic reticulum (ER) protein involved in the quality control of misfolded proteins.[28] This remarkable coincidence suggests that the combination of these alterations results in a higher propensity to lose resistance, although this hypothesis needs further study. Except for a single ChrA duplication found in one strain, most ANIFLZ samples showed mutational signatures similar to those acquired during sequential exposure to the two drugs (AinF and FinA; Figure 6). This observation suggests that the genetic basis driving the acquisition of resistance to each of the drugs is similar when the two drugs are in combination.

A remarkable finding of our experiment is the cross-resistance to flz found in a significant fraction of ANI samples (see above). Whole-genome sequencing of 7 of these strains revealed that all of them carried alterations in *ERG3*, which encodes the C-5 sterol desaturase of the ergosterol biosynthesis pathway

(Figure 6). This association was further explored by Sanger-based target sequencing of the *ERG3* gene in the remaining ani-evolved strains, which showed that all 21 ani-evolved strains showing cross-resistance to flz (MIC > 256 μg/mL) carried alterations in *ERG3* (Table S2). Accordingly, we detected a significant association between *ERG3* ns mutations and flz resistance in ANI samples (Figures 7C, 7D, and S7A–S7D). Interestingly, these samples showed lower levels of flz resistance when compared to FLZ samples (Figure S7C). This finding indicates that the quantitative contribution of *ERG3* mutations to flz resistance differs from that of *PDR1* or *ERG11* alterations and suggests different mechanisms of resistance in FLZ and ANI samples. When *ERG3*-mutated strains were subsequently exposed to flz (AinF), three of them did not acquire additional mutations in *PDR1* or *ERG11*, nor did they present ChrE duplications, thereby suggesting that their *ERG3* mutations were sufficient for their survival in flz. In support of this notion, the levels of flz resistance of these three AinF samples and their respective ANI parents were similar (Figure 7A). However, the relationship between *ERG3* alterations and cross-resistance to flz was incomplete and mutation dependent. We found that of 28 ANI samples harboring *ERG3* mutations, 6 carrying premature stop (3), missense (2), and frameshift (1) mutations retained WT levels of susceptibility. The absence of resistance in strains carrying *ERG3* mutations leading to truncated proteins is compatible with earlier work showing that *ERG3* deletion in *C. glabrata* does not affect flz susceptibility.[29] Consistent with some *ERG3* alterations being selected under exposure to ani, 2 ANIFLZ and 6 FinA samples bearing *ERG3* changes additional to *PDR1* and/or *ERG11* mutations were detected (Figure 6). Incidentally, another FinA sample carried a deletion in the gene immediately upstream of *ERG3* (*CAGL0F01815 g*, of unknown function), which we speculate may result in regulatory alterations of *ERG3* through disruption of the promoter (Figure S3A; STAR Methods). To investigate the relationship between *ERG3* mutations and flz resistance further, we re-introduced one of the *ERG3* mutations (D122Y) into two WT strains (CBS138 and EB0911) and an ani-evolved and flz-susceptible progeny of EB0911-3H_ANI. In addition, we reverted *ERG3* to the WT sequence in one strain (3B_ANI, progeny of EB0911) originally harboring *ERG3* D122Y mutation. We then assayed the susceptibility phenotype of these transformants and the original strains. Our results (Figures S2A and S2B) show that the introduction of the D122Y mutation in *ERG3* led to increased resistance to flz, and that the reversion of the mutation had the opposite effect, confirming the link of *ERG3* and flz susceptibility. We noted that the effect of this mutation was stronger in an ani-resistant background as compared to a WT background, where growth on flz was observed at a later time point. Our results support a dual role of *ERG3* alterations in the adaptation to ani and in the development of cross-resistance to flz in *C. glabrata*.

To gain mechanistic insight into these *ERG3* alterations, we performed various experiments. We tested whether the introduced *ERG3* alterations were associated with altered response to various stresses. Our results (Figure S2C) suggest no major effects, with the exception of a lower tolerance to membrane (SDS) and oxidative (H_2O_2) stresses restricted to a particular ani-resistant background strain. In addition, we traced the order of appearance of *ERG3* and *FKS* mutations along intermediate generations in ANI strains and found equal numbers of cases (2 each) in which either *ERG3* or *FKS* mutations predated the other one, and 5 cases in which both mutations are traced to the same intermediate generation (Table S3). These data suggest that one mutation does not necessarily predate the other one. Resistance to flz is often spontaneously acquired in *C. glabrata* by partial or total loss of mitochondrial DNA, rendering a so-called petite phenotype.[30] However, we can discard this effect in the identified *ERG3* mutants due to the absence of deletions in the mtDNA (Figure 6; Data S3) and the absence of a petite phenotype (Figure S7E). We further analyzed competitive fitness between ani-resistant strains with and without *ERG3* mutations using *in vitro* and *in vivo* (*G. mellonella*) competition assays (STAR Methods). Both assays provided similar results (Figures 7E and 7F), supporting a competitive disadvantage of the *ERG3* mutants in the absence of the drug. However, when the *in vitro* competition experiment was performed in the presence of ani, the *ERG3* mutant outcompeted the WT. These results support the selection for *ERG3* mutations only during ani treatment and point to a possible explanation for the lack of clinical cases showing this alteration.

DISCUSSION

Our study adds support to the suitability of *in vitro* approaches to study the evolutionary acquisition of resistance to antifungal drugs,[23,31,32] and contributes to a better understanding of the mechanisms of drug adaptation in *C. glabrata*. Given the high number of replicates and the drug-specific patterns we consistently observed, we can conclude that the discussed mutations are likely related to the specific drug exposure and not to the experimental setting. Our results show that *C. glabrata* exhibits a remarkable capacity to acquire resistance to the tested drugs, independently of the phylogenetic background of the strain.[20] This is also true for the case of serial exposure to the two drugs, to which all strains and replicates adapted. However, the combined exposure to both drugs prevented adaptation in a significant fraction of the cases, with two strains from two different clades showing an inability to develop resistance in this scenario. Our results show that neither phylogenetic clade nor the presence of non-synonymous mutations in *MSH2* are good predictors of the ability to develop MDR, which is pervasive in *C. glabrata*. Whole-genome sequencing revealed a relatively limited catalog of a few genes that are commonly affected upon sustained adaptation to antifungal drugs. We observed the appearance of commonly reported alterations in *FKS*, *PDR1*, and *ERG11* genes, which indicates that our experiment reflects processes that also occur in the clinics. However, 5 other genes (*CDR1*, *CNE1*, *EPA13*, *ERG3*, and *ERG4*) were recurrently mutated in our experiments. This finding indicates that alternative mechanisms may be concomitantly used to achieve a stable resistant phenotype. Alterations in the promoter region of the efflux pump *CDR1* have already been reported in azole-resistant strains,[33,34] and our results suggest that alterations of the protein product may also contribute to flz adaptation. We propose that the observed *CDR1* mutations increase azole efflux and thus decrease flz susceptibility. As discussed, *CNE1* is involved in the quality control of misfolded proteins in the ER. *EPA13* is a sub-telomerically encoded lectin-like adhesin with a role in cell

adhesion, whose potential role in drug resistance is unknown. Altered adhesion has been linked to azole resistance in *C. glabrata*,[35] which may explain why *EPA13* deletions could be adaptive under exposure to both azoles and echinocandins in our experiments. *ERG4* is another gene involved in the ergosterol biosynthesis pathway, which, similar to *ERG3*, may influence resistance to flz. Future experiments should help determine the order of appearance of these mutations and their specific roles in drug resistance or adaptation. In addition, our results suggest that GRs and CNVs around these genes are related to drug resistance, as previously proposed in *C. albicans*.[36] This indicates that the traditional focus on SNPs is underpowered to understand the genomic drivers of drug resistance. Finally, our results suggest that although the set of genes altered during the process of adaptation may be limited, the diversity of possible resistance-conferring mutations in each of the affected genes is very large.

An important result from our experiment is the observation that adaptation to ani often results in cross-resistance to flz (but not the other way around). This result was unexpected, given the different modes of action of the two drugs, in which ani affects the cell wall in a fungicidal manner and flz affects the cell membrane, causing growth arrest. This observation is of high relevance, given the expanding MDR in *C. glabrata* and also considering that some recent guidelines (e.g., from the Infectious Disease Society of America[37]) recommend an echinocandin-based initial therapy against most invasive *Candida* spp. infections. Importantly, these findings are consistent with a recent report of flz cross-resistance in ani-adapted *C. glabrata* isolates.[38] We consider that our results can inform future clinical trials or therapy guidelines. For instance, our data suggest that flz resistance may be common after the failure of ani therapy, so that flz treatment following ani may also result in therapy failure. Thus, monitoring of flz resistance after ani therapy, or the use of a different drug as a second line of therapy may be recommended. Similarly, our results point to the absence of cross-resistance to ani when flz is used as a first therapy or to a high clearance potential of the concomitant use of flz and ani, which may be considered in specific cases. Importantly, many flz-resistant strains were susceptible to vrz, which could be a promising therapeutic alternative. However, this observation was drawn from a few samples and requires further research. The scarcity of sequenced genomes for MDR clinical strains and the lack of information of the treatment regime they were exposed to (STAR Methods) prevented us from assessing how commonly this cross-resistance mechanism occurs in the clinics, something that deserves further investigation. We studied the possible molecular basis of such cross-resistance and found compelling evidence of the involvement of *ERG3* mutations. In our experiment, alterations in this gene often appeared under ani exposure and were retained in subsequent flz exposure, sometimes without any further mutation being acquired that would explain the acquisition of resistance to flz. In addition, *ERG3* mutations were always present in ani-evolved strains that showed cross-resistance to flz, and we confirmed the causative association of flz resistance of the *ERG3* alteration by reintroducing it in a flz-sensitive background. Competition assays between strains carrying the WT and the mutated *ERG3* allele showed a competitive disadvantage of *ERG3* mutants in the absence of drug treatment, but an advantage in the presence of ani. This underscores the complex fitness trade-offs of resistance-conferring mutations and suggests that the frequency of resistance-conferring alleles is likely to fluctuate after treatment. An intriguing possibility is that clones carrying resistance-conferring mutations and causing therapy failure may be missed during the process of strain identification, as blood cultures and colony isolation is generally performed in the absence of drug exposure. Such phenomenon could partly explain the observed discrepancies between resistance levels of clinical isolates and therapy failure.[39]

Importantly, the link between *ERG3* and cross-resistance may not be restricted to *C. glabrata* as *ERG3* mutations leading to the depletion of ergosterol and the accumulation of less toxic sterols when *ERG11* is inhibited have been implicated in cross-resistance between azoles and polyenes in *S. cerevisiae* and *C. albicans*[40–43] and between echinocandins and azoles in *C. parapsilosis*.[44,45] In addition, acquisition of *ERG3* mutations upon echinocandin exposure has also been described in *C. auris*.[46] Why *ERG3* mutations are often acquired under exposure to ani and how they contribute to resistance to flz remain unclear and need further attention. A speculative scenario is that certain *ERG3* mutations lead to alterations in the membrane composition in a way that partially compensates cell-wall alterations induced by ani exposure. In this regard, it has been reported that cell membrane modifications related to changes in ergosterol production affect the structure and composition of the cell wall.[47]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Microbe strains
  - Galleria mellonella
- METHOD DETAILS
  - *In vitro* evolution
  - Susceptibility tests
  - Fitness and susceptibility measurements
  - DNA extraction
  - Target *FKS* and *ERG3* sequencing
  - Petite phenotype in ani adapted mutants
  - Whole genome sequencing
  - Small variant calling and interpretation
  - Identification of large aneuploidies, segmental duplications and deletions
  - Analysis of genomic rearrangements
  - Analysis of clinical isolates' sequencing datasets
  - CRISPR-Cas9 based genetic modifications
  - Fitness competition
  - Virulence assays
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - rAUC, MIC and fitness measurements
  - Correlation analyses

- Association between *ERG3* mutations and flz resistance
- Comparing continuous distributions
- Variant calling from sequencing data
- Competitive fitness measurements
- Estimating the overlap between drug resistance mutations among samples

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.09.084.

ACKNOWLEDGMENTS

The authors thank Ester Saus, Jesse Willis, and Cinta Pegueroles for their help and technical assistance with some of the analyses. M.A.S.-T. received a predoctoral fellowship from the "Caixa" Foundation (LCF/BQ/DR19/11740023). The T.G. group acknowledges support from the Spanish Ministry of Science and Innovation grant no. PGC2018-099921-B-I00, cofounded by the European Regional Development Fund (ERDF); from the CERCA Programme/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR423; the European Union's Horizon 2020 research and innovation program under grant agreement no. ERC-2016-724173; and the Marie Skłodowska-Curie grant agreement no. H2020-MSCA-IF-2017-793699. The group also receives support from an INB grant (PT17/0009/0023-ISCIII-SGEFI/ERDF). The Bioactive Microbial Metabolites research platform (BiMM) is supported by grants K3-G-2/026-2013 and COMBIS/ LS16005, both funded by the Lower Austria Science and Education Fund (NfB).

AUTHOR CONTRIBUTIONS

E.K., R.B., and J.C.N.-R. conducted the experiments; M.A.S.-T. performed the computational analysis; E.K. and M.A.S.-T. performed the statistical analysis; C.S. and T.G. supervised the project. All of the authors interpreted the results. E.K., M.A.S.-T., and T.G. wrote the manuscript, with input from all of the authors. T.G. conceived the study. All of the authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 9, 2021
Revised: July 20, 2021
Accepted: September 29, 2021
Published: October 25, 2021

REFERENCES

1. Bongomin, F., Gago, S., Oladele, R.O., and Denning, D.W. (2017). Global and Multi-National Prevalence of Fungal Diseases-Estimate Precision. J. Fungi (Basel) *3*, 57.
2. Gabaldón, T.; Consortium OPATHY (2019). Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. FEMS Microbiol. Rev. *43*, 517–547.
3. Fisher, M.C., Hawkins, N.J., Sanglard, D., and Gurr, S.J. (2018). Worldwide emergence of resistance to antifungal drugs challenges human health and food security. Science *360*, 739–742.
4. Arastehfar, A., Gabaldón, T., Garcia-Rubio, R., Jenks, J.D., Hoenigl, M., Salzer, H.J.F., Ilkit, M., Lass-Flörl, C., and Perlin, D.S. (2020). Drug-Resistant Fungi: An Emerging Challenge Threatening Our Limited Antifungal Armamentarium. Antibiotics (Basel) *9*, 877.
5. Ksiezopolska, E., and Gabaldón, T. (2018). Evolutionary Emergence of Drug Resistance in Candida Opportunistic Pathogens. Genes (Basel) *9*, 461.
6. Pfaller, M.A., Diekema, D.J., Turnidge, J.D., Castanheira, M., and Jones, R.N. (2019). Twenty Years of the SENTRY Antifungal Surveillance Program: Results for *Candida* Species From 1997-2016. Open Forum Infect. Dis. *6* (*Suppl 1*), S79–S94.
7. Gabaldón, T., and Carreté, L. (2016). The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in Candida glabrata. FEMS Yeast Res. *16*, fov110.
8. Vallabhaneni, S., Cleveland, A.A., Farley, M.M., Harrison, L.H., Schaffner, W., Beldavs, Z.G., Derado, G., Pham, C.D., Lockhart, S.R., and Smith, R.M. (2015). Epidemiology and Risk Factors for Echinocandin Nonsusceptible Candida glabrata Bloodstream Infections: Data From a Large Multisite Population-Based Candidemia Surveillance Program, 2008-2014. Open Forum Infect. Dis. *2*, ofv163.
9. Perlin, D.S. (2015). Echinocandin Resistance in Candida. Clin. Infect. Dis. *61* (*Suppl 6*), S612–S617.
10. Pristov, K.E., and Ghannoum, M.A. (2019). Resistance of Candida to azoles and echinocandins worldwide. Clin. Microbiol. Infect. *25*, 792–798.
11. Arendrup, M.C., and Patterson, T.F. (2017). Multidrug-Resistant Candida: Epidemiology, Molecular Mechanisms, and Treatment. J. Infect. Dis. *216* (*Suppl_3*), S445–S451.
12. Heimark, L., Shipkova, P., Greene, J., Munayyer, H., Yarosh-Tomaine, T., DiDomenico, B., Hare, R., and Pramanik, B.N. (2002). Mechanism of azole antifungal activity as determined by liquid chromatographic/mass spectrometric monitoring of ergosterol biosynthesis. J. Mass Spectrom. *37*, 265–269.
13. Perlin, D.S. (2007). Resistance to echinocandin-class antifungal drugs. Drug Resist. Updat. *10*, 121–130.
14. Lupetti, A., Danesi, R., Campa, M., Del Tacca, M., and Kelly, S. (2002). Molecular basis of resistance to azole antifungals. Trends Mol. Med. *8*, 76–81.
15. Sanglard, D., Ischer, F., Calabrese, D., Majcherczyk, P.A., and Bille, J. (1999). The ATP binding cassette transporter gene CgCDR1 from Candida glabrata is involved in the resistance of clinical isolates to azole antifungal agents. Antimicrob. Agents Chemother. *43*, 2753–2765.
16. Perlin, D.S. (2015). Mechanisms of echinocandin antifungal drug resistance. Ann. N Y Acad. Sci. *1354*, 1–11.
17. Cowen, L.E., and Steinbach, W.J. (2008). Stress, drugs, and evolution: the role of cellular signaling in fungal drug resistance. Eukaryot. Cell *7*, 747–764.
18. Healey, K.R., Jimenez Ortigosa, C., Shor, E., and Perlin, D.S. (2016). Genetic Drivers of Multidrug Resistance in *Candida glabrata*. Front. Microbiol. *7*, 1995.
19. Biswas, C., Marcelino, V.R., Van Hal, S., Halliday, C., Martinez, E., Wang, Q., Kidd, S., Kennedy, K., Marriott, D., Morrissey, C.O., et al. (2018). Whole Genome Sequencing of Australian *Candida glabrata* Isolates Reveals Genetic Diversity and Novel Sequence Types. Front. Microbiol. *9*, 2946.
20. Carreté, L., Ksiezopolska, E., Pegueroles, C., Gómez-Molero, E., Saus, E., Iraola-Guzmán, S., Loska, D., Bader, O., Fairhead, C., and Gabaldón, T. (2018). Patterns of genomic variation in the opportunistic pathogen Candida glabrata suggest the existence of mating and a secondary association to the human host. Curr. Biol *28*, 15–27.e7.
21. Carreté, L., Ksiezopolska, E., Gómez-Molero, E., Angoulvant, A., Bader, O., Fairhead, C., and Gabaldón, T. (2019). Genome Comparisons of *Candida glabrata* Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations. Front. Microbiol. *10*, 112.
22. Singh-Babak, S.D., Babak, T., Diezmann, S., Hill, J.A., Xie, J.L., Chen, Y.-L., Poutanen, S.M., Rennie, R.P., Heitman, J., and Cowen, L.E. (2012). Global analysis of the evolution and mechanism of echinocandin resistance in Candida glabrata. PLoS Pathog. *8*, e1002718.
23. Duxbury, S.J.N., Bates, S., Beardmore, R.E., and Gudelj, I. (2020). Evolution of drug-resistant and virulent small colonies in phenotypically diverse populations of the human fungal pathogen *Candida glabrata*. Proc. Biol. Sci. *287*, 20200761.

24. Cavalheiro, M., Costa, C., Silva-Dias, A., Miranda, I.M., Wang, C., Pais, P., Pinto, S.N., Mil-Homens, D., Sato-Okamoto, M., Takahashi-Nakaguchi, A., et al. (2019). A Transcriptomics Approach To Unveiling the Mechanisms of Evolution towards Fluconazole Resistance of a Clinical Isolate. Antimicrob. Agents Chemother. *63*, e00995-18.

25. Shields, R.K., Nguyen, M.H., Press, E.G., Cumbie, R., Driscoll, E., Pasculle, A.W., and Clancy, C.J. (2015). Rate of FKS Mutations among Consecutive Candida Isolates Causing Bloodstream Infection. Antimicrob. Agents Chemother. *59*, 7465–7470.

26. vanden Bossche, H., Marichal, P., Odds, F.C., Le Jeune, L., and Coene, M.C. (1992). Characterization of an azole-resistant Candida glabrata isolate. Antimicrob. Agents Chemother. *36*, 2602–2610.

27. Bing, J., Hu, T., Zheng, Q., Muñoz, J.F., Cuomo, C.A., and Huang, G. (2020). Experimental Evolution Identifies Adaptive Aneuploidy as a Mechanism of Fluconazole Resistance in Candida auris. Antimicrob. Agents Chemother. *65*, e01466-20.

28. Molinari, M., Eriksson, K.K., Calanca, V., Galli, C., Cresswell, P., Michalak, M., and Helenius, A. (2004). Contrasting functions of calreticulin and calnexin in glycoprotein folding and ER quality control. Mol. Cell *13*, 125–135.

29. Geber, A., Hitchcock, C.A., Swartz, J.E., Pullen, F.S., Marsden, K.E., Kwon-Chung, K.J., and Bennett, J.E. (1995). Deletion of the Candida glabrata ERG3 and ERG11 genes: effect on cell viability, cell growth, sterol composition, and antifungal susceptibility. Antimicrob. Agents Chemother. *39*, 2708–2717.

30. Kaur, R., Castaño, I., and Cormack, B.P. (2004). Functional genomic analysis of fluconazole susceptibility in the pathogenic yeast Candida glabrata: roles of calcium signaling and mitochondria. Antimicrob. Agents Chemother. *48*, 1600–1613.

31. Anderson, J.B., Sirjusingh, C., Parsons, A.B., Boone, C., Wickens, C., Cowen, L.E., and Kohn, L.M. (2003). Mode of selection and experimental evolution of antifungal drug resistance in Saccharomyces cerevisiae. Genetics *163*, 1287–1298.

32. Cowen, L.E., Sanglard, D., Calabrese, D., Sirjusingh, C., Anderson, J.B., and Kohn, L.M. (2000). Evolution of drug resistance in experimental populations of Candida albicans. J. Bacteriol. *182*, 1515–1522.

33. Tsai, H.-F., Krol, A.A., Sarti, K.E., and Bennett, J.E. (2006). Candida glabrata PDR1, a transcriptional regulator of a pleiotropic drug resistance network, mediates azole resistance in clinical isolates and petite mutants. Antimicrob. Agents Chemother. *50*, 1384–1392.

34. Looi, C.Y., D' Silva, E.C., Seow, H.F., Rosli, R., Ng, K.P., and Chong, P.P. (2005). Increased expression and hotspot mutations of the multidrug efflux transporter, CDR1 in azole-resistant Candida albicans isolates from vaginitis patients. FEMS Microbiol. Lett. *249*, 283–289.

35. Vale-Silva, L.A., Moeckli, B., Torelli, R., Posteraro, B., Sanguinetti, M., and Sanglard, D. (2016). Upregulation of the Adhesin Gene EPA1 Mediated by PDR1 in Candida glabrata Leads to Enhanced Host Colonization. mSphere *1*, e00065-15.

36. Todd, R.T., and Selmecki, A. (2020). Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. eLife *9*, e58349.

37. Pappas, P.G., Kauffman, C.A., Andes, D.R., Clancy, C.J., Marr, K.A., Ostrosky-Zeichner, L., Reboli, A.C., Schuster, M.G., Vazquez, J.A., Walsh, T.J., et al. (2015). Clinical Practice Guideline for the Management of Candidiasis: 2016 Update by the Infectious Diseases Society of America. Clin. Infect. Dis. *62*, e1–e50.

38. Hatwig, C., Balbueno, E.A., Bergamo, V.Z., Pippi, B., Fuentefria, A.M., and Silveira, G.P. (2019). Multidrug-resistant Candida glabrata strains obtained by induction of anidulafungin resistance in planktonic and biofilm cells. Braz. J. Pharm. Sci. *55*, https://doi.org/10.1590/s2175-97902019000218025.

39. Kartsonis, N., Killar, J., Mixson, L., Hoe, C.-M., Sable, C., Bartizal, K., and Motyl, M. (2005). Caspofungin susceptibility testing of isolates from patients with esophageal candidiasis or invasive candidiasis: relationship of MIC to treatment outcome. Antimicrob. Agents Chemother. *49*, 3616–3623.

40. Cowen, L.E., Sanglard, D., Howard, S.J., Rogers, P.D., and Perlin, D.S. (2014). Mechanisms of Antifungal Drug Resistance. Cold Spring Harb. Perspect. Med. *5*, a019752.

41. Kelly, S.L., Lamb, D.C., Kelly, D.E., Manning, N.J., Loeffler, J., Hebart, H., Schumacher, U., and Einsele, H. (1997). Resistance to fluconazole and cross-resistance to amphotericin B in Candida albicans from AIDS patients caused by defective sterol delta5,6-desaturation. FEBS Lett. *400*, 80–82.

42. Martel, C.M., Parker, J.E., Bader, O., Weig, M., Gross, U., Warrilow, A.G.S., Rolley, N., Kelly, D.E., and Kelly, S.L. (2010). Identification and characterization of four azole-resistant erg3 mutants of Candida albicans. Antimicrob. Agents Chemother. *54*, 4527–4533.

43. Morio, F., Pagniez, F., Lacroix, C., Miegeville, M., and Le Pape, P. (2012). Amino acid substitutions in the Candida albicans sterol Δ5,6-desaturase (Erg3p) confer azole resistance: characterization of two novel mutants with impaired virulence. J. Antimicrob. Chemother. *67*, 2131–2138.

44. Rybak, J.M., Dickens, C.M., Parker, J.E., Caudle, K.E., Manigaba, K., Whaley, S.G., Nishimoto, A.T., Luna-Tapia, A., Roy, S., Zhang, Q., et al. (2017). Loss of C-5 Sterol Desaturase Activity Results in Increased Resistance to Azole and Echinocandin Antifungals in a Clinical Isolate of Candida parapsilosis. Antimicrob. Agents Chemother. *61*, e00651-17.

45. Papp, C., Bohner, F., Kocsis, K., Varga, M., Szekeres, A., Bodai, L., Willis, J.R., Gabaldón, T., Tóth, R., Nosanchuk, J.D., et al. (2020). Triazole Evolution of Candida parapsilosis Results in Cross-Resistance to Other Antifungal Drugs, Influences Stress Responses, and Alters Virulence in an Antifungal Drug-Dependent Manner. mSphere *5*, e00821-20.

46. Carolus, H., Pierson, S., Muñoz, J.F., Subotić, A., Cruz, R.B., Cuomo, C.A., and Van Dijck, P. (2020). Genome-wide analysis of experimentally evolved Candida auris reveals multiple novel mechanisms of multidrug-resistance. mBio *12*, e03333-20.

47. Lesage, G., and Bussey, H. (2006). Cell wall assembly in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. *70*, 317–343.

48. Schwarzmüller, T., Ma, B., Hiller, E., Istel, F., Tscherner, M., Brunke, S., Ames, L., Firon, A., Green, B., Cabral, V., et al. (2014). Systematic phenotyping of a large-scale Candida glabrata deletion collection reveals novel antifungal tolerance genes. PLoS Pathog. *10*, e1004211.

49. Lawless, C., Wilkinson, D.J., Young, A., Addinall, S.G., and Lydall, D.A. (2010). Colonyzer: automated quantification of micro-organism growth characteristics on solid agar. BMC Bioinformatics *11*, 287.

50. Hovhannisyan, H., Hafez, A., Llorens, C., and Gabaldón, T. (2020). CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies. Bioinformatics *36*, 925–927.

51. Modi, A., Vai, S., Caramelli, D., and Lari, M. (2021). The Illumina Sequencing Protocol and the NovaSeq 6000 System. Methods Mol. Biol. *2242*, 15–42.

52. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

54. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

55. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178.

56. Garrison, E., Kronenberg, Z.N., Dawson, E.T., Pedersen, B.S., and Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. bioRxiv. https://doi.org/10.1101/2021.05.21.445151.

57. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

58. Sievert, C. (2019). Interactive web-based data visualization with R, plotly, and shiny. https://plotly-r.com.

59. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

60. Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., and Papenfuss, A.T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res *27*, 2050–2060.

61. Schröder, J., Wirawan, A., Schmidt, B., and Papenfuss, A.T. (2017). CLOVE: classification of genomic fusions into structural variation events. BMC Bioinformatics *18*, 346.

62. Bódi, Z., Farkas, Z., Nevozhay, D., Kalapis, D., Lázár, V., Csörgő, B., Nyerges, Á., Szamecz, B., Fekete, G., Papp, B., et al. (2017). Phenotypic heterogeneity promotes adaptive evolution. PLoS Biol. *15*, e2000644.

63. Vale-Silva, L., Beaudoing, E., Du, T., Tran, V., and Sanglard, D. (2017). Comparative Genomics of Two Sequential Candida glabrata Clinical Isolates. G3 (Bethesda) *7*, 2413–2426.

64. Arendrup, M.C., Cuenca-Estrella, M., Lass-Flörl, C., and Hope, W.; EUCAST-AFST (2012). EUCAST technical note on the EUCAST definitive document EDef 7.2: method for the determination of broth dilution minimum inhibitory concentrations of antifungal agents for yeasts EDef 7.2 (EUCAST-AFST). Clin. Microbiol. Infect. *18*, E246–E247.

65. Zomorodian, K., Bandegani, A., Mirhendi, H., Pakshir, K., Alinejhad, N., and Poostforoush Fard, A. (2016). In Vitro Susceptibility and Trailing Growth Effect of Clinical Isolates of Candida Species to Azole Drugs. Jundishapur J. Microbiol. *9*, e28666.

66. Rueda, C., Puig-Asensio, M., Guinea, J., Almirante, B., Cuenca-Estrella, M., and Zaragoza, O.; CANDIPOP Project from GEIH-GEMICOMED (SEIMC) and REIPI (2017). Evaluation of the possible influence of trailing and paradoxical effects on the clinical outcome of patients with candidemia. Clin. Microbiol. Infect. *23*, 49.e1–49.e8.

67. Marcos-Zambrano, L.J., Escribano, P., Sánchez-Carrillo, C., Bouza, E., and Guinea, J. (2016). Scope and frequency of fluconazole trailing assessed using EUCAST in invasive Candida spp. isolates. Med. Mycol. *54*, 733–739.

68. Thompson, G.R., 3rd, Wiederhold, N.P., Vallor, A.C., Villareal, N.C., Lewis, J.S., 2nd, and Patterson, T.F. (2008). Development of caspofungin resistance following prolonged therapy for invasive candidiasis secondary to Candida glabrata infection. Antimicrob. Agents Chemother. *52*, 3783–3785.

69. Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M., and Sherlock, G. (2017). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. *45* (D1), D592–D596.

70. Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., and Papenfuss, A.T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. *27*, 2050–2060.

71. Ferrari, S., Ischer, F., Calabrese, D., Posteraro, B., Sanguinetti, M., Fadda, G., Rohde, B., Bauser, C., Bader, O., and Sanglard, D. (2009). Gain of function mutations in CgPDR1 of Candida glabrata not only mediate antifungal resistance but also enhance virulence. PLoS Pathog. *5*, e1000268.

72. Tsai, H.-F., Sammons, L.R., Zhang, X., Suffis, S.D., Su, Q., Myers, T.G., Marr, K.A., and Bennett, J.E. (2010). Microarray and molecular analyses of the azole resistance mechanism in Candida glabrata oropharyngeal isolates. Antimicrob. Agents Chemother. *54*, 3308–3317.

73. Spettel, K., Barousch, W., Makristathis, A., Zeller, I., Nehr, M., Selitsch, B., Lackner, M., Rath, P.-M., Steinmann, J., and Willinger, B. (2019). Analysis of antifungal resistance genes in Candida albicans and Candida glabrata using next generation sequencing. PLoS ONE *14*, e0210397.

74. Grahl, N., Demers, E.G., Crocker, A.W., and Hogan, D.A. (2017). Use of RNA-Protein Complexes for Genome Editing in Non-*albicans Candida* Species. MSphere *2*, https://doi.org/10.1128/mSphere.00218-17.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Anidulafungin	CYMIT QUIMICA S.L.	Cat# 3D-FA16270-10
Fluconazole	SIGMA-ALDRICH QUIMICA S.L.	Cat# F8929-100MG
Caspofungin diacetate	SIGMA-ALDRICH QUIMICA S.L.	Cat# SML0425-5MG
Voriconazole	SIGMA-ALDRICH QUIMICA S.L.	Cat# PZ0005-5MG
Amphotericin B from Streptomyces sp.	SIGMA-ALDRICH QUIMICA S.L.	Cat# A4888-100MG
Flucytosine	SIGMA ALDRICH	Cat# PHR1659
Chloramphenicol	Merck Life Science S.L.U.	Cat# C1919-25G
Pfu Mix	DongSheng Biotech	Cat# P2022
Taq Mix, 1mlx5	DongSheng Biotech	Cat# P2012
Fluorescent Brightener 28 - Calcofluor White (1 g)	SIGMA-ALDRICH QUIMICA S.L.	Cat# F3543-1G
Congo Red	SIGMA-ALDRICH QUIMICA S.L.	Cat# C6277-25G
Hydrogen peroxide solution	SIGMA-ALDRICH QUIMICA S.L.	Cat# 16911-250ML-F
DTT, DL-DITHIOTHREITOL	Thermo Fisher Scientific	Cat# R0861
Sodium chloride, for molecular biology	PANREAC	Cat# A2942,1000
Sodium docecyl sulfate, SDS	PANREAC	Cat# A2263,0100
Methanol (Reag. Ph. Eur.) for analysis, ACS, ISO	PANREAC QUIMICA SLU	Cat# 1310911211
DMSO (Dimethyl sulfoxide), Sterile	Werfen España S.A.U.	Cat# 16712611S
MOPS	SIGMA-ALDRICH QUIMICA S.L.	Cat# M3183
RPMI-1640 (without HEPES and Sodium bicarbonate; with L-glutamine and phenol red)	SIGMA-ALDRICH QUIMICA S.L.	Cat# 51800035
DMSO (Dimethyl sulfoxide) for EUCAST	SIGMA-ALDRICH QUIMICA S.L.	Cat# W387520
Glucose monohydrate	Carl Roth GmbH + Co. KG	Cat# 6780.4
EtOH (Supelco)	MERCK	Cat# 1.00983.1011
Glycerin anhydrous/GLYCEROL 100% Molecular Biology grade	PANREAC	Cat# A2926,1000
T4 DNA polymerase	New England Biolabs	Cat# M0201L
dATP	New England Biolabs	Cat# N0440S
3′ −5′ -exo- Klenow fragment	New England Biolabs	Cat# M0212L
T4 DNA ligase	New England Biolabs	Cat# M0202L
Phusion DNA polymerase	Finnzymes	Cat# F530S
Sorbitol	SIGMA ALDRICH	Cat# S1876-500G
Tris hydrochloride	PANREAC	Cat# A3452
Lithium acetate	SIGMA ALDRICH	Cat# L4158
EDTA	SIGMA ALDRICH	Cat# E5134-500G
Critical commercial assays		
MasterPure Yeast DNA Purification Kit (200 Purif.)	BIONOVA CIENTIFICA S.A.	Cat# MPY80200
Genomic DNA clean & concentrator	ZYMO RESEARCH	Cat# D4011
QIAquick PCR purification kit	QIAGEN	Cat# 50928106
MinElute PCR Purification Kit	QIAGEN	Cat# 28004
Agilent High Sensitivity DNA Kit	AGILENT	Cat# 5067-4626
NEBNext Ultra II DNA library prep kit for Illumina	New England Biolabs	Cat# E7645L

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NEBNext® Multiplex Oligos for Illumina	New England Biolabs	Cat# E7335L
Qubit® dsDNA BR Assay Kit	INVITROGEN	Cat# Q32850
Qubit® dsDNA HS Assay Kit	INVITROGEN	Cat# Q32851
Deposited data		
Sequence data	This study	https://www.ncbi.nlm.nih.gov/sra/PRJNA635652
Experimental models: Organisms/strains		
Candida glabrata CST109	[20]	CST109
*Candida glabrata*CST 34	[20]	CST 34
Candida glabrata EB0911	[20]	EB0911
Candida glabrata CST78	[20]	CST78
Candida glabrata M12	[20]	M12
Candida glabrata EF1237	[20]	EF1237
Candida glabrata EF1620	[20]	EF1620
Candida glabrata F15	[20]	F15
Candida glabrata reference genome CBS138	[20]	CBS138
Candida glabrata P35_2	[20]	P35_2
Candida glabrata BG2	[20]	BG2
Candida glabrata SLL2 glab	This study	SLL2 glab
Recombinant DNA		
vector pTS50 with *NAT1* (Karl Kuchler lab)	[48]	pTS50
Software and algorithms		
qfa package (v0.0-44), R package	[49]	http://qfa.r-forge.r-project.org/
Crossmapper	[50]	https://github.com/Gabaldonlab/crossmapper
NovaSeq 6000 RTA 3.4.4	[51]	https://www.illumina.com
Burrows-Wheeler Alignment (v0.7.17)	[52]	http://bio-bwa.sourceforge.net/bwa.shtml
samtools (v1.9)	[53]	http://samtools.sourceforge.net/
fastqc (v0.11.8)	N/A	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
trimmomatic (v0.38)	[54]	http://www.usadellab.org/cms/?page=trimmomatic
picard (v2.18.26)	N/A	http://broadinstitute.github.io/picard/
GATK Haplotype Caller (v4.1.2)	[55]	https://github.com/broadinstitute/gatk
freebayes (v1.3.1)	N/A	https://arxiv.org/abs/1207.3907
bcftools (v1.9)	N/A	https://github.com/samtools/bcftools
vcfallelicprimitives from vcflib (v1.0.0)	[56]	https://github.com/vcflib/vcflib
ensembl Variant Effect Predictor (v96.3)	[55,57]	https://useast.ensembl.org/info/docs/tools/vep/index.html
python plotly package (v2.7)	[58]	https://plotly-r.com
Pipeline for small variant and CNV calling	This study	https://github.com/Gabaldonlab/VarCall_Cglabrata_IVevolution.
mosdepth (v0.2.6)	[59]	https://github.com/brentp/mosdepth
gridss (v2.8.1)	[60]	https://github.com/PapenfussLab/gridss
clove (v0.17)	[61]	https://www.github.com/PapenfussLab.
perSVade pipeline (v0.0)	N/A	https://github.com/Gabaldonlab/perSVade
python scipy.stats (v1.5.2)	N/A	http://www.scipy.org

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Optimase Protocol Writer	N/A	http://www.mutationdiscovery.com/md/MD.com/screens/optimase/OptimaseInput.html?action=none
Libre Office (v6.0.7.3)	N/A	https://www.libreoffice.org
Graphpad Prism (v8.4.2)	N/A	https://www.graphpad.com
Oligonucleotides		
Oligonucleotides used in this study—see Table S5	N/A	N/A
Other		
Sandwich cover	Enzyscreeen BV	Cat# CR1296
MegaBlock 96 Well 2.2 ml Plates	Sarsted	Cat# 82.1972.002
Nunc OmniTray	Life Technologies	Cat# 242811
3mm glass beads	SIGMA-ALDRICH QUIMICA S.L.	Cat# 1040150500
Microplate, 96 well, PS, F-BOTTOM, clear, sterile, 2 PCS./BAG	Greiner Bio-One North America, Inc.	Cat# 655161
Lid, PS, High Profile (9 MM), clear, sterile, single packed	Greiner Bio-One North America, Inc.	Cat# 656161

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Toni Gabaldon (toni.gabaldon@bsc.es).

Materials availability

Material generated in this study is available upon request from the lead contact.

Data and code availability

The raw sequencing data of the whole genomes have been deposited in the Short Read Archive (SRA) database, with accession number PRJNA635652 (SRA: PRJNA635652) and are publicly available as of the date of publication. The DOI is listed in the key resources table.

All the code for calling small and structural variants can be found in the repositories https://github.com/Gabaldonlab/VarCall_Cglabrata_IVevolution and https://github.com/Gabaldonlab/perSVade and are publicly available. The DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported here is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Microbe strains

The 12 strains of *C. glabrata* used in this study are listed in the key resources table. Eleven clinical strains had been previously analyzed for several phenotypic characteristics, including susceptibility to various drugs.[20] In addition, they have been shown to belong to seven genetically distinct clades. The remaining strain (SLL2_glab) was isolated from an oral wash of a healthy individual from Spain, and can thus be considered commensal. SLL2 glab was sequenced within this project and assigned to clade 7.

Galleria mellonella

Unsexed *Galleria mellonella* larvae were purchased from DNAT ecosistemas (https://www.dnatecosistemas.es).

METHOD DETAILS

In vitro evolution

We conducted experimental evolution experiments using a batch serial transfer approach[62] (Figure 1). Wild-type (WT) strains were collected from glycerol stocks, plated, left to grow until single colonies could be detected and re-plated again for an overnight culture (YPD agar plate at 37°C). A few colonies were suspended in sterile water and diluted to 2.5 × 10^5 colony forming units per milliliter (CFU/mL). A 96 deep-well plate (2.2mL) with 450 μL of YPD – the master plate - was inoculated with 50 μL of the cell suspension in four

replicates for each strain. To ensure lack of cross contamination the inoculations were organized using a checkerboard design (Figure 1) and visually inspected for unwanted growth in non-inoculated wells. Each well of the deep well plate also included a glass bead to ensure proper oxygen transfer and prevent the samples from sedimentation. The master plate was covered with a sandwich cover (Enzyscreeen BV) to ensure optimal oxygenation and limit evaporation. It was then shaken at 300 rpm, and incubated at 37°C for 72 h. Afterward, 50 μL of each culture was transferred to a fresh 450 μL of YPD medium and left again to grow in the same conditions. Next, 50 μL of samples from the master plate were distributed into four independent 96-well plates containing 450 μL of YPD medium supplemented with the following: 1) an echinocandin: anidulafungin (drug: ani, outcome samples: ANI); 2) an azole: fluconazole (flz, FLZ); 3) anidulafungin and fluconazole (aniflz, ANIFLZ); or 4) no drug (YPD). Adaptation to the drugs involved passages of the (50 μL) samples to a fresh (450 μL) medium every 3 days, and in every second passage the concentrations of flz and ani were gradually increased from 4 μg/mL and 0.016 μg/mL to 192 μg/mL and 4 μg/mL, respectively (Table S4), except YPD where no change in the composition of the medium was applied. For each passage the medium with antifungals was freshly made on the same day using a frozen stock of the drugs. Before each increase in drug concentration, part of the culture was frozen and stored at −80°C (100 μL of the sample in 100 μL of 50% glycerol). All in all, the experiment involved 6 days of adaptation to the same conditions before increasing the stress, and further adaptation. Starting with 4 μg/mL flz and 0.016 μg/mL ani, the experiments finished after 54 days, 18 passages with drugs, and 9 increments in drug concentrations. We estimate this period to involve between 60 to 500 generations (assuming a minimum of three doublings per passage in a 1:10 dilution and a maximum of 5-10 generations/day based on earlier studies[63]). From the last passage we selected, stored and analyzed single colonies that were picked from agar plates and regrown on liquid medium supplemented with the last concentrations of the drugs used in each condition. In the second part of the experiment, we repeated the evolution experiment, this time evolving ANI isolates in flz (AinF), and FLZ isolates in ani (FinA), using the same regimes as explained above. Due to the inability to re-grow two samples (1 ANI and 1 FinA) from the glycerol stock, and several extinct populations in the simultaneous treatment with 2 drugs, the total number of analyzed samples was as follows: 48 FLZ, 47 FinA, 47 ANI, 48 AinF, 21 ANIFLZ and 48 YPD. The growth of the samples was visually assessed by their capacity to grow at the last drug concentration(s) after 4 × 3-day long passages in YPD medium without drugs.

Susceptibility tests

Susceptibility to flz and ani was studied in a high-throughput manner using a robot, and recording not only the endpoints but also the growth curves of all drug dilution assays over at least 18h. Susceptibility tests were performed in at least three replicates following the EUCAST E.DEF 7.3.1. protocol.[64] Briefly, isolates were pinned on agar containing RPMI with 2% glucose buffered with MOPS (3-(N-morpholino)propanesulfonic acid) and grown at 37°C. Fresh overnight cultured strains were adjusted to 2-10 x10^5 CFU/mL in distilled water. Next 50 μL of broth was then added to 150 μL antifungal solution (in RPMI /w MOPS) and incubated at 37°C. OD600nm was measured every 60 - 90 min and growth was evaluated after around 18h. The range of concentrations tested was 16-0.016 μg/mL for ani 256-0.25 μg/mL for flz, following EUCAST guidelines .

Fitness and susceptibility measurements

For each sample at each drug concentration, fitness was measured as the area under the time-versus-optical density curve (hereafter referred as fAUC, calculated with the qfa package (v0.0-44 http://qfa.r-forge.r-project.org/). Minimum Inhibitory Concentration 50 (MIC_{50}) values were calculated as the minimum concentration where the fAUC relative to the no-drug control was below 50%. If 50% of the inhibition was not met within the tested concentration range, then MIC was set to twice the maximum assayed concentration for numerical analyses in Figures 6 and S1. We also define rAUC as the area under the drug concentration-versus-relative fitness curve (AUC), normalized by the maximum AUC_{MAX} where there is no change in fitness across the entire range of concentrations (Figure 2A). rAUC was used as a proxy for the quantitative levels of resistance for each sample. To filter out experimental artifacts, we kept the three technical replicates that were closest to the median for each sample and measure (fitness, relative fitness, MIC and rAUC).

To correct for intraspecific fitness differences,[20] we based our fitness analysis (see Results) on the log_2-ratio between the fAUC of each sample and the unevolved WT strain. This value was used as a proxy for fitness changes occurring during the experiment. Under the same reasoning, we defined strains with acquired resistance as those where the MIC was more than 2 times the WT MIC. This threshold separated our samples clearly into susceptible and resistant strains (Figures S1C and S1D). All the fitness and susceptibility measurements are in Data S1. Doubling rate per hour was inferred from the maximum slope in the time-versus-log_2 (OD) data using bins of 3 time points for the analysis of EF1620_7B_ANI (see below).

Analysis of MIC and rAUC measures of antifungal drug resistance

As discussed in the main text, both MIC and rAUC measurements were correlated (Figures S1A and S1B). However, they presented several important differences that we discuss here in more detail. First of all, MIC values presented clearer increments and a bimodal distribution, making it easier to define thresholds for resistant versus susceptible samples as compared to rAUC (Figures 2B and S1D). Accordingly, we used MIC values to define resistant samples. In addition, although measurement errors are similar (Figures S1A and S1B) MIC is more consistent across independently evolved strains of the same condition (Figures 2B and S1C). However, rAUC values provided a continuous estimate of resistance, which is better suited for quantitative analyses (such as those of Figures 2E and 7C). Importantly, rAUC was not affected by the trailing effect. This effect occurs when total growth inhibition is not achieved with increasing concentration of the drug, but rather cell densities are maintained. This effect has been reported with azoles and *Candida* species.[65–67] We observed this effect occurring in most (8/10) ANI samples with *ERG3* mutations, leading to high MIC values

that were in the range of FLZ samples (Figure S7C). The rAUC values, however, were not affected by the trailing effect and these strains presented flz rAUC values intermediate between flz non-resistant ANI and flz-resistant FLZ samples (Figures 2B and 7C). Conversely, there is one sample (BG2_11H_ANI) bearing an *ERG3* premature termination codon and presenting a mismatch between flz MIC and rAUC. Although MIC is in the WT range, visual inspection of the flz concentration-versus-fitness curve showed a trailing effect around 50% of growth (Figure S7C), implying increased resistance. This is consistent with the observed high rAUC (Figures 7C and S7A). Taken together, these examples suggest that rAUC captures better the quantitative landscape of drug resistance.

Finally, we found another sample (EF1620_7B_ANI) where neither MIC nor rAUC captured the true nature of flz resistance. This sample shows a non-monotonic relationship between flz concentration and relative fitness (Figures S7C and S7D). This motivated us to analyze this sample under another fitness estimate, the doubling rate per hour (DR), in addition to fAUC. We found that this sample had low fitness (by both fAUC and DR) in the absence of the drug, with a small increase in the lower flz concentrations. This low level of basal fitness results in high relative fitness at low drug concentrations (as compared to other samples) (Figures S7C and S7D). This analysis suggests that this non-monotonic relationship (if present) is very weak in terms of absolute fitness. This example illustrates how MIC and rAUC values can be misleading in strains with very low basal fitness.

DNA extraction

A modified protocol from the MasterPure Yeast DNA Purification Kit was used to extract DNA. In brief, samples were grown overnight in liquid YPD at 37°C. Cells were pelleted and lysed with RNase treatment at 65°C for 15 min. After 5 min of cooling down on ice, samples were purified by the kit reagent by mixing, centrifugation and removal of the debris as described in the kit protocol. Further, samples were left at −20°C with absolute ethanol for at least 2 h after which the DNA was precipitated for 30 min at 4°C. The pellet was washed in 70% ethanol and left to dry. TE buffer was used to resuspend the DNA. The Genomic DNA Clean & Concentrator kit (Zymo Research) was used for the final purification.

Target *FKS* and *ERG3* sequencing

All ani-exposed samples (ANI, ANIFLZ and FinA) were examined for mutations in one region of *FKS1* and two regions of *FKS2* encompassing echinocandin resistance mutational HSs.[5] Three samples without mutations in the above-mentioned HSs were also inspected in the HS2 of *FKS1*. All the new *FKS* mutations are in Data S2. We used PCR primers described earlier[68] (Table S5). ANI samples not subjected to WGS were also amplified by two PCRs with two sets of primers (Table S5) to obtain *ERG3* sequences. PCRs were carried out by using Taq DNA polymerase from DongShengBio. The reaction mixture included primers of concentration of 0.4 μM, 20 μL Taq DNA polymerase, 1 μL liquid sample grown for 24-48 h in YPD and water up to a final volume of 40 μL. Optimase ProtocolWriter was used to develop conditions for each primer set.

We tested for the possible trajectories of final *FKS* and *ERG3* mutations in the 10 ANI samples subjected to WGS and presenting *ERG3* alterations to infer which might have appeared first in the evolution. We selected and analyzed single colonies from our glycerols stocks of stored populations after the 2nd passage at 0.032, 0.064, 0.128 and 0.256 ug/ml ani (beginning of the adaptation). PCRs were carried out as described above.

Petite phenotype in ani adapted mutants

10 ANI samples that underwent WGS and show changes in *ERG3* gene, CBS138 WT and *Saccharomyces cerevisiae* petite control were inspected for presenting a petite phenotype. Samples were grown on YPD (1% yeast extract, 2% bactopeptone, 2% glucose) and YPG (1% yeast extract, 2% bactopeptone, 2% glycerol) for 24h-48h.

Whole genome sequencing

Evolved mutants: Genome sequences were obtained at the Ultra-sequencing core facility of the CRG, using Illumina HiSeq 2500 sequencing machines, and as previously described.[20] In brief, libraries of paired-end, 125 bases-long reads were prepared. The DNA was fragmented by nebulization or in Covaris to a final size of ~600 bp. After shearing, the ends of the DNA fragments were blunted with T4 DNA polymerase and the Klenow fragment (New England Biolabs). DNA was purified using QIAquick PCR purification kit (QIAGEN). 3′-adenylation was performed by incubation with dATP and the 3′-5′-exo-Klenow fragment (New England Biolabs). DNA was purified using MinElute spin columns (QIAGEN) and double-stranded Illumina paired-end adapters were ligated to the DNA using rapid T4 DNA ligase (New England Biolabs). After another purification step, adaptor-ligated fragments were enriched, and adapters were extended by selective amplification in an 18-cycle PCR reaction using Phusion DNA polymerase (Finnzymes). Libraries were quantified and loaded into Illumina flow-cells at concentrations of 7–20 pM. Cluster generation was performed in an Illumina cluster station. Sequence runs of 2 × 100 cycles were performed on the sequencing instrument. Base calling was performed using Illumina pipeline software. In multiplexed libraries, we used 4 bp internal indexes (5 indexed sequences). De-convolution was performed using the CASAVA software (Illumina). Sequence data of the genomes have been deposited in the Short Read Archive (SRA) database, with accession number PRJNA635652 (SRA: PRJNA635652).

The genome of the CRISPR 3H_ANI with *ERG3*(D122Y) sample was pooled with two genomes from divergent species (*Candida albicans* and *Candida parapsilosis*), after confirming with Crossmapper[50] the absence of read cross-mapping in the chosen sequencing design. Sequencing libraries were made at the Functional Genomics Core Facility at the IRB and genome sequences were obtained at the sequencing core facility of the CNAG. 500-1,000 ng of genomic DNA dissolved in a final volume of 50 ul TE buffer were sheared with a Bioruptor sonicator (Diagenode) using the following settings: temperature 4-10°C; intensity: high; cycles: 3; cycle

time: 5 minutes; cycle program: 30 s pulse and 30 s rest time. At the end of each sonication cycle samples were centrifuged at 4°C and the water tank was refilled with pre-cooled water. DNA fragmentation was quality controlled using the Bioanalyzer 2100 and its DNA High Sensitivity chip (Agilent) and quantified using the Qubit fluorometer and its dsDNA HS assay (Invitrogen). NGS libraries were prepared from 250 ng of fragmented DNA using the NEBNext Ultra II DNA library prep kit for Illumina (New England Biolabs). Adaptor-ligated DNA were size-selected using the provider-recommended settings to obtain an insert size distribution of 300-400 bp. After purification, libraries were amplified through five PCR cycles using the NEBNext multiple oligos for Illumina (New England Biolabs). The final libraries were quantified on Qubit and quality controlled in the Bioanalyzer. An equimolar pool was prepared with the six libraries and submitted for sequencing at the Centre Nacional d'Anàlisi Genòmica (CRG-CNAG).The libraries were sequenced on NovaSeq 6000 (Illumina) with a paired-end read length of 2x150 bp. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (NovaSeq 6000 RTA 3.4.4). To select the *C. glabrata* sequencing reads we used *Burrows-Wheeler Alignment* (bwa v0.7.17) *mem* (http://bio-bwa.sourceforge.net/bwa.shtml) to align the reads to a concatenated reference genome including the three pooled species. We took the reference genomes from the Candida Genome Database[69] (v_s02-m07-r35 for *C. glabrata* and haplotype A of v_s07-m01-r110 for *C. albicans*) and the NCBI (sequence GCA_000182765.2 for *C. parapsilosis*). We next separated the reads uniquely mapping to *C. glabrata* with *samtools* (v1.9[53]), which yielded the final whole-genome sequencing dataset.

Small variant calling and interpretation

For each library, we first performed quality control of the reads with *fastqc* (v0.11.8, https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and trimming with *trimmomatic* (v0.38[54]). The trimmed reads were aligned against the reference *C. glabrata* genome (the latest version by 12/03/2019, which is v_s02-m07-r35 from the Candida Genome Database[69] (CGD: v_s02-m07-r35)) using *Burrows-Wheeler Alignment* (bwa v0.7.17) *mem* (http://bio-bwa.sourceforge.net/bwa.shtml). In addition, indexing of the genome and construction of a sequence dictionary was performed with *samtools* (v1.9[53]) and *picard* (v2.18.26 http://broadinstitute.github.io/picard/), respectively. We next used three different algorithms (*GATK Haplotype Caller* (HC) (v4.1.2[55]), freebayes (FB) (v1.3.1 https://arxiv.org/abs/1207.3907) and bcftools (BT) (v1.9, https://github.com/samtools/bcftools) to call and filter Single Nucleotide Polymorphisms (SNP) and small insertions/deletions (IN/DEL) in both haploid and diploid configurations. We defined as high-confidence (PASS) variants those with read depth above 20, with extra filters for HC and FB. For HC, we kept as PASS variants those where 1) there were less than four additional variants within 20 bases; 2) the mapping quality was above 40; 3) the confidence based on depth was above 2; 4) the phred-scaled p value was below 60; 5) the MQRankSum was above −12.5 and 6) the ReadPosRankSum was above −8. For FB, we kept as PASS variants those where 1) quality was above 1 or alternate allele observation count was above 10; 2) strand balance probability of the alternate allele was above 0; 3) number of observations in the reverse strand was above 0; and 4) number of reads placed to the right/left of the allele were above 1. We further used *vcfallelicprimitives* from *vcflib*[56] (v1.0.0 https://github.com/vcflib/vcflib) to uniformize the called variants across the three algorithms, and the *ensembl Variant Effect Predictor* (v96.3[57]) to annotate the potential functional effect of each variant in both coding and non-coding regions. In addition, we developed a tool to visualize (and better interpret) the genomic location of each variant across multiple samples using the *python plotly* package[58] (v2.7). This pipeline is ready to use for any paired-end short-read sequencing library at https://github.com/Gabaldonlab/VarCall_Cglabrata_IVevolution.

We considered PASS variants to be those SNPs that passed the filtering of the three algorithms and those INDELs that passed both HC/FB filters (which were shown to have highest overlap). For each sample evolved in drug conditions, we defined variants newly-acquired during the experiment to be those that were not called in any of the corresponding WT and YPD samples. We ran this variant calling pipeline in both haploid and diploid configurations for all samples. Diploid variants may have appeared in regions that are under whole-chromosome duplications. We keep only as true ''heterozygous'' or ''homozygous'' diploid variants as those that appear to be like this by all the programs tested and within a duplicated chromosome (see below). All the new small variants are found in Data S3. In addition, Table S1 includes the variants shared between CST109 and M12 and absent in the other representatives of their clades.

Identification of large aneuploidies, segmental duplications and deletions

To detect genes affected by CNV, we calculated the read depth for each gene relative to the median read depth per gene across all nuclear chromosomes that did not have signs of large duplications (see Results) (hereafter referred to as relative coverage). The read depth was calculated using *mosdepth* (v0.2.6[59]). We then defined deleted genes as those with > 50% of their length not covered by reads. To keep only gene deletions appearing during the experiment we further filtered out genes that were also lost in the corresponding WT or with a relative coverage below 0.1 in YPD-evolved sample (which may suggest a loss also in the WT or in the YPD). We manually curated the deletion list to find regions potentially deleted in a previous sample of the evolution experiment, which was the case of a small region in chromosome D (including *CNE1*, with a relative coverage below 0.1 in EF1620_7B_ANI) and the *S. cerevisiae GPB2* ortholog (with a relative coverage below 0.1 in EF1620_7B_ANI). Importantly, these two genes were lost in a single genomic rearrangement (see below, Figure S3).

CNV was defined by calculating the $\log_2$ ratio between the relative coverage of each sample against the matching YPD (log2cov_vsYPD). Copy-number (CN) increase refers to log2cov_vsYPD above 1 and a relative coverage above 1.8, while CN decrease refers to log2cov_vsYPD below −1 and a relative coverage of the corresponding YPD above 1.8. The rationale of this filtering was to detect genes lost and under CNV during drug exposure, correcting for intrinsic biases in per-gene coverage. As noted in other studies, we

found that relative coverage was correlated with the distance to the telomere (hereafter referred as "smiley-pattern"), which may be an artifact of library preparation and/or sequencing, with this effect varying across samples. We hypothesize that this is partially why most of the CNV was found in subtelomeric regions (defined here as the first and last 50 genes of a chromosome). We thus filtered out any CNV call that was not supported by equivalent genomic rearrangements (see below). In addition, chromosomes with large aneuploidies were defined as those where we consistently observe genes with increased CN and relative coverage around 2x across a region spanning at least 10% of the non-subtelomeric chromosome (Figure 5A).

Analysis of genomic rearrangements

To identify GR we implemented an algorithm that uses split-reads, discordantly aligned read-pairs and *de novo* assembly evidence to call genomic breakpoints and interpret the resulting GRs and CNVs. Breakpoints were called using gridss (v2.8.1[70]) and integrated into complex structural variation with clove (v0.17[61]). The straightforward implementation of this pipeline was challenging because of the lack of established parameters for yeast genomes, and the "simley-pattern" bias (see above) impeding the use of a single read-depth threshold for filtering deletions and tandem duplications (used by clove). We thus chose the running and filtering parameters from a simulation-based optimization implemented in the *perSVade* pipeline (v0.0, https://github.com/Gabaldonlab/perSVade).

GR appearing during the experiment were defined as those where none of the breakends (each of the ends of a breakpoint) matched a breakend in any of the parents (with an overlap of less 200 bp), in a way that resembles the small variant calling (see above). This is an extremely conservative approach (as most called breakends in the parents may be false positives) to ensure high confidence in our final set of variants. In addition, we defined "haploid breakends" as those with an allele frequency (AF) above 0.75 and "heterozygous breakends" as those with an AF > 0.25. We also filtered out tandem duplications, inversions and deletions where any of the breakends was not haploid, as these variants can not yield heterozygous breakends in haploid chromosomes. Note that we did not detect any such heterozygous events in the aneuploid chromosomes. Furthermore, we manually curated the results to identify errors in the summarization of breakpoints into complex rearrangements. This approach yielded one sample (P35_10E_FinA) with two reciprocal inverted interchromosomal breakpoints between close positions (less than 200 bp apart) of chromosome (Chr) G (breaking the CDS of *FKS1*) and ChrM. These were called as two independent unbalanced translocations, but we interpret them as an inverted balanced translocation between the two chromosomes. The coverage "smiley-pattern" was also consistent with this model.

To focus on resistance-conferring events, we examined genes with ns mutations or nearby GR (within less than 2kb) appearing recurrently (at least twice) in our experiment. These included *ERG3*, *FKS1* and the ortholog of *S. cerevisiae CNE1*, mentioned in the main text (see Results). We confirmed all these rearrangements through PCR (see below). Regarding *ERG3*, we found one ANIFLZ sample with a deletion at the beginning of the CDS and a FinA sample with a deletion in the 5′ region (potentially spanning the promoter, and related to the loss of *CAGL0F01815 g* (see Results). Both of these were associated with low relative coverage (< 0.01) spanning the breakpoint, which further confirmed these deletions (Figure S3A). These are additional *ERG3* mutations potentially related to ani exposure. We also found an inter-chromosomal breakpoint between ChrD and ChrL in EF1620_7B_ANI with the orientation of a deletion breakpoint. Importantly, the WT strain underwent a balanced translocation between these chromosomes (as compared to the reference genome), which means that the alteration appearing upon drug exposure was actually a deletion event (also confirmed by coverage). The deleted region included *CNE1*, which may be related to ani adaptation (see Results) (Figure S3B). This also constitutes an example of how the rearrangements found in each strain modulate the interpretation of breakpoints appearing during the experiment. Finally, we found two FinA samples with GR breaking the *FKS1* coding region, including one deletion at the beginning of the coding sequence (with relative coverage < 0.01) and one balanced inverted translocation between ChrG and ChrM (Figure S3C). Both samples carried *FKS2* mutations (potentially conferring ani resistance), suggesting that these rearrangements are complementary *FKS1* alterations with a similar impact as the truncating small variants mentioned in the main text.

On another note, we attempted to infer the precise events leading to partial aneuploidies during the experiment (Figures 5A and 6). We found an unbalanced translocation explaining the partial duplication of ChrE in CBS138_9F_FLZ. Our GR-detection method predicted that the right arm of ChrE (matching the aneuploid region (Figure 5A) was duplicated and attached to ChrJ, replacing the left-end at the breakpoint. This region showed low coverage after the breakpoint (supporting the unbalanced translocation call), but not until the end of the chromosome (which would be expected from such an event). Interestingly, the deleted region was found between the unbalanced translocation breakpoint and a location with low WT coverage. We propose that this configuration is the result of a pre-existing rearrangement in the WT strain, which explains why the deleted region does not span the entire left-end of the chromosome. Accordingly, the ChrE breakend was called heterozygous, while the ChrJ was haploid (Figure S3D). Conversely, we could only find an inverted heterozygous breakpoint matching the start of the aneuploid region of ChrE in CST34_2A_AinF, which was not enough to explain the source of the duplication. Finally, we found that the (apparently) partial duplications of ChrI in the EB0911 samples are actually whole-chromosome aneuploidies. The WT EB0911 depicted balanced translocations between Chr D, I and L, generating three (mixed) chromosomes from the successive fusions. We found one of these mixed chromosomes with 2x coverage in both samples with aneuploidies (Figure S3E). Interestingly, this chromosome is much shorter than the reference ChrI, perhaps resulting in a lower fitness cost of this aneuploidy. We speculate that this is the reason why this aneuploidy is found only in this strain. Taken together, these results suggest that complex structural variation may contribute to drug resistance. They also show how breakpoint calling can explain the precise events leading to CNV and aneuploidies.

Presence of all the GRs discussed in the text was confirmed with PCR using primers specifically designed to provide amplicons only in the presence of the GR (translocations) or with a different size (deletions). Results are presented in Figure S3F. All events were positively confirmed. Primers used for each GR validation are presented in Table S5. PCRs were performed using Taq DNA

polymerase from DongShengBio. The reaction mixture included primers of concentration of 0.4 μM, 15 μL Taq DNA polymerase, 1 μL liquid sample grown for 24h in YPD and water up to a final volume of 30 μL. Optimase ProtocolWriter was used to develop conditions for each primer set.

Analysis of clinical isolates' sequencing datasets

We obtained all the variant calling files for publicly available whole genome sequences of *Candida* clinical isolates from the CandidaMine database (v1, https://candidamine.org, publication in progress). The MIC values for each sample were obtained by manual curation of the associated literature, when available.

In *Candida glabrata,* we could find these data in 126/393 clinical isolates, including resistance to fluconazole (flz 126/126), posaconazole (pos 84/126), voriconazole (vrz 91/126), isavuconazole (ivz 37/126), micafungin (mif 42/126), anidulafungin (ani 9/126) and caspofungin (cas 91/126). Some of these drugs lack established clinical resistance breakpoints, which did not allow a direct identification of resistant isolates. We thus, defined the resistance breakpoint for each drug as 2x the maximum MIC reported in a set of susceptible isolates (from Carreté et al.[20]). Ani susceptibility was not measured for these isolates, so that we took the standard EUCAST breakpoint to define ani resistance. This data is sparse, so that we do not always know the MIC values for all drugs in a given isolate. We thus, focused our analysis on "azole" or "echinocandin" resistance instead of splitting by individual drugs. In order to achieve this, we defined an isolate to be "resistant" to a given class of drugs if it was resistant to all the measured drugs of that class. This yielded 41/126 and 19/91 isolates resistant to all tested azoles or echinocandins, respectively. We could find two samples with resistance to both classes of drugs. In *Candida albicans,* we could find MIC data for 187/478 clinical isolates. We could define the resistance breakpoints according to EUCAST for all tested drugs but caspofungin. We defined an isolate to be cas-resistant if the MIC was above the percentile 90. This yielded 39/186 and 9/150 isolates resistant to all tested azoles or echinocandins, respectively. We could find one sample with resistance to both classes of drugs. Given the low numbers of samples with resistance to both drugs, we conclude that the available data is insufficient to perform analysis of cross-resistance or multidrug resistance. In order to assess whether the mechanisms driving single-drug resistance *in vitro* are clinically relevant, we first analyzed these publicly available sequences of *Candida* clinical isolates. We assessed how many of the drug resistance variants described in this work were also found in these clinical isolates, which yielded little or no overlap depending on the gene (Figure S5A). We hypothesized that the underlying reason is that several mutations in the same gene can explain drug resistance (Figure 6). In order to test this we calculated the overlap between CandidaMine variants and two datasets of previously described drug resistance-mutations: the SENTRY database[6] and a set of described *PDR1* mutations from the literature.[71–73] This yielded low overlaps as well, comparable to those found in our work (Figure S5A).

In addition, we inferred the expected overlap between different mutation datasets through a randomization strategy on our samples. We divided the samples carrying mutations in a given gene into two random subsets. For each subset, we calculated the number of mutations only in the subset or also found in the other subset. This process was repeated 100 times, and the results (Figure S5B) show that the overlap is comparable to the observed between datasets of different works.

We conclude that it is difficult to measure the clinical impact of the mutations described here because most of them cannot be found in the currently available isolates. However, this low overlap is expected and comparable to other datasets of well-known resistance-conferring mutations.

CRISPR-Cas9 based genetic modifications

Donor DNAs

Short fragment of *ERG3* with D122Y (G364T) was ordered from Integrated DNA Technologies, Inc.. This fragment of the gene also contained additional synonymous mutations in PAM region (short NGG sequence that follows the DNA region targeted for cleavage by the CRISPR system) to bypass recutting by the Cas9 once the donor DNA is integrated, hence to improve the number of positive transformations. A large donor DNA containing *ERG3* mutation (D122Y) was also amplified from 3B_ANI evolved sample by FL1_FWD and FL2_REV primers. All primers and the ordered sequence can be found in the supplementary information (Table S5). Two approaches were used to introduce *ERG3* mutations. The first approach involved the transformation of a fragment containing the *ERG3* alteration and assumption that the positive transformants would exhibit increased resistance to fluconazole, hence the transformation was followed by selection on agar plates containing fluconazole. Second approach involved creating a DNA construct containing *ERG3* gene fused with *NAT1* gene (upstream) as a selection marker. *NAT1* was amplified from a vector pTS50 (a kind gift from Karl Kuchler). Two of such donors were used. One contained *NAT1* fused with wild-type *ERG3* (amplified from DNA extracted from wild-type *Candida glabrata* strain) and second contained *NAT1* with *ERG3* bearing the mutation (D122Y, amplified from DNA of fluconazole and anidulafungin resistant evolved mutant (3B_ANI)). The first donor was used to examine the influence of the presence of *NAT1* on flz susceptibility as well to eliminate the mutation acquired during the evolution and check for the reversion of the phenotype. In this approach, *ERG3* with downstream region was amplified by PCR using FLKI_ERG3 set of primers from 2 strains: one containing wild-type *ERG3* and one containing the mutation (3B_ANI). Upstream *ERG3* region was amplified by FLKII set of primers. *NAT1* was amplified from a vector pTS50 by PCR and 'NAT1(for DNA donor constructs))' set of primers. All primers contain additional homologous sequences to ensure the fusion FLKI_ERG3:NAT1:FLKII. The fused fragments were gel purified and correct fusion was confirmed by PCR with internally placed primers –inside_ERG3_FWD and inside_NAT1_REV and inside_NAT1_FWD with flank_ERG3_REV.

CRISPR-based mutagenesis

CRISPR-based mutagenesis was performed using ribonucleoproteins (RNPs) and following a previously described method by Grahl et al.[74] RNPs were created using the Alt-R CRISPR-Cas9 system bought from Integrated DNA Technologies, Inc.). The CRISPR machinery included: purified Alt-R S.p. Cas9 Nuclease V3, and guide RNA containing universal transactivating Alt-R CRISPR-Cas9 tracrRNA and target specific crRNA (Table S5).

The synthetic *ERG3* fragment as well as the large donor DNA containing the *ERG3* mutation were transformed into 3H_ANI sample and selected on 64ug/ml flz. The same trial of transformations was done on CBS138 WT strain but the selection of positive transformants was unsuccessful. One of the positive transformants was subjected to whole genome sequencing to infer the presence of only inserted *ERG3* mutations and absence of additional protein altering mutations or CNVs (which could explain the resistance). In parallel, an alternative approach with improved selection was conducted.

ERG3 with *NAT1* were transformed into wild-type *Candida glabrata* strains CBS138, EB0911 and its anidulafungin resistant progenies: 3H_ANI and 3B_ANI mutants. The positive transformants were selected on YPD with 200 μg/ml nourseothricin. To ensure that the DNA donors were transformed in the correct place in the genome a PCR with a REV primer that falls outside of the designed constructs and a FWD primer that falls inside *NAT1* gene was performed – inside_NAT1_FWD with out_REV. Additionally, the insertion of the *NAT1* was examined by amplification of longer fragment when performing a PCR with primers surrounding the place of the insertion (inside_ERG3_FWD with flank_ERG3_REV). The PCR conditions were designed with OptimaseProtocol. Presence and absence of *ERG3* mutations was confirmed by Sanger sequencing.

All transformations were performed by electroporation of competent cells prepared using lithium acetate (LiAc). Overnight cultures were diluted to an optical density at 600nm (OD600) of 0.3 in 50 mL YPD and left to grow to obtain OD600 of approximately 1.6 to 2.2. Then cells were pelleted, washed once with 25ml of sterile water and resuspended in 10ml of a transformation buffer (100 mM LiAc, 10 mM Tris-HCl, 1 mM EDTA) and incubated with shaking for 1h. The cells were further incubated with shaking for 30min with 1ml of 1M dithiothreitol (DTT), washed twice with 40ml ice-cold water and once with 5ml ice-cold 1 M sorbitol before resuspension in 200 μL of ice-cold 1 M sorbitol.

CrRNAs and tracrRNA were first dissolved in RNase-free distilled water (dH2O) at 100 μM and stored at –20°C. The guide RNA was created by mixing equimolar concentrations (4 μM final) of the gene-specific crRNA and tracrRNA (to obtain a final volume of 3.6 μl per transformation) and incubating at 95°C for 5 min, followed by cooling down to room temperature. The Cas9 nuclease (60 μM stock from IDT) was diluted to 4 μM in dH2O to a volume of 3 μl per transformation. RNPs were assembled by mixing 3.6 μL of guide RNAs with 3 μl of diluted Cas9 protein, followed by incubation at room temperature for 5 min. Transformation of cells was carried out by electroporation of cell suspension containing: 40 μl of cells, 6.6 μl of RNP and 1 μg of repair constructs.

Electroporation was performed using an 0.2-cm electroporation cuvette and electroporated with a manual 1.8 pulse (Bio-Rad MicroPulser). Following the transformation, 1 mL ice-cold 1 M sorbitol was added to the cuvette. The cell suspension was then transferred to an eppendorf and the cells were gently pelleted (3 min, 3,000 rpm) before resuspension in 1 mL of YPD. Cells were recovered for 3 to 4 h at 30°C while gently shaking. After recovery, cells were pelleted and resuspended in 200 μL liquid YPD and the aliquots were spread plated onto YPD plates with 200 μg/ml nourseothricin and incubated at 37°C for 2 days.

Validation of phenotypes

Spot tests were performed to visualize changes that the transformations exert on antifungal drugs susceptibilities. Briefly, overnight cultures were set to the OD = 0.5 and serially diluted 10-fold and 10ul was spotted on YPD agar plates supplemented with antifungal drugs (Figure S2).

Fitness competition

In vitro competitive fitness was tested between ani resistant strains (containing *FKS* mutations) with and without *ERG3* mutations and in rich medium as well as in rich medium supplemented with 0.5ug/ml of anidulafungin. To be able to distinguish the strains, we used CRISPR transformants containing *NAT1* as a selection marker. Two pairs of ani resistant strains were used: 1:ERG3(D122Y) versus ERG3(WT)+*NAT1* and 2:ERG3(D122Y)+*NAT1* versus ERG3(WT). Two pairs were used to assure no fitness effect of the presence of *NAT1*. The competition test was conducted following a protocol described by Duxbury et al.[23] Briefly, all 4 strains were grown overnight and adjusted to 6.49×10^6 cells/ml prior to mixing and subsequent two fold dilution in the growth media. The first pair of strains was mixed in two different ratios (50:50 and 75:25), while the second pair in one (50:50). Each pool of mixed strains along with the strains alone were inoculated in wells of a 96 well plate in triplicates and incubated at 37°C with shaking for 24h hours. Cells at the beginning of the experiment, after 24h growth in YPD medium as well as in YPD + 0.5 ug/ml anidulafungin were diluted and plated on YPD agar plates and YPD agar plates with 200 μg/ml nourseothricin (each at least in duplicates). The number of cells of the strains that lack *NAT1* were obtained by subtracting the cells obtained from YPD+nourseothricin plate (average of the plated replicates) from the total number of cells observed on YPD plates (average of the plated replicates). Since we observed that strains containing *NAT1* were growing in lower abundance on the antibiotic than on YPD plates alone, the total number of cells were accounted for this discrepancy. *In vivo* fitness competition between the strain containing *ERG3* mutation (D122Y) and ERG3(WT)+*NAT1* was also tested in *Galleria mellonella*. For that, the overnight grown strains were adjusted to 2.5×10^8 CFU/ml in PBS, mixed in the ratio 50:50 and 10ul of the cell suspension was injected into at least 3 larva and left 24h at 37°C. To determine the fungal burden, 3 larvas per mix were briefly washed in 70% ethanol followed by sterile water, and then placed into screw-cap tubes with 3 sterile glass beads and 1ml of PBS. The tissue was then homogenized through 3 rounds of shaking for 20 s at 4 m/s in a Fastprep-24 (MP Biomedicals). The suspensions were serially diluted and inoculated into YPD+chloramphenicol (100 μg/ml) and YPD+chloramphenicol

(100 μg/ml) + nourseothricin (200 μg/ml). The real ratios of the cells at the beginning and end of the experiment were obtained as described *in vitro* competition experiment.

Virulence assays

The differences in virulence between strains with and without chromosomal duplications were tested in *Galleria mellonella*. Three strains were used: EB0911, parental WT, and its two flz evolved progenies 3B_FLZ and 3H_FLZ, where the second presents chromosomal duplications (ChrE and ChrI). Groups of 20 healthy larvae were injected with 10 μl of cell suspension, equivalent to 7.5×10^6 CFU, into the haemocoel with a Hamilton syringe through the last left pro-leg. Control set of larvae were injected with 10ul of PBS. Following infection, larvae were incubated at 37°C and survival, based on response to physical stimulation, was monitored daily for 6 days. The survival plots were created by Graphpad Prism 8.4.2.

QUANTIFICATION AND STATISTICAL ANALYSIS

rAUC, MIC and fitness measurements

We calculated the MIC, rAUC and fitness values for all evolved samples as explained in the STAR Methods section 'Fitness and susceptibility measurements'. For each evolved strain and drug concentration, we measured between three to five technical replicates, and kept the three replicates that were closest to the median for each measure (fitness, relative fitness, MIC and rAUC). We used the median across these three replicates and the central estimate for several analyses (Figures 2, 6, 7, S1, and S6; Data S1). In addition, we calculated the median absolute deviation across technical replicates as a measurement of dispersion (Figures 2, 7, S1, and S6; Data S1). All these measurements were performed with python (v3.7.8) and the packages pandas (v1.1.1) and scipy.stats (v1.5.2).

Correlation analyses

We calculated the spearman correlation between fitness and drug resistance (Figure 2E), between ani resistance and the distance to the *FKS* hotspot (HS) in samples with no-HS mutations (see Results; Data S2), between the number of different genes with acquired mutations and fitness/susceptibility (see Results) and between the rAUC and MIC (Figure S1). All the fitness and susceptibility measurements were taken as the median across technical replicates (see the STAR Methods section Fitness and susceptibility measurements). We defined as significant correlations those with a $p < 0.05$. We used the python package scipy.stats (v1.5.2) to perform all these analyses.

Association between *ERG3* mutations and flz resistance

We used a Fisher's exact test to evaluate the correlation between the presence of *ERG3* non synonymous mutations and flz resistance in anidulafungin-evolved samples (Figure 7D). We defined this as a significant association because of the $p < 0.05$. We used the python package scipy.stats (v1.5.2) to perform this analysis.

Comparing continuous distributions

We implemented a Kolmogorov-Smirnov test to assess the equality of two distributions in a non-parametric manner, thus not assuming normal distributions. We used this test to compare the relative fitness levels in each evolution condition and the YPD samples (see Results), the flz rAUC levels of anidulafungin-evolved samples with and without *ERG3* mutations (Figure 7C), the intra-strain versus inter-strain differences in fitness/susceptibility (Figure S1E) and the fitness/susceptibility levels of samples with different mutations (Figure S6). We used $p < 0.05$ as threshold for significant differences. We used the python package scipy.stats (v1.5.2) to perform all these analyses.

Variant calling from sequencing data

We identified the genomic variants (SNPs, INDELs, CNVs and genomic rearrangements) appearing during the *in vitro* evolution from whole-genome sequencing data (Figures 4, 5, 6, 7, S3–S7; Table S2; Data S3) as described in the STAR Methods sections 'Small variant calling and interpretation', **'**Identification of large aneuploidies, segmental duplications and deletions**'** and **'**Analysis of genomic rearrangements**'**. In addition, we used the python package scipy.stats (v1.5.2) to calculate the mean and standard deviation of the number of new small variants across replicates of the same strain and condition (Figure 4).

Competitive fitness measurements

In vitro fitness competition (Figure 7E) was performed in three replicates of each mixed ratio of the samples. Each of these mixed replicates was plated on agar plates at least twice to obtain an average number of the growing cells. Further, we calculated mean values of growing cells and standard deviation between them. *In vivo* fitness competition (Figure 7F) was performed in two separate competition experiments and both in three *Galleria mellonella* larvae. Mean number of growing cells and standard deviations were calculated from the averaged number of cells obtained from each of the homogenized larvae separately (plated on at least two agar plates). All calculations were done in Libre Office (v 6.0.7.3).

Estimating the overlap between drug resistance mutations among samples

We inferred an expected overlap between drug resistance mutations among different samples of the same condition (Figure S5B) using python (v3.7.8) and the package pandas (v1.1.1).

Current Biology, Volume *31*

Supplemental Information

Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen *Candida glabrata*

Ewa Ksiezopolska, Miquel Àngel Schikora-Tamarit, Reinhard Beyer, Juan Carlos Nunez-Rodriguez, Christoph Schüller, and Toni Gabaldón

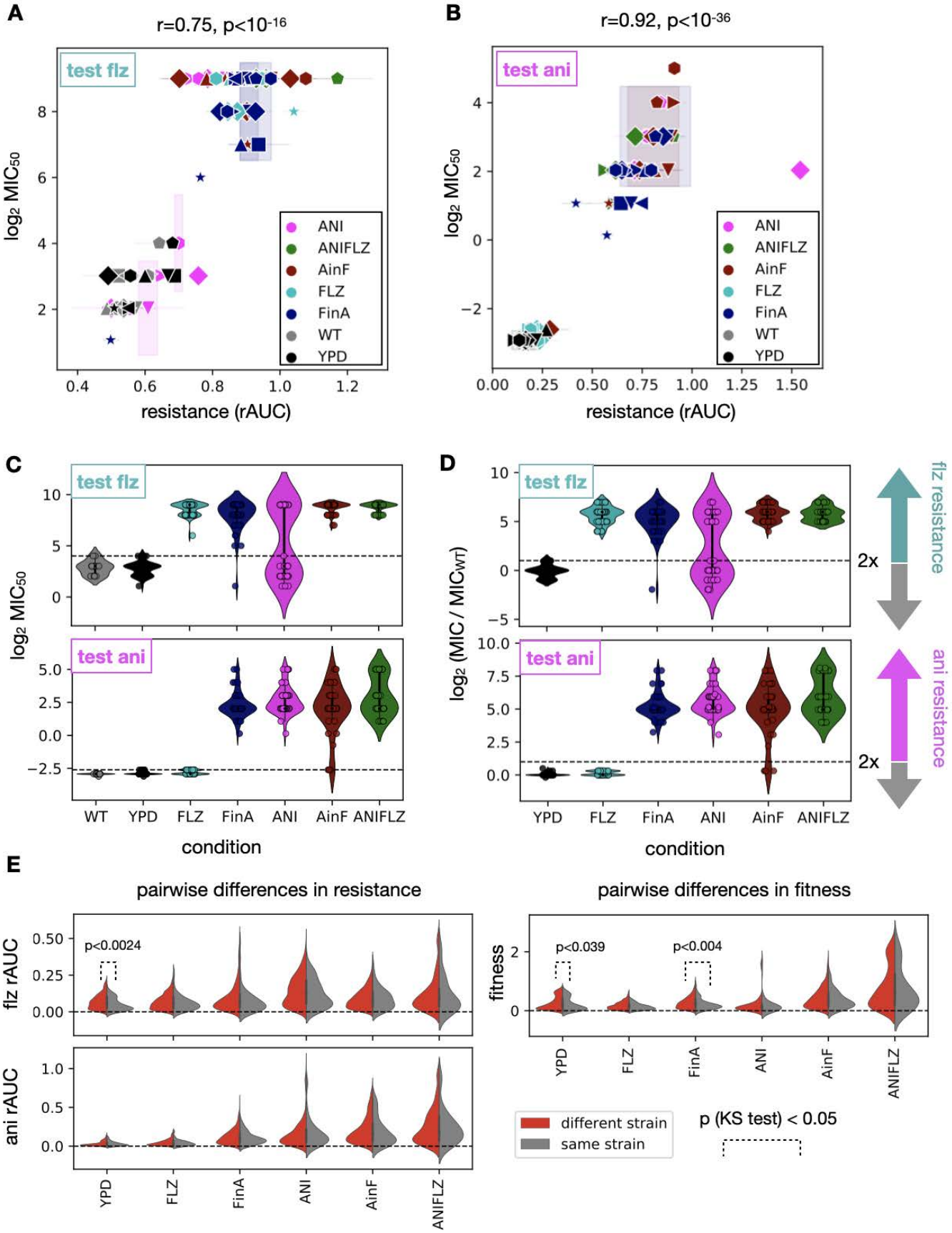

A
r=0.75, p<10-16
test flz
log2 MIC50
resistance (rAUC)
ANI
ANIFLZ
AinF
FLZ
FinA
WT
YPD
B
r=0.92, p<10-36
test ani
log2 MIC50
resistance (rAUC)
C
test flz
test ani
log2 MIC50
WT YPD FLZ FinA ANI AinF ANIFLZ
condition
D
test flz
test ani
log2 (MIC / MICWT)
2x
flz resistance
ani resistance
YPD FLZ FinA ANI AinF ANIFLZ
condition
E
pairwise differences in resistance
flz rAUC
p<0.0024
ani rAUC
YPD FLZ FinA ANI AinF ANIFLZ
pairwise differences in fitness
fitness
p<0.039
p<0.004
different strain
same strain
p (KS test) < 0.05

Figure S1. Comparing Minimum Inhibitory Concentration (MIC) and rAUC. Related to Figure 2.

(A) We compared the flz resistance levels estimated from rAUC and MIC_{50}. The Spearman correlation coefficients and p-values are shown. Each point corresponds to a biological replicate and the error bars reflect the median absolute deviation across technical replicates. **(B)** The same as in (A) but for ani resistance. **(C)** MIC_{50} for flz (top) and ani (bottom) was measured for all samples, presented here as single points. The dashed line indicates the maximum observed value in a YPD sample. **(D)** The increase in MIC relative to WT was calculated as the log2 ratio of MIC of the sample and MIC of WT. Resistant samples are defined as those having a MIC twice as high as the corresponding WT (dashed line). **(E)** Investigating the repeatability of our *in vitro* evolution experiment. We measured the pairwise differences in flz susceptibility (top left), ani susceptibility (bottom left) and fitness (right) between evolved samples of the same (gray) or different (red) strains. The quantitative phenotypes shown in the y axis are similar to **Figure 2B,D**. The x axis shows the evolution condition. In order to test whether the evolution of these phenotypes is particularly repeatable across samples of the same strain we compared the distribution of different-strain (red) vs same-strain (gray) pairwise differences in each condition. This yielded significant differences ($p<0.05$ in a Kolmogorov-Smirnov (KS) test) for some comparisons, indicated with dashed lines.

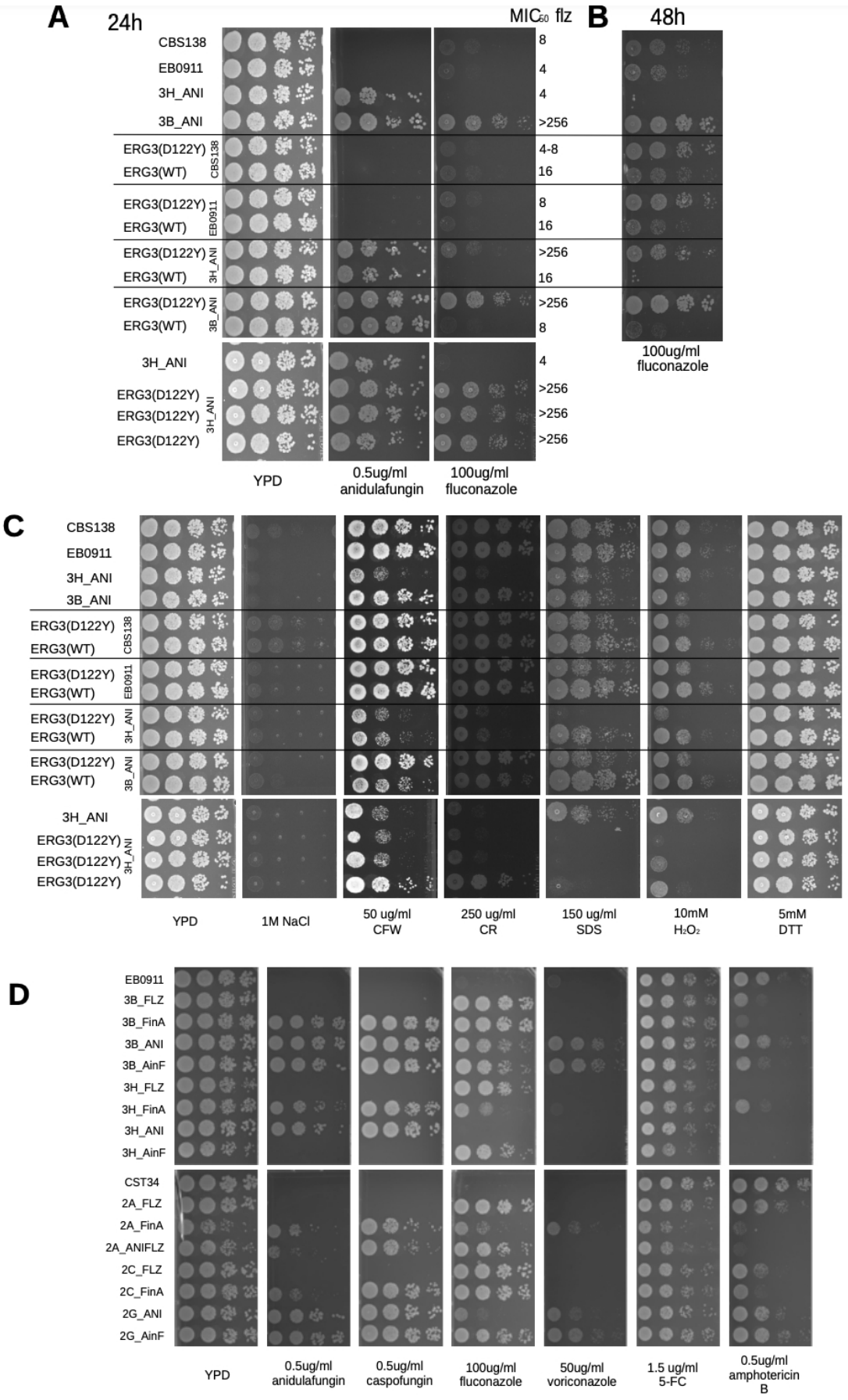

A
24h
MIC50 flz
CBS138
EB0911
3H_ANI
3B_ANI
8
4
4
>256
ERG3(D122Y)
ERG3(WT)
CBS138
4-8
16
ERG3(D122Y)
ERG3(WT)
EB0911
8
16
ERG3(D122Y)
ERG3(WT)
3H_ANI
>256
16
ERG3(D122Y)
ERG3(WT)
3B_ANI
>256
8
3H_ANI
ERG3(D122Y)
ERG3(D122Y)
ERG3(D122Y)
3H_ANI
4
>256
>256
>256
YPD
0.5ug/ml anidulafungin
100ug/ml fluconazole
B
48h
100ug/ml fluconazole
C
CBS138
EB0911
3H_ANI
3B_ANI
ERG3(D122Y)
ERG3(WT)
CBS138
ERG3(D122Y)
ERG3(WT)
EB0911
ERG3(D122Y)
ERG3(WT)
3H_ANI
ERG3(D122Y)
ERG3(WT)
3B_ANI
3H_ANI
ERG3(D122Y)
ERG3(D122Y)
ERG3(D122Y)
3H_ANI
YPD
1M NaCl
50 ug/ml CFW
250 ug/ml CR
150 ug/ml SDS
10mM H2O2
5mM DTT
D
EB0911
3B_FLZ
3B_FinA
3B_ANI
3B_AinF
3H_FLZ
3H_FinA
3H_ANI
3H_AinF
CST34
2A_FLZ
2A_FinA
2A_ANIFLZ
2C_FLZ
2C_FinA
2G_ANI
2G_AinF
YPD
0.5ug/ml anidulafungin
0.5ug/ml caspofungin
100ug/ml fluconazole
50ug/ml voriconazole
1.5 ug/ml 5-FC
0.5ug/ml amphotericin B

Figure S2. Spot tests. Related to the samples with re-introduced *ERG3* mutations and susceptibility to a wide panel of drugs and STAR Methods.

(**A**) Susceptibility of transformants carrying introduced changes in *ERG3* gene. Spot tests demonstrate changes in susceptibility (on a rich medium YPD supplemented with 100ug/ml flz and 0.5ug/ml ani) followed by EUCAST test after 24h incubation. The first four strains are the background strains used for the transformation: wild type *Candida glabrata* CBS138 and EB0911 as well as ani evolved progenies of EB0911: 3H_ANI (susceptible to flz) and 3B_ANI (bearing *ERG3* D122Y mutation and resistant to flz). The following spots represent the susceptibility of transformants carrying: *ERG3*(WT) or *ERG3*(D122Y) alleles fused with a *NAT1* selection marker. The bottom panel shows three independent transformants carrying *ERG3*(D122Y) mutation inserted into an ani resistant strain (3H_ANI) - 1. transformed with a long fragment with *ERG3* and crRNA_ERG3_1 and 2. and 3. are 2 different colonies obtained from a transformation with synthetic *ERG3* fragment and crRNA_ERG3_2. These transformants do not contain *NAT1* gene and were selected on flz. Note that **Table S5** includes the list of used oligos. (**B**) presents a spot test of CRISPR transformants grown on flz and incubated for 48h. (**C**) shows susceptibility of *ERG3* CRISPR transformants to NaCl, Calcofluor White (CFW), Congo Red (CR), SDS, H_20_2 and DTT. (**D**) presents susceptibility of selected evolved mutants to anidulafungin, caspofungin, fluconazole, voriconazole, flucytosine (5-FC) and amphotericin B. Used concentrations are indicated in the figure.

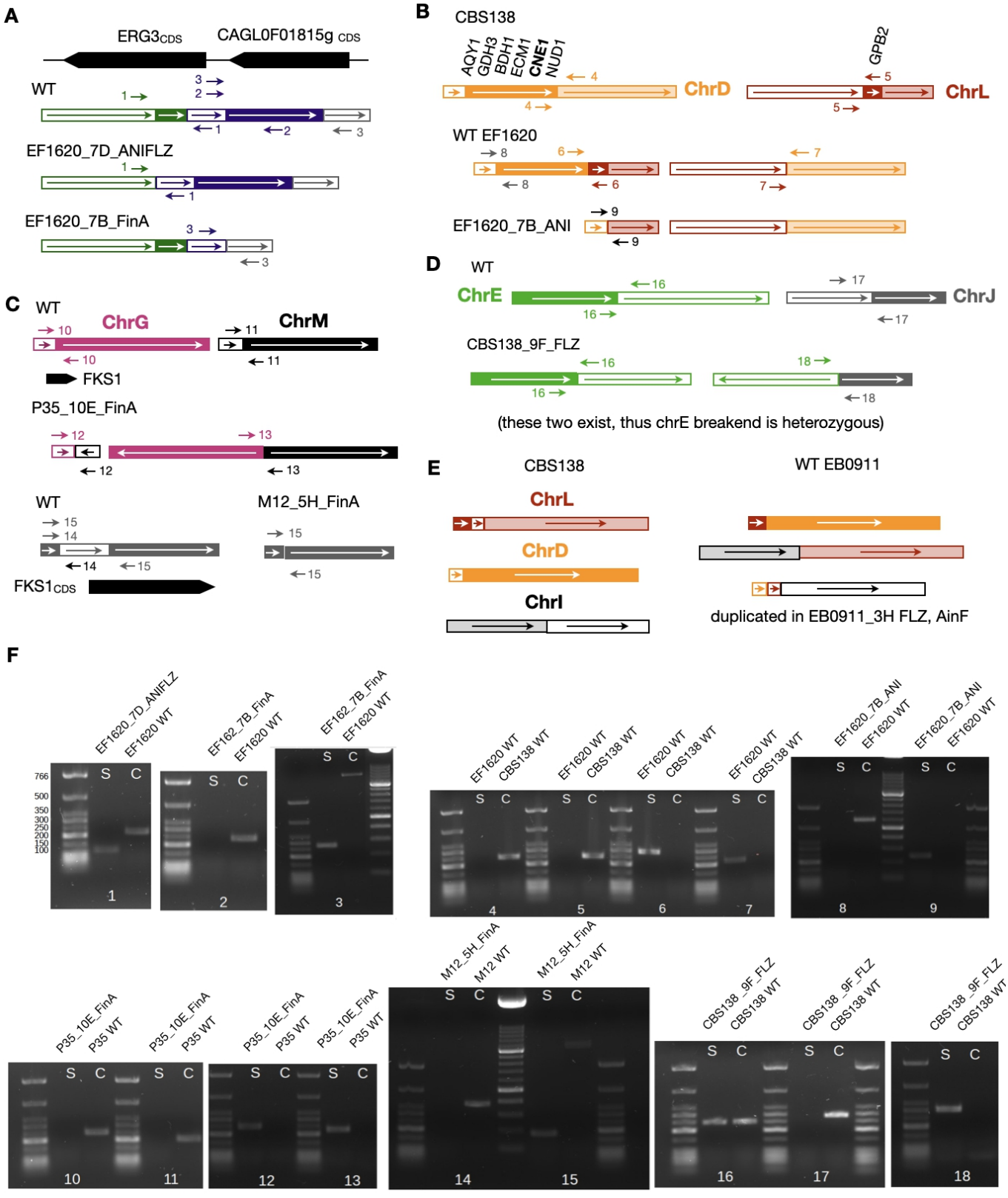


Figure S3. Genomic rearrangements that appear during evolution in antifungal drugs. Related to Figure 6.

(A) We found two samples with a deletion in the *ERG3* CDS (medium) and upstream region (bottom), respectively. The browser represents the genomic coordinates of *ERG3* and the upstream gene *CAGL0F01815g*. The boxes represent the WT regions that are rearranged in each sample. We confirmed these rearrangements with three PCRs on these samples (using primer pairs 1, 2 and 3). The results are

shown in (F), with the numbers matching the primer pairs of each PCR. **(B)** *CNE1* and *GBP2* were lost due to a single deletion rearrangement in EF1620_7B_ANI. The representation is analogous to (A), showing a EF1620 WT balanced translocation between Chr D and L which, in addition to the deletion-like breakpoint appearing in ANI, generates a loss of the region between the two breakpoints. **(C)** Two FinA samples carried rearrangements breaking the *FKS1* CDS (black box). P35_10E_FinA had an inverted balanced translocation between Chrs G and J (top), and M12_5H_FinA carried a partial deletion (bottom). **(D)** Genomic rearrangements can explain the partial Chr E aneuploidy in CBS138_9F_FLZ **(Figure 5A)**. This sample carried an unbalanced translocation between Chr E and J. Both Chr E breakends were heterozygous, while the Chr J breakend was haploid. **(E)** The apparent partial duplication of Chr I **(Figure 5A)** is actually a complete aneuploidy in two EB0911 samples. We found WT balanced translocations between these chromosomes that result in three mixed chromosomes in this strain (bottom). We found that two EB0911_3H samples had one of these mixed chromosomes duplicated (bottom), including mostly half of the reference Chr I. **(F)** We performed PCRs using primer pairs around the rearrangements (1 to 18 in **(A)** - **(D)**) to confirm them. These primers can be found in **Table S5.** Each PCR was carried on a given sample and the corresponding control. We note that we could obtain bands with the expected sizes in all samples.

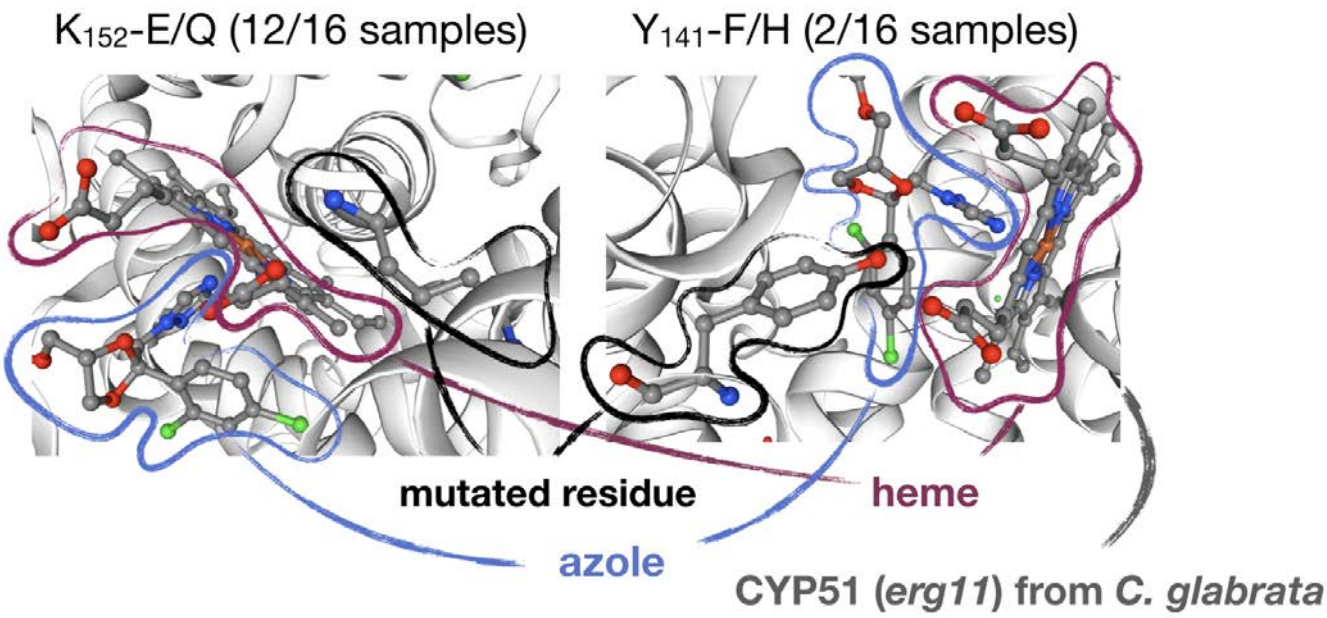


Figure S4. Structural localization of frequent *ERG11* mutations. Related to Figure 6.

Given the availability of a characterized 3D structure for Erg11p in contact with azoles (pdb id: 5JLC) we inspected the location of recurrently mutated residues and found that they are close to the azole binding pocket. The structure (pdb id: 5JLC) was visualized using SWISS MODEL[S1]. A screenshot of the two residues in the context of itraconazole and a heme group is shown. The basic group of K152 is close to an acid group in heme, potentially establishing an electrostatic interaction that is important for stability. Importantly, Y141 is conserved with Y132, a position that has been mutated in various other azole resistant *Candida* species[S2-S4]. As a possible mechanism of resistance, we hypothesize that the substitution by E or Q destabilizes this interaction, thereby impairing the binding of azoles.

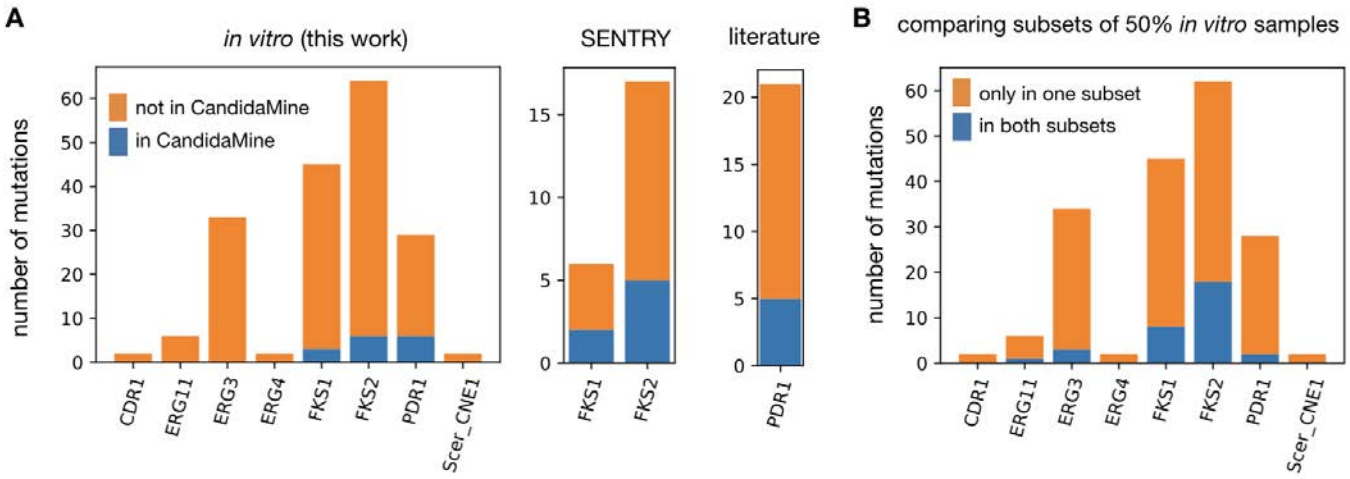


Figure S5. The overlap between drug resistance-conferring mutations from different studies in *Candida glabrata*. Related to the analysis of clinical isolates' sequencing datasets in STAR Methods.

(A) We compared the drug resistance variants described in this work (left), the SENTRY database (middle[S5]) and a set of described *PDR1* mutations (right[S6-S8]) against those in clinical isolates with available whole genomes (393 in total) (see **STAR Methods**). Shown is the number of mutations that are found in each study and in some (blue) or no (orange) clinical isolates. **(B)** In order to estimate the expected overlap between drug resistance mutations among different samples, we implemented a randomised strategy from our own experiments. We divided the samples carrying mutations in a given gene into two random subsets. For each subset, we calculated the number of mutations only in the subset or also found in the other subset. This process was repeated 100 times, and shown is the median number of mutations not shared (orange) or shared (blue) across subsets.

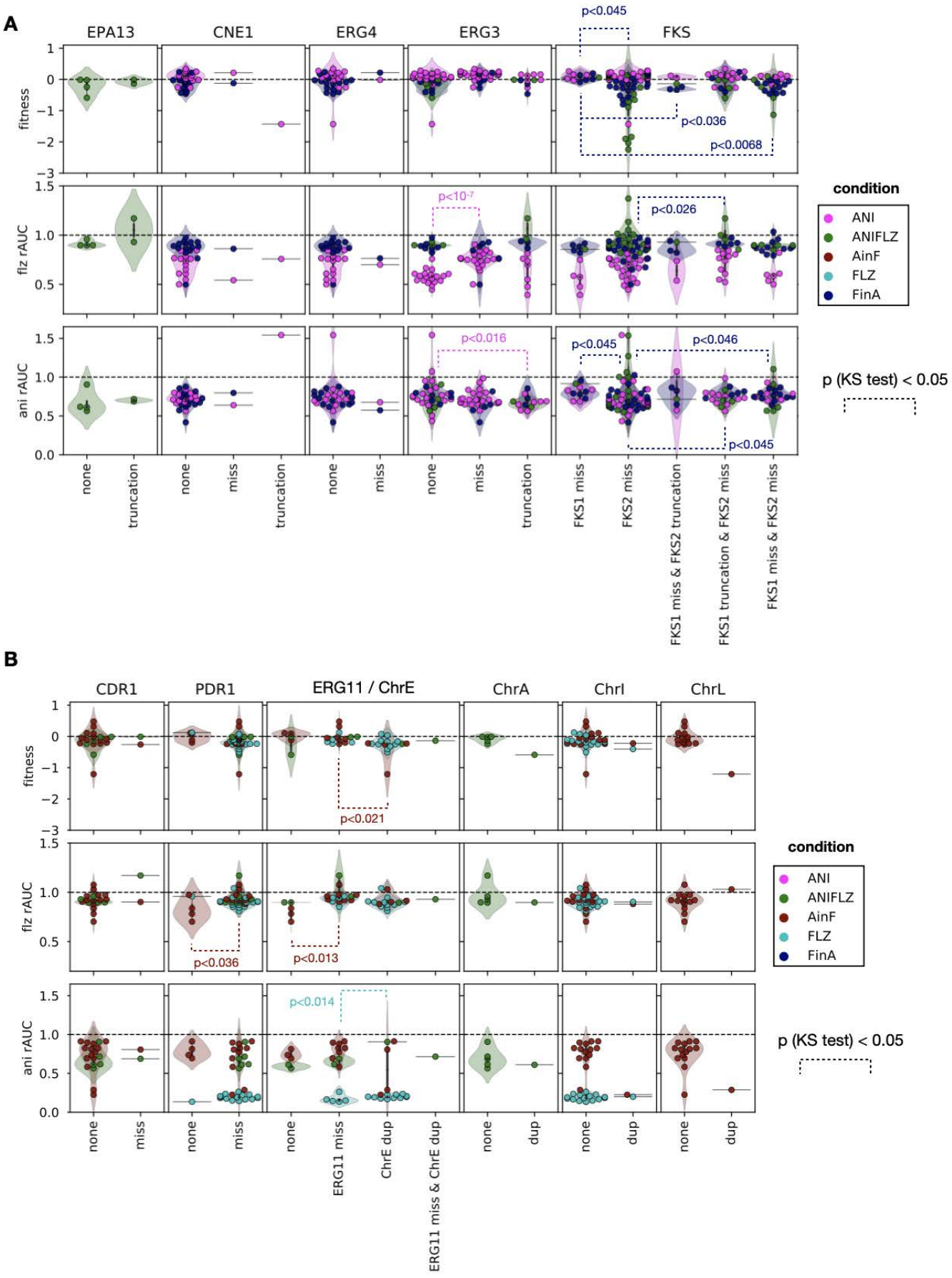

A
EPA13
CNE1
ERG4
ERG3
FKS
p<0.045
p<0.036
p<0.0068
fitness
flz rAUC
ani rAUC
p<10-7
p<0.026
p<0.016
p<0.045
p<0.046
p<0.045
condition
ANI
ANIFLZ
AinF
FLZ
FinA
p (KS test) < 0.05
none
truncation
miss
FKS1 miss
FKS2 miss
FKS1 miss & FKS2 truncation
FKS1 truncation & FKS2 miss
FKS1 miss & FKS2 miss
B
CDR1
PDR1
ERG11 / ChrE
ChrA
ChrI
ChrL
p<0.021
p<0.036
p<0.013
p<0.014
ERG11 miss
ChrE dup
ERG11 miss & ChrE dup
dup

Figure S6. Genotype-phenotype relationship in the evolved samples. Related to Figures 2 and 6.

(A) Similar mutations in genes altered during evolution in ani seldom modulate fitness (top), flz susceptibility (medium) or ani susceptibility (bottom). The y axis shows each quantitative phenotype as in **Figure 2B,D**. Each point represents one *in vitro*-evolved sample and the color indicates the condition. The x axis shows whether each sample has no mutations ("none"), missense mutations ("miss") or truncating mutations ("truncation") in the given gene. In addition, we separate the samples by *FKS1/FKS2* mutation status (right panel) in order to show how different combinations of mutations in these genes may affect each phenotype. We compared the phenotypes for each of the condition/mutation type combinations in a pairwise manner with a Kolmogorov-Smirnov (KS) test to find significant differences between the groups. The dashed lines correspond to comparisons with a $p<0.05$. **(B)** The same as in (A), but for genes / chromosomes mutated during evolution in flz. The "ERG11 / ChrE" panels shows these phenotypes for different combinations of *ERG11* missense mutations and Chr E duplications ("ChrE dup"). The samples in the rightmost three panels are separated by the absence ("none") or presence ("dup") of chromosomal duplications.

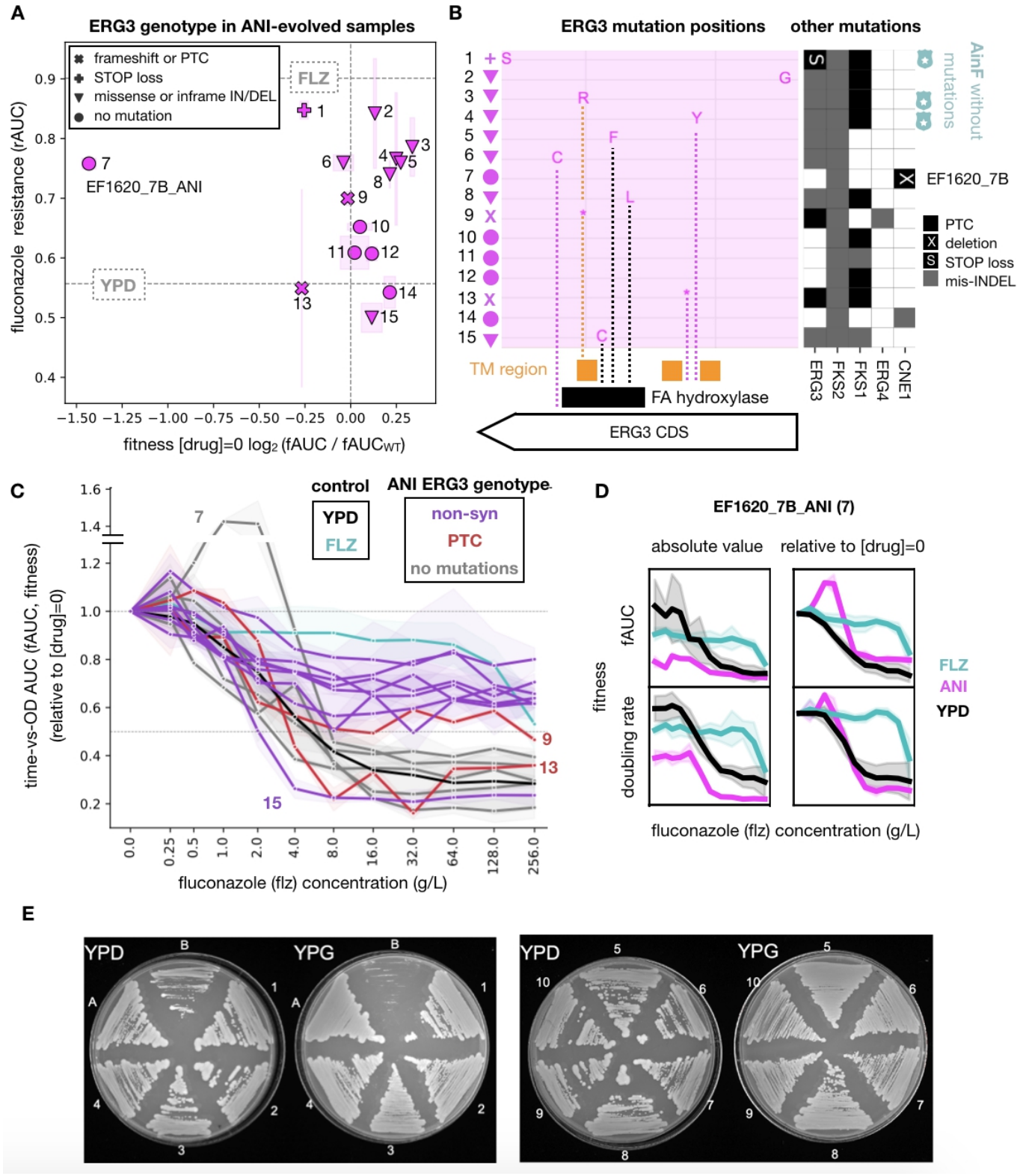


Figure S7. Acquisition of *ERG3* mutations in ANI samples and fluconazole cross-resistance. Related to Figure 7.

(A) Fitness (relative to the WT as in **Figure 2D**) is high in most ANI-evolved samples (EF1620_7B_ANI is an exception), while flz-resistance (shown as rAUC) is variable. The symbols correspond to different types of *ERG3* protein-altering mutations. The dashed lines correspond to the median flz rAUC for all the FLZ and YPD samples. Each point represents the median across technical replicates for a given sample, while the boxes show the median absolute deviation. The numbers are related to the order of flz-

resistance used to show the relationship of each sample to panels (B), (C) and (D). **(B)** *ERG3* amino acid mutations are scattered throughout the coding region of the gene. The boxes in the bottom represent annotated protein domains (see **STAR Methods**), where the "catalytic domain" is the Fatty acid (FA) hydroxylase superfamily (PF04116) and TM are transmembrane regions. Three samples with no additional mutations nor increase in flz resistance in subsequent flz treatment (AinF) are marked with blue shields. PTC and '*' indicate Premature Termination and S indicates the loss of the STOP codon. **(C)** Growth of the ANI samples (with colored *ERG3* genotype) at increasing concentrations of flz shown as fAUC and compared to all FLZ (blue) and YPD (black) samples. Purple lines indicate samples with non-synonymous alterations, red - with protein termination codon (PTC) and gray - no *ERG3* changes. Samples 9 and 13 bear a PTC but the former showed improved growth at higher flz concentrations. Although assessed as susceptible based on MIC, sample 9 presented a growth curve more similar to that of resistant samples, and maintained a relative growth around ~50% across increasing concentrations (see **Analysis of MIC and rAUC measures for antifungal drug resistance in STAR methods**). Sample 15 bears the only ns mutation that did not result in increased resistance to flz by rAUC, MIC or shape and position of the growth curve. The points and error bars correspond to the median and median absolute deviation for each assayed concentration in each sample, respectively. The numbers (7, 15, 9, 13) correspond to those in (A) and (B). **(D)** EF1620_7B_ANI (number 7 in this figure) was found to be susceptible to flz according to our MIC-based thresholding (**Figure S1C,D**) but depicted an rAUC in the range of resistant samples (A). To understand this mismatch, we studied the quantitative relationship between flz concentration and several fitness estimates (the doubling rate per hour (bottom) and fAUC (top)) in both absolute (left) and relative to no drug (right) representations. The median values across all FLZ and YPD EF1620 samples are shown for comparison. **(E)** Petite phenotype assessment. Growth of ANI evolved mutants (1. 2G_ANI, 2. 3B_ANI, 3. 5F_ANI, 4. 7D_ANI, 5.7F_ANI, 6. 9F_ANI, 7. 9H_ANI, 8. 10G_ANI, 9. 11G_ANI, 10. 11H_ANI), CBS138 (A) and petite *S. cerevisiae* mutant (B), on YP medium supplemented with glucose (YPD) and glycerol (YPG).

Systematic name	Standard name or ortholog	CST109	M12	description
CAGL0C05313g	-	N547H	M206V	Protein of unknown function
CAGL0B01166g	SWI6	A246T	R414K	Ortholog(s) have DNA-binding transcription activator activity, RNA polymerase II-specific, RNA polymerase II proximal promoter sequence-specific DNA binding, transcription coactivator activity
CAGL0E03564g	Scer_CDC3	K383R	I278M	Ortholog(s) have GTP binding, phosphatidylinositol-4-phosphate binding, phosphatidylinositol-5-phosphate binding, structural molecule activity
CAGL0G01430g	Scer_LAP2	N469S	P222S	Ortholog(s) have aminopeptidase activity, epoxide hydrolase activity and role in cellular lipid metabolic process, protein catabolic process, protein localization by the NVT pathway
CAGL0H03179g	**Scer_MAD1**	**Y390H**	**D387N**	**Ortholog(s) have protein-containing complex binding activity**
CAGL0H09130g	Scer_MNN4	R573Stop	P734A	Ortholog(s) have enzyme activator activity and role in fungal-type cell wall polysaccharide biosynthetic process, protein N-linked glycosylation, protein O-linked glycosylation
CAGL0H02255g	Scer_RSN1	E709G	I787M	Ortholog(s) have role in Golgi to plasma membrane transport and membrane localization
CAGL0J07326g	Scer_SQS1	Q359L	D533E	Ortholog(s) have role in mRNA splicing, via spliceosome, maturation of SSU-rRNA, positive regulation of ATPase activity, positive regulation of helicase activity

Table S1. List of shared polymorphisms found in CST109 (clade 1) and M12 (clade 3) that were not found in other representatives of their respective clades - CST34 and CST78 for clade 1 and 3, respectively. Related to the Results section '*Candida glabrata* has a widespread ability to acquire drug and multidrug resistance' and STAR Methods.

We highlight the ortholog of *Saccharomyces cerevisiae MAD1* for which polymorphisms in CST109 and M12 were found to affect nearby residues in the protein sequence (390 and 387, respectively). Dysfunction of this gene has been previously related to chromosome instability in *S. cerevisiae*[S9]. Thus, these polymorphisms might be associated with higher chromosome instability resulting in lower capacity to preserve long-term drug resistance.

					genome sequencing	sanger sequencing
Mutant	**Condition**	**Clade**	**Strain**	**Replicate**	***ERG3***	
TGL00051	ANI	1	CST109	1B		-
TGL00052	ANI	1	CST109	1D		-
TGL00053	ANI	1	CST109	1F		-
TGL00054	ANI	1	CST109	1H		-
TGL00055	ANI	1	CST34	2A	-	
TGL00056	ANI	1	CST34	2C		mis|p.213|L/S
TGL00057	ANI	1	CST34	2E		ins|c.215|TC, ins|p.77|
TGL00058	ANI	1	CST34	2G	mis|p.207|P/L	
TGL00059	ANI	2	EB0911	3B	mis|p.122|D/Y	
TGL00060	ANI	2	EB0911	3D		-
TGL00061	ANI	2	EB0911	3F		-
TGL00062	ANI	2	EB0911	3H	-	
TGL00063	ANI	3	CST78	4A		-
TGL00065	ANI	3	CST78	4E		mis|p.1|M/L
TGL00066	ANI	3	CST78	4G		PTC|p.67|Y/*
TGL00067	ANI	3	M12	5B		-
TGL00068	ANI	3	M12	5D		mis|p.265|N/K
TGL00069	ANI	3	M12	5F	mis|p.9|D/G	-
TGL00070	ANI	3	M12	5H	-	
TGL00071	ANI	4	EF1237	6A		-
TGL00072	ANI	4	EF1237	6C		mis|p.302|Q/K
TGL00073	ANI	4	EF1237	6E		-
TGL00074	ANI	4	EF1237	6G		-
TGL00075	ANI	4	EF1620	7B	-	
TGL00076	ANI	4	EF1620	7D	mis|p.267|W/R	
TGL00077	ANI	4	EF1620	7F	mis|p.243|Y/C	
TGL00078	ANI	4	EF1620	7H		-
TGL00079	ANI	5	F15	8A		-
TGL00080	ANI	5	F15	8C		mis|p.224|T/A
TGL00081	ANI	5	F15	8E		-
TGL00082	ANI	5	F15	8G		mis|p.135|Q/R
TGL00083	ANI	5	CBS138	9B		mis|p.128|H/Y
TGL00084	ANI	5	CBS138	9D		partial deletion
TGL00085	ANI	5	CBS138	9F	PTC|p.135|Q/*	
TGL00086	ANI	5	CBS138	9H	STOP|c.1094|tAg/tCg	
TGL00087	ANI	6	P352	10A		PTC|p.239|Q/*
TGL00088	ANI	6	P352	10C		mis|p.71|P/L
TGL00089	ANI	6	P352	10E	-	
TGL00090	ANI	6	P352	10G	mis|p.228|S/F	
TGL00091	ANI	7	BG2	11B	mis|p.300|Y/C	
TGL00092	ANI	7	BG2	11D		mis|p.87|R/I
TGL00093	ANI	7	BG2	11F		mis|p.300|Y/C
TGL00094	ANI	7	BG2	11H	PTC|p.267|W/*	
TGL00096	ANI	7	SLL2glab	12A		mis|p.225|P/T
TGL00095	ANI	7	SLL2glab	12C		mis|p.301|G/D
TGL00097	ANI	7	SLL2glab	12E		mis|p.225|P/T
TGL00098	ANI	7	SLL2glab	12G		PTC|p.203|W/*

Table S2. *ERG3* mutations. Related to Figure 7.

Columns indicate, in this order: mutant name, evolution media, clade, strain, replicate, mutations in *ERG3* gene from genome and sanger sequencing. The variants are encoded as “type of mutation” / “molecule affected” . “position” | “reference allele” / “alternative allele”. The “type of mutation” can be: mis - missense variant, del - inframe deletion, PTC – Premature Termination Codon, FS - frameshift, ins – inframe insertion, lostSTOP – lost STOP codon, lostATG - lost START codon. The “molecule affected” can be “p” for protein and “c” for cDNA. The “reference” and “alternative” alleles correspond to amino acids or codons for proteins or cDNA alterations, respectively.

evolution media	ANI			ANI			ANI		ANI			ANI			ANI			ANI			ANI			ANI			ANI		
clade	1			2			3		4			4			5			5			6			4			7		
strain	CST34			EB0911			M12		EF1620			EF1620			CBS138			CBS138			P352			BG2			BG2		
replicate	2G			3B			5F		7D			7F			9F			9H			10G			11B			11H		
tested gene/fragment	FKS2 HS1	ERG3		FKS1 HS1	FKS2 HS1	ERG3	FKS2 HS1	ERG3	FKS1 HS1	FKS2 HS1	ERG3	FKS1 HS1	FKS2 HS1	ERG3	FKS1 HS1	FKS2 HS1	ERG3	FKS2 HS1		ERG3	FKS2 HS1		ERG3	FKS2 HS1		ERG3	FKS2 HS2	ERG3	
final mutation	F659-	P207L	E139*	S652*	F659-	D122Y	F659-	D9G	W650*	F659-	W267R	D632Y	F659L	Y243C	W611*	L662W	Q135*	F659-	F659L	lost STOP	F659-	R665G	S228F	A651V	S663P	Y300C	R1378H	W267*	G301V
			new!																new!										new!
ani (ug/ml)																													
0.032	--			--	--		--		--	--		--	--		--	--		--	--		--	--		A651V	--	--	--		
0.064	--			--	--		--		--	--	--	--	--	--	--	L662W		--	--		--	--		--	--		--		
0.126	--	--	E139*	--	--	--	--	--	--	F659-	W267R	--	F659L	Y243C	--	L662W		--	--	lost STOP	--	--	S228F	A651V	--		--		
0.256	F659-	P207L	--	--	F659-	D122Y	F659-	D9G	--	F659-		--	F659L		--	L662W	--	--	F659L	lost STOP	--	R665G	S228F	A651V	--		--	--	G301V

Table S3. Trajectory of final *FKS* and *ERG3* mutations. Related to Figure 7.

Rows indicate, in this order: evolution media, clade, strain, replicate, tested gene/fragment, final mutation, and concentrations of anidulafungin (μg/ml) corresponding to intermediate glycerol stocks (isolated single colonies) of tested trajectories. Mutations that were not found at the finalization of the evolution experiment are marked as 'new'.

passages	drug increase	fluconazole (μg/mL)	anidulafungin (μg/mL)
		0	0
		0	0
1		4	0.016
2	1	4	0.016
3		8	0.032
4	2	8	0.032
5		16	0.064
6	3	16	0.064
7		32	0.128
8	4	32	0.128
9		64	0.256
10	5	64	0.256
11		96	0.512
12	6	96	0.512
13		128	1.024
14	7	128	1.024
15		160	2.048
16	8	160	2.048
17		192	4.096
18	9	192	4.096

Table S4. Information on drugs' concentrations used in the evolution experiments. Related to Figure 1.

Columns indicate, in this order: number of passages, number of drug increases and corresponding fluconazole and anidulafungin concentrations (μg/mL).

	name	sequence
genome rearrangements	ChrF_1_F	GTAGGACAAAGAGGCGGTGA
	ChrF_1_R	TCTACGCTGCTGCATGAGAC
	ChrF_2_F	CCCAGACAATGGGATGAAAT
	ChrF_2_del_R	TATCATGTGACAGCGTCTGC
	ChrF_3_R	GTGTTGGGCAAAGGTGACTT
	ChrD_1_F	CACCAAAGGAAAGGACAAGG
	ChrD_1_R	CCCTGTTGGTGGTCATTTTT
	ChrL_1_F	TCGCATATGCATTTCATCGT
	ChrL_1_R	AACTGCCTCCAACACTTTCG
	ChrDL_1_F	CAGGTCAAATACGTTTCCCATAA
	ChrDL_1_del_R	TTTCATTTGTTATTGAATATCTTTGC
	ChrDL_2_R	CCAGCAGGAACTCTATCAAGG
	ChrG_1_F	GAAGGTATCGCTAAGATTGTCTTC
	ChrG_1_R	GACCAATTGTTGATAGTTGTGTG
	ChrM_1_F	TTGCGATAGAAGCTTTCCTACA
	ChrM_1_R	TCCGATGTGCCATCAATCTA
	revChrG_1_F	CCAATTGTTGATAGTTGTGTGTG
	revChrM_1_R	TCGATGAGTCCATGAAAAGAAA
	ChrG_2_F	AAGAGGTGAGGGAGGGAGAA
	ChrG_2_del_R	GGGACTAAGCTGATACACGAAGA
	ChrG_3_R	GGCTTGACCATTCTGTTGGT
	ChrE_1_F	TCTGCACCACGGTAGAAAG
	ChrE_1_R	GATGATTGCAAGGAAGAAGAA
	ChrJ_1_F	CTGAATAAGGGTTGCGTGCT
	ChrJ_1_R	ATGAGGGCCCCTGTCTTTAC
	revChrE_1_F	ATGAGGGCCCCTGTCTTTAC
ERG3	ERG3_1_FWD	TTGCATTTCAGATAACCTACAGC
	ERG3_1_REV	CAGTGCAGCCATCTGTGAG
	ERG3_2_FWD	TCCCTCTTGACTGTCCCTTG
	ERG3_2_REV	AAAGTAATGTGTGCGCGAGA
FKS	FKS1_HS1_FWD	CCATTGGGTGGTCTGTTCACG
	FKS1_HS1_REV	GATTGGGCAAAGAAAGAAATACGAC
	FKS1_HS2_FWD	GGTATTTCAAAGGCTCAAAAGGG
	FKS1_HS2_REV	ATGGAGAGAACAGCAGGGCG
	FKS2_HS1_FWD	GTGCTCAACATTTATCTCGTAGG
	FKS2_HS1_REV	CAGAATAGTGTGGAGTCAAGACG
	FKS2_HS2_FWD	CGTAGACCGTTTCTTGACTTC
	FKS2_HS2_REV	CTTGCCAATGTGCCACTG

CRISPR		
ERG3	crRNA_ERG3_1	/AltR1/rGrA rArArA rCrGrU rArGrG rArCrA rArArG rArGrG rGrUrU rUrUrA rGrArG rCrUrA rUrGrC rU/AltR2/
	crRNA_ERG3_2	/AltR1/rUrC rUrGrU rCrGrA rArGrA rCrGrA rArArA rCrGrU rGrUrU rUrUrA rGrArG rCrUrA rUrGrC rU/AltR2/
	donor_ERG3	/Alt-R-HDR1/T*G* GTT CTT CAA GTA TTT TGG ATG GTT GAA GAT AGT TCT GTA GAA GAC GAA AAC GTA TGA CAA AGA GGC GGT GAT CAG GTA CAA TAG CAG ACC GAA GA*C* G/Alt-R-HDR2/
LongFragmentERG3	FL1_FWD	TCCTCGACCAACAGACCATC
	FL1_REV	TGTTCGAGACTAGTAGCGGG
1. FLKI_ERG3	1_flank_ERG3_FWD	TCCTCGACCAACAGACCATC
	1_REV	gtcgacctgcagcgtacgAATGAGAACCCAGGTCAGCAC
2. NAT1(for DNA donor constructs)	2_NAT1_FWD	GTGCTGACCTGGGTTCTCATTcgtacgctgcaggtcgac
	2_NAT1_REV	TTAATTTGTTGCCATAAAAAATctacgagaccgacaccg
3. FLKII	3_FLKII_FWD	cggtgtcggtctcgtagATTTTTTATGGCAACAAATTAA
	3_FLKII_REV	TGTACTGGCACTTCGACCTT
check the fusion	inside_ERG3_FWD	TCCCTCTTGACTGTCCCTTG
	inside_NAT1_REV	caaccacaaatgaccagcac
	inside_NAT1_FWD	gtgatttggctggtttcgtt
	flank_ERG3_REV	GTGGAGGCGAGGAGTAGAAA
after the transformation – in the correct place	out_REV	GGTAGTCAGCAAGGTCTCGT
	inside_NAT1_FWD	gtgatttggctggtttcgtt
double check if NAT inside, primers down and upstream of the *NAT1*	inside_ERG3_FWD	TCCCTCTTGACTGTCCCTTG
	flank_ERG3_REV	GTGGAGGCGAGGAGTAGAAA

Table S5. Information about all the oligos used in the study. Related to Figures: 3, 7, S2, S3 and STAR Methods.

The table includes primers used to confirm the GR, investigate *ERG3* gene and *FKS1* and *FKS2* fragments as well as crRNAs, ordered *ERG3* fragment and primers used in CRISPR-Cas9 transformations. Lowercase letters in primers used in CRISPR Cas9 transformations indicate the sequences in *NAT1* gene.

Supplemental References

S1. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. *46*, W296–W303.

S2. Lockhart, S.R., Etienne, K.A., Vallabhaneni, S., Farooqi, J., Chowdhary, A., Govender, N.P., Colombo, A.L., Calvo, B., Cuomo, C.A., Desjardins, C.A., et al. (2017). Simultaneous Emergence of Multidrug-Resistant Candida auris on 3 Continents Confirmed by Whole-Genome Sequencing and Epidemiological Analyses. Clin. Infect. Dis. *64*, 134–140.

S3. Berkow, E.L., Manigaba, K., Parker, J.E., Barker, K.S., Kelly, S.L., and Rogers, P.D. (2015). Multidrug Transporters and Alterations in Sterol Biosynthesis Contribute to Azole Antifungal Resistance in Candida parapsilosis. Antimicrob. Agents Chemother. *59*, 5942–5950.

S4. Tan, J., Zhang, J., Chen, W., Sun, Y., Wan, Z., Li, R., and Liu, W. (2015). The A395T mutation in ERG11 gene confers fluconazole resistance in Candida tropicalis causing candidemia. Mycopathologia *179*, 213–218.

S5. Pfaller, M.A., Diekema, D.J., Turnidge, J.D., Castanheira, M., and Jones, R.N. (2019). Twenty Years of the SENTRY Antifungal Surveillance Program: Results for Candida Species From 1997–2016. Open Forum Infectious Diseases *6*, S79–S94.

S6. Ferrari, S., Ischer, F., Calabrese, D., Posteraro, B., Sanguinetti, M., Fadda, G., Rohde, B., Bauser, C., Bader, O., and Sanglard, D. (2009). Gain of function mutations in CgPDR1 of Candida glabrata not only mediate antifungal resistance but also enhance virulence. PLoS Pathog. *5*, e1000268.

S7. Tsai, H.-F., Sammons, L.R., Zhang, X., Suffis, S.D., Su, Q., Myers, T.G., Marr, K.A., and Bennett, J.E. (2010). Microarray and molecular analyses of the azole resistance mechanism in Candida glabrata oropharyngeal isolates. Antimicrob. Agents Chemother. *54*, 3308–3317.

S8. Spettel, K., Barousch, W., Makristathis, A., Zeller, I., Nehr, M., Selitsch, B., Lackner, M., Rath, P.-M., Steinmann, J., and Willinger, B. (2019). Analysis of antifungal resistance genes in Candida albicans and Candida glabrata using next generation sequencing. PLoS One *14*, e0210397.

S9. Zhu, J., Pavelka, N., Bradford, W.D., Rancati, G., and Li, R. (2012). Karyotypic determinants of chromosome instability in aneuploid budding yeast. PLoS Genet. *8*, e1002719.

S10. Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M., and Sherlock, G. (2017). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 45, D592–D596.

3.4

Genome-wide signatures of recent selection and drug resistance across *Candida* opportunistic pathogens

Genome-wide signatures of recent selection and drug resistance across *Candida* opportunistic pathogens

Miquel Àngel Schikora-Tamarit[1,2] , Toni Gabaldón[1,2,3,4,*]

1) Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell 1-3, 08034 Barcelona, Spain.

2) Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.

3) Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

4) Centro Investigación Biomédica En Red de Enfermedades Infecciosas, Barcelona, Spain.

*Author for correspondence: toni.gabaldon@bsc.es

ABSTRACT

Understanding how pathogens adapt to drug treatment, to the host and to clinical environments is key to improving current therapies and diagnosis. This is particularly relevant for opportunistic pathogens, which alternate between host and environment. Here, we analyzed small, copy-number and structural variants across ~2,000 genomes of isolates from six major opportunistic *Candida* species and identified multiple genes under recent selection, suggesting that clinical adaptation is diverse and complex. These involve species-specific but also convergently selected processes, such as adhesion, which may underlie core adaptive mechanisms. Using genome-wide association analysis we identify known drivers of drug resistance alongside potentially novel players. Finally, our analyses reveal an important role of the generally neglected complex genomic variants, and suggest an unexpected involvement of (para)sexual recombination in the spread of resistance mechanisms. Altogether, our results provide novel insights on how opportunistic pathogens adapt to human-related environments and unearth candidate genes that deserve future attention.

INTRODUCTION

Fungal infections pose a serious health threat, affecting >1 billion people and causing ~1.5 million deaths each year[1,2]. The problem is growing due to insufficient diagnostic and therapeutic options[3,4], increased number of susceptible patients[1,5], expansion of pathogens partly linked to climate change[6,7], and the alarming rise of antifungal drug resistance[4,8,9]. *Candida* species are a major cause of severe hospital-acquired infections[1], prompting the classification of some species (*Candida auris, Candida albicans, Candida glabrata, Candida tropicalis* and *Candida parapsilosis*) as critical or high priority targets by the WHO[2].

A promising strategy to improve current therapies is to understand the evolutionary mechanisms of adaptation to antifungal drugs and to the human host. *Candida* pathogens have highly dynamic genomes (both within-species[10–12] and within-patient[13,14]), which likely underlie these adaptive processes[13,15–18]. For example, *in vitro* evolution studies have pinpointed genome-wide changes underlying drug resistance[19–21]. In addition, analysis of serial clinical isolates[13,14], Genome Wide Association Studies (GWAS)[22,23] and population genomics research[11,12,24] have clarified partially the clinical relevance of resistance mechanisms. Similarly, directed evolution experiments in mice[25–27], analysis of paired clinical isolates[13] and population genomics studies[12,28] have explored host adaptation mechanisms involving virulence, adhesion or filamentous growth. Furthermore, some studies used ratios between non-synonymous and synonymous variation (i.e. π_N/π_S) to infer signatures of selection, useful to predict genes involved in clinical adaptation where the relevant phenotypes (i.e. drug susceptibility or cell adhesion within a patient) are not measurable[12,29–31].

However, our understanding about how *Candida* species adapt in a clinical context is limited due to many reasons. First, most clinical studies include small sample sizes and/or lack rigorous statistical testing of the associations between genotypes and adaptive changes. Second, most studies involve only *C. albicans*, leaving open questions in other species[2]. Third, despite the importance of structural variants[32–34], their contribution to clinically-relevant adaptation remains largely unexplored. Fourth, similarities in adaptation mechanisms across species remain elusive because most studies focus on only one species and use different methods. This is key to understanding the epidemiology of these pathogens and enabling personalized treatments and prevention strategies. Fifth, many exploratory clinical studies focus only on known adaptive mechanisms (i.e. known drug resistance genes, as discussed in[23]), meaning that there may be unexplored factors. Sixth, current studies of selection consider all variants within a gene, which may reflect ancient adaptation unrelated to the clinics, and it may be important to focus on more recent signs of selection (as done in[35]).

To address these gaps, we used ~2,000 available genomes from major *Candida* species to investigate two open questions in clinical adaptation. First, we used phylogenetics and π_N/π_S-inspired tools to infer the genes with signatures of recent, clinically-relevant, selection in *C. glabrata*, *C. auris*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis*. Second, we used convergence-based GWAS to infer the genomic drivers of resistance towards echinocandins, polyenes and azoles in *C. glabrata*, *C. auris* and *C. albicans*. In both cases we measured the contribution of various variant types, including structural variants. Our analysis reveals both expected and novel adaptive mechanisms, including those convergently acting in several species.

RESULTS AND DISCUSSION

Publicly available sequencing data allows studying recent evolution of *Candida* pathogens at unprecedented scales

To identify genes under recent selection in *Candida* pathogens we retrieved all publicly available short-read whole-genome sequencing data for pure isolates of six major species, and identified four variant types: Single Nucleotide Polymorphisms (SNPs), small Insertions and Deletions (INDELs), Structural Variants (SVs) and Copy-Number Variants (CNVs) (**Online Methods**, **Figure 1, Figure S1 A,B**). We enriched genomic information with strain metadata from the literature, including isolation source and antifungal drug susceptibility where available (**Figure 1B, Table S1**). This dataset, comprising 1987 high-quality samples available at https://candidamine.org, is unprecedented in terms of types of variants and number of strains considered[11,12,24,28,36].

To provide a phylogenetic framework to our analysis we inferred a strain tree (**Figure 1B, Table S1**) and used a systematic approach to identify genetically divergent, monophyletic clades in each species (**Online Methods**, **Figure S2A**). Comparison with previously-defined clades (**Online Methods**) revealed an overall consistency, underscoring the validity of our clade-definition approach, but also showed that our dataset encompasses a higher intraspecific diversity. In summary, we generated a dataset with unprecedented power to study the signs of selection and drug resistance mechanisms in major *Candida* pathogens.

Generally overlooked structural re-arrangements underlie significant intra-specific genomic variation in *Candida* pathogens

To determine the relevance of considering different variant types in subsequent analyses, we quantified their relative contribution to genetic diversity. Such comparative analysis across *Candida* species is lacking, as most previous studies focus on SNPs and use specific methodologies. For each variant type we measured the genetic distance (# variants/kb) between all pairs of isolates within a given species. We found that most species span high levels of genetic diversity, close to 10 SNPs/kb (1% divergence) or higher between distant conspecific strains (**Figure 2**). In some species, this could be attributed to their hybrid nature (i.e. *C. orthopsilosis*[36] and *C. albicans*[10]). For non-hybrid species (*C. glabrata* and *C. auris*), this indicates that their diversification predates human colonization, which must have occurred in parallel in divergent clades for each species. *C. parapsilosis* is an exception to this trend, pointing to a more recent origin of this lineage.

Regarding non-SNP variants, we found that SV and INDEL diversity correlate to SNP diversity (**Figure 2A**), suggesting they are accurately called. CNVs, however, displayed distinct diversity patterns, which may result from overspliting given the difficulty of defining their precise boundaries. As expected, SNPs are quantitatively the most common variant type, followed by INDELs, one order of magnitude less prominent, and then SVs and CNVs, at much lower frequencies (**Figure 2B**). Despite their lower abundance, SVs and CNVs can affect a significant fraction of protein-coding genes (**Figure 2C**), highlighting their relevance. We investigated mechanisms underlying the formation of SVs and CNVs and found that most variants are unrelated to repetitive elements or rearrangements derived from homologous recombination (**Figure S1 C,D**). This suggests that Non-Homologous End-Joining DNA repair pathways[37,38] could be the main driver of SVs/CNVs in *Candida* species. In summary, we find that all variant types are quantitatively important and therefore should not be overlooked in subsequent analyses.

Signatures of recent selection reveal known and novel mechanisms of adaptation

To infer signatures of recent, clinically-relevant selection we took advantage of the predominance of clinical strains in our collection. We reasoned that recently acquired variants in clinical isolates are likely enriched in those acquired in a clinical context, and could therefore inform on selective pressures related to adaptation to human-related environments. The standard approach of quantifying π_N/π_S ratios[12,31,39,40] is not suitable for our aim for the following reasons. First, we focus on recently-acquired variants, and π_N/π_S considers all mutations in a gene, thereby also detecting ancient selection. Second, considering only recent variants poses a statistical challenge to reliably calculate π_N/π_S, since many genes have few recent variants and thus a π_S of 0. Third, π_N/π_S cannot be applied to INDELs, SVs and CNVs, which we deem important.

To overcome these drawbacks, we developed a π_N/π_S-inspired method that detects genes with an excess of recent functionally-relevant variants (either nonsynonymous SNPs (nsyn_SNP), in-frame INDELs (if_INDEL), gene duplications (DUP) or truncations (DEL)) (**Online Methods**, **Figure 3A, S3**). DUPs could be SVs or CNVs, and DELs may be nonsense SNPs, frameshifting INDELs, SVs or CNVs. To focus on recent variants, we identified monophyletic clusters comprising only clinical strains with high genetic relatedness (**Figure S2 B,C**), and only considered variants inferred to have appeared within the cluster. These clusters likely represent clonally propagating lineages that evolved in human-associated environments (as they are closely related and recurrently isolated from patients), and therefore recently-appeared mutations may reflect selective pressures related to adaptation to the host, hospital environments, or antifungal drugs. We used these variants to define 'genes under recent selection' as those showing an excess of functionally-relevant variants.

Our approach detected many recently selected genes belonging to 879 / 7499 Orthologous Groups (OGs, a proxy for gene families) (**Figure 3B, Table S2**). The low numbers in *C. orthopsilosis* and *C. parapsilosis* likely reflect reduced statistical power due to few strains or low intraspecific diversity, respectively. Most OGs are affected by a single variant type, with few exceptions that suggest evolutionary antagonistic effects among OG members (**Figure 3B**, **Supplementary Results**). Supporting the validity of our approach, we find several expected genes related to virulence and drug resistance (**Figure 3B, Table S2**). Some examples include *ALS* genes in *C. albicans* (implicated in adhesion and biofilm formation[41]), *TAC1b / ERG11 / MRR1* in *C. auris* (related to azole resistance[11,21,42,43]), *PDR1* in *C. glabrata* (implicated in azole resistance[19,44]), *EPA* genes in *C. glabrata* (related to adhesion[32,45]), a drug exporter in *C. orthopsilosis* (gene *CORT_0G00240*) or filamentous growth proteins in *C. tropicalis* (genes *CTRG_00655* or *CTRG_03085*). We find signs of selection on all variant types in most species, suggesting that considering SVs/CNVs is relevant. This gene catalog constitutes a valuable resource to validate the clinical relevance of evolutionary mechanisms inferred from future non-clinical studies (i.e. *in vitro* evolution[19,21], virulence in animal models[27] or high-throughput genotype-phenotype screenings[46,47]).

Only 68/879 OGs have a gene affected by selection in different species, suggesting that each species has unique signatures of selection (**Figure 3C**). This is consistent with generally different infection mechanisms in each species which is also reflected in mostly non-overlapping transcriptional profiles upon host interactions[48,49]. However, in many instances the number of shared OGs is higher than expected by chance ($p<0.05$) (**Online Methods**, **Figure 3C, Table S2**), pointing to convergent adaptive mechanisms in *Candida* pathogens. Some examples include *ALS* genes from *C. albicans* and *C. auris*, *OPT2/OPT3* genes (transporters related to pseudohyphal growth and fluconazole presence) in *C. albicans* and *C. tropicalis*, *MRR1a* in *C. auris* and *C. tropicalis* (related to drug resistance), *FLO8/MSS11* genes (related to pseudohyphal growth) in *C. glabrata* and *C. auris*, *MDS3* (virulence factor) in *C. albicans* and *C. auris*, *CST6* (associated to azole resistance[22]) in *C. glabrata* and *C. auris* and *WOR4* (related to phenotype switching) in *C. albicans* and *C. auris*.

We performed enrichment analyses on functional annotations and found 1074 domains, 151 GO terms, 5 MetaCyc and 3 Reactome pathways enriched across all gene sets (**Figure 4, S4, Table S2**), including hyphal growth, biofilm formation, transcriptional regulation, response to temperature, cell adhesion, carbohydrate metabolism, cell wall and membrane regions (**Figure 4**). Most enriched functional groups are unique to a single species (991/1074 domains, 143/151 GO terms, and all Metacyc and Reactome pathways), suggesting that each species has unique signatures of recent selection also at the pathway and domain level (**Figure 4**). However, there are several convergently affected pathways, which may reflect conserved adaptive mechanisms (**Figure 4, S4**)**.** Relevant examples include a Zn-dependent transcription factor domain in *C. tropicalis*, *C. albicans* and *C. auris*, disordered regions in *C. tropicalis*, *C. albicans* and *C. glabrata* or

hyphally-regulated cell-wall proteins in *C. tropicalis*, *C. albicans* and *C. auris*. We also find several GO terms related to adhesion ('biological process involved in symbiotic interaction', 'adhesion of symbiont to host' and 'cell-cell adhesion') enriched in genes with selected deletions from *C. tropicalis*, *C. albicans* and *C. glabrata*, suggesting recurrent rewiring of these functions (**Figure 4**). Further research is needed to associate these functions with possible adaptive advantages. For instance, disordered proteins can generate new traits in yeast[50], and the deletion of adhesion genes could modulate host attachment, biofilm formation or immune evasion[51–54], therefore improving survival. We also evaluated species-specific functional enrichments, which may underlie particular adaptations. In *C. glabrata*, these include response to oxygen-containing compounds, regulation of filamentous growth, regulation of actin nucleation, pigment catabolism, tRNA methylation, bud-site selection, G-protein signaling and nuclear membrane proteins. In *C. auris*, rRNA binding, the TTT complex (regulating the phosphatidylinositol-3-related kinase) and the host cellular component. In *C. tropicalis*, flocculation, regulation of transcription (related to the histone deacetylase complex), oligosaccharide synthesis, glucan degradation and protein mannosylation. In *C. albicans*, response to temperature stimulus, inositol phosphate dephosphorylation, cell wall and extracellular region. Finally, in *C. orthopsilosis* we find carbohydrate metabolism (including hexose transport) functions. In summary, our results suggest hundreds of gene families (~10% of all families) and pathways under recent selective pressure, often in a single species. This may be explained by the natural niche of these pathogens being massively different to the human host. In addition, we find convergently selected families and pathways, which may be at the core of recent adaptation and constitute interesting therapeutic targets. Future experiments should validate these results and pinpoint the most important drivers of recent adaptation.

Convergence GWAS suggests drivers of azole, polyene and echinocandin resistance in *C. auris*, *C. glabrata* and *C. albicans*

Drug susceptibility is a measurable phenotype which is available for a sizable fraction of our strains (**Table S1, Figure 5A**), which motivated us to find genomic changes underlying the drug resistance phenotype in clinical isolates. For this, we performed a convergence-based Genome Wide Association Study (GWAS), which uses ancestral state reconstruction (ASR) to find variant changes that are significantly associated with transitions in drug resistance phenotypes in their reconstructed evolutionary histories[55,56]. Given the peculiarities of our dataset we developed a custom pipeline, inspired by hogwash's synchronous algorithm[57] (**Online Methods** and **Figure 5B**). In addition, to take into account that different variants may drive drug resistance by altering the same feature (i.e. a gene, or a pathway), we tested the association between groups of collapsed variants and the phenotype. To focus on key associations we only analyzed species-drug combinations with at least 5 sharp transitions (from high susceptibility to high resistance, or vice versa) (**Online Methods** and **Figure S5**). This resulted in 12 species-drug datasets including seven compounds from

all main classes (azoles, echinocandins, and polyenes) and covering most clades of *C. albicans*, *C. glabrata* and *C. auris* (**Table S1, Figures 1B, 5A**). To ensure high-confidence hits we used a conservative approach that minimized the false positives expected from such multiple testing and chose the GWAS algorithm parameters and filtering criteria based on prior expectations of resistance genes (**Online Methods**, **Figure S6**). To remove redundancy we kept the strongest, most specific association among overlapping high-confidence variants/genes/domains/pathways (**Online Methods** and **Table S3**). As an example of a significant association, we find that small variants affecting *PDR1* (drug efflux regulator[44]) are correlated with voriconazole resistance in *C. glabrata* (**Figure 5C, Table S3**). In **Supplementary Results** we discuss results that do not meet this stringent selection but that we deem interesting.

Unexpectedly, in some cases, the Manhattan plots showing variant-phenotype correlations suggested the existence of linked variants, i.e. distant variants jointly segregating with the phenotype (**Figure S7**, **Supplementary Results**). Such a distribution may be explained by recent inter-strain recombination partly underlying the emergence of drug resistance. This is consistent with previous studies suggesting sexual (or parasexual) cycles in these species[12,28,58], and points to a possible role of (para)sexual recombination in the spread of antifungal resistance. A possible role of recombination makes the detection of causal variants slightly more difficult, as they may be linked to passenger variants unrelated to the phenotype. We therefore focused on protein-altering variants, which are more likely to underlie changes in drug resistance[19,59,60]. When considering all types of groupings, we find 227 non-redundant significant associations (hits) affecting 130 OGs and 38 pathways across all 12 datasets, with variations across datasets likely reflecting differences in sample size (**Table S3, Figure 6**). Close examination of these hits underscored the importance of considering SVs/CNVs and domain/pathway grouping of variants (**Supplementary Results**).

In summary, our multi-species genotype-phenotype association study helps reveal genome-wide determinants of drug resistance to all major drug classes. Beyond our analysis, this is a valuable resource to validate that the resistance mechanisms found in future studies are meaningful in clinical isolates, as we illustrate for a recent *in vitro* evolution study[19] (**Supplementary Results**).

GWAS analysis suggests novel and known drivers of drug resistance

To validate our strategy and gain insights into known mechanisms of antifungal drug resistance we checked the GWAS results for expected driver genes (**Supplementary Results, Figure S8**). Our analysis confirm that *ERG11* (target of azoles[61]) is associated with *C. albicans* fluconazole and *C. auris* fluconazole/voriconazole resistance, *TAC1b* (drug efflux regulator[59]) underlies pan-azole resistance in *C. auris*, *FKS* (echinocandin target[62]) mutations are likely drivers of strong pan-echinocandin resistance in *C. auris* and *C. glabrata,* and

PDR1 underlies pan-azole resistance in *C. glabrata*. Conversely, *ERG11* may be unrelated to resistance towards some azoles in *C. auris*.

Beyond these 'known genes' our results hint to other players. To focus on the most relevant, potentially conserved mechanisms we considered OGs associated with resistance in more than one drug-species combination (**Table S3**). These include *PDR1*, *ERG11* and 13 other OGs, which are often (12/13 OGs) related to 'core' resistance mechanisms towards multiple drugs of the same species. We find six such OGs in *C. glabrata* related to various azoles and micafungin resistance, including four adhesin families (*CAGL0J01727g*, *PWP4*/*AWP13*, *AWP4*/*AWP9* and *EPA19*/*EPA11*), the ortholog of *Saccharomyces cerevisiae NET1* (putative chromatin-silencing rRNA regulator), and *CAGL0K07502g* (a protein with unknown function). The link between adhesins and resistance could be explained by their role in biofilm formation, a known resistance mechanism[63,64]. In addition, the role of *NET1* is consistent with studies linking chromatin silencing with azole resistance in *C. glabrata*[65], and with the observation that its deletion in *S. cerevisiae* increases sensitivity to some compounds[66,67]. Similarly, we find six 'core' OGs in *C. auris*, including *B9J08_005550* (with RNA binding activity) related to fluconazole and voriconazole resistance, *B9J08_004248*/*B9J08_004896* (putative RNA-dependent DNA polymerases) related to amphotericin B and multiple azole resistance, *B9J08_004249*/*B9J08_005494* (putative Zn-binding TFs) associated to amphotericin B and fluconazole resistance, and the ortholog of *S. cerevisiae MRPS35* (mitochondrial ribosomal protein) related to itraconazole and voriconazole resistance. These results suggest that different aspects of gene regulation (transcription and RNA life cycle regulation) are key for multi-drug resistance in *C. auris*. In addition, the role of *MRPS35* is consistent with the observations that its deletion decreases resistance to some compounds in *S. cerevisiae*[67], and that mitochondrial regulation is linked to drug efflux in *C. albicans*[68]. On another note, we find one OG related to fluconazole resistance in both *C. glabrata* and *C. auris*, affecting the orthologs of *S. cerevisiae NRG1* and *NRG2*, respectively, both transcriptional repressors. These *NRG1*/*NRG2* convergent associations suggest that this is a conserved drug resistance mechanism across species. This is consistent with the fact that both *NRG1*/*NRG2* null mutants impact azole resistance in *S. cerevisiae*[69,70]. We next considered pathways significant in multiple datasets, and we found one such Reactome annotation in *C. auris* voriconazole and posaconazole (Miscellaneous transport and binding events) (**Table S3**), which may also underlie a core resistance mechanism. In summary, we find several lesser known gene families associated with resistance in multiple datasets, which illuminate core and conserved functions related to antifungal drug resistance. These results may guide future confirmatory experimental work, which could inform about the most important drivers and suggest relevant therapeutic targets.

CONCLUSIONS

Understanding human-associated adaptation in pathogens is a long-standing question because it underlies virulence, hospital transmission and drug resistance mechanisms. Our current knowledge is limited due to insufficient sampling, lack of multi-species studies and exclusive focus on SNPs and on specific genes. We have addressed these gaps in six major *Candida* species by analyzing the public genomes and phenotypes of ~2,000 (mostly clinical) strains. Our collection is a valuable resource due to its unprecedented size, the common analysis framework in multiple species, the consideration of complex variants (SVs and CNVs) and the availability of phenotypes. This underscores the value of depositing genomic and clinical data in public repositories that can be mined to generate new knowledge.

First, we used the generated variants to find genes affected by recent, clinically-relevant selection. We found hundreds of affected gene families and pathways, mostly species-specific, suggesting highly variable, multifactorial adaptive mechanisms. In addition, we predicted novel conserved adaptive processes involving drug resistance and cell adhesion functions, which are interesting pan-fungal therapeutic targets. We next analyzed the variants, genes, and pathways associated with clinical resistance towards all major antifungal drugs in three *Candida* species. Beyond confirming the implication of known drivers of resistance, which validates our approach, our results identified potential novel players related to adhesion, biofilm formation and transcriptional regulation. These novel mechanisms involve genes underlying cross-resistance towards multiple drugs of the same species and also gene families driving resistance in multiple species. Beyond the general trends discussed here, our catalog of selection signatures and drivers of drug resistance is valuable to validate gene functions inferred from non-clinical studies (i.e. drug resistance genes predicted from *in vitro* evolution). Finally, our analyses reveal an important role of the generally neglected complex variants (CNV and SV), and suggest an unexpected involvement of (para)sexual recombination in the spread of resistance mechanisms.

All in all, we provide novel insights and valuable resources that improve our understanding about selection and drug resistance across major *Candida* pathogens. Our findings may guide future confirmatory experiments, which could improve therapeutic and diagnostic options.

ONLINE METHODS

1. Generation of the filtered variant calling dataset for each *Candida* species

We used the SRA toolkit (v2.10.9) (https://github.com/ncbi/sra-tools) to download all paired-end whole-genome re-sequencing (WGS) datasets for the NCBI taxon IDs[71] related to each species (*C. albicans, C. auris, C. glabrata, C. tropicalis, C. orthopsilosis* and *C. parapsilosis*) from the SRA database[72] as of 09/06/2020. For each run we used fastQC (v0.11.9) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and trimmomatic (v0.38)[73] with default parameters to remove adapters and trim the reads. Finally, we ran perSVade (v0.6)[74] to align (with bwa mem (v0.7.17) (http://bio-bwa.sourceforge.net/bwa.shtml)) the trimmed reads to the reference genome (included in **Table S1**) and calculate the coverage per windows (using mosdepth (v0.2.6)[75])**.** We filtered out low quality runs having a read depth <40x or covering <90% of the reference.

We next used the aligned reads to call variants with perSVade (v0.6)[74] which calls and functionally annotates Single Nucleotide Polymorphisms (SNPs), small Insertions and Deletions (IN/DELs), copy number variants (CNVs) and structural variants (SVs). SVs are complex variants where we could find the precise underlying rearrangements (i.e.: tandem duplications, inversions, or balanced translocations). Conversely, CNVs are variants generating large (>600 bp) duplications and deletions (inferred from changes in read-depth) with unknown underlying rearrangements. Technically, CNVs are a type of SV, but we differentiate them because the method used to infer them is different, and some CNV-like SVs (i.e. tandem duplications) may be detectable with the coverage-based method but not with the SV-detection method. By considering these two types of variants, we provide a comprehensive characterization of structural variants. Note that any CNV that had an equivalent SV was not considered.

The small variant calling pipeline integrates the results of three callers (GATK Haplotype Caller (v4.1.2)[76], freebayes (v1.3.1)[77] and bcftools (v1.9) (https://github.com/samtools/bcftools)). The CNV calling pipeline detects deletions and duplications from coverage alterations using two algorithms (HMMcopy (v1.32.0)[78] and AneuFinder (v1.18.0)[79]). The SV calling pipeline finds rearrangements with gridss (v2.9.2)[80] (which uses split reads, discordantly paired reads and *de novo* assembly signatures) and summarizes them into actual SVs with clove (v0.17)[81]. The called SVs are tandem duplications, deletions, inversions, translocations, copy-paste insertions, cut-paste insertions, inverted copy-paste insertions, inverted cut-paste insertions, inverted translocations and unclassified breakpoints (see **Figure S1**). In addition, perSVade automatically selects the optimal gridss / clove filtering parameters for each sample based on simulations of SVs, which is useful for *Candida* species (where SV-callers have not been tested extensively). PerSVade also integrates SVs

and CNVs, which may be partially redundant, so that any CNV overlapping an equivalent SV would be discarded. Finally, this pipeline uses VEP (v100.2)[82] to annotate the functional effect of each variant, and RepeatModeler (v2.0.1)[83] followed by RepeatMasker (v4.0.9)[84] to annotate which variants overlap repeats. Note that, for the functional annotation we used the gff files corresponding to each genome (included in **Table S1**) with the exception of *C. tropicalis* and *C. parapsilosis* (which lacked annotations of the mtDNA). For these two species we generated the mtDNA annotations with augustus (v3.2.3)[85] using default parameters and 'candida_albicans' as the train species.

We ran perSVade with custom parameters adapted to either haploid species (*C. glabrata* and *C. auris*) or diploid species (*C. albicans, C. tropicalis, C. parapsilosis* and *C. orthopsilosis*). For small variant calling, we used '--ploidy 1 --run_ploidy2_ifHaploid' (for haploid species, which runs the calling in both haploid and diploid mode), '--ploidy 2' (for diploid species) and '--coverage 12' (to discard positions with <12x read depth). Note that we ran the variant calling in diploid mode for haploids to take into account that they may have heterozygous variants in duplicated regions. For CNV calling, we used '--window_size_CNVcalling 300' (to call CNVs based on windows of 300bp) and '--min_CNVsize_coverageBased 600' (to discard CNVs <600bp). For SV calling, we used '--min_chromosome_len 100000' (to use only large chromosomes for SV simulations), '--simulation_ploidies auto' (which results in parameter optimization based on haploid SVs (for haploid species) or heterozygous SVs (for diploid species)) and '--range_filtering_benchmark theoretically_meaningful_NoFilterRepeats' (to run parameter optimization without filtering out repetitive elements). In addition, we used a custom function from perSVade's source code (function 'get_integrated_SV_CNV_df_severalSamples' (v0.6)[74]) to integrate the CNVs and SVs from different samples in a way that equivalent variants get the same ID. This is not a trivial task, since the used algorithms often lack single-bp resolution, so that the same variant in different samples may get slightly different coordinates. To solve this, the function 'get_integrated_SV_CNV_df_severalSamples' from perSVade uses bedmap from the bedops suite (v2.4.39)[86] to cluster variants from the same type that reciprocally overlap by >75% of their total length and where their breakpoints are <50bp from each other. In addition, note we ran perSVade with custom NCBI translation codes (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) to perform functional variant annotations. We set the gDNA code to either 1 for *C. glabrata* (standard code) or 12 for *C. albicans, C. tropicalis, C. parapsilosis*, *C. auris* and *C. orthopsilosis*. We set the mtDNA code to 4 for *C. albicans, C. tropicalis, C. parapsilosis* and *C. orthopsilosis* or 3 for *C. auris* and *C. glabrata*. This procedure yielded the raw variant calls and their corresponding functional annotations. We discarded all runs where any of these steps (read trimming, alignment or variant calling) could not be performed due to file truncation or incompatible file formats.

To get the high-confidence variants we applied some extra filterings to discard artifacts. For small variants, we kept variants that passed the filters in at least two callers and where the fraction of reads covering the variant was ≥0.9 (for haploid configuration) or ≥0.25 (for diploid configuration). For CNVs, we filtered variants based on both the predicted relative copy number (CN, which in a diploid may be 0 for a homozygous loss, 0.5 for a heterozygous loss, 1.5 for a trisomy and 2.0 for a tetrasomy) and the relative coverage (measured as the ratio between the median coverage of the region under CNV and the median coverage across the whole gDNA). For deletions, we required CN=0.0 and relative coverage≤0.1 (for haploid species) or CN≤0.5 and relative coverage≤0.6 (for diploid species). For duplications, we required CN≥2.0 and relative coverage≥1.7 (for haploid species) or CN≥1.5 and relative coverage≥1.3 (for diploid species). For SVs, we calculated the variant allele frequency (VAF, as in https://github.com/PapenfussLab/gridss/issues/234) for each breakend forming each SV to discard variants with low VAF that may not be real haploid / diploid events. We kept SVs fulfilling two criteria: 1) they should have at least one breakend with VAF≥0.8 (for haploid species) or VAF≥0.3 (for diploid species) and 2) all breakends should have VAF≥0.2 (for haploid species) or VAF≥0.1 (for diploid species). These filters yielded the high-confidence variant calls used in this paper. Note that for haploid species we used the small variants called in haploid configuration in all analyses described below (unless specifically indicated).

2. Strain-tree generation

To reconstruct a phylogenetic tree for all strains of a given species we used a different approach depending on the species ploidy. For haploid species we generated a pseudo-genome sequence for each strain based on the reference genome but substituting the reference sequences according to filtered haploid SNPs. To avoid the biases introduced by CNVs and INDELs, these pseudo-genomes only included positions matching the following criteria in all strains: 1) coverage≥12x, 2) absence of INDELs and 3) absence of heterozygous SNPs. In addition, we only considered variable positions. We used Biopython (v1.78)[87] and bedmap to obtain the aligned pseudo-genomes, with 285,345 sites in *C. auris* and 311,174 sites in *C. glabrata*. We then obtained the unrooted tree with iqtree (v2.1.2)[88] from these aligned pseudo-genomes using '-m TEST+ASC', to use default automatic model selection and ascertainment bias correction (which is necessary to calculate meaningful branch lengths). We then used midpoint rooting to get the final tree, which has support values from 1000 bootstraps. Note that we visually inspected the heterozygous SNP patterns (along the tree) in these haploids to pinpoint runs that may be mixed or contaminated strains (which are expected to have many heterozygous SNPs). We found two such *C. auris* samples that had heterozygous SNPs with a VAF ~ 35%, which were discarded from subsequent analyses.

For diploid species it was not possible to use an analogous method due to high heterozygosity in *C. albicans*[28], *C. tropicalis*[24] and *C. orthopsilosis*[36]. We implemented a tree-generation method, inspired by [24,89],

to take into account both homozygous and heterozygous SNPs. We generated 100 pseudo-genome sequences for each strain based on the reference genome, but substituting the reference sequences according to filtered SNPs (only those that had defined heterozygous or homozygous genotype calls). These pseudo-genomes only included positions matching the following criteria in all strains: 1) coverage≥12x and 2) absence of INDELs. Each of these 100 pseudo-genomes included all homozygous SNPs and a random selection of heterozygous SNPs (each heterozygous SNP with a probability of 0.5 to be included). We then obtained one unrooted tree for each of these 100 aligned pseudo-genomes (with only variable positions) with iqtree using '-m GTR+F+ASC+G4' (equivalent to the "GTRGAMMA" model used in [24]), required to have a consistent model and ascertainment bias correction. The pseudo-genomes had 319,439-320,188 sites for *C. albicans,* 765,044-766,422 sites for *C. tropicalis,* 11,627-11,827 sites for *C. parapsilosis* and 575,685-576,053 sites for *C. orthopsilosis*. We rooted all 100 trees with midpoint rooting, and generated a final consensus tree with branch lengths using iqtree (-con argument) followed by the consensus.edges function from phytools (v0.7_90)[90]. Note that the branch support for this consensus tree was derived from the number of re-sampled trees including a given branch. **Table S1** includes all the used trees in newick format.

3. Clade definition

To define meaningful clades in each tree we first identified potential 'clade-qualifying' nodes as those having support≥95 and long subtending branches (above a 'min_relative_branch_length' threshold). For a given 'min_relative_branch_length' threshold, the clades would be 'clade-qualifying' nodes where none of the children are also 'clade-qualifying' nodes. We defined the 'relative_branch_length' for each node of each tree as the actual branch length normalized by the farthest distance between any two nodes. Thus, the 'min_relative_branch_length' was the minimum 'relative_branch_length' required for 'clade-defining' nodes. Note that the choice of a meaningful value for 'min_relative_branch_length' was not trivial, and some values may leave out many strains without an assigned clade. To identify a reasonable 'min_relative_branch_length' for each tree we tried a range of values (between 0.001 and 0.2) and calculated, for each value, the total number of clades and the fraction of samples assigned to some clade. We defined as the final 'min_relative_branch_length' a value that maximized the number of samples with a clade and minimized the total number of clades. We could find such optimal values, resulting in 4-24 clades (depending on the species) and >90% of strains within some clade for all species (**Figure S2A**).

To evaluate our clade definition we compared it with previous population genomics studies for *C. albicans*[28], *C. auris*[11], *C. glabrata*[12], *C. tropicalis*[24] and *C. orthopsilosis*[36] (**Figure 1B**). We found that most clades (21/21 in *C. albicans*, 22/24 in *C. glabrata*, 4/5 in *C. orthopsilosis*, 2/4 in *C. auris* and 2/3 in *C. tropicalis*) were either new (the strains within the clade were not included in the previous study) or had a 1-to-1 strain correspondence with the previous study. To verify the absence of artifactual clades we manually inspected

the inconsistencies (see **Table S1**). We found that our clades 15 and 8 from *C. glabrata* were grouped into clade 5 in [12], but our larger dataset provides higher resolution supporting the split of this clade in two. This is consistent with previous reports suggesting that clade 5 from [12] is polyphyletic [29]. In addition, we found that one *C. auris* strain (SRR10852068) was assigned to clade 3 in [11] but it appears as clade 2 in our analysis (clade 1 from [11]), suggesting misclassification in [11]. This means that our clade definition in *C. auris* is fully consistent with [11] except for this strain. Furthermore, we found that our tree topology in *C. orthopsilosis* is different around clade 4 (as compared to [36]), resulting in some unclassified samples. Finally, we describe three highly divergent clades in *C. tropicalis* (**Figure 1B, 2A**), whereas [24] only assigned clades for one of them (our clade 3). This explains the inconsistency in our clade assignment. Taken together, these findings suggest that our clade assignments are largely consistent with previous findings. **Table S1** lists all current and former clade assignments.

4. Generation of the strain metadata and definition of drug resistance

To obtain relevant metadata information (type of isolate and drug susceptibility information) for all datasets with variant calls we compiled two types of information. We first used either the BioSampleParser package (https://github.com/angelolimeta/BioSampleParser) or Entrez-Direct utilities (v13.9)[91] (only if BioSampleParser failed) to get the BioSample annotations (http://www.ncbi.nlm.nih.gov/biosample/) for each sequencing dataset. This provided the already accessible, machine-ready metadata, including the strain IDs. We then manually curated the literature associated with each of these strains to get the information about the type of strain and the available drug susceptibility information. From a total of 1987 samples, we could find 1705 clinical isolates, 30 environmental strains, 49 genome-engineered strains, 201 strains from directed evolution experiments and 2 reference samples. We could find Minimum Inhibitory Concentrations (MIC) or reports (statements in the literature) about the susceptibility towards amphotericin B (AMB, 464 strains), beauvericin (BVN, 5 strains), 5-flucytosine (5FC, 162 strains), terbinafine (TRB, 1 strain), miconazole (MIZ, 11 strains), ketoconazole (KET, 69 strains), isavuconazole (IVZ, 47 strains), voriconazole (VRC, 250 strains), posaconazole (POS, 214 strains), itraconazole (ITR, 151 strains), fluconazole (FLC, 796 strains), micafungin (MIF, 462 strains), caspofungin (CAS, 463 strains) and anidulafungin (ANI, 141 strains). To define discrete susceptibility profiles for each strain (susceptibility (S), intermediate susceptibility (I) or resistance (R)) we relied on either breakpoints for MIC data or direct reports of R/S (when MIC data was not available). We defined the breakpoints (BPs) for MICs based on either EUCAST recommendations (v10.0) (https://www.eucast.org/), previous work[11,92,93] or manually curated breakpoints based on our data (**Figure S5**). If MIC was available, we defined each strain into R ($MIC \geq 2 \cdot BP$), S ($MIC \leq BP/2$) or I ($BP/2 < MIC < 2 \cdot BP$). Note that **Table S1** includes all this metadata.

5. Diversity analysis

To measure the pairwise genetic distance (number of variants / kb) across all pairs of isolates in a given species we counted the filtered variants unique to each strain of the pair. To measure the number of genes with protein-altering variants between each pair of isolates we calculated the number of proteins altered by these unique variants (according to perSVade's functional annotation). For small variants, we considered either haploid mutations (for haploid species) or both homozygous and heterozygous variants (for diploid species). For SVs and CNVs we considered all variants.

We calculated the Minor Allele Frequencies (MAF) for each haploid SNP, SV and CNV as $MAF = (\#\ strains\ w/\ var)/(\#\ strains)$. This may be an over-simplification for SVs and CNVs, but we considered it appropriate since we could not get precise genotype calls for such complex variants. For each diploid small variant, we calculated it as:

$$MAF = \left(\sum_{i=1}^{n} GT_i\right)/(\#\ strains)$$

Where *n* is the number of strains with the variant, *i* refers to the strain (from 1 to *n*) and GT_i is either 0.5 (for heterozygous calls) or 1.0 (for homozygous variants). Note that we only considered diploid small variants with a genotype call (homozygous or heterozygous) that was consistent across all algorithms that identified a given variant. In addition, we only considered MAFs for variants with a MAF<0.5. **Figure S1B** includes the MAF distributions.

6. Investigating mechanisms of structural variant formation

To understand the mechanisms of SV and CNV formation we first investigated whether each variant overlaps RepeatMasker annotations[84]. We extracted the regions under SV/CNV (duplicated, inverted, deleted or translocated) and ran RepeatMasker on them using standard libraries and species-specific RepeatModeler[83] libraries. We used the module 'infer_repeats' of perSVade[74] to run these programs. If ≥ 10% of the altered region (duplicated, inverted, deleted or translocated) was covered by a RepeatMasker annotation, this was considered as the formation mechanism. These included insertions of transposable elements and expansions / contractions of tRNAs, rRNAs or simple repeats. We could not find such overlaps for most variants (**Figure S1 C,D**) suggesting that other mechanisms are essential for SV/CNV formation. For all remaining variants we investigated the role of homologous regions in SV formation, which could be relevant[37,38]. We checked whether each variant had breakpoints with either exact microhomology (2-10bp are identical), inexact microhomology (2-10bp are similar), exact homology (>10bp are equal) or inexact

homology (>10bp are homologous) between the breakends. Variants with microhomology may have been generated by Microhomology-Mediated End Joining (MMEJ, a Double Strand Break (DSB) repair pathway), and variants with long homology could be attributable to meiotic Non-Allelic Homologous Recombination (NAHR)[37]. If none of these signatures was found we classified the variant as 'other', which may be related to Non-Homologous End Joining (NHEJ) to repair DSBs[37]. Note that we did not consider variants potentially biased by overlapping simple repeats and low complexity regions for this analysis. For CNVs, such variants were those with simple repeats of low complexity regions spanning ≥25% of the CNV (inferred with RepeatMasker), which may affect coverage calculations. For SVs, these were variants where at least one breakend was overlapping any such repetitive elements, inferred with bedmap**. Figure S1 C,D** includes the results of this analysis.

7. Gene annotations

We obtained broad gene annotations (gene name, type of gene, location, description and *Saccharomyces cerevisiae* orthologs) from the Candida Genome Database (CGD) chromosomal feature files[94] (available at **Table S1**). The gene length was calculated from the gff annotations, considering UTRs if available. To get protein functional annotations we first obtained the protein sequences by 1) retrieving spliced transcripts from each gff with gffread (v0.12.1)[95] and 2) translating these transcripts with Biopython. We next ran Interproscan (v5.52-86.0)[96] on these proteins with the arguments '-appl Pfam,ProSitePatterns,ProSiteProfiles,PANTHER,TIGRFAM,SFLD,SUPERFAMILY,Gene3D,Hamap,Coils,SMART,CDD,PRINTS,PIRSR,MobiDBLite,PIRSF' (to run several annotation modules), --pathways (to get Metacyc and Reactome annotations) and -goterms (to get automatic Gene Ontology (GO) annotations). To get information about orthologous groups (hereafter referred to as 'gene families') we ran orthofinder (v2.5.2)[97] (with arguments '-M dendroblast -S diamond') on the proteomes of all *Candida* species. To get the set of GO annotations shown in all the tables we mixed annotations from both Interproscan and CGD (see **Table S1**).

To get the pathway annotations for GWAS and enrichment analyses (see below) we applied some extra steps. To map each gene to the complete set of Metacyc pathways we took all annotations from Interproscan, and added the parent pathways (using Pathway Tools (v25.0)[98]). We discarded Metacyc pathways where the taxonomic range did not include Ascomycota. Similarly, to map each gene to the set of Reactome pathways we tool the Interproscan annotations and added the parents (using the files ReactomePathways.txt and ReactomePathwaysRelation.txt from https://reactome.org/download/current/ at 04/10/2021). Since Reactome has several mammalian-specific pathways, we only kept annotations under these groups: 'Metabolism of proteins', 'Autophagy', 'Transport of small molecules', 'Gene expression (Transcription)', 'Cellular responses to stimuli', 'Reproduction', 'Digestion and absorption', 'Signal Transduction', 'Extracellular matrix organization', 'DNA Repair', 'Chromatin organization', 'Cell Cycle',

'Metabolism', 'Organelle biogenesis and maintenance', 'DNA Replication', 'Programmed Cell Death', 'Vesicle-mediated transport', 'Metabolism of RNA', 'Cell-Cell communication', 'Protein localization' and 'DNA replication and repair'. In addition, we only considered pathways annotated for 'Saccharomyces cerevisiae' and 'Schizosaccharomyces pombe'. Finally, to map each gene to all GO terms we used both annotations from CGD and Interproscan, and added all the parent terms (using goatools (v1.1.6)[99] and the obo file from http://purl.obolibrary.org/obo/go/go-basic.obo at 30/06/2021). In addition, to ensure that the annotated terms are meaningful in each species, we only kept GO terms that were defined in some gene of the CGD-curated dataset (see **Table S1**).

8. Measuring signatures of recent selection

Measuring selection in such population genomic data is often achieved through the use of sweep detection-based or π_N/π_S-based (similar to dN/dS but for population genomic data[12,39]) methods[40]. *Candida* species mostly propagate clonally, which suggests that a π_N/π_S-based method (where synonymous SNPs reflect near-neutral evolution and can be useful to correct biases in mutation rates across genes) is more suitable to detect signatures of selection. However, standard approaches were unfit for our question because we wanted to measure recent selection for various variant types. Thus, to understand the signatures of recent positive selection we developed a custom method to identify genes that recently acquired nonsynonymous or functional variants in a highly recurrent manner (variants appearing often in different parts of the tree). The sections below explain this method in detail.

8.1. Obtaining recent variants

To only consider recent variants we defined monophyletic clusters of (likely) clonally-propagating strains with a recent common ancestor (they should be under nodes with support≥95 where all leaf strains have ≤ 1 SNP/kb to each other). Pairwise SNPs/kb was calculated using the approach described in 'Diversity analysis' but discarding positions with coverage<12x in any strain (using mosdepth and bedmap). This 1 SNP/kb threshold was not trivial to set, since a high threshold may group together very divergent strains, and a low threshold may leave many strains without a cluster and thus not considered by our analysis. We tested this tradeoff for several thresholds and found that 1 SNP/kb was a reasonable value, where most strains were into some cluster (98% in *glabrata*, 99% in *auris*, 78% in tropicalis, 59% in albicans, 100% in *parapsilosis* and 36% in *orthopsilosis*) (**Figure S2 B,C**). Note that the large fraction of unassigned *orthopsilosis* samples (64%) may limit our power to detect selection in this species. We then ran Ancestral State Reconstruction (ASR) on all variants to define those that appeared after the diversification of each clonal cluster. For this, we ran pastml (v1.9.34)[100] with '--prediction_method ALL' (to use the six available ASR methods) on each variant independently using the strains tree generated as described in 'Strain-tree

generation'. To avoid having branches with 0 length, we added a pseudocount to each branch length (10% of the shortest leaf with non-0 branch length) for the ASR using ete3 (v3.1.2)[101]. We considered that a variant is 'recent' in a given strain if it was not predicted to be present in the common ancestor of the clonal cluster by any of the ASR methods implemented in pastml.

8.2. Defining functional types of variants

To measure selection by different variant types we grouped these recent SNPs, INDELs, CNVs and SVs into functionally equivalent categories according to the effects on coding regions (taken from perSVade's 'Consequence' field). Non-synonymous SNPs (nsyn_SNPs) were SNPs with 'stop_lost' or 'missense_variant' consequences. Synonymous SNPs (syn_SNPs) were SNPs with 'synonymous_variant' or 'stop_retained_variant' consequences. In-frame INDELs (if_INDELs) were INDELs with 'start_retained_variant', 'inframe_deletion' or 'inframe_insertion' consequences. Duplications (DUPs) were SVs / CNVs with 'transcript_amplification' consequence. Deletions (DELs) were either truncating small variants (with 'stop_gained', 'protein_altering_variant', 'frameshift_variant', 'start_lost' or 'coding_sequence_variant' consequences), gene-deleting SVs / CNVs (with 'transcript_ablation' consequence) or transcript-breaking SVs (with 'frameshift_variant', 'inframe_deletion', 'start_retained_variant', 'inframe_insertion', 'start_lost', 'stop_lost', 'coding_sequence_variant', 'protein_altering_variant', 'stop_gained', '5_prime_UTR_variant', '3_prime_UTR_variant', 'splice_region_variant' or 'intron_variant' consequences). Our selection detection method identified genes with either an excess of recurrent nsyn_SNPs (using syn_SNPs to correct for neutral evolution) or with particularly high numbers of recurrent if_INDELs, DUPs and DELs (see below). We thus only considered protein-coding genes with no pseudogene annotation (according to the chromosomal feature files from CGD (see 'Gene annotations')). In addition, we discarded all variants potentially biased by overlapping simple repeats and low complexity regions for this analysis. For CNVs, such variants were those with simple repeats of low complexity regions spanning $\geq$25% of the CNV (inferred with RepeatMasker), which may affect coverage calculations. For SVs and small variants, these were variants where some part of the variant was overlapping any such repetitive elements, as inferred with bedmap.

8.3. Finding genes under selection by non-synonymous SNPs

To find genes under selection by nsyn_SNPs we implemented a selection-detection method inspired by the π_N/π_S (ratio between non-synonymous (π_N) and synonymous (π_S) diversity) approach (see above), where synonymous SNPs reflect neutral evolution and can be useful to correct biases in mutation rates across genes. We define as 'genes under recent selection' those that have $\pi_N > \pi_S$ in a high number of strains and clusters (higher than expected under an empiric model of neutral evolution (**Figure S3 A-C**)). For each gene, we define as 'strains under selection' those with a $\pi_N > \pi_S$, which suggests accelerated evolution and

potentially positive selection[35]. We then calculated a 'selection score' *S* for each gene as the harmonic mean between the fraction of strains under selection ($\pi_N > \pi_S$) and the fraction of clusters that have a strain under selection. We used the harmonic mean ($h(x, y) = (2 \cdot x \cdot y)/(x + y)$) because it is a value between 0 and 1 that is only high if both values are high. This ensures that genes with high *S* values have $\pi_N > \pi_S$ in several strains and clusters, suggesting that they bear the strongest signatures of recent selection. In addition, by considering both the number of strains and the number of divergent clusters we correct possible stochastic errors derived from biased sampling of some clades and/or recent clonal population expansions could be unlinked to selection. We calculated diversity (π_N or π_S) for each gene in each sample as:

$$diversity\ (\pi) = n_{recent,\ gene} / (c \cdot f)$$

Where $n_{recent,\ gene}$ is the number of recent SNPs (either non-synonymous for π_N or synonymous for π_S), c the length of the CDS that does not overlap repeats or low complexity regions and f is either 0.75 for π_N or 0.25 for π_S. Note that f is a normalization parameter to take into account that synonymous variants are less likely to happen, and we set the f as done in [12]. We used bedtools (v2.30.0)[102] 'subtract' and 'merge' modules to calculate CDS lengths. Note that we considered that diploids have two copies of each gene (c is twice the annotated CDS length), so that heterozygous SNPs add 1 to $n_{recent,\ gene}$ and homozygous SNPs add 2.

One of the biases for *S* calculation is that, since we considered only recent variants, the π_N and π_S values could be low or 0 for some genes, leading to high *S* values due to stochastic biases from low variant counts. To provide a statistical framework and find genes with significantly high *S* we calculated the empiric probability (p) that a gene has a *S* equal or above the observed one under a neutral model of evolution. To do this we obtained a distribution of *S* values generated randomly (on the same strains used to calculate the real *S*) by a model considering the neutral mutation rate of each gene. We used synonymous SNPs as a proxy for such a neutral mutation rate. To calculate a synonymous SNP mutation rate (r_S) we used information from all the synonymous SNPs (not only recent variants) present in each strain, so that r_S is defined (for each gene) as:

$$synonymous\ mutation\ rate\ (r_S) = mean\left[n_{all,\ gene} / n_{all,\ all} \right]$$

This reflects a mean mutation rate across strains, where $n_{all,\ gene}$ is the number of all synonymous SNPs in the gene for a given strain and $n_{all,\ all}$ is the number of all synonymous SNPs in any gene. For calculating r_S

in each gene, we only used strains with $n_{all,\,gene} \geq 1$ and $n_{all,\,all} \geq 10$ (good strains), and we filtered out genes with < 3 good strains. We assume that the synonymous mutation rate per gene is similar across all strains and between recent and ancestral variants (those that appeared before the cluster diversification). Under these assumptions, r_S represents the probability of having a synonymous SNP in the gene for each synonymous SNP in any gene. In addition, assuming that non-synonymous SNPs are three times more frequent than synonymous SNPs, we defined a $r_N = 3 \cdot r_S$, which represents (under neutral evolution) the probability of having a non-synonymous SNP in the gene for each synonymous SNP in any gene.

We used these probabilities to generate random numbers of recent SNPs (expected by neutral evolution) from a binomial distribution where $n_{recent,\,all}$ (for a given strain, the total number of recent SNPs in any gene) is the 'number of tries' and r is the 'probability of SNP for each try'. For each gene and 10,000 samples we generated, in each strain:

$$random\ neutral\ diversity\ (\pi_{R,\,i}) \ = \ binomial\left(n_{recent,\,all},\ r\right) / \ (c \cdot f)$$

Where i reflects the sample index (from 1 to 10,000), r is r_N for non-synonymous random neutral diversity ($\pi_{N,\,R,\,i}$) or r_S for synonymous random neutral diversity ($\pi_{S,\,R,\,i}$), c is the length of the CDS that does not overlap repeats or low complexity regions and f is either 0.75 for $\pi_{N,\,R,\,i}$ or 0.25 for $\pi_{S,\,R,\,i}$. We then calculated, for each gene and each sample, a random neutral selection score $S_{R,\,i}$ as the harmonic mean between the fraction of strains under 'selection' ($\pi_{N,\,R,\,i} > \pi_{S,\,R,\,i}$) and the fraction of clusters that have a strain under 'selection'. We calculated the final empirical probability $p(S)$, which indicates how likely is to observe a given *S* under neutral evolution, as:

$$p(S) = \left(\sum_{i=1}^{10,000} 1\ if\ \left(S_{R,\,i} \geq S\right) \right) / \ 10,000$$

To validate this neutral model we reasoned that the observed π_S values (considering recent variants) should fall within the neutral distribution of $\pi_{S,\,R,\,i}$. We thus calculated, for each strain, whether the observed π_S is extreme in the neutral distribution (there are >95% of samples with $\pi_{S,\,R,\,i} > \pi_S$ or >95% of samples with $\pi_{S,\,R,\,i} < \pi_S$). We find that most strains in the majority of genes have non-extreme π_S (**Figure S3C**), suggesting that the null model is generally reasonable. To discard possible biases we filtered out genes

where ≥10% of strains had such extreme π_S values. In addition, to discard genes with low variability we only considered genes with $\pi_N > \pi_S$ in ≥2 clusters and ≥3 strains.

Finally, we define as genes under recent positive selection by non-synonymous SNPs those that had an FDR-corrected $p(S)$ value < 0.05.

8.4. Finding genes under selection in frame INDELs, duplications and deletions

To find genes where if_INDELs, DUPs and DELs are selected we implemented a different approach, since the concept of synonimity does not apply here. For each gene and variant type (if_INDEL, DUP or DEL), we calculate a 'selection score' *S* as the harmonic mean between the fraction of strains that have a recent variant and the fraction of clusters that have a strain with a variant. Genes with high *S* values are likely the ones with the most frequent recurrent variants, suggesting selection on them. To discard genes with low variability we only considered genes with recent variants in ≥2 clusters and ≥3 strains. Finally, we define as 'genes under selection' by these variants those with an *S* above the percentile 90% of considered genes (**Figure S3E**). A limitation of this approach is that recurrent variant acquisition could be sometimes unrelated to selection, since some genomic regions may have higher mutation rates for these types of variants. However given our focus on functional variants, we consider it a valid proxy to identify genes potentially under recent selection.

8.5. Strain filtering

We filtered out some strains to get meaningful selection score *S* calculations. To ensure that the inferred genes may be under clinically-relevant selective processes (like adaptation to the host, hospital environments or antifungal drugs) we only considered clinical isolates. To avoid biases derived from low coverage and pseudogenization, we filtered out some strains for each gene. For nsyn_SNPs and if_INDELs we wanted to discard strains were the gene may be broken, so that we required the following criteria to accept strains: 1) median coverage ≥24x, 2) ≥95% of the gene is covered and 3) absence of truncating small variants or transcript-breaking SVs /CNVs (defined above). For DELs and DUPs, we wanted to consider strains where the cluster's ancestor had the intact gene, so that we required the absence of truncating small variants or transcript-breaking SVs / CNVs in the ancestor (by all ASR methods used here). In addition, to ensure that all used strains had some degree of divergence to measure *S* we only considered strains that acquired at least one synonymous SNP in any gene after the cluster diversification.

The list of genes under positive selection by different variant types is available at **Table S2**, and **Figure 3A** includes a cartoon that explains how selection was calculated. Note that **Table S2** includes both the genes under selection and also the *S* selection scores and p values for all genes in which *S* could be calculated.

9. Calculating the significance of the overlaps between orthologous groups

We used an empirical approach to calculate the significance of the overlap between orthologous groups (OGs) with genes under selection between either pairs of species or pairs of variant types of a given species (**Figure 3, Supplementary Results**). We tried to answer the following question: if we observe O overlapping OGs between two sets of n, m genes (i.e. n genes under selection in *C. glabrata* and m genes under selection in *C. auris*), what is the empirical probability ($p(O)$) to have an overlap higher or equal than O when randomly sampling genes? To answer this question for each pair of n, m gene sets (to compare) we generated 10,000 sets of randomly-sampled n_i, m_i genes. For each pair of random gene sets we obtained the corresponding OGs and calculated the number of overlapping groups O_i. We then calculated $p(O)$ as:

$$p(O) = \left(\sum_{i=1}^{10,000} 1 \; if \left(O_i \geq O \right) \right) / \; 10,000$$

For example, there are 25 genes ($n = 25$) under selection by DELs (from 21 OGs) and 92 genes ($m = 92$) under selection by SNPs (from 90 OGs) in *C. glabrata* (**Figure 3B**, **Table S2**). There are 6 OGs with genes under selection by both SNPs and DELs ($O = 6$), and the probability $p(6)$ of having 6 or more overlapping OGs when taking 25 and 92 random genes is 0.0001. We consider this overlap significant because p<0.05.

10. Functional enrichment of genes recent selection

To get the domains and pathways enriched in genes under selection we ran a Fisher's exact test on each gene set (selected in each species, by each variant type) for all relevant (see above) Gene Ontology (GO) terms, Reactome, Metacyc pathways and Interproscan annotations (a proxy for domains). We defined as enriched groups (pathways or domains) those with a raw p<0.05, a FDR-corrected p<0.05 and an odds-ratio $\geq$ 2. Note that we ran the FDR correction independently for the following sets of groupings: domains, Reactome, Metacyc pathways, GO Biological Process (BP), GO Molecular Function (MF) and GO Cellular Component (CC) terms. We used statsmodels (v0.11.1)[103] to do the Fisher tests and the FDR correction. **Table S2** includes the results of these enrichments. For all pathway types (Metacyc, Reactome and GO), we discarded very general annotations (they are in 25% of genes).

To visualize the enriched groups across (**Figure 4, S4**) we performed some clustering of the groups for easier interpretation. For domains, Reactome and Metacyc pathways we mapped each gene to the OGs, and we hierarchically clustered the groups (i.e. domains in **Figure S4**) according to the Jaccard distance between OG sets in different species.

To visualize only a subset of representative GO terms (out of significant terms in all species) (**Figure 4**) we performed a redundancy reduction step inspired by REVIGO's algorithm[104]. To define these representatives we iterated through all pairs of terms with a Lin semantic similarity[99] $\geq$ 0.5 (pairs sorted by descending similarity). For each pair of terms, we defined a 'rejected' (non-representative) term following a hierarchical algorithm. If one term was very general (the median % of genes with that term (across species) was $\geq$ 5%) and the other was not, we rejected the general term. Alternatively, if the terms had clearly different p values (the median p across species of one term was < half of the other's median p) we rejected the term with the highest p. Alternatively, if one term was a parent of the other we rejected the child unless both terms were similar (the Jaccard index between the children of both terms was $\geq$ 0.75). If none of these conditions were met we rejected the second term after numeric sorting of the GO identifiers. At the end, we defined as 'representative terms' those that were not rejected in any pairwise comparisons. For each non-representative term, we assigned the representative term as the closest representative term (in terms of Lin's semantic similarity). The output of this process is shown in **Figure 4**, where each row is one representative term (hierarchically clustered by semantic similarity), and the p value is the lowest across all significant terms (in each species-type variant) mapped to that representative. This visualization ensures that similar significant GO terms appear in the same row, improving interpretation of the shared functions under selection in different species and types of variant. Note that the key change from the original REVIGO is that our algorithm can select representatives across sets of GO terms from different species.

11. Convergence-based GWAS

To find the variants underlying resistance we performed a convergence-based Genome Wide Association Study (GWAS). In brief, we used Ancestral State Reconstruction (ASR) on each variant and the drug resistance phenotype to find nodes of the strain tree with variant and/or phenotype transitions (**Figure 5A**). Nodes with variant / phenotype transitions are those that acquired or lost the variant / resistance phenotype as compared to the parent node. We identified variants whose transition is statistically correlated with the transition in the drug resistance phenotype. The following sections describe in detail how we ran this analysis.

11.1. Selecting strains and building a tree for each species and drug

To maximize our power to detect variant-phenotype associations we treated drug resistance as a dichotomous trait, and we only analyzed strains with either strong resistance (R strains) or strong susceptibility (S strains) (see section 4 from **Online Methods** and **Figure S5**), discarding intermediate phenotypes. In addition, to make sure that the associations are clinically-relevant we only considered clinical isolates. We only ran the GWAS pipeline for drugs with $\geq$ 5 R and $\geq$ 5 S clinical strains in a given species, which we could find for *C. albicans*, *C. glabrata* and *C. auris*. To have a balanced set of R/S isolates

and reduce redundancy we first pruned the strains' tree to keep only R/S strains, and then selected three representative isolates for each monophyletic node (where all strains are either R or S). To select these representatives we performed a Multidimensional Scaling (MDS) representation of all strains within a monophyletic node based on pairwise branch distances (with sklearn (v0.24.2)[105]), and we chose the three strains that are closest to uniformly spaced along the first axis of the MDS. This strategy ensured that the representative strains included the highest diversity possible of each monophyletic node. We then built one tree for each species-drug combination considering only the representative R or S samples using the same pipeline as described in the section 'Strain-tree generation' above. This tree was used to do the convergence-based GWAS.

11.2. Defining groups of variants for collapsed GWAS running

To define a set of variants for GWAS we took all the filtered SNPs, INDELs, SVs and CNVs found in each sample. In addition, we defined aneuploidies (whole chromosome loss or deletion) as additional variants. To identify aneuploidies we used bedmap to find chromosomal windows (5,000 bp each) under duplication (if the median copy number based on called CNVs (CN) was $\geq$ 1.8) or deletion (if the CN was $\leq$ 0.2). We defined as aneuploid chromosomes those that had $\geq$ 50% of windows under deletion or duplication. For small variants we used a different set of variants depending on the ploidy of each species. For diploids we kept both homozygous and heterozygous calls. For haploids we kept all haploid variants and the diploid heterozygous variants from duplicated regions (positions with a copy number $\geq$2 according to perSVade's outputs).

To consider that different variants may drive similar resistance by altering the same genomic feature we wanted to collapse variants into groups. This allowed us to test the association between the transition in any variant of a group and the phenotype transition. We collapsed variants taking into account 1) the type of variant ('all variants', 'small variants', 'CNVs', 'SVs', 'SVs and CNVs', 'small variants and SVs' or 'small variants and CNVs'), 2) the type of mutation ('all mutations', 'non-synonymous', 'truncating' or 'non-synonymous that are not truncating') and 3) the type of collapsing (at the level of 'domains', 'genes', 'Reactome', 'GO' or 'Metacyc'). We ran one GWAS for each combination of 'variant type', 'type of mutation' and 'type of collapsing' with the exception of domain and pathway-level collapsing, where we only considered types of mutations that were either 'non-synonymous', 'truncating' or 'non-synonymous that are not truncating'. Note that for the domain and pathway-level collapsing we only considered protein coding genes because these are the ones that we could map to such functional annotations. Finally, we ran a total of 113 GWAS analyses for each species and drug: 1 for the non-collapsed variants (where we tested each variant individually) and 112 for each combination of collapsing modes. For example, one of these GWAS analyses involved collapsing truncating SVs and small variants into genes (the type of variant would be 'small variants and SVs', the type of mutation would be 'truncating' and the type of collapsing would be

‘genes’), testing whether the truncation of each gene by small variants or SVs was correlated to the resistance. To avoid redundancy with the single-variant uncollapsed GWAS runs we only considered groups with ≥2 variants.

To define this ‘type of mutation’ we used perSVade’s functional annotations of each variant in each gene. We defined as ‘truncating’ variants those that had at least one of the following consequences on each gene: 'stop_gained', 'protein_altering_variant', 'frameshift_variant', 'start_lost', 'coding_sequence_variant_BND', 'intron_variant_BND', 'non_coding_transcript_exon_variant_BND', 'transcript_ablation', 'non_coding_transcript_variant_BND', 'coding_sequence_variant'. We defined as ‘non-synonymous variants’ those that had at least one of the following consequences on each gene: 'stop_gained', 'protein_altering_variant', 'frameshift_variant', 'start_lost', 'coding_sequence_variant_BND', 'intron_variant_BND', 'non_coding_transcript_exon_variant_BND', 'transcript_ablation', 'non_coding_transcript_variant_BND', 'inframe_insertion', 'coding_sequence_variant', 'missense_variant', 'inframe_deletion', 'stop_lost', 'transcript_amplification'. We defined as ‘non-synonymous that are not truncating’ variants those that had non-synonymous consequences but no truncating consequence in a given gene.

To define the ‘type of collapsing’ we considered the gene, domain and pathway annotations as described in the section ‘Gene annotations’ from above. For domain collapsing we grouped variants overlapping each Interproscan annotation and also each window of either 10, 25, 50 or 100 amino acids from all proteins. We defined that a variant altered a domain if it was overlapping it by at least 1 bp according to bedmap’s output. For example, we grouped together all variants affecting a given domain from a gene, and we tested whether the transition in any of these variants was correlated to the phenotype transition. For gene collapsing we grouped variants according to the consequences on genes annotated in perSVade’s outputs. We thus tested whether the transition in any variant from a given gene was correlated to the phenotype transition. Finally, for pathway collapsing we extended the gene collapsing to the GO, Reactome and Metacyc annotations. To avoid having too general variant groupings we discarded pathways involving ≥5% of all genes in each species. For example, we grouped together all variants affecting any gene from a given pathway, and we tested whether the transition in any of these variants was correlated to the phenotype transition.

11.3. Running the GWAS analysis

To measure the association of each group of variants (or single variants without grouping) to the resistance phenotype towards each drug in each species we used a custom pipeline, inspired by hogwash’s synchronous algorithm[57]. For simplicity, the paragraphs below mention ‘groups’ to indicate both groups of variants (i.e. the ones that belong to a gene) or single variants.

One of the challenges of this analysis was that there are no studies in *Candida* species using similar convergence-GWAS methods, suggesting that previous methods (designed for bacteria like hogwash) may not be directly transferable. For example, hogwash used a maximum likelihood (ML) method to run ASR, but maybe using maximum parsimony (MP) would be more accurate in some of our datasets. To address this we ran the analysis using different parameters, changing the ASR methods, the branch support thresholds and the methods to calculate empirical p values (see below). This allowed us to define the optimal parameters for our datasets, as described in the section 'Filtering GWAS results' below. The following paragraphs describe how we measured the associations by different parameter combinations.

The first step to run convergence-based GWAS was to infer ancestral states for all variants and resistance phenotypes. To do this we used the same ASR pipeline as described in the section 'Obtaining recent variants', but using the strains tree generated for each drug and species. This yielded, for different ASR methods, a state of 1 (presence of the variant or phenotype in the node), 0 (absence of the variant or phenotype) or NA (unknown state due to uncertain ASR results) in each node. To test the effect of different ASR methods (implemented in pastml[100]) we considered the results from either 1) the MPPA ML method, 2) The DOWNPASS MP method and 3) the consensus between the ML and MP methods. We defined the ML/MP consensus state as 1 (if both ML/MP were 1, ML was 1 and MP was NA or ML was NA and MP was 1), 0 (if both ML/MP were 0, ML was 0 and MP was NA or ML was NA and MP was 0) or NA if none of these conditions were met. In addition, to discard lowly supported branches we set to NA states all nodes with a branch support below a 'min_branch_support' threshold (either 50 and 70). This means that for each group we ran 6 different association measurements using either the ML, MP or ML/MP ASR methods and a 'min_branch_support' of 50 or 70.

To measure the association of each group to the resistance we identified the following types of nodes:

- Genotype transition nodes, where at least one variant has a 1 state in the node and a 0 state in the parent (or vice versa).
- Genotype no-transition nodes, where all the variants have the same state (0 or 1) in the parent and the node.
- Phenotype transition nodes, where the phenotype has a 1 state in the node and a 0 state in the parent (or vice versa).
- Phenotype no-transition nodes, where the phenotype has the same state (0 or 1) in the parent and the node.

Note that many nodes were not assigned to any of these types due to low support or uncertain ASR results (which generated NA states). We only ran the analysis on nodes assigned to one of these types for both

genotypes and phenotypes. In addition, to avoid biases from considering nodes with long branches we discarded branches longer than 25% of the sum of all branch lengths in the tree (similar to hogwash's approach). To calculate the association of genotype and phenotype we considered the following two-by-two table indicating the number of nodes belonging to each type:

	Geno. transition nodes (n_{Gt})	Geno. no-transition nodes (n_{Gnt})
Pheno. transition nodes (n_{Pt})	$n_{Gt,Pt}$	$n_{Gnt,Pt}$
Pheno. no-transition nodes (n_{Pnt})	$n_{Gt,Pnt}$	$n_{Gnt,Pnt}$

For example, $n_{Gt,Pt}$ indicates the number of nodes that are both genotype transition and phenotype transition nodes. To measure the strength of the association for each group we considered the epsilon statistic (as defined in hogwash):

$$\varepsilon = 2 \cdot n_{Gt,Pt} / \left(n_{Gt} + n_{Pt}\right)$$

This is a value between 0 and 1 summarizing 1) how often the transition in the phenotype is explained by a transition in the genotype and 2) how often the transition in the genotype underlies a transition in the genotype. If $\varepsilon == 1$, the association is complete, meaning that there can't be a genotype transition without a phenotype change and *vice versa*.

To measure the statistical significance of the association we calculated the probability (p, either parametric or empirical) of having an association as strong as the observed one (or stronger) by chance. To obtain parametric p values we used scipy.stats (v1.5.2)[106] to calculate the Fisher's exact p_{FISHER} of each tested group. To infer empiric p values we considered either $n_{Gt,Pt}$ or the chi-square of the two-by-two table $\left(X^2\right)$ (calculated with scipy.stats) as test statistics measuring the strength of the association. To generate a null distribution of test statistics for a given group we generated 10,000 trees with randomly re-shuffled phenotypes and real genotypes, only considering nodes with clear transition states for both genotypes and phenotypes. We then calculated, for each random sample i, the two-by-two association matrix and the corresponding ${X^2}_i$ and $n_{Gt,Pt,i}$ statistics. We defined two empiric p values as:

$$p\left(X^2\right) = \left(\sum_{i=1}^{10,000} 1\ if\ \left({X^2}_i \geq X^2\right)\right) / 10,000$$

$$p\left(n_{Gt,Pt}\right) = \left(\sum_{i=1}^{10,000} 1\ if\ \left(n_{Gt,Pt,i} \geq n_{Gt,Pt}\right)\right) / 10,000$$

To obtain each set of null phenotypes we reshuffled the original per-strain resistance and ran ASR and phenotype state inference to define null phenotype transition or no-transition nodes. Finally, we used the bonferroni-corrected $p\,(X^2)$, $p\left(n_{Gt,Pt}\right)$ or p_{FISHER} as indicators of significance (with statsmodels.stats.multitest).

A limitation of using such p values is that bonferroni correction can be conservative since there is no independence between groups due to linkage between variants. This is also true for other widely used multiple-testing correction algorithms such as the False Discovery Rate method used in hogwash. To address this we calculated additional p values using the empiric $maxT$ method, which has been proposed to be useful in GWAS[107,108]. In brief, we first calculated the maximum X^2 and ε (across all groups) for each random phenotype sample i (1000 samples in total from the 10,000 mentioned above). This yielded a distribution of $max(X^2)_i$ and $max(\varepsilon)_i$ null statistics, which we used to calculate the $maxT$ p values for each group as:

$$p\,(X^2)(maxT) = \left(\sum_{i=1}^{1000} 1\ if\ \left(max(X^2)_i \geq X^2\right)\right) / 1000$$

$$p\,(\varepsilon)(maxT) = \left(\sum_{i=1}^{1000} 1\ if\ \left(max(\varepsilon)_i \geq \varepsilon\right)\right) / 1000$$

Note that these p values are already corrected for multiple testing because the null distribution of statistics considers all the tested groups.

There are four differences with hogwash's approach to calculate p values. First, hogwash only uses $n_{Gt,Pt}$, which is not a statistic *per se* (it could be inadequate in some cases), so that we also considered the X^2 because it is a common statistic to measure associations from two-by-two tables. Similarly, we calculated the p_{FISHER}, which is not considered in hogwash. Second, hogwash uses genotype reshuffling which may be biased in trees with highly variable branch lengths (as we discussed in https://github.com/katiesaund/hogwash/issues/87), which motivated us to use phenotype reshuffling. Third, hogwash uses False Discovery Rate correction (instead of bonferroni) on p values, which may give misleading results in our dataset where there is high dependence between groups. Fourth, we calculate parametric and $maxT$ p values, which are not considered in hogwash. All in all, this means that for each group, min_branch_support and ASR method we obtained five association p values that may define significantly associated hits.

To maximize computational efficiency we implemented several steps (some of them are improvements as compared to hogwash). First, to focus on relevant groups we only tested associations for groups with

$n_{Gt,Pt} \geq 2$, $n_{Gnt,Pnt} \geq 1$ and the odds-ratio (of the two-by-two table) ≥ 1 (similar to hogwash's approach). Second, to optimize resource consumption we parallelized many steps. Third, to avoid redundancy in variant ASR we grouped variants into sets of fully linked variants and ran ASR for only one representative of each group. Fourth, to avoid redundancy in association tests we merged the groups that have the same variants to only run the association test on one representative group. Fifth, to minimize the burden of p value inference we first calculated empiric p values on 1,000 null samples, and we only used 10,000 samples if the p based on 1,000 samples was below 0.1. All the computational optimization steps were necessary to run the analysis on such a high amount of species, drugs and parameter combinations.

In summary, we applied a custom GWAS pipeline on each species and drug, resulting in an association p value in each group for each ASR method, min_branch_support and type of p value. Our approach is more comprehensive than current implementations like hogwash because we use more ASR methods, we consider different types of p values and we optimize many steps of the process. This pipeline can be used as a standalone software on any input dataset (see section 'Data and Code Availability'). The following sections explain how we chose the optimal parameters (ASR method, p value type and min_branch_support) to define the high-confidence, non-redundant set of groups underlying drug resistance.

11.4. Filtering GWAS results

To get enough power to detect associations we only considered datasets (one for each species and drug) with at least 5 resistance transitions according to the consensus ML/MP ASR methods and using a min_branch_support of 70. This resulted in 12 analyzed species-drug pairs, comprising seven antifungal drugs (fluconazole (FLC), itraconazole (ITR), posaconazole (POS), voriconazole (VRC), anidulafungin (ANI), micafungin (MIF) and amphotericin B (AMB)) in the three species (**Table S1, Figure 5**).

To find a meaningful filtering strategy we evaluated the significantly associated genes yielded by different parameter/filter combinations. We considered combinations of varying ASR methods (ML, MP or ML/MP), min_branch_support (50 or 70), types of p value (bonferroni $p\ (X^2)$, bonferroni $p\left(n_{Gt,Pt}\right)$, bonferroni p_{FISHER}, $p\ (X^2)(maxT)$ and/or $p\ (\varepsilon)(maxT)$), minimum ε (0, 0.1, 0.2, 0.3, 0.4 or 0.5) and minimum $n_{Gt,Pt}$ (2 or 3). For example, the most conservative parameter/filter combination would be using the GWAS results based on the ML ASR method and a min_branch_support of 70, and defining as significant groups those that have a p<0.05 by all five types of p values and $\varepsilon \geq 0.5$ and $n_{Gt,Pt} \geq 3$. To obtain significant genes we applied each set of parameters/filters to the raw GWAS results from both the single-variant analysis (only for non-synonymous variants) and the collapsing of non-synonymous variants at the gene and domain level. Any gene affected by significant variants or domains would also be considered as a gene yielded by the given parameter/filter combination.

We reasoned that 'appropriate' sets of parameters/filters should meet two criteria. First, they should yield <100 significant genes to skip overly permissive parameters. Second, appropriate parameters should minimize the false positive burden derived from multiple testing. To test if a given parameter/filter set addressed this burden we applied it to the single-variant GWAS results (yielding N significant variants), and calculated the empirical probability of having $\geq N$ significant variants ($p(N)$) in a null dataset with random phenotypes (which lack true associations). To calculate $p(N)$ we generated, for each species/drug combination, 50 datasets with randomly reshuffled phenotypes, and then ran a per-variant GWAS analysis on each set as described in 'Running the GWAS analysis'. For each random dataset i we used the tested parameters/filters on the raw per-variant GWAS results and obtained N_i significant variants, which allowed us to calculate $p(N)$ as:

$$p\,(N) = \left(\sum_{i=1}^{50} 1\ if\ \left(N_i \geq N \right) \right) / 50$$

Parameters addressing the multiple testing burden should have a $p(N) < 0.05$, indicating that the observed number of significant associations was higher than what would be expected solely by random multiple testing. Note that this analysis implied a high computational cost, which is why we only used 50 re-samples. In addition, note that different combinations of species and drugs may require different parameters because the underlying trees and drug resistance evolution modes may be different. After analyzing this tradeoff for 2232 filter combinations we could find 'appropriate' parameters yielding at least one significant gene for most datasets (11/12, all except *C. glabrata*-POS), suggesting that our parameter range yields meaningful GWAS hits (**Figure S6A**).

We find several 'appropriate' filters for a given dataset (**Figure S6A**), suggesting that additional criteria were necessary to select the final, optimal parameters. We reasoned that the presence of known resistance genes (*ERG11* in *C. albicans*' azoles, *ERG11* and *TAC1b* in *C. auris*' azoles, *PDR1* in *C. glabrata*' azoles, *FKS1* and *FKS2* in *C. glabrata*'s echinocandins and *FKS1* in *C. auris*' echinocandins) among the list of significant hits could be such a criteria. To understand if this is the case we analyzed how often the 'appropriate' filters yield such expected genes in 11 datasets (all except *C. auris*-AMB, where we could not define expected genes). We found such expected genes in 5 datasets, but not in the other 6 (**Figure S6A**). To understand whether this lack of expected genes was due to limited power we investigated if omitting multiple testing considerations (p value corrections and $p(N)$ constraints) would yield the expected genes (**Figure S6C**). We find that in 4/6 datasets (POS, ITR and *C. albicans*-FLC) omitting multiple testing considerations was sufficient to yield the expected *ERG11*/*PDR1*/*TAC1b* genes (**Figure S6C**). This suggests that the expected genes may have mild associations to resistance, but we don't have enough power to detect them without risking false positives derived from multiple testing. Conversely, none of the parameter combinations

yielded the expected *FKS* genes in the other 2/6 datasets (*C. auris*-ANI and *C. glabrata*-MIF) (**Figure S6 A,C**), suggesting that association is likely absent in our dataset (see **Supplementary Results** for further discussion on these datasets). In summary, expected genes may be useful to select the final filters in 5/12 datasets, but not in the others due to power limitations and lack of expected associations. We thus define as 'potentially good' filters those that either yielded expected genes (in these 5/12 datasets) or that yielded some significant gene (in the remaining 7/12 datasets).

To choose the optimal parameters for each dataset we first defined a rationally-designed 'base' set of parameters: using the GWAS results based on the ML/MP ASR and a min_branch_support of 70, and defining as significant groups those that have a $p\ (X^2)(maxT) < 0.05$, $p\ (\varepsilon)(maxT) < 0.05$, $\varepsilon \geq 0.1$ and $n_{Gt,Pt} \geq 2$. Then, for each dataset we define as the optimal set of filters those that were 'potentially good' for that dataset while having the least number of changes as compared to the 'base' filters (**Figure S6B**). This ensured sets of optimal parameters that were similar to one another while adapted to each dataset, suggesting that they are useful to detect relevant associations. Note that in *C. glabrata*-POS we used the 'base' parameters because we could not find any 'potentially good' filters, likely due to power limitations. All in all, these were the parameters/filters used to define high-confidence GWAS results.

11.5. Removing redundancy in filtered GWAS results

Since we collapsed variants into partially overlapping groups (i.e. each variant may be in several groups) these high-confidence significant hits were expected to be highly redundant. For example, if a variant is associated with resistance we expect the genes, domains and pathways related to the variant to also be significant. To remove redundancy and keep only the relevant associations we implemented a filtering strategy to always keep the strongest and most specific results among clusters of redundant GWAS hits. In addition, to prioritize functional associations we only focused on protein-altering variants. The following paragraphs describe our redundancy-removal algorithm for any set of input GWAS hits.

To define a list of non-redundant hits (NR hits) for a set of input hits we iterated through all the relevant variants (those that belong to a significant group), sorted by maximum ε (across all groups that contain the variant) in a non-ascending way. For each variant we identified all the (redundant) hits that involve the variant and we selected one representative NR hit (the one with the strongest, most specific association). To ensure proper redundancy reduction, in each iteration we discarded (redundant) hits with variants related to some already-defined NR hit. To find each NR hit we sorted hierarchically the redundant hits by ε, odds-ratio, specificity of the type of collapsing, type of variant, type of mutation and number of variants related to the hit. For ε and odds-ratio we prioritized the largest values to keep the strongest associations. For the type of collapsing we prioritized uncollapsed variants, then domains, then genes, then MetaCyc

pathways, then GO terms and then Reactome annotations. For the type of variant we prioritized single variant types (i.e. 'SVs') over combinations of types (i.e. 'SVs and CNVs'). For the type of mutation we prioritized more specific types i.e. ('truncating') over more general ones (i.e. 'non-synonymous'). Finally, for the number of variants we prioritized hits with the smallest numbers of variants to increase specificity.

In some cases we found that these criteria were insufficient to get a single representative NR hit, since multiple hits had the same ε, odds-ratio, number of variants and grouping specificity levels. In these cases we applied additional hierarchical sorting taking into account different parameters for each type of collapsing. For gene-level collapsing we considered the conservation across *Candida* (prioritizing genes with orthologs in the highest number of species), whether the gene had a defined name, whether the gene had orthologs in *S. cerevisiae*, the number of annotated GO terms in CGD (prioritizing the largest) and the gene length (prioritizing shorter genes). For domain-level collapsing we considered the type of annotation (prioritizing domain-like signatures (i.e. Pfam or PANTHER) over biochemical-like annotations (i.e. MobiDBLite)), the range of the protein covered (prioritizing the smallest), the start of the domain (prioritizing more N-terminal annotations), the domain annotation description lengths (prioritizing annotations with longer descriptions in cases where they cover the same protein coordinates) and the alphabetical order of the description text (in few cases where two redundant domains had an equally-lengthed description). For Reactome collapsing we considered the fraction of genes with a given annotation (prioritizing annotations found in less genes), the source species of the pathway (prioritizing *S. cerevisiae* over *S. pombe* annotations), the number of parent pathways (prioritizing those with more parents), the length of the pathway description (prioritizing longer descriptions) and the alphabetical order of the description text (as with domains). For MetaCyc collapsing we considered the fraction of genes with a given annotation (prioritizing annotations found in fewer genes), the number of parent pathways (prioritizing those with more parents), the length of the pathway description (prioritizing longer descriptions) and the alphabetical order of the description text. For GO collapsing we considered the fraction of genes with a given annotation (prioritizing annotations found in less genes), the namespace (prioritizing biological process, then cellular component, then molecular function), the number of children terms (prioritizing those with less children), the level and depth of terms (prioritizing higher values), the length of the pathway description (prioritizing longer descriptions) and the alphabetical order of the description text.

To generate the final list of high-confidence NR hits (found in **Table S3**) we applied this redundancy-reduction algorithm to different subsets of all significant GWAS hits. To define a set of NR hits covering all involved genes we applied the redundancy-reduction pipeline to each group of hits affecting a given gene (through either gene / domain collapsing or single-variant analysis). Next, to define NR significant pathways we first discarded significant pathways that were based on variants already considered

in the significant genes. In addition, we applied the redundancy-reduction pipeline to all remaining hits grouped by each type of collapsing (Reactome, GO and MetaCyc). This generated our final list of NR GWAS hits, which includes (mostly) one hit for each significant gene and also one hit for each significant NR pathway that does not involve significant genes.

11.6. Generating a set of comprehensive (low-confidence) non-redundant GWAS hits

The previous sections describe how we obtained the list of high-confidence, non-redundant (NR) GWAS hits, analyzed in the main text and shown in **Figure 6**. We also generated additional sets of NR GWAS hits based on more relaxed filters (low-confidence hits) (**Supplementary Results**). We generated six such low-confidence sets, one for each combination of ASR method (ML, MP or ML/MP) and min_branch_support (50 or 70), defining as significant groups those that have an (uncorrected) $p\ (X^2) < 0.05$, $\varepsilon \geq 0$ and $n_{Gt,Pt} \geq 2$. After applying these filters, we obtained the set of NR hits using the same algorithm as described in 11.5. These datasets likely include some false positives and may be unsuited for exploratory analysis, but they could be useful (as an example) to validate hypotheses about specific genes (where the burden of multiple testing is less prominent). In **Supplementary Results** we provide some examples of such hypotheses, that can only be tested using the low-confidence datasets. All the low confidence NR sets of GWAS hits are found in **Table S3.**

DATA AND CODE AVAILABILITY

All the code and software environments used to generate the datasets, results, tables and figures presented here are in https://github.com/Gabaldonlab/Candida_Selection_DrugResistance. Of note, this github repository contains the convergence-GWAS standalone pipeline, which may be useful beyond this project. In addition, this repository also contains the csv versions of the supplementary tables. On another note, the sequencing datasets from the SRA analyzed are in **Table S1**.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

TG group acknowledges support from the Spanish Ministry of Science and Innovation for grants PID2021-126067NB-I00, CPP2021-008552, PCI2022-135066-2, and PDC2022-133266-I00, cofounded by ERDF "A way of making Europe"; from the Catalan Research Agency (AGAUR) SGR01551; from the European Union's Horizon 2020 research and innovation programme (ERC-2016-724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742); from the "La Caixa" foundation (Grant LCF/PR/HR21/00737), and from the Instituto de Salud Carlos III (IMPACT Grant IMP/00019 and CIBERINFEC CB21/13/00061-ISCIII-SGEFI/ERDF). MAST received a Predoctoral Fellowship from "Caixa" Foundation (LCF/BQ/DR19/11740023).

ACKNOWLEDGEMENTS

The authors thank all the members of the Gabaldón group for key support during this work. In particular, Hrant Hovhannisyan, Valentina del Olmo, Diego Fuentes, Matteo Schiavinato, Marina Marcet, Guifré Torruella, Juan Carlos Núñez, Uciel Chorostecki and Ewa Księżopolska provided useful feedback, which was key for the project development. In addition, we thank Marina Lleal for useful feedback in designing the figures.

AUTHOR'S CONTRIBUTIONS

MAST performed all bioinformatic analysis. MAST and TG conceived the study, interpreted the results, and wrote the manuscript. TG supervised the project and provided resources.

FIGURES

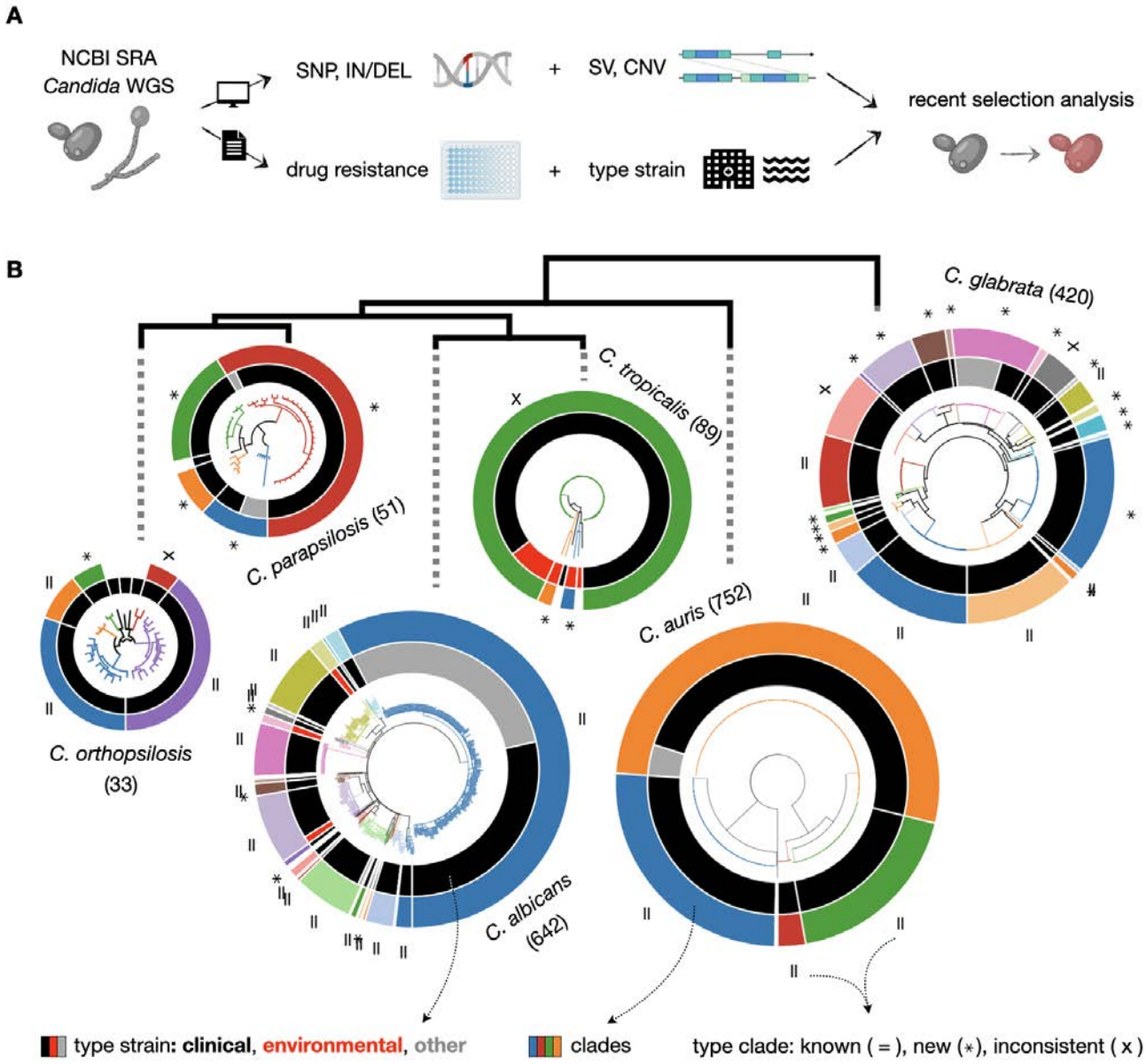

A
NCBI SRA
Candida WGS
SNP, IN/DEL
+
SV, CNV
drug resistance
+
type strain
recent selection analysis
B
C. glabrata (420)
C. tropicalis (89)
C. parapsilosis (51)
C. orthopsilosis
(33)
C. auris (752)
C. albicans
(642)
type strain: clinical, environmental, other
clades
type clade: known (=), new (*), inconsistent (x)
C
Structural Variants (SVs)
deletion
copy-paste insertion
tandem
duplication
cut-paste insertion
Inversion
translocation
Copy Number Variants (CNVs)
duplication
deletion
coverage
genome
Small variants
SNP
small IN/DEL

Figure 1. A genome dataset to study recent evolution across major *Candida* species. (A) Overview about the data-generation process. To study the genome-wide signs of recent selection and drug resistance we processed available whole-genome sequencing datasets from the Sequence Read Archive for *C. glabrata*, *C. auris*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis*. We used these data to identify SNPs, IN/DELs, CNVs and SVs in each strain. In addition, we manually curated the associated literature to get antifungal drug susceptibility data and information about the type of strain (i.e. clinical or environmental). **(B)** SNP-based trees for all strains of each species (see **Online Methods**). The size of each tree is proportional (in log scale) to the number of strains (indicated in parenthesis). Different colors in branches and outer strips represent the clades inferred here, and the '=', '*', 'x' symbols indicate how each clade overlaps with clades defined in other recent population studies (*C. albicans*[28], *C. auris*[11], *C. glabrata*[12], *C. tropicalis*[24] and *C. orthopsilosis*[36]): one-to-one match ('='), it is a new clade ('*') or it is inconsistent with previous clade definitions ("x") (see **Online Methods**). **Table S1** includes all the clade definitions and the trees in newick format. The inner strip represents the type of strain, where the 'other' refers to strains with engineered genomes or strains resulting from directed evolution experiments. In this inner strip, the width of each color (black for clinical, red for environmental and gray for other) indicates the number of strains of each type in each clade, but they are not displayed in the order of the tree. Branches with support<95 are collapsed. The species tree on the top was obtained with orthofinder. **(C)** Variant types identified in this study. SVs are complex rearrangements identified with a breakpoint-detection algorithm, while CNVs are variants generating large duplications and deletions inferred from changes in coverage.

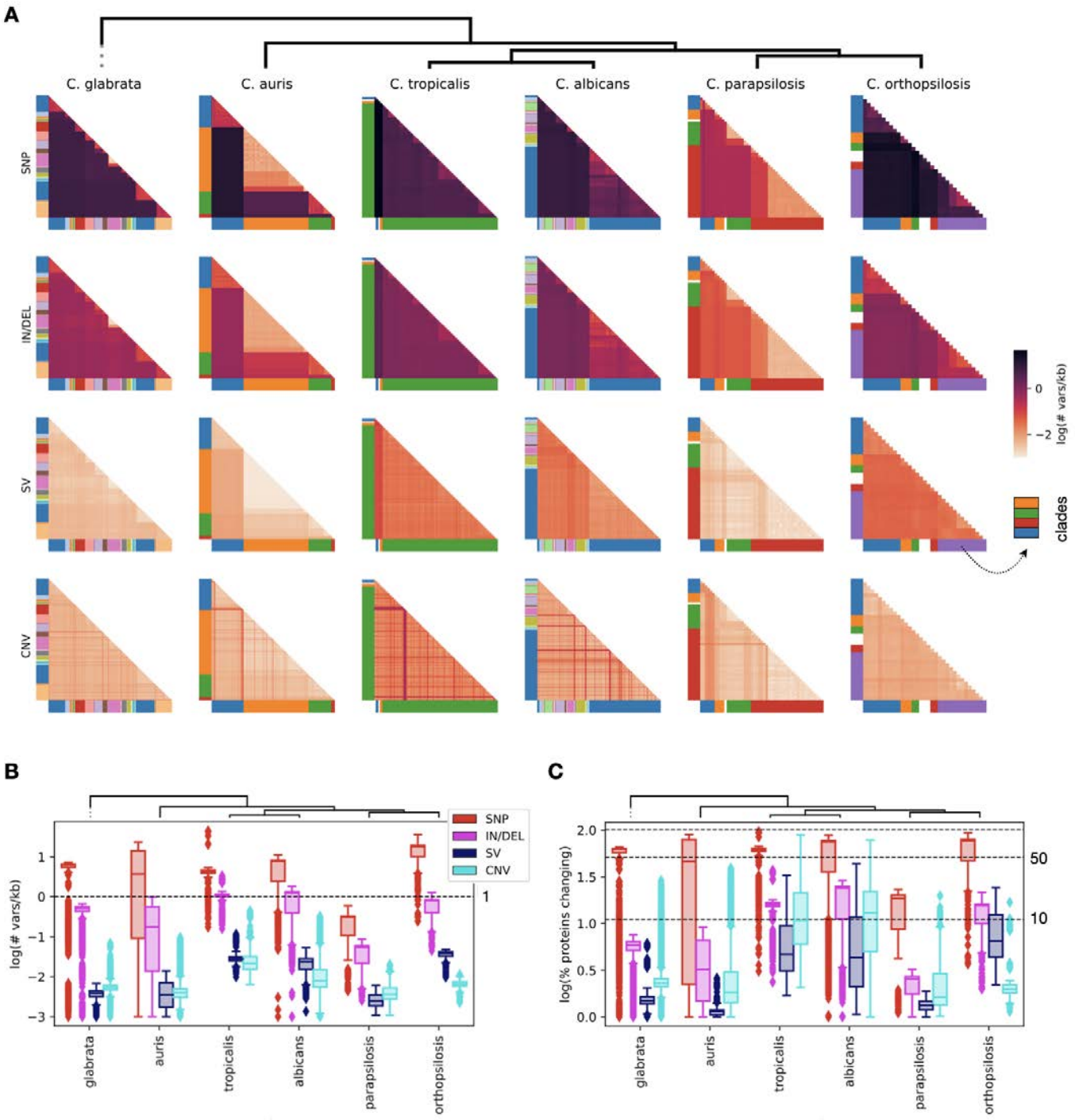


Figure 2. All variant types contribute significantly to intraspecific diversity. (A) Overview about the genetic distance (# variants / kb) patterns across all species generated by each variant type. Each row and column represents a strain ordered as in the strains tree and colored by clade (see **Figure 1B**), and each cell corresponds to the genetic distance in log scale between all pairs of strains. We added a pseudocount of 0.001 variants / kb for the log calculations. **(B)** The same as in (A), but as a boxplot. Each cell in (A) corresponds to one point in the distributions shown here. **(C)** Distribution of the predicted percentage of proteins that are altered by the different variant types, across all pairs of strains. Each point of the distribution corresponds to a pair of strains. We added a pseudocount of 1% of genes affected for the log calculations.

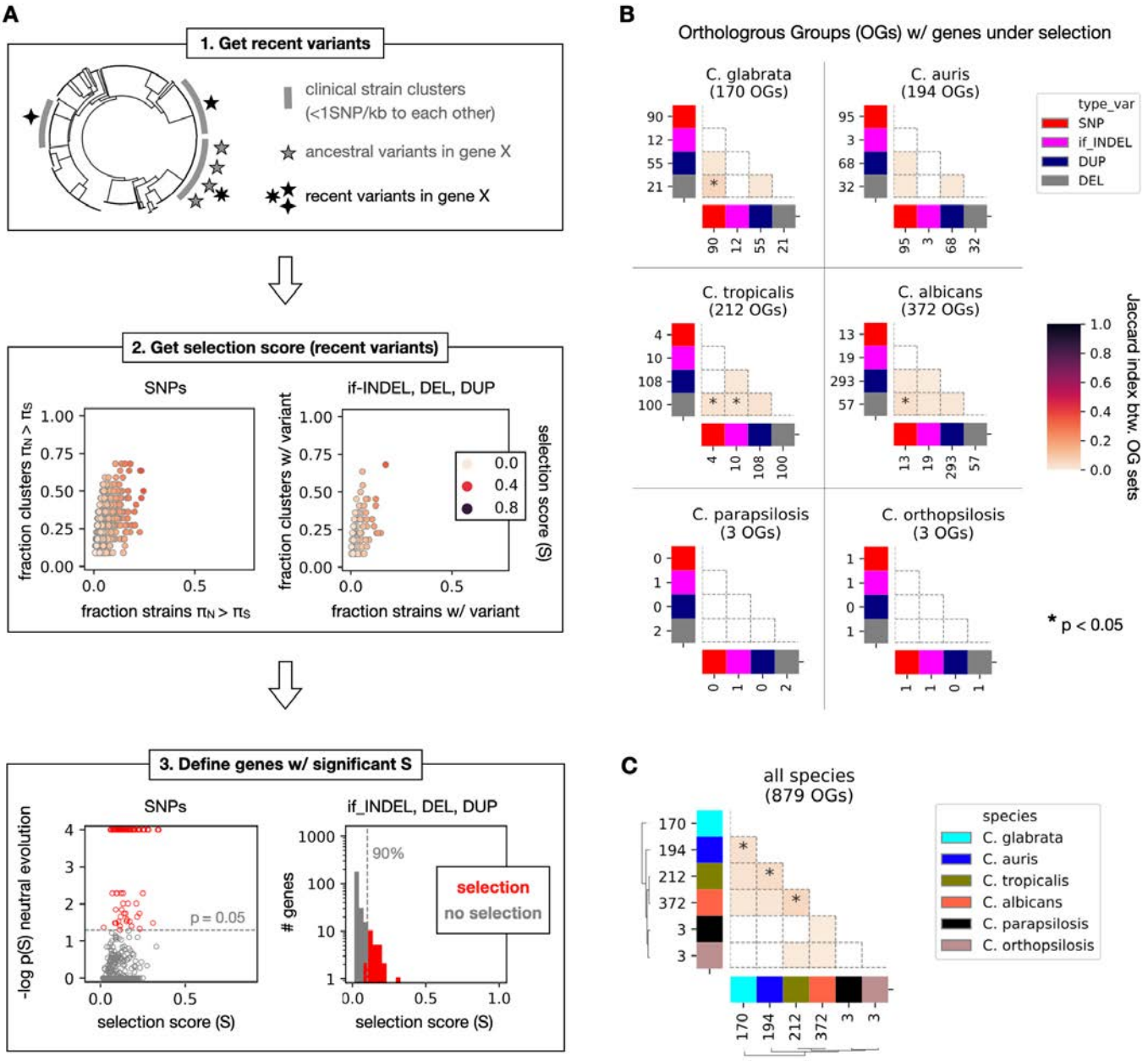


Figure 3. Genome-wide signatures of recent selection in clinical isolates of *Candida* species. (A) Schematic representation of our pipeline to measure recent selection for each gene by different variant types, using *C. glabrata* as an example. First (top box), we defined recently-appeared variants as those that were acquired after the diversification of monophyletic clusters of close clinical strains (where all strains have ≤1 SNP/kb to each other). Shown is an example for 'gene X' which has three variants, including some recently-acquired ones (in black). The gray stripes represent the relevant strain clusters for this gene. Second (middle box), we calculated a selection score (*S*), which measures whether a gene (each point) has an excess of recurrent, recent functionally-relevant variants (either nonsynonymous SNPs, in-frame INDELs (if_INDEL), gene duplications (DUP) or gene truncations (DEL)). For SNPs (left plot), *S* takes into account which strains have a typical hallmark of positive selection: a π_N (non-synonymous diversity) > π_S (synonymous diversity). Thus, we defined *S* as the harmonic mean between the fraction of strains with $\pi_N > \pi_S$ (x axis) and the fraction of

clusters with at least one strain that has $\pi_N > \pi_S$ (y axis). In the scatter plots we show these values for *C. glabrata* genes. For the other variant types (if_INDEL, DEL, DUP) (right plot) we defined *S* as the harmonic mean between the fraction of strains with a variant in that gene (x axis) and the fraction of clusters with at least one strain that has a variant (y axis). *S* measures 'excess of recurrent variants' in these variant types. The example shows the results of DEL variants in *C. glabrata*. Finally (lower box), we defined as 'genes under selection' those that had a significantly high *S.* For SNPs (left plot), we defined as 'genes under selection' those that had a low empirical probability of observing *S* under a neutral model of evolution (FDR-corrected probability p(*S*)<0.05, see **Online Methods**). The scatterplot shows, for each *C. glabrata* gene, the *S* and $-\log_{10}$ p(*S*) (FDR-corrected) values, so that significant genes under selection are in red. For other variant types (right histogram), we defined as 'genes under selection' those that had an *S* value above the percentile 90 of all genes (shown in red). The list of genes and OGs under selection are in **Table S2**. In addition, **Figure S3** shows these distributions for all species and types of variants. **(B)** Distribution of the number of gene families (Orthologous Groups, OGs) with genes under selection by different variant types across species. All shown numbers reflect the numbers of such OGs. The heatmaps show the overlap between the set of involved OGs, measured as the Jaccard distance, and the '*' symbols indicate whether the observed overlaps are significantly high. To infer the significance of having a given number *n* of overlapping orthogroups (OGs) across genes under selection by different variant types we calculated the empiric probability (p) of having a *n* or more overlapping OGs when taking random genes from each set of compared genes (see **Online Methods**). The asterisks show the comparisons where p<0.05. For example, there are 25 genes under selection by DELs (from 21 OGs) and 92 genes under selection by SNPs (from 90 OGs) in *C. glabrata* (upper-left plot). There are 6 OGs with genes under selection by both SNPs and DELs, and the probability of having 6 or more overlapping OGs when taking 25 and 92 random genes is 0.0001. **(C)** Distribution of the numbers of OGs with genes under selection (by any variant type) across species. The heatmap shows the overlaps between such OGs as in (B), and the '*' symbols indicate whether the observed overlaps are significantly high (see (B)).

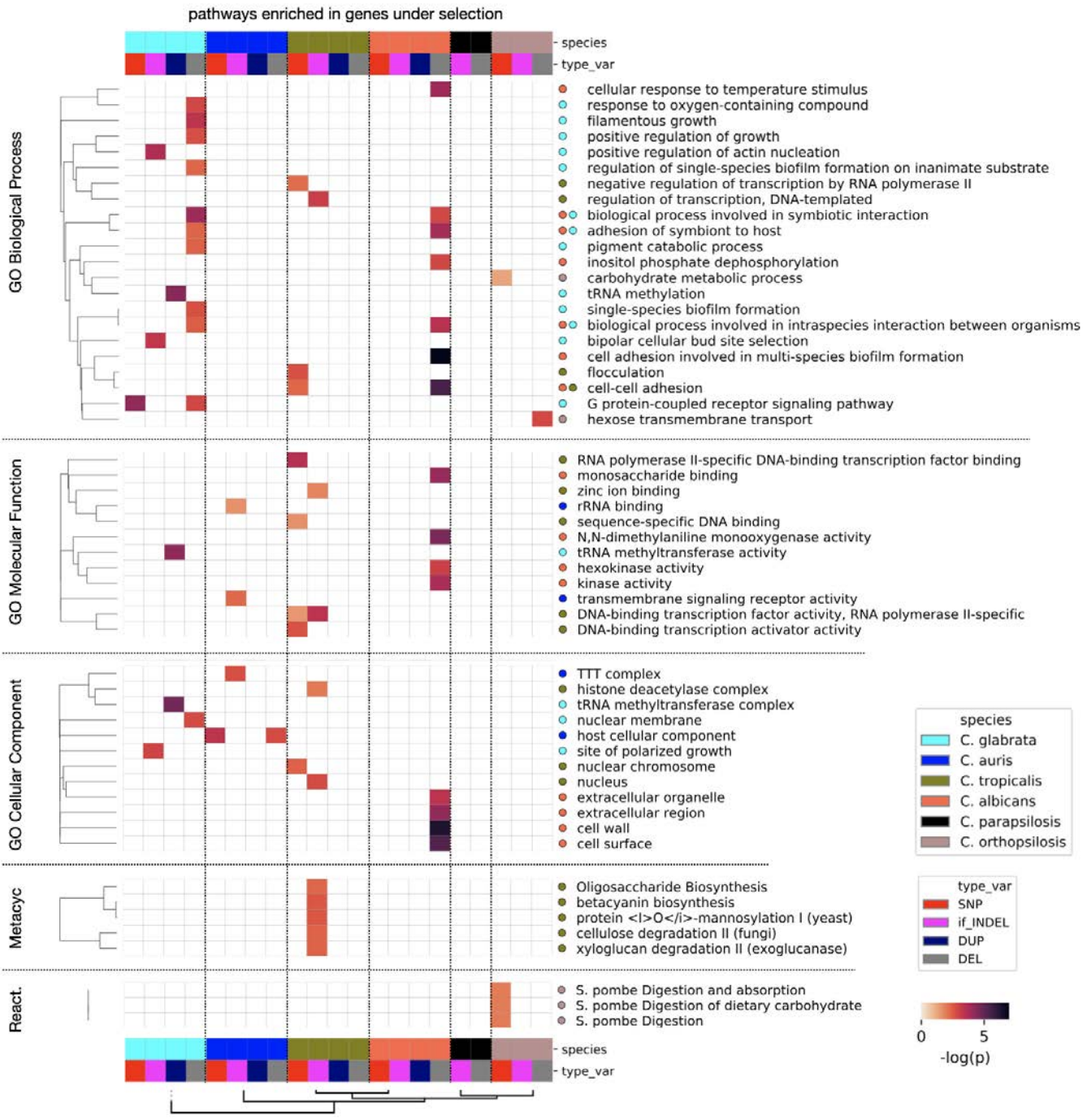


Figure 4. Species-specific and conserved functions are enriched among genes under recent selection. This heatmap represents the Gene Ontology (GO) terms, Metacyc and Reactome annotations enriched in genes under recent selection in different species by different variant types. Only pathways with a FDR-corrected p<0.05 are considered as significant and shown here, and this p value is shown in the color map. To ease comparisons, the GO terms are clustered by Lin's semantic similarity. In addition, we ran a REVIGO-like redundancy reduction algorithm to only keep representative terms for this plot (see Online Methods). Conversely, the Reactome/Metacyc pathways are clustered according to the Jaccard distance between the Orthologous GroupsOGs affected in different sets of genes. The circles represent the species where each pathway was found to be enriched, which is useful to see functions altered in multiple taxa. **Table S2** contains all the related enrichments.

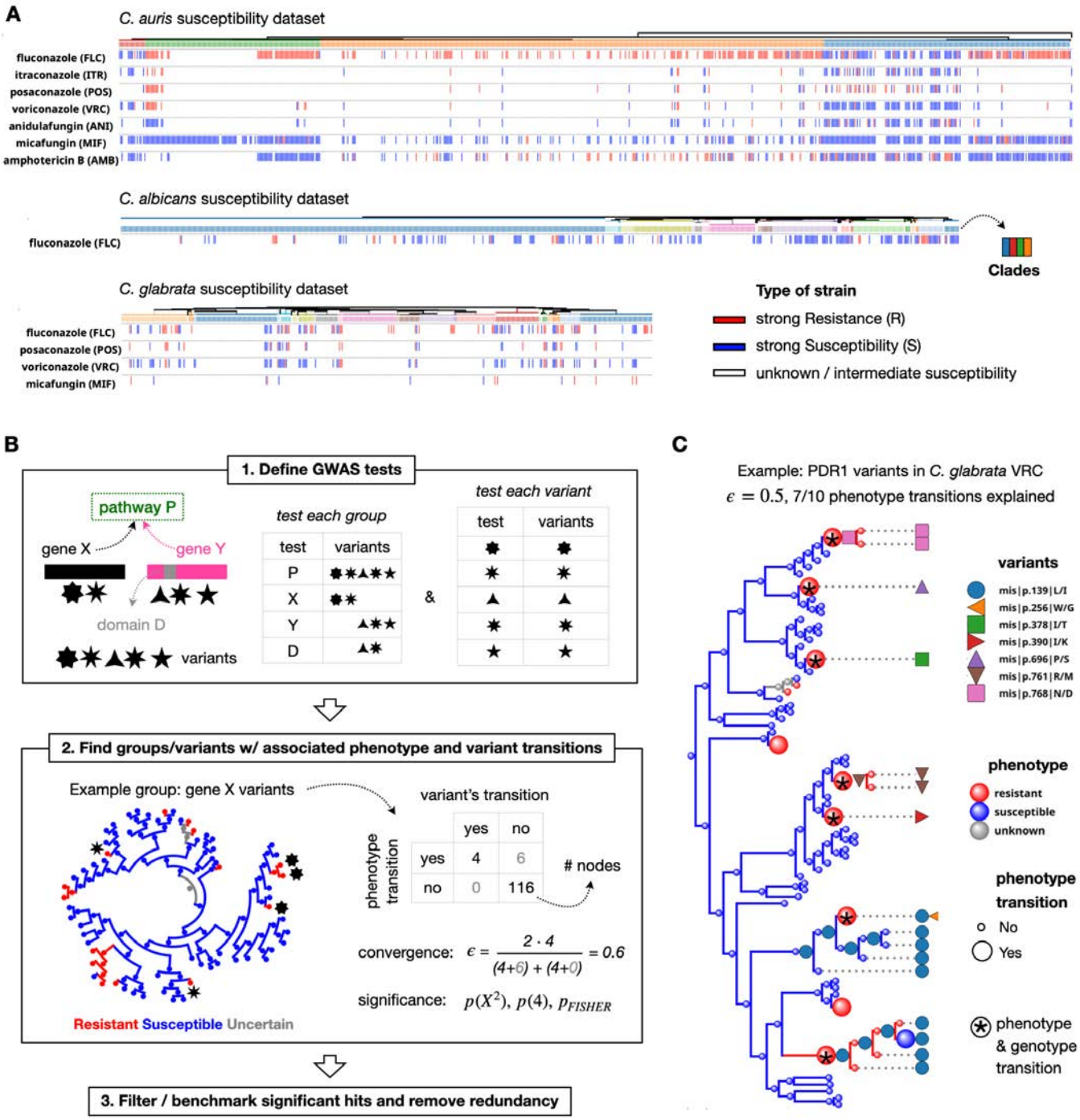


Figure 5. Genome-wide genotype-phenotype associations underlying resistance towards antifungal drugs. **(A)** Distribution of the available drug susceptibility data across the tree of each species for which we performed a convergence Genome-Wide Association Study (GWAS). We only considered strains with either strong susceptibility (S) or strong resistance (R), and we discarded those with intermediate susceptibility or unavailable data. We only performed a GWAS on these datasets because we could find ≥5 transitions from S to R or vice versa in the evolutionary history of these strains. The colors indicate the clades (as in **Figure 1B**), which show how each dataset covers the diversity of each species. **Table S1** includes all these data. **(B)** Schematic view of the GWAS pipeline. First (top box), we defined the GWAS tests to be performed, which include one test for each variant and one test for different groups of collapsed variants (to take into account

that different variants may drive resistance by altering the same gene, domain or pathway). Second (middle box), we tried to find groups (or single variants) where transitions in the variants are significantly associated with phenotype transitions. Shown is an example group, 'gene X', which has two variants (black stars) associated with changes in voriconazole resistance in *C. glabrata*. In the tree, the colors (equivalent to (A)) represent the resistance state of each node of (inferred with ancestral state reconstruction). To measure the strength and significance of the association, we generated a two-by-two table with the number of nodes that have a transition in the resistance phenotype and/or a transition in any of the variants of the group ('gene X' in this case). In this example, there are 4 nodes with both a transition in the phenotypes and in some variants. The strength of the association was approximated with the convergence statistic ε, and the significance was inferred with various p values for each group, such as $p(X^2)$, $p(4)$ or p_{FISHER}. For example, $p(4)$ is the empiric probability of having ≥4 nodes with both variant and phenotype transitions by chance (see **Online Methods**). Finally (lower box), we used information about known drug resistance genes to choose a filtering strategy for each dataset (i.e. which p values to consider), resulting in the final set of high-confidence GWAS associations (hits). In addition, we kept only non-redundant hits (see **Online Methods** and **Table S3**). **(C)** Visual representation of an example high-confidence hit: variants in the gene *PDR1* that are correlated to *C. glabrata*'s voriconazole resistance. At each node, the spheres represent the resistance phenotype (resistant, susceptible or unknown), and the circles / squares / rectangles indicate the presence of different variants (all missense mutations). The size of the sphere indicates whether the node has a phenotype transition (so that the phenotype in the node is different than the parent phenotype), and the '*' indicates phenotype-transition nodes that also have a transition in the variants. For clarity, only *PDR1* variants that are correlated to resistance in some nodes are shown.

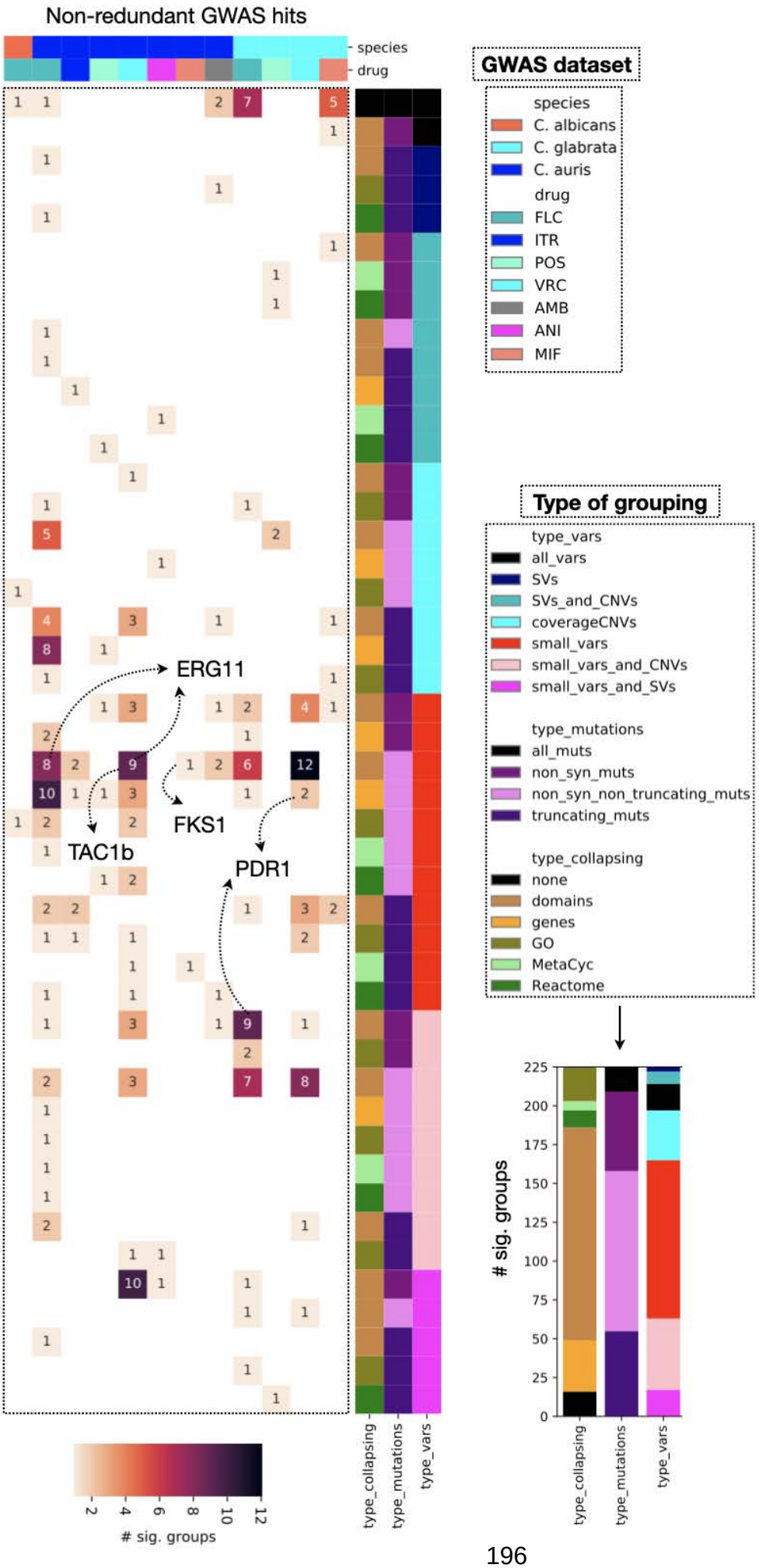

Non-redundant GWAS hits
species
drug
GWAS dataset
species
C. albicans
C. glabrata
C. auris
drug
FLC
ITR
POS
VRC
AMB
ANI
MIF
Type of grouping
type_vars
all_vars
SVs
SVs_and_CNVs
coverageCNVs
small_vars
small_vars_and_CNVs
small_vars_and_SVs
type_mutations
all_muts
non_syn_muts
non_syn_non_truncating_muts
truncating_muts
type_collapsing
none
domains
genes
GO
MetaCyc
Reactome
ERG11
TAC1b
FKS1
PDR1
sig. groups
type_collapsing
type_mutations
type_vars

Figure 6. Hundreds of GWAS hits underlie known and potentially novel mechanisms of drug resistance. The heatmap shows the number of high-confidence, non-redundant GWAS hits (or groups) obtained on each dataset (columns) when using different variant grouping strategies (rows). To consider different ways of grouping variants, we performed one 'grouped' GWAS for different combinations of variant types (SVs, CNVs, small variants and any combinations of them, see 'type_vars'), mutation types (non-synonymous, non-synonymous non-truncating, and truncating, see 'type_mutations') and collapsing levels (domains, genes or pathways (GO, Reactome, MetaCyc), see 'type_collapsing'). For example, in one of these GWAS we tested the genotype-phenotype association for each gene (type_collapsing=genes), considering truncating (type_mutations=truncating_muts) small variants and SVs (type_vars=small_vars_and_SVs). We thus ran a total of 113 GWAS analyses for each species and drug: one for the single variants (type_vars=all_vars, type_collapsing=none) and 112 for different combinations of collapsing modes. Each row in the heatmap corresponds to one of these GWAS analyses, restricted to those that yield some high-confidence hits. These grouping strategies yielded redundant results (i.e. a significant variant may drive a significant association in the genes affected by that variant), so that we only kept (and show here) the strongest, most specific association among sets of redundant hits. For example, if we have a gene that is significant when considering either small variants (with ε=0.3) or small variants and SVs (with ε=0.4), we'd keep the hit that considers small variants and SVs, as it has the highest ε. Similarly, if there is a significant gene (with ε=0.3) and a significant variant altering that gene (with ε=0.3), we'd keep the variant, since it is more specific. This redundancy reduction ensures that the numbers of hits by different collapsing strategies are informative (i.e. hits involving SVs around a gene will only appear here if they yield stronger associations than the hits which only consider small variants in the same gene). On another line, the small inset plot on the right summarizes the number of unique hits (i.e. if a gene is found in two datasets it will only count as one hit here) obtained when considering different grouping strategies, which informs about the most important ones. In addition, the arrows point to hits involving known drug resistance genes.

REFERENCES

1. Bongomin, F., Gago, S., Oladele, R. O. & Denning, D. W. Global and Multi-National Prevalence of Fungal Diseases-Estimate Precision. *J. Fungi Basel Switz.* **3**, 57 (2017).
2. Division, W. H. O. A. R., Diseases, W. H. O. C. of N. T., Partnership, W. H. O. G. C. and & Alastruey-Izquierdo, A. WHO fungal priority pathogens list to guide research, development and public health action. https://repisalud.isciii.es/handle/20.500.12105/15113 (2022).
3. Consortium OPATHY & Gabaldón, T. Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS Microbiol. Rev.* **43**, 517–547 (2019).
4. Brown, G. D. *et al.* Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **4**, 165rv13-165rv13 (2012).
5. Raut, A. & Huy, N. T. Rising incidence of mucormycosis in patients with COVID-19: another challenge for India amidst the second wave? *Lancet Respir. Med.* **9**, e77 (2021).
6. Nnadi, N. E. & Carter, D. A. Climate change and the emergence of fungal pathogens. *PLoS Pathog.* **17**, e1009503 (2021).
7. Wu, X., Lu, Y., Zhou, S., Chen, L. & Xu, B. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environ. Int.* **86**, 14–23 (2016).
8. Arastehfar, A. *et al.* Drug-Resistant Fungi: An Emerging Challenge Threatening Our Limited Antifungal Armamentarium. *Antibiot. Basel Switz.* **9**, 877 (2020).
9. Denning, D. W. Antifungal drug resistance: an update. *Eur. J. Hosp. Pharm. Sci. Pract.* **29**, 109–112 (2022).
10. Mixão, V. & Gabaldón, T. Genomic evidence for a hybrid origin of the yeast opportunistic pathogen Candida albicans. *BMC Biol.* **18**, 48 (2020).
11. Chow, N. A. *et al.* Tracing the Evolutionary History and Global Expansion of Candida auris Using Population Genomic Analyses. *mBio* **11**, e03364-19 (2020).
12. Carreté, L. *et al.* Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans. *Curr. Biol. CB* **28**, 15-27.e7 (2018).
13. Ni, Q. *et al.* CgPDR1 gain-of-function mutations lead to azole-resistance and increased adhesion in clinical Candida glabrata strains. *Mycoses* **61**, 430–440 (2018).
14. Barber, A. E. *et al.* Comparative Genomics of Serial Candida glabrata Isolates and the Rapid Acquisition of Echinocandin Resistance during Therapy. *Antimicrob. Agents Chemother.* **63**, e01628-18 (2019).
15. Perlin, D. S. Echinocandin Resistance in Candida. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **61 Suppl 6**, S612-617 (2015).
16. Pristov, K. E. & Ghannoum, M. A. Resistance of Candida to azoles and echinocandins worldwide. *Clin. Microbiol. Infect.* **25**, 792–798 (2019).
17. Ferrari, S. *et al.* Gain of function mutations in CgPDR1 of Candida glabrata not only mediate antifungal

resistance but also enhance virulence. *PLoS Pathog.* **5**, e1000268 (2009).

18. Schikora-Tamarit, M. À. & Gabaldón, T. Using genomics to understand the mechanisms of virulence and drug resistance in fungal pathogens. *Biochem. Soc. Trans.* **50**, 1259–1268 (2022).
19. Ksiezopolska, E. *et al.* Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen Candida glabrata. *Curr. Biol. CB* **31**, 5314-5326.e10 (2021).
20. Avramovska, O., Smith, A. C., Rego, E. & Hickman, M. A. Tetraploidy accelerates adaptation under drug selection in a fungal pathogen. *Front. Fungal Biol.* **3**, (2022).
21. Carolus, H. *et al.* Genome-Wide Analysis of Experimentally Evolved Candida auris Reveals Multiple Novel Mechanisms of Multidrug Resistance. *mBio* **12**, e03333-20 (2021).
22. Guo, X. *et al.* Understand the genomic diversity and evolution of fungal pathogen Candida glabrata by genome-wide analysis of genetic variations. *Methods* **176**, 82–90 (2020).
23. Sanglard, D. Finding the needle in a haystack: Mapping antifungal drug resistance in fungal pathogens by genomic approaches. *PLoS Pathog.* **15**, e1007478 (2019).
24. O'Brien, C. E. *et al.* Population genomics of the pathogenic yeast Candida tropicalis identifies hybrid isolates in environmental samples. *PLoS Pathog.* **17**, e1009138 (2021).
25. Forche, A. *et al.* Rapid Phenotypic and Genotypic Diversification After Exposure to the Oral Host Niche in Candida albicans. *Genetics* **209**, 725–741 (2018).
26. Forche, A., Magee, P. T., Selmecki, A., Berman, J. & May, G. Evolution in Candida albicans Populations During a Single Passage Through a Mouse Host. *Genetics* **182**, 799–811 (2009).
27. Tso, G. H. W. *et al.* Experimental evolution of a fungal pathogen into a gut symbiont. *Science* **362**, 589–595 (2018).
28. Ropars, J. *et al.* Gene flow contributes to diversification of the major fungal pathogen Candida albicans. *Nat. Commun.* **9**, 2253 (2018).
29. Helmstetter, N. *et al.* Population genetics and microevolution of clinical Candida glabrata reveals recombinant sequence types and hyper-variation within mitochondrial genomes, virulence genes, and drug targets. *Genetics* **221**, iyac031 (2022).
30. Muñoz, M., Wintaco, L. M., Muñoz, S. A. & Ramírez, J. D. Dissecting the Heterogeneous Population Genetic Structure of Candida albicans: Limitations and Constraints of the Multilocus Sequence Typing Scheme. *Front. Microbiol.* **10**, (2019).
31. Zhang, N. *et al.* Selective Advantages of a Parasexual Cycle for the Yeast Candida albicans. *Genetics* **200**, 1117–1132 (2015).
32. Marcet-Houben, M. *et al.* Chromosome-level assemblies from diverse clades reveal limited structural and gene content variation in the genome of Candida glabrata. *BMC Biol.* **20**, 226 (2022).
33. Sekizuka, T. *et al.* Clade II Candida auris possess genomic structural variations related to an ancestral strain. *PloS One* **14**, e0223433 (2019).
34. Todd, R. T. & Selmecki, A. Expandable and reversible copy number amplification drives rapid adaptation

to antifungal drugs. *eLife* **9**, e58349 (2020).

35. Chiner-Oms, Á., López, M. G., Moreno-Molina, M., Furió, V. & Comas, I. Gene evolutionary trajectories in Mycobacterium tuberculosis reveal temporal signs of selection. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2113600119 (2022).
36. Schröder, M. S. *et al.* Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species. *PLoS Genet.* **12**, e1006404 (2016).
37. Bahrambeigi, V. *et al.* Distinct patterns of complex rearrangements and a mutational signature of microhomeology are frequently observed in PLP1 copy number gain structural variants. *Genome Med.* **11**, 80 (2019).
38. Zhang, F., Carvalho, C. M. B. & Lupski, J. R. Complex human chromosomal and genomic rearrangements. *Trends Genet. TIG* **25**, 298–307 (2009).
39. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
40. Derbyshire, M. C. Bioinformatic Detection of Positive Selection Pressure in Plant Pathogens: The Neutral Theory of Molecular Sequence Evolution in Action. *Front. Microbiol.* **11**, 644 (2020).
41. Roudbarmohammadi, S. *et al.* ALS1 and ALS3 gene expression and biofilm formation in Candida albicans isolated from vulvovaginal candidiasis. *Adv. Biomed. Res.* **5**, 105 (2016).
42. Chybowska, A. D., Childers, D. S. & Farrer, R. A. Nine Things Genomics Can Tell Us About Candida auris. *Front. Genet.* **11**, 351 (2020).
43. Li, J., Coste, A. T., Bachmann, D., Sanglard, D. & Lamoth, F. Deciphering the Mrr1/Mdr1 Pathway in Azole Resistance of Candida auris. *Antimicrob. Agents Chemother.* **66**, e0006722 (2022).
44. Vermitsky, J.-P. *et al.* Pdr1 regulates multidrug resistance in Candida glabrata: gene disruption and genome-wide expression studies. *Mol. Microbiol.* **61**, 704–722 (2006).
45. Gabaldón, T. *et al.* Comparative genomics of emerging pathogens in the Candida glabrata clade. *BMC Genomics* **14**, 623 (2013).
46. Gao, J. *et al.* LncRNA DINOR is a virulence factor and global regulator of stress responses in Candida auris. *Nat. Microbiol.* **6**, 842–851 (2021).
47. Horton, B. N. & Kumar, A. Genome-wide synthetic genetic screening by transposon mutagenesis in Candida albicans. *Methods Mol. Biol. Clifton NJ* **1279**, 125–135 (2015).
48. Pekmezovic, M. *et al.* Candida pathogens induce protective mitochondria-associated type I interferon signalling and a damage-driven response in vaginal epithelial cells. *Nat. Microbiol.* **6**, 643–657 (2021).
49. Pais, P. *et al.* Microevolution of the pathogenic yeasts Candida albicans and Candida glabrata during antifungal therapy and host infection. *Microb. Cell Graz Austria* **6**, 142–159 (2019).
50. Chakrabortee, S. *et al.* Intrinsically Disordered Proteins Drive Emergence and Inheritance of Biological Traits. *Cell* **167**, 369-381.e12 (2016).
51. Gale, C. A. *et al.* Linkage of Adhesion, Filamentous Growth, and Virulence in Candida albicans to a Single Gene, INT1. *Science* **279**, 1355–1358 (1998).

52. ZHAO, X., OH, S.-H. & HOYER, L. L. Deletion of ALS5, ALS6 or ALS7 increases adhesion of Candida albicans to human vascular endothelial and buccal epithelial cells. *Med. Mycol. Off. Publ. Int. Soc. Hum. Anim. Mycol.* **45**, 429–434 (2007).
53. Naglik, J. R., Moyes, D. L., Wächtler, B. & Hube, B. Candida albicans interactions with epithelial cells and mucosal immunity. *Microbes Infect.* **13**, 963–976 (2011).
54. Cavalheiro, M. & Teixeira, M. C. Candida Biofilms: Threats, Challenges, and Promising Strategies. *Front. Med.* **5**, (2018).
55. San, J. E. *et al.* Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front. Microbiol.* **10**, 3119 (2019).
56. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45**, 1183–1189 (2013).
57. Saund, K. & Snitkin, E. S. Hogwash: three methods for genome-wide association studies in bacteria. *Microb. Genomics* **6**, mgen000469 (2020).
58. Wang, Y. & Xu, J. Population genomic analyses reveal evidence for limited recombination in the superbug Candida auris in nature. *Comput. Struct. Biotechnol. J.* **20**, 3030–3040 (2022).
59. Rybak, J. M. *et al.* Mutations in TAC1B: a Novel Genetic Determinant of Clinical Fluconazole Resistance in Candida auris. *mBio* **11**, e00365-20 (2020).
60. Morio, F. *et al.* Precise genome editing using a CRISPR-Cas9 method highlights the role of CoERG11 amino acid substitutions in azole resistance in Candida orthopsilosis. *J. Antimicrob. Chemother.* **74**, 2230–2238 (2019).
61. Heimark, L. *et al.* Mechanism of azole antifungal activity as determined by liquid chromatographic/mass spectrometric monitoring of ergosterol biosynthesis. *J. Mass Spectrom. JMS* **37**, 265–269 (2002).
62. Perlin, D. S. Resistance to echinocandin-class antifungal drugs. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **10**, 121–130 (2007).
63. Vale-Silva, L. A. *et al.* Upregulation of the Adhesin Gene EPA1 Mediated by PDR1 in Candida glabrata Leads to Enhanced Host Colonization. *mSphere* **1**, e00065-15 (2016).
64. Taff, H. T., Mitchell, K. F., Edward, J. A. & Andes, D. R. Mechanisms of Candida biofilm drug resistance. *Future Microbiol.* **8**, 10.2217/fmb.13.101 (2013).
65. Orta-Zavalza, E. *et al.* Local silencing controls the oxidative stress response and the multidrug resistance in Candida glabrata. *Mol. Microbiol.* **88**, 1135–1148 (2013).
66. Nicastro, R. *et al.* Indole-3-acetic acid is a physiological inhibitor of TORC1 in yeast. *PLoS Genet.* **17**, e1009414 (2021).
67. Hoepfner, D. *et al.* High-resolution chemical dissection of a model eukaryote reveals targets, pathways and gene functions. *Microbiol. Res.* **169**, 107–120 (2014).
68. Hossain, S. *et al.* Mitochondrial perturbation reduces susceptibility to xenobiotics through altered efflux in Candida albicans. *Genetics* **219**, iyab095 (2021).

69. Vandenbosch, D. *et al.* Genomewide screening for genes involved in biofilm formation and miconazole susceptibility in Saccharomyces cerevisiae. *FEMS Yeast Res.* **13**, 720–730 (2013).
70. Guan, M., Xia, P., Tian, M., Chen, D. & Zhang, X. Molecular fingerprints of conazoles via functional genomic profiling of Saccharomyces cerevisiae. *Toxicol. In Vitro* **69**, 104998 (2020).
71. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).
72. Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).
73. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).
74. Schikora-Tamarit, M. À. & Gabaldón, T. PerSVade: personalized structural variant detection in any species of interest. *Genome Biol.* **23**, 175 (2022).
75. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinforma. Oxf. Engl.* **34**, 867–868 (2018).
76. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 Preprint at https://doi.org/10.1101/201178 (2018).
77. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://doi.org/10.48550/arXiv.1207.3907 (2012).
78. Shah, S. P. *et al.* Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinforma. Oxf. Engl.* **22**, e431-439 (2006).
79. Bakker, B. *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* **17**, 115 (2016).
80. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 202 (2021).
81. Schröder, J., Wirawan, A., Schmidt, B. & Papenfuss, A. T. CLOVE: classification of genomic fusions into structural variation events. *BMC Bioinformatics* **18**, 346 (2017).
82. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
83. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
84. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **5**, 4.10.1-4.10.14 (2004).
85. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
86. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinforma. Oxf. Engl.* **28**, 1919–1920 (2012).
87. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.* **25**, 1422–1423 (2009).

88. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268 (2015).
89. Lischer, H. E. L., 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 (2014).
90. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
91. Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492-496 (2010).
92. McTaggart, L. R., Cabrera, A., Cronin, K. & Kus, J. V. Antifungal Susceptibility of Clinical Yeast Isolates from a Large Canadian Reference Laboratory and Application of Whole-Genome Sequence Analysis To Elucidate Mechanisms of Acquired Resistance. *Antimicrob. Agents Chemother.* **64**, e00402-20 (2020).
93. Sanglard, D., Ischer, F., Calabrese, D., Majcherczyk, P. A. & Bille, J. The ATP binding cassette transporter gene CgCDR1 from Candida glabrata is involved in the resistance of clinical isolates to azole antifungal agents. *Antimicrob. Agents Chemother.* **43**, 2753–2765 (1999).
94. Skrzypek, M. S. *et al.* The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* **45**, D592–D596 (2017).
95. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. Preprint at https://doi.org/10.12688/f1000research.23297.1 (2020).
96. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* **30**, 1236–1240 (2014).
97. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
98. Karp, P. D. *et al.* Pathway Tools Management of Pathway/Genome Data for Microbial Communities. *Front. Bioinforma.* **2**, (2022).
99. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
100. Ishikawa, S. A., Zhukova, A., Iwasaki, W. & Gascuel, O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol. Biol. Evol.* **36**, 2069–2085 (2019).
101. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635 (2016).
102. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
103. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. *Proc. 9th Python Sci. Conf.* 92–96 (2010) doi:10.25080/Majora-92bf1922-011.
104. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene

ontology terms. *PloS One* **6**, e21800 (2011).

105. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. Preprint at https://doi.org/10.48550/arXiv.1309.0238 (2013).

106. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

107. Gundlach, S., Kässens, J. C. & Wienbrandt, L. Genome-wide Association Interaction Studies with MB-MDR and maxT Multiple Testing Correction on FPGAs. *Procedia Comput. Sci.* **80**, 639–649 (2016).

108. Swiel, Y. *et al.* FPGA Acceleration of GWAS Permutation Testing. Preprint at https://doi.org/10.1101/2022.03.11.483235 (2022).

This is the supplementary material to the paper 'Genome-wide signatures of recent selection and drug resistance across *Candida* opportunistic pathogens'. This document includes Supplementary Results and Discussion, Supplementary Figures and Supplementary Table legends. Below is the content of this file:

SUPPLEMENTARY RESULTS AND DISCUSSION

Evolutionary interactions between paralogs in recent selection

To detect complex evolutionary recent selective interactions between in-paralogs of the same gene family we evaluated, for each species, the overlap between OGs affected by selection in different variant types (nsyn_SNPs, if_INDELs, DELs and DUPs) (**Figure 3B**). We find that most OGs are affected by a single variant type, as only 29/879 OGs have a gene affected by selection on different types. However, the overlap between OGs shaped by selection on nsyn_SNPs or DELs was higher than expected by chance ($p<0.05$, see **Online Methods**) in three species (**Figure 3B, Table S2**), suggesting that these variants may yield complex evolutionary outcomes in some families. For example, DELs are selected in adhesins *EPA6*, *EPA7* and *EPA2* in *C. glabrata*, while nsyn_SNPs are selected in *EPA1* and *EPA15*. Similarly, some paralogs from the *FGR51* group in *C. albicans* (involved in filamentous growth) have selected DELs, while others have selected nysn_SNPs. This could be explained by an antagonistic effect among family members, where some paralogs are intrinsically more adaptable (and thus acquire gain of function (GoF) nsyn_SNPs), while the others tend to be deleted. In addition, both DELs and nsyn_SNPs were selected in *MRR1a* (a MDR gene) from *C. auris*, *MIT1* (related to pseudohyphal growth), *IRA1* (encoding a GTPase-activating protein) and *GPR1* (related to invasive growth) from *C. glabrata* and *RIO2* (involved in caspofungin sensitivity) from *C. albicans*. These could be instances where loss of function (LoF) nsyn_SNPs were selected, which may be equivalent to DELs. Overall, these results suggest that gene loss can be a major driver of recent adaptation.

Performing convergence-based GWAS in *Candida* species

To understand how to best do GWAS in our datasets we evaluated the types of groupings (in terms of types of variants and various collapsing levels) that yield significant groups (**Figure 6**). We performed one 'collapsed' GWAS for different combinations of 'variant type' (SVs, CNVs and/or small variants), 'mutation type' (non-synonymous and/or truncating) and 'functional level' (domains, genes or pathways (GO, Reactome, MetaCyc)) (**Online Methods**, **Figure 6**). For example, in one of these GWAS we tested the genotype-phenotype association for each gene (functional level), considering truncating (mutation type) small variants and SVs (variant type). In doing so, we gained several novel insights about how different GWAS strategies work. First, we often find stronger associations if we include SVs and CNVs in addition to small variants (55% of hits consider SVs/CNVs), suggesting that such complex variants partially underlie resistance changes. Second, in some tests we find stronger associations if we consider only truncating variants (24% of hits consider only truncations), indicating that gene truncation is a major driver of drug resistance. Third, collapsing variants at the gene, domain and/or pathway level yields most of our significant

results (60.4% of hits involve domains, 14.5% involve genes, 17.2% involve pathways and 7.9% involve no collapsing). This suggests that collapsing is essential to detect convergence that would be missed if only testing single variants (**Figure 6**). Of note, pathway collapsing can be essential to find associations in datasets with insufficient strains for typical variant-focused or gene-based collapsing (i.e. *C. glabrata* posaconazole) (**Figure 6**). Similarly, domain collapsing is key to find regions of proteins underlying resistance (**Figure 6**).

To understand the landscape of the genotype-phenotype associations we evaluated the Manhattan plots that show the correlation of each variant with the drug resistance phenotype (**Figure S7 A-C**). We find that there are significant variants in some datasets (1 in *C. albicans*/fluconazole, 65 in *C. glabrata*-fluconazole, 52 in *C. glabrata*/micafungin, 6 in *C.auris*/amphotericin B, 43 in *C.auris*/fluconazole, 66 in *C.auris*-voriconazole), with a distribution that is consistent with recent genome-wide recombination partly underlying the emergence of drug resistance. To further validate this role of recombination we checked whether significant SNPs are linked to each other. We find that 83.3%-100% of them (depending on the dataset) are linked to at least one other significant SNPs (two SNPs are considered linked if they are together and may underlie resistance transitions in >1 strain). In addition, visual examination of the SNPs in the tree (**Figure S7 D-F**) suggests that there are linked haplotypes of significant SNPs. As mentioned in the main text, these results support the idea that recombination has played a role in the emergence of drug resistance.

Finally, to validate our GWAS filtering strategy and gain insights into known mechanisms of antifungal drug resistance we inspected GWAS results for genes that are known drivers of resistance (see **Table S3**). Variants in *ERG11* (the target of azoles[1]) are associated to resistance in *C. auris* fluconazole (non-synonymous small variants in the PTHR24286 PANTHER signature) and *C. auris* voriconazole (non-synonymous small variants in the protein region 101-150). In addition, variants in *C. glabrata*'s *PDR1* (a transcription factor (TF) driving expression of the *CDR1* efflux pumps[2]) are associated to fluconazole resistance (non-synonymous small variants and CNVs (including truncations) in the protein region 901-1000) and voriconazole resistance (non-synonymous small variants). Similarly, variants in *TAC1b* (a TF driving expression of azole efflux pumps[3]) are associated with voriconazole resistance in *C. auris* (non-synonymous small variants in the protein region 211-235). Finally, variants in *FKS1* (the target of echinocandins) are associated to micafungin resistance in *C. auris* (non-synonymous small variants in the protein region 580-679 in gene *B9J08_000964*, which includes a 'hotspot' region that has been previously associated to resistance[4,5]). These findings suggest that our GWAS pipeline yields relevant results and confirm these genes as important drivers of resistance. However, we missed expected genes in some datasets: *ERG11* in *C. albicans* fluconazole and *C. auris* itraconazole/posaconazole, *PDR1* in *C. glabrata* posaconazole, *TAC1b* in *C.auris* fluconazole/itraconazole/posaconazole and *FKS* genes in *C. glabrata* micafungin and *C. auris* anidulafungin. Since our filtering strategy was conservative to limit false positives (see **Online Methods**), we hypothesize

that the lack of such genes may result from limited statistical power in some datasets impeding the finding of all true associations. To test this and to better dissect these lack of associations we evaluated whether different sets of relaxed filters would yield the significant genes (**Online Methods** and **Figure S6C**), and we also checked the corresponding set of low-confidence GWAS hits (**Table S3**). We find that this is sometimes true, since some filters yielded the previously missed *ERG11* in *C. albicans* fluconazole, *PDR1* in *C. glabrata* posaconazole and *TAC1b* in *C.auris* fluconazole/itraconazole/posaconazole. This suggests that the lack of high-confidence associations around the expected genes is sometimes related to limited statistical power, but that these genes actually play a role in resistance. However, we could not find relaxed filters yielding significant hits involving *ERG11* in *C. auris* itraconazole/posaconazole nor *FKS* genes in *C. glabrata* micafungin and *C. auris* anidulafungin, suggesting a true lack of association. All in all, these results suggest that *ERG11* is key in *C. albicans* fluconazole and *C. auris* fluconazole / voriconazole resistance, *TAC1b* drives pan-azole resistance in *C. auris*, *FKS* mutations are drivers of micafungin resistance in *C. auris* and *PDR1* underlies pan-azole resistance in *C. glabrata*. Conversely, *ERG11* may be unrelated to resistance towards some azoles in *C. auris*, while *FKS* mutations could be less important in our *C. glabrata* micafungin and *C. auris* anidulafungin datasets.

This lack of correlation between *FKS* mutations and resistance in *C. glabrata*/micafungin and *C. auris*/anidulafungin was puzzling because there is strong evidence that this is the major driver of echinocandin resistance[4]. To further understand this observation we visualized the trees and the actual MICs (**Figure S8**). In both species/drug datasets there is a mix between highly-resistant strains (MICs of 0.5-1 in *C. glabrata*/micafungin and 8-16 in *C. auris*/anidulafungin) and (more prevalent) intermediately-resistant strains (MICs of 0.06-0.1 in *C.glabrata*/micafungin and 2-4 in *C. auris*/anidulafungin) (**Figure S8**). As expected, the highly resistant samples often have canonical hotspot *FKS* mutations, but there are not enough such samples (only 1 in each drug) to drive a significant association. However, the intermediately-resistant samples lack *FKS* mutations, suggesting that some degree of echinocandin resistance can exist without these mutations. We can find other variants and groups correlated to echinocandin resistance in these drugs (13 hits in *C. glabrata* micafungin and 4 in *C. auris* anidulafungin), which may be driving the resistance phenotype (**Table S3**). For *C. glabrata*/micafungin the top hit is a G993S variant in the putative rRNA regulator *CAGL0H02783g* (ortholog of *NET1* in *S. cerevisiae*) (**Figure S8A**). For *C. auris*/anidulafungin the top hit are variants (the most important is a SV) around the putative filamentation-related glycoprotein *B9J08_003526* (ortholog of *MUC1* in *C. albicans*) (**Figure S8B**). These examples suggest that non-*FKS* functions like filamentous growth or ribosomal function could be important for intermediate echinocandin resistance. In addition, this illustrates a common limitation of GWAS analyses: resistance phenotypes are not 100% dichotomous which may complicate the interpretation of results. In addition, this example shows how, despite only considering strains with strong resistance/susceptibility, the strains with ‘strong resistance’ (in the context of the whole MIC distribution

(**Figure S5**)) could be further stratified in various levels of resistance, each entailing diverse underlying evolutionary mechanisms.

Analysis of low-confidence GWAS hits can be useful to test specific hypotheses about drug resistance genes

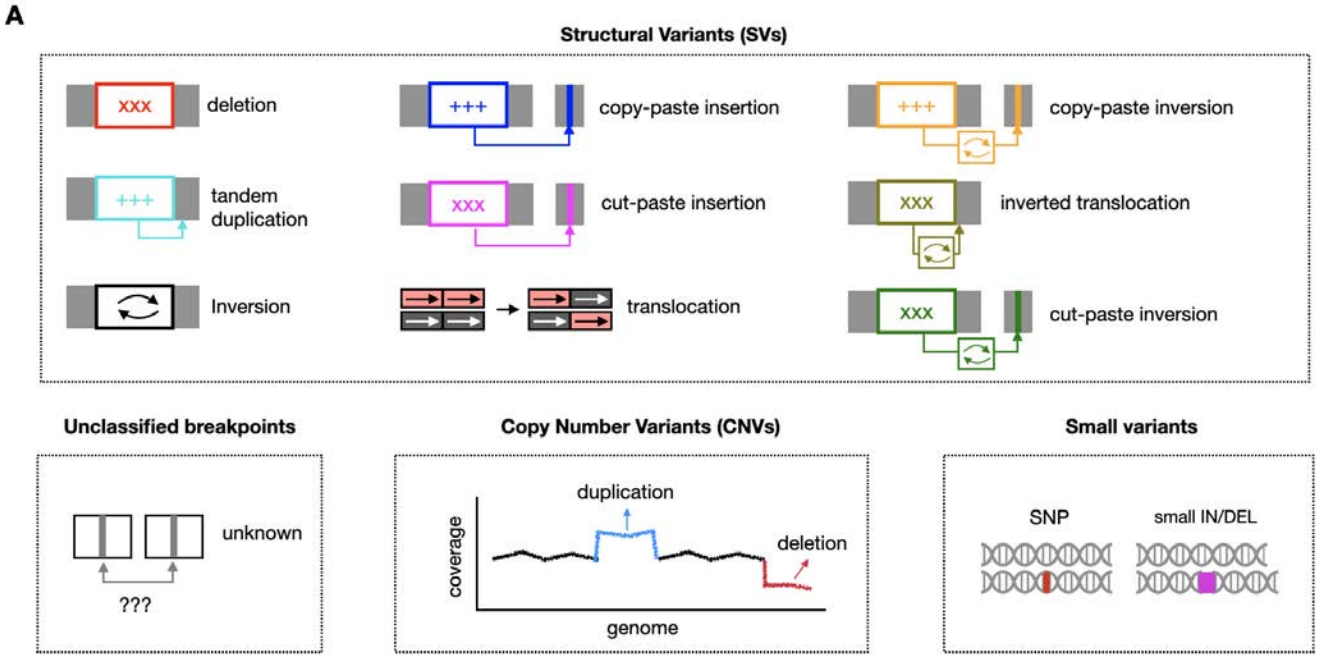
In the main text and **Figure 6** we describe the high-confidence non-redundant (NR) GWAS hits, which provide relevant exploratory insights. However, since our filtering strategy was rather conservative to minimize false positives, it can miss true associations due to limited statistical power, as shown, for instance, by the absence of high-confidence *ERG11* hits in *C. albicans* fluconazole (see above). One of the aims of this work is to provide a comprehensive GWAS dataset useful to validate the clinical importance drug resistance mechanisms inferred in other studies (i.e. *in vitro* evolution approaches). This motivated us to also provide additional sets of NR GWAS hits based on more relaxed filters (low-confidence hits), obtained as described in **Online Methods**, and available at **Table S3**. As an example, these datasets could be useful to validate hypotheses about specific genes (where the burden of multiple testing is less prominent).

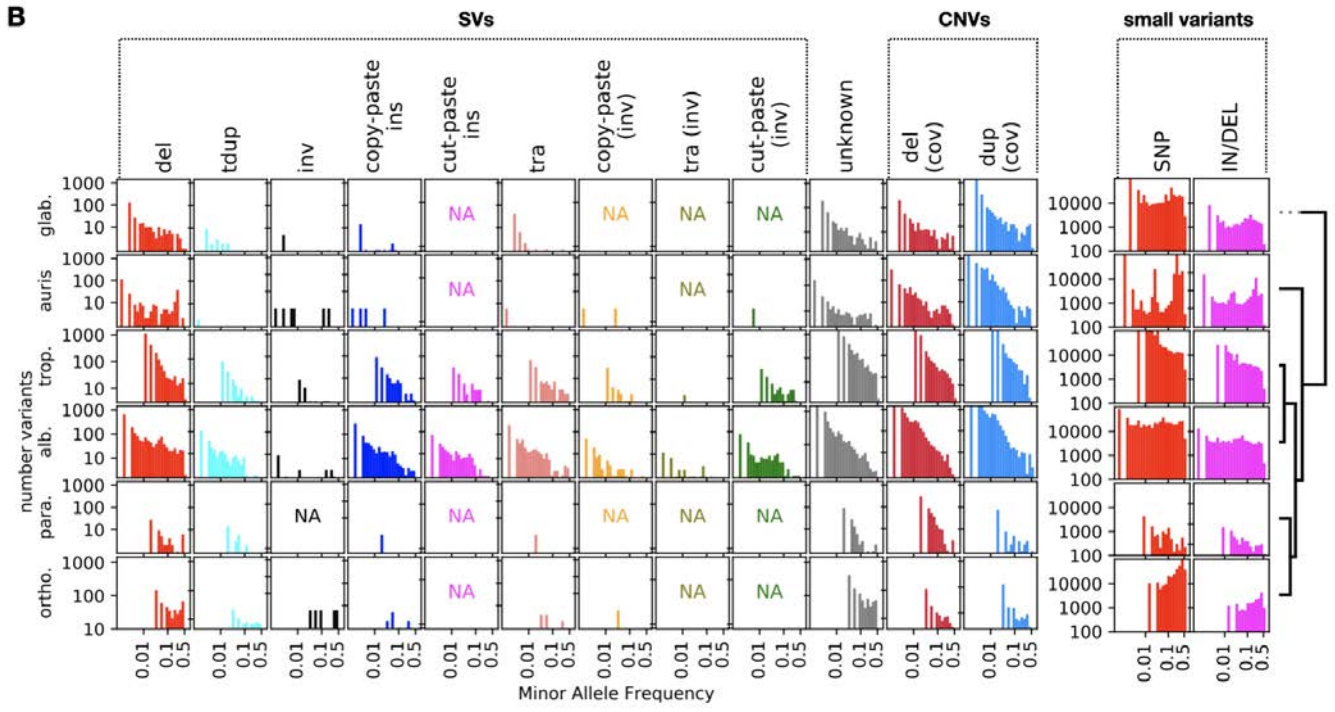
To illustrate this we tested the clinical validity of the findings reported in a recent *in vitro*-evolution study on *C. glabrata*, which suggested that chromosome E duplications and mutations in *ERG11*, *PDR1*, *CDR1*, *CNE1*, *EPA13*, *FKS1*, *FKS2*, *ERG3*, *ERG4* are related to fluconazole and anidulafungin resistance[6]. As mentioned above, *PDR1* mutations are correlated to all azoles tested in *C. glabrata*. *CDR1* had low-confidence hits (non-truncating small variants and CNVs) in voriconazole (region 401-500 of the protein) and fluconazole (ABC transporter signature in the protein region 165-325). *CNE1* had low-confidence hits in fluconazole, involving non-synonymous small variants (some truncating) in protein region 380-383. *EPA13* had low-confidence hits involving various protein regions, small variants and CNVs (some truncating) in fluconazole and posaconazole. These hits suggest that *PDR1, CDR1, CNE1* and *EPA13* could be important for clinical azole resistance, as predicted from the *in vitro* experiment. However, we could not find any hits involving *ERG11* nor chromosome E (containing *ERG11*) duplications, suggesting that this gene may not be related to azole resistance in the clinics (as previously reported[5,7]). This lack of expected *ERG11* hits motivated us to understand whether mutations in other genes of the ergosterol biosynthetic pathway could have an analogous role to the *in vitro* effects of *ERG11* changes. Accordingly, the GO term 'ergosterol biosynthetic process' is a high-confidence hit in voriconazole. In addition, we find low-confidence hits affecting *UPC2A* (a transcription factor regulating azole resistance), *ERG4* and *ERG13* (enzymes implicated in ergosterol biosynthesis, regulated by *UPC2A*[8]) in fluconazole, posaconazole and/or voriconazole. Based on these results, we speculate that, while *ERG11* may be unrelated to clinical azole resistance, these other members of the ergosterol biosynthesis pathway do, resulting in similar outcomes (i.e. higher ergosterol production which compensates azole inhibition). A possible explanation for this difference is that *ERG11*

mutations may yield a higher fitness cost during human infection as compared to mutations in other members of ergosterol biosynthesis. This could explain why such mutations drive resistance *in vitro*, but not in clinical isolates. On another note, we could not find echinocandin-related hits involving *ERG3*, *ERG4, FKS1* or *FKS2*, likely because our *C. glabrata* micafungin dataset does not include enough strains with a strong resistance (mentioned above).

All in all, by using our GWAS dataset we could validate the clinical relevance of antifungal drug resistance mechanisms inferred from *in vitro* studies in *C. glabrata*. Beyond this example, our dataset will be useful to validate future findings in other species and drugs.

SUPPLEMENTARY FIGURES





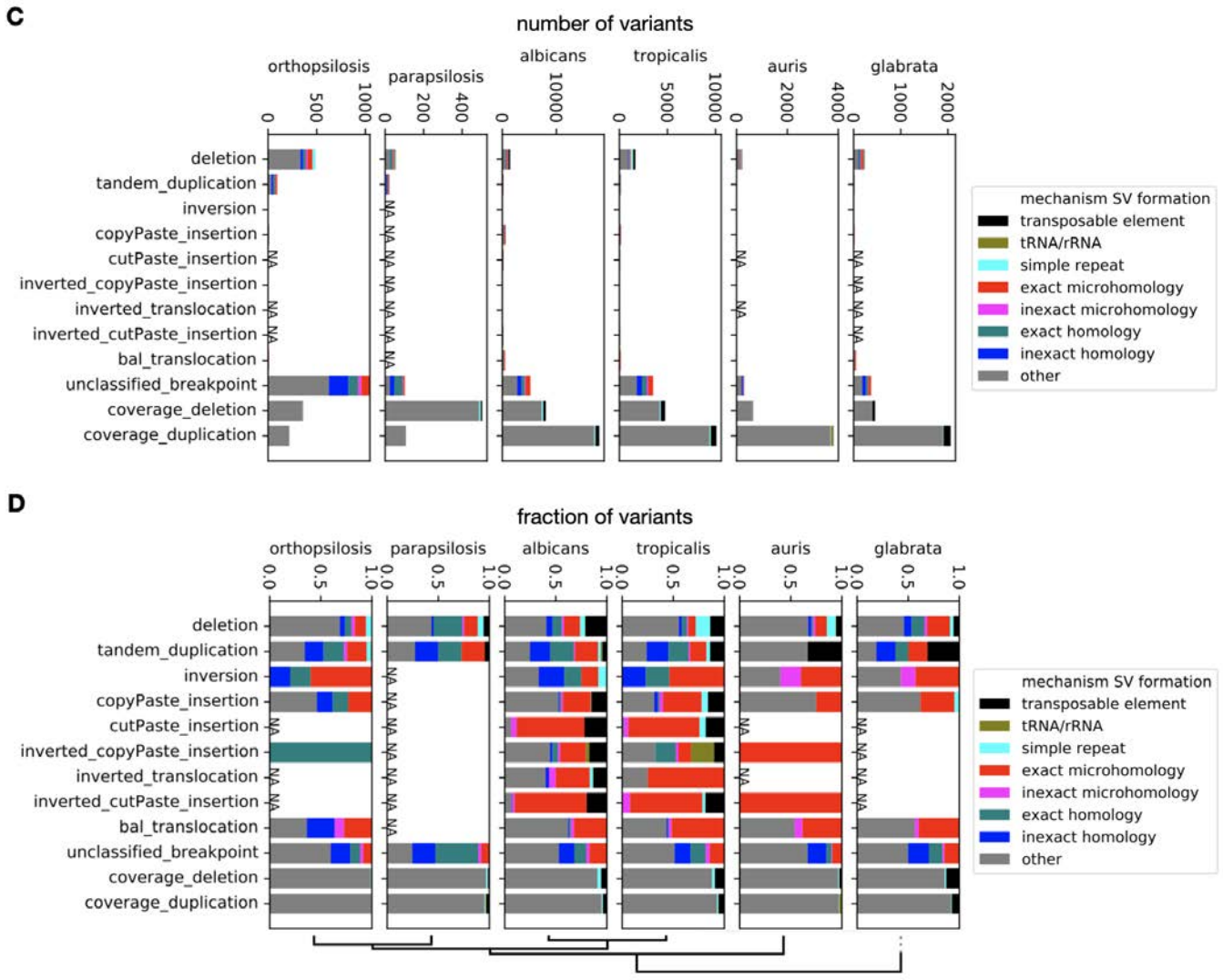


Figure S1. Our dataset includes different variant types. (A) Representation of the types of variants identified in this work. SVs are complex variants where we could find the precise underlying rearrangements and breakpoint positions. Unclassified breakpoints are also a type of SV where we do not know the exact type of underlying rearrangement, either because it is an unknown type of SV or because we missed other breakpoints that could explain the SV. Conversely, CNVs are variants generating large (>600 bp) duplications and deletions (inferred from changes in coverage) with unknown underlying rearrangements. **(B)** Distribution of Minor Allele Frequencies across all variant types and species. 'NA' indicates that a given variant type could not be found in that species. **(C,D)** Proportion of SVs and CNVs attributable to different mechanisms of formation. (C) shows the absolute number of variants, and (D) the fraction of variants relative to each species and type of variant. Variants potentially biased by simple repeats were discarded, including CNVs largely overlapping simple repeats and SVs where the breakpoints are around such repeats. This explains why there are some variants classified as 'simple repeats' (i.e. deletions that involve a region with simple repeats but where the breakpoints do not overlap them). Variants classified as 'other' could not be assigned to any of the above.

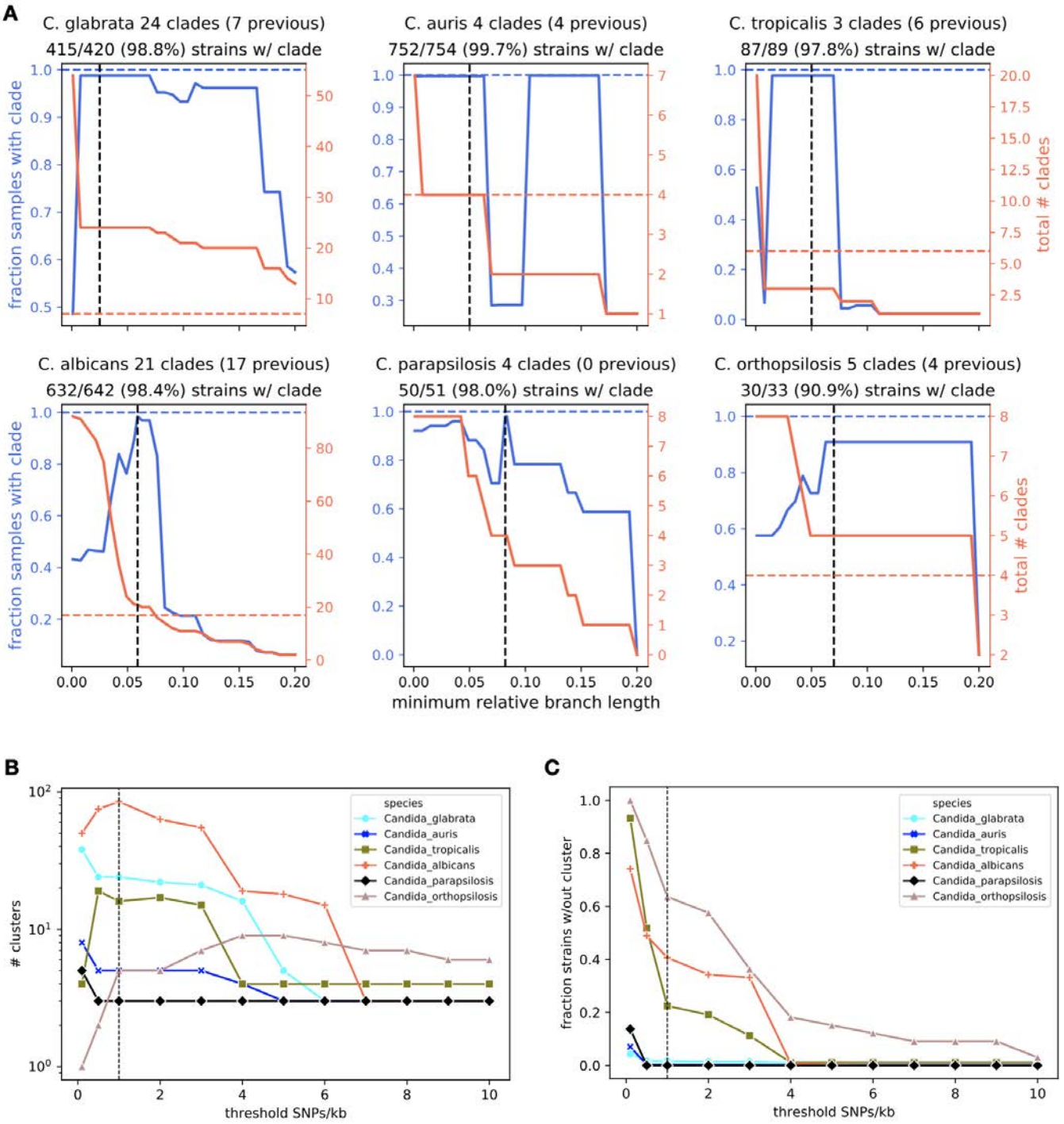
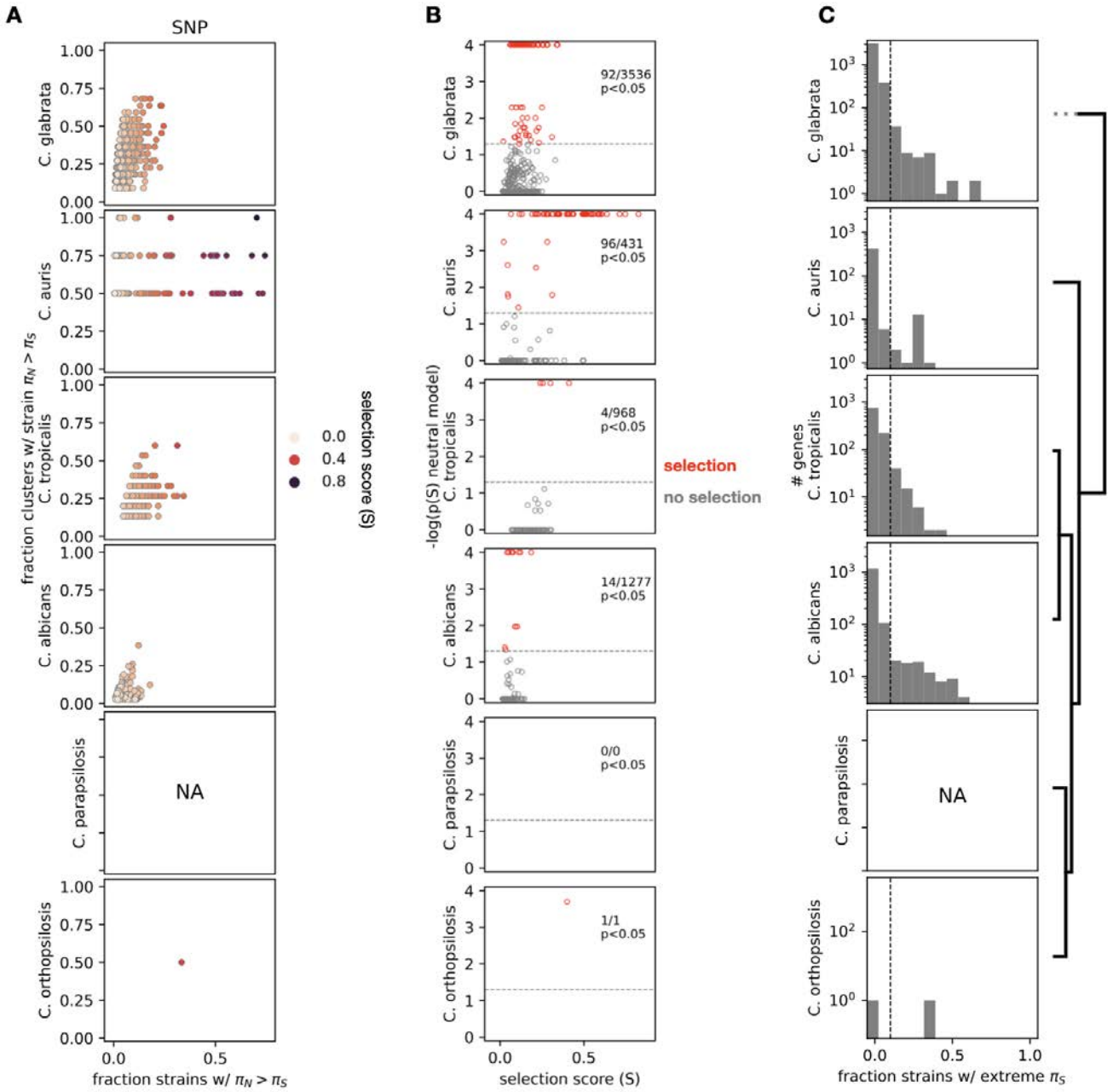


Figure S2. Defining clades and clusters of strains in a systematic manner. (A) To define clades in a systematic manner we had to choose a minimum branch length threshold (see **Online Methods**). These plots show the relationship between the minimum branch length that defines a clade (see **Online Methods** and **Figure 1B**) and the fraction of strains assigned to some clade (blue) or the total number of clades (red). To define the final set of clades, we set a minimum branch length threshold that maximized the number of strains with a clade and minimized the total number of clades (dashed vertical lines). The title of each subplot indicates the number of clades resulting from the selected thresholds, the number of clades defined in previous studies (also in the red dashed line) and the fraction of strains within some clade. **(B,C)**

To define clusters of clonal clinical strains in the analysis of selection (**Online Methods**) we had to define a threshold for the maximum number of SNPs/kb allowed between two strains of a given cluster. These plots show the relationship between different thresholds and the number of clusters (B) or the fraction of strains that can be assigned to a cluster (C). We chose 1 SNP/kb (black dashed line) as a reasonable value because most strains were into some cluster without a very high divergence threshold. We added a pseudocount of 1 cluster to all points of (B) to show log scales.


A
SNP
C. glabrata
C. auris
C. tropicalis
C. albicans
C. parapsilosis
C. orthopsilosis
NA
fraction clusters w/ strain $\pi_N > \pi_S$
fraction strains w/ $\pi_N > \pi_S$
selection score (S)
0.0
0.4
0.8
B
92/3536
p<0.05
96/431
p<0.05
4/968
p<0.05
14/1277
p<0.05
0/0
p<0.05
1/1
p<0.05
-log(p(S) neutral model)
selection
no selection
selection score (S)
C
genes
fraction strains w/ extreme π_S

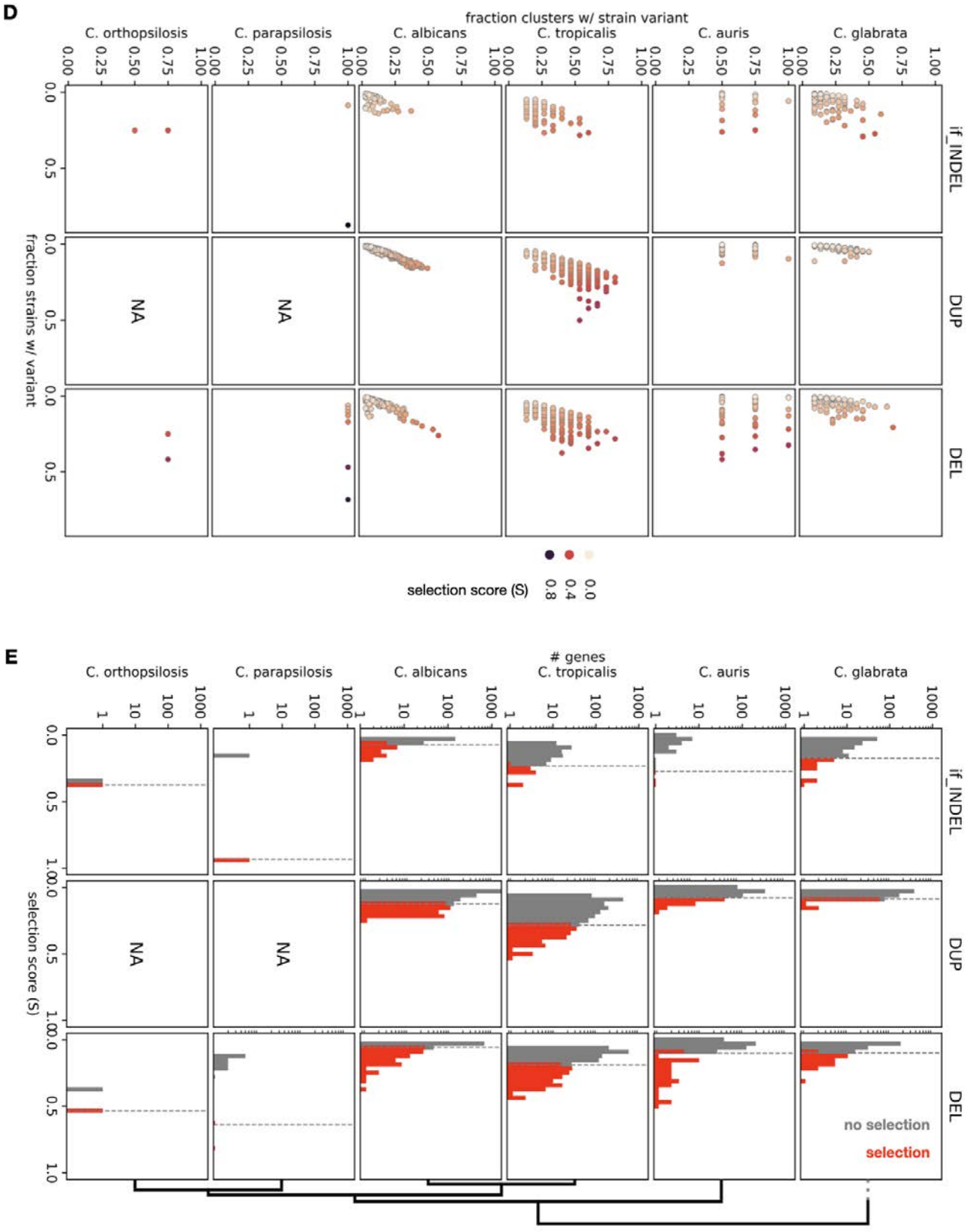

D
fraction clusters w/ strain variant
C. orthopsilosis
C. parapsilosis
C. albicans
C. tropicalis
C. auris
C. glabrata
if_INDEL
DUP
DEL
NA
fraction strains w/ variant
selection score (S)
E
genes
selection score (S)
no selection
selection

Figure S3. Distribution of selection scores by different variants across *Candida* species. (A, B, D, E) Distribution of the selection score *S* and definition of genes with significantly high *S*, shown as in **Figure 3A**, for all species and types of variants. In (B), the text inset shows the number of genes that are considered to be 'under selection', out of all the genes for which we could measure *S.* We only measured *S* values for genes with enough variability to make calculations, which explains why some distributions have few or zero (plots with 'NA') genes. For SNPs (A,B), we only considered genes with $\pi_N > \pi_S$ in ≥2 clusters and ≥3 strains, and we discarded genes where the null model may not be reasonable (see (C)). For DEL, DUP and if_INDEL (D, E), we only considered genes with recent variants in ≥2 clusters and ≥3 strains. Note that **Table S2** includes all these measurements. **(C)** Distribution of samples with extreme π_S according to the neutral evolution model. To validate this neutral model, used to define p(*S*) in (B), we reasoned that the real, observed π_S should fall within the simulated, empirical distribution of neutral π_S generated by the model. This panel shows the distribution across genes of the fraction of strains with extreme π_S according to the neutral model (see **Online Methods**). We only considered genes with a fraction < 0.1 (dashed line) for all analyses, including the data from (A) and (B).

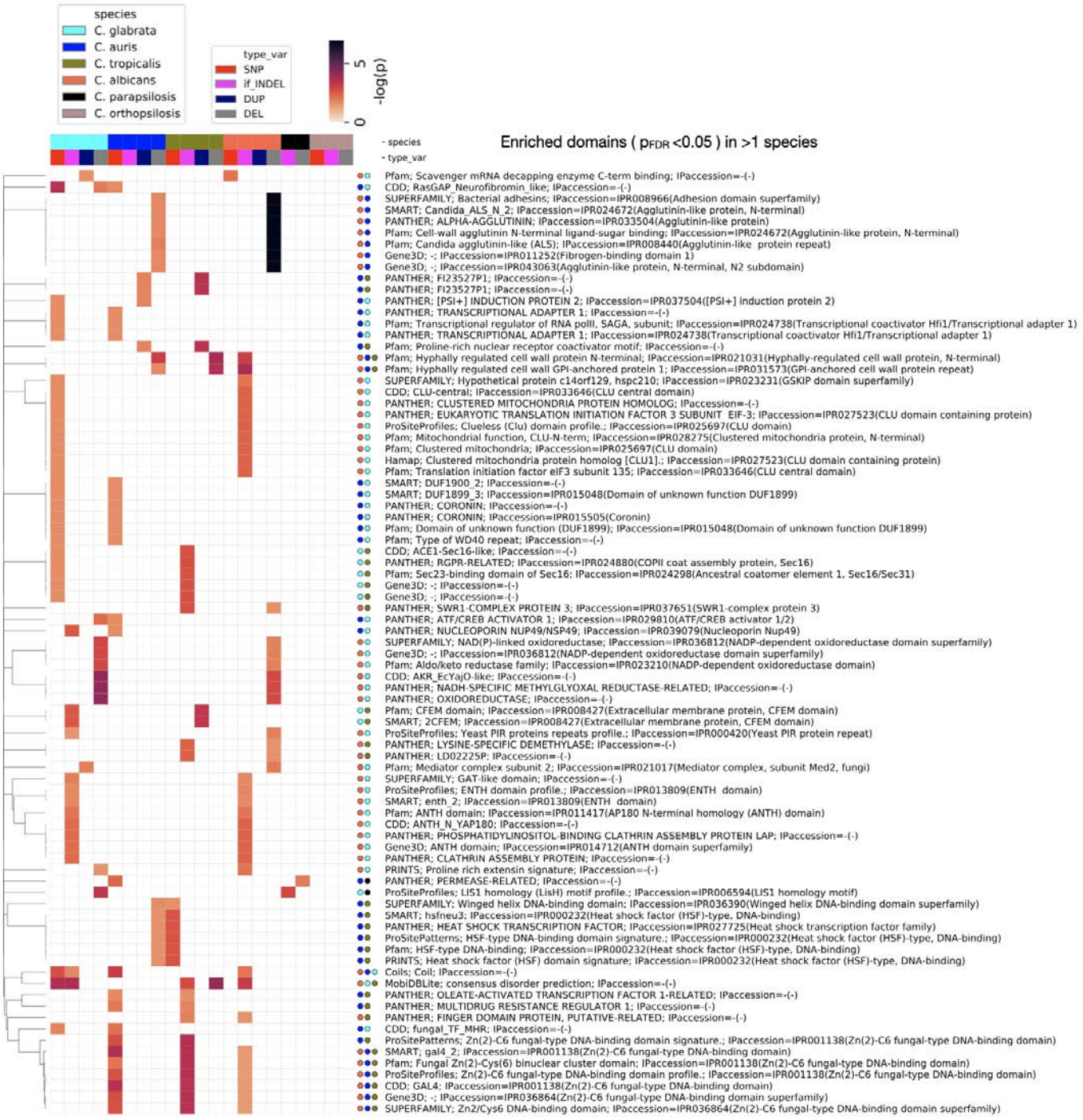


Figure S4. Many Interproscan annotations are enriched in multiple species across genes under selection. This heatmap represents the Interproscan annotations enriched in genes under selection in >1 species. The representation is equivalent to **Figure 4**. The domains are clustered according to the Jaccard distance between the OGs affected in different sets of genes. Note that **Table S2** includes all the enrichments.

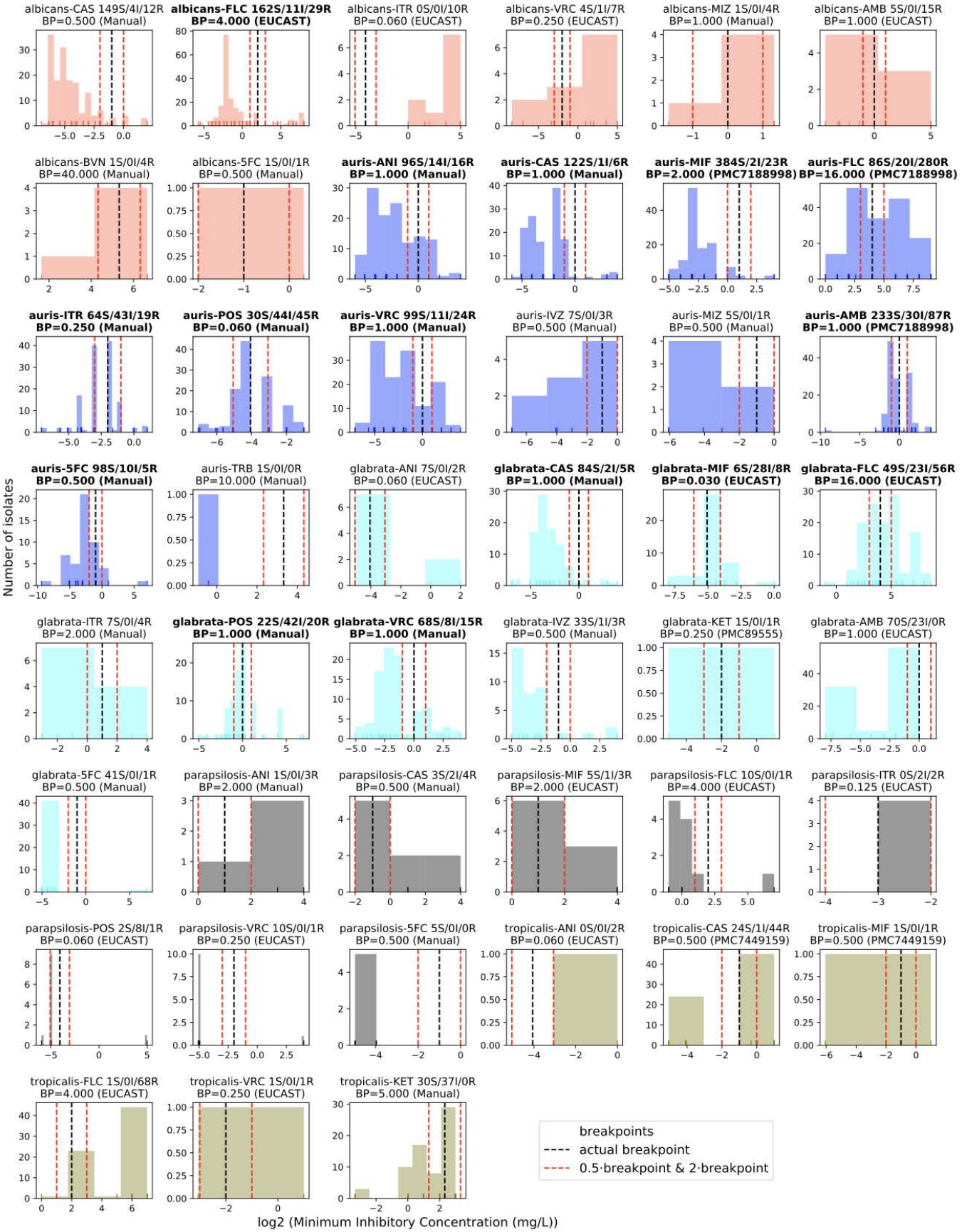
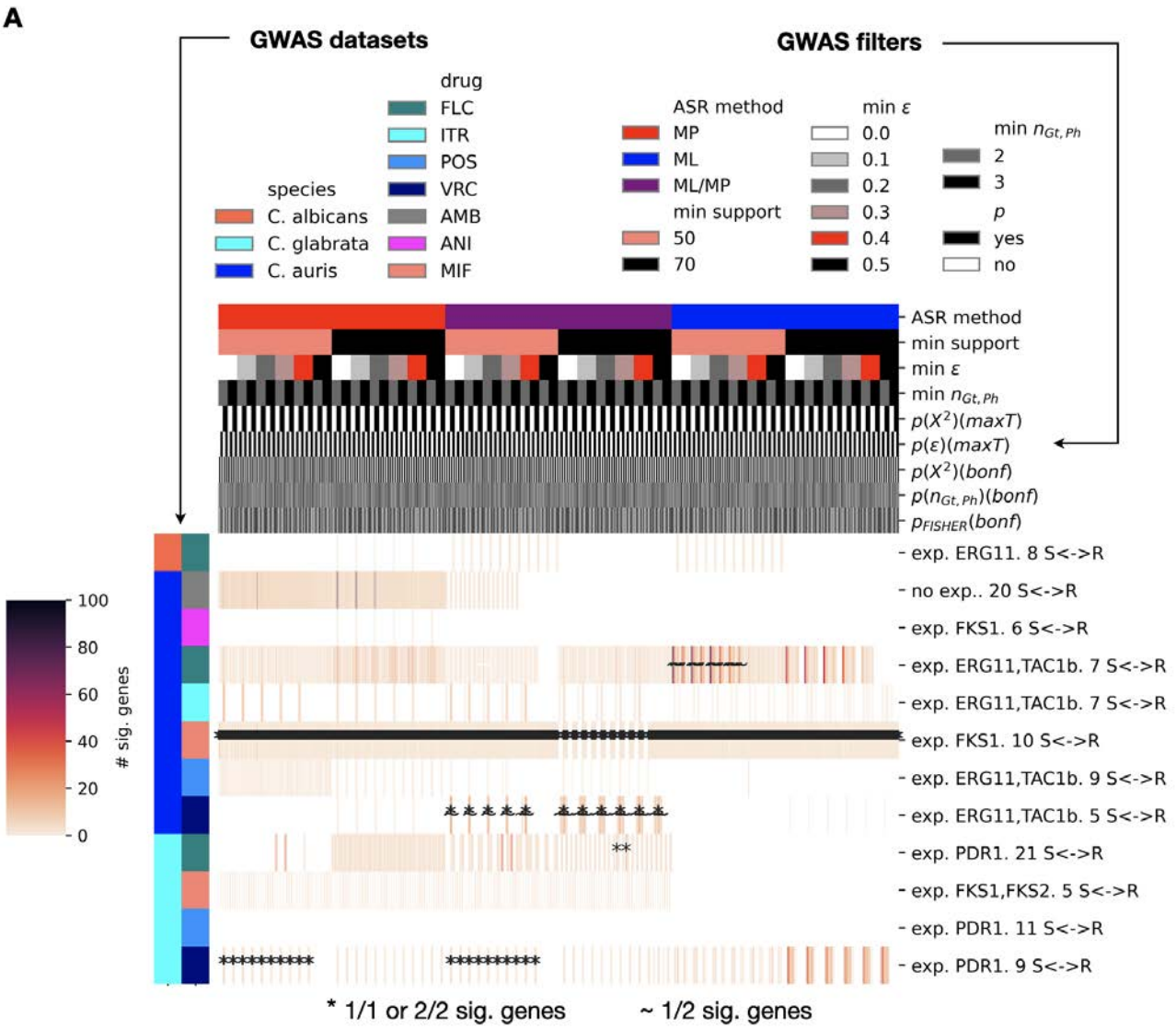

albicans-CAS 149S/4I/12R
BP=0.500 (Manual)
albicans-FLC 162S/11I/29R
BP=4.000 (EUCAST)
albicans-ITR 0S/0I/10R
BP=0.060 (EUCAST)
albicans-VRC 4S/1I/7R
BP=0.250 (EUCAST)
albicans-MIZ 1S/0I/4R
BP=1.000 (Manual)
albicans-AMB 5S/0I/15R
BP=1.000 (EUCAST)
albicans-BVN 1S/0I/4R
BP=40.000 (Manual)
albicans-5FC 1S/0I/1R
BP=0.500 (Manual)
auris-ANI 96S/14I/16R
BP=1.000 (Manual)
auris-CAS 122S/1I/6R
BP=1.000 (Manual)
auris-MIF 384S/2I/23R
BP=2.000 (PMC7188998)
auris-FLC 86S/20I/280R
BP=16.000 (PMC7188998)
auris-ITR 64S/43I/19R
BP=0.250 (Manual)
auris-POS 30S/44I/45R
BP=0.060 (Manual)
auris-VRC 99S/11I/24R
BP=1.000 (Manual)
auris-IVZ 7S/0I/3R
BP=0.500 (Manual)
auris-MIZ 5S/0I/1R
BP=0.500 (Manual)
auris-AMB 233S/30I/87R
BP=1.000 (PMC7188998)
auris-5FC 98S/10I/5R
BP=0.500 (Manual)
auris-TRB 1S/0I/0R
BP=10.000 (Manual)
glabrata-ANI 7S/0I/2R
BP=0.060 (EUCAST)
glabrata-CAS 84S/2I/5R
BP=1.000 (Manual)
glabrata-MIF 6S/28I/8R
BP=0.030 (EUCAST)
glabrata-FLC 49S/23I/56R
BP=16.000 (EUCAST)
glabrata-ITR 7S/0I/4R
BP=2.000 (Manual)
glabrata-POS 22S/42I/20R
BP=1.000 (Manual)
glabrata-VRC 68S/8I/15R
BP=1.000 (Manual)
glabrata-IVZ 33S/1I/3R
BP=0.500 (Manual)
glabrata-KET 1S/0I/1R
BP=0.250 (PMC89555)
glabrata-AMB 70S/23I/0R
BP=1.000 (EUCAST)
glabrata-5FC 41S/0I/1R
BP=0.500 (Manual)
parapsilosis-ANI 1S/0I/3R
BP=2.000 (Manual)
parapsilosis-CAS 3S/2I/4R
BP=0.500 (Manual)
parapsilosis-MIF 5S/1I/3R
BP=2.000 (EUCAST)
parapsilosis-FLC 10S/0I/1R
BP=4.000 (EUCAST)
parapsilosis-ITR 0S/2I/2R
BP=0.125 (EUCAST)
parapsilosis-POS 2S/8I/1R
BP=0.060 (EUCAST)
parapsilosis-VRC 10S/0I/1R
BP=0.250 (EUCAST)
parapsilosis-5FC 5S/0I/0R
BP=0.500 (Manual)
tropicalis-ANI 0S/0I/2R
BP=0.060 (EUCAST)
tropicalis-CAS 24S/1I/44R
BP=0.500 (PMC7449159)
tropicalis-MIF 1S/0I/1R
BP=0.500 (PMC7449159)
tropicalis-FLC 1S/0I/68R
BP=4.000 (EUCAST)
tropicalis-VRC 1S/0I/1R
BP=0.250 (EUCAST)
tropicalis-KET 30S/37I/0R
BP=5.000 (Manual)
Number of isolates
breakpoints
actual breakpoint
0.5·breakpoint & 2·breakpoint
log2 (Minimum Inhibitory Concentration (mg/L))

Figure S5. **Definition of drug resistant (R) and susceptible (S) strains.** Per-strain distribution of Minimum Inhibitory Concentrations (MICs), in $\log_2$ space, for various species and drugs in which this data was available. The colors represent different species. We defined as strains with high resistance (R) those that had a MIC $\geq$2·breakpoint, and as strains with high susceptibility (S) those with a MIC $\leq$0.5·breakpoint. The vertical black lines indicate the actual breakpoints, and the red lines indicate these thresholds above and below the breakpoint. We set the breakpoints (BP in the title) either using EUCAST recommendations (https://www.eucast.org/), based on previous studies (PMC7449159[9], PMC89555[10] and PMC7188998[11]) or by manually looking at the distribution. We performed GWAS on datasets with at least 5R and 5S strains (plots with bold titles), except in the *C. tropicalis* data because the MIC inference was non-standard[12]. In the title, S/I/R indicatestrains belonging to each susceptibility type ('I' means intermediate susceptibility). Note that these S/I/R numbers also include strains in which MIC was not available, but instead we found explicit reports of resistance/susceptibility.


A
GWAS datasets
GWAS filters
drug
FLC
ITR
POS
VRC
AMB
ANI
MIF
species
C. albicans
C. glabrata
C. auris
ASR method
MP
ML
ML/MP
min support
50
70
min ε
0.0
0.1
0.2
0.3
0.4
0.5
min $n_{Gt,Ph}$
2
3
p
yes
no
ASR method
min support
min ε
min $n_{Gt,Ph}$
$p(X^2)(maxT)$
$p(\varepsilon)(maxT)$
$p(X^2)(bonf)$
$p(n_{Gt,Ph})(bonf)$
$p_{FISHER}(bonf)$
100
80
60
40
20
0
sig. genes
- exp. ERG11. 8 S<->R
- no exp.. 20 S<->R
- exp. FKS1. 6 S<->R
- exp. ERG11,TAC1b. 7 S<->R
- exp. ERG11,TAC1b. 7 S<->R
- exp. FKS1. 10 S<->R
- exp. ERG11,TAC1b. 9 S<->R
- exp. ERG11,TAC1b. 5 S<->R
- exp. PDR1. 21 S<->R
- exp. FKS1,FKS2. 5 S<->R
- exp. PDR1. 11 S<->R
- exp. PDR1. 9 S<->R
* 1/1 or 2/2 sig. genes
~ 1/2 sig. genes

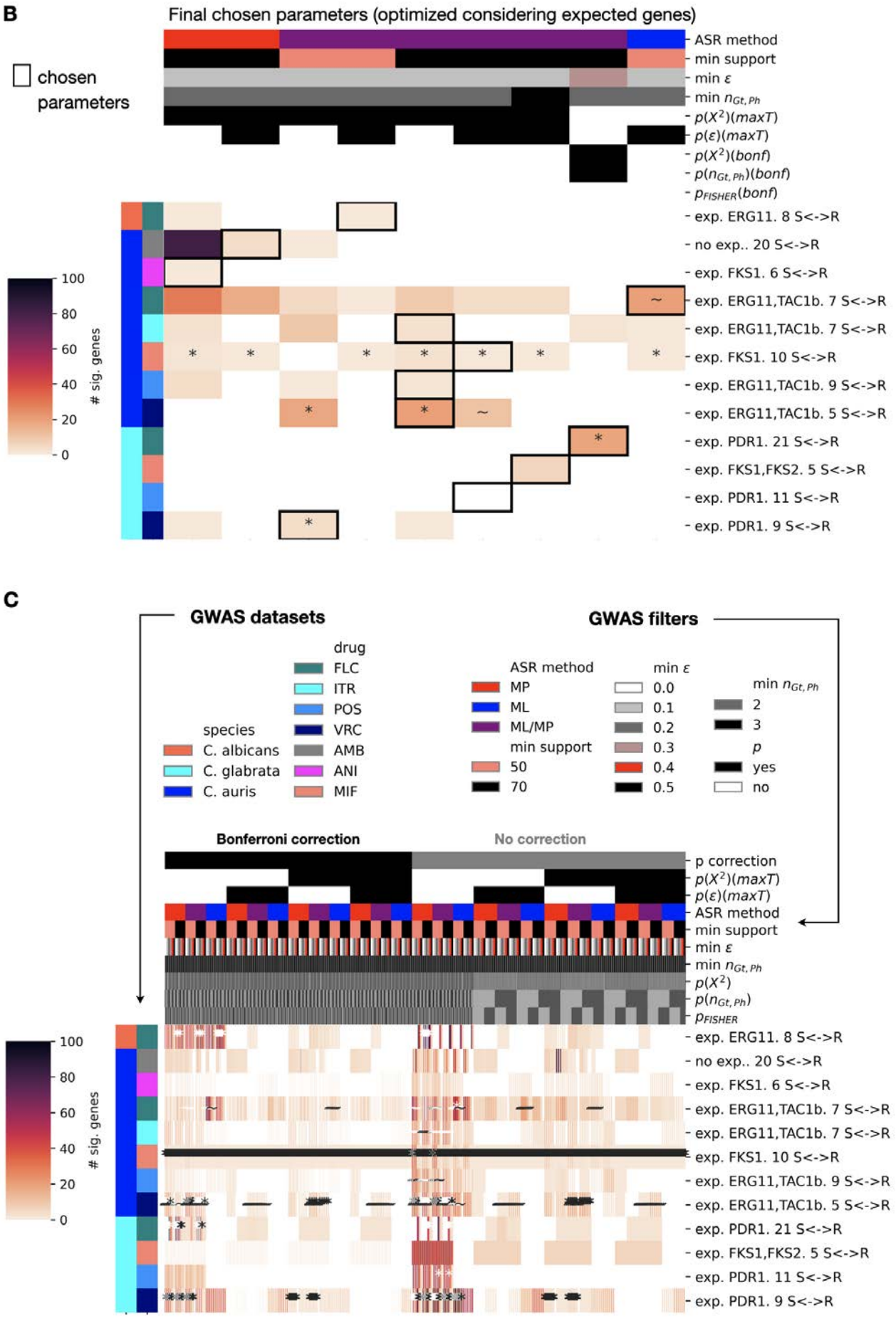
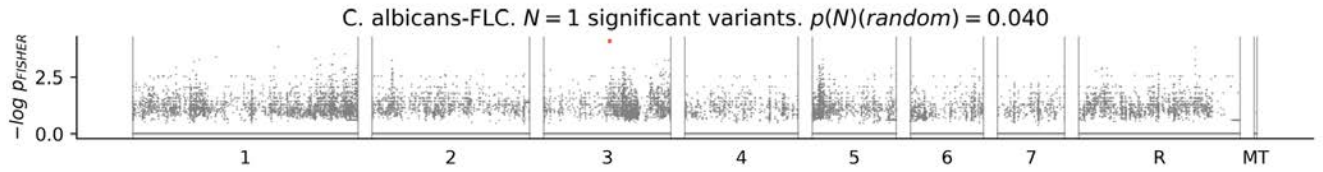
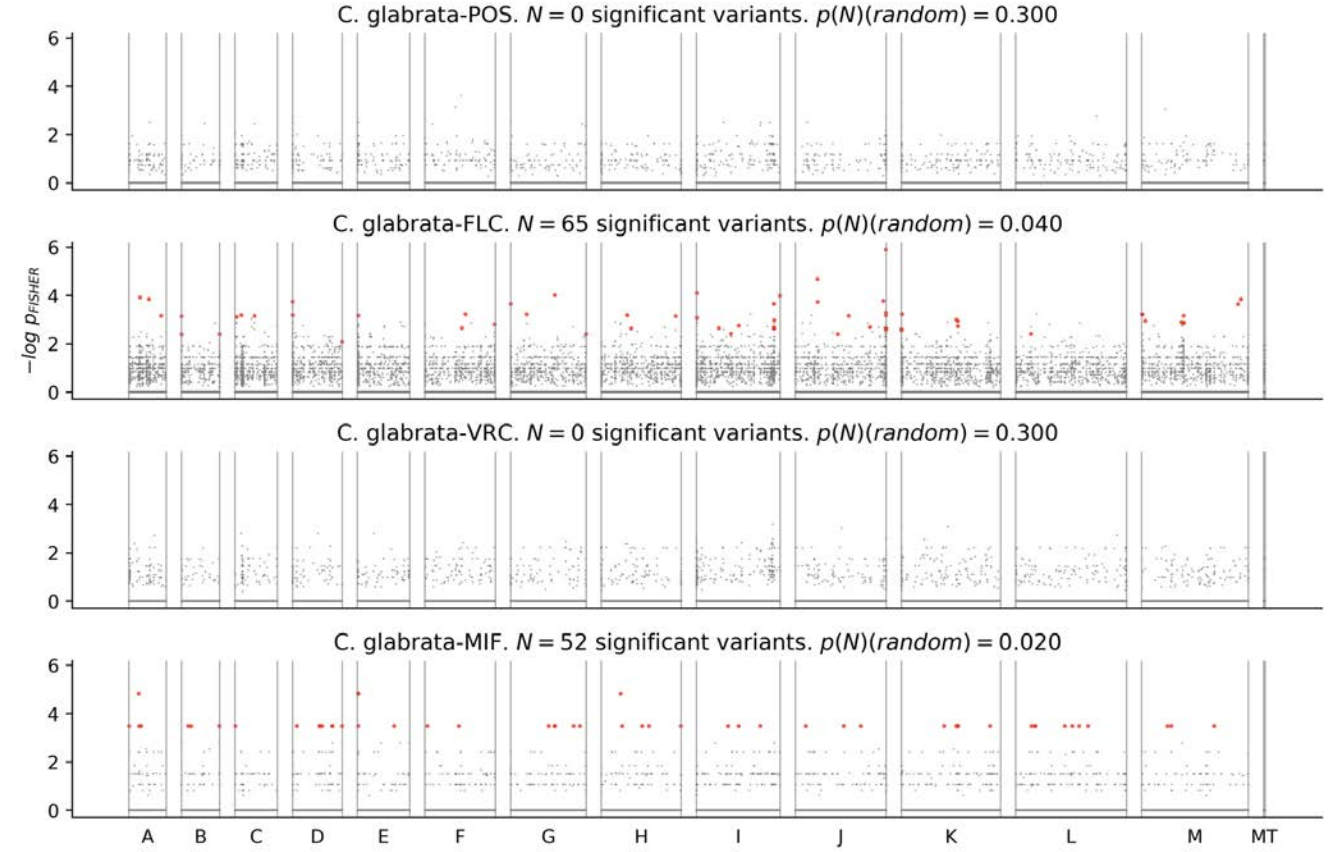

B
Final chosen parameters (optimized considering expected genes)
chosen parameters
ASR method
min support
min ε
min $n_{Gt,Ph}$
$p(X^2)(maxT)$
$p(\varepsilon)(maxT)$
$p(X^2)(bonf)$
$p(n_{Gt,Ph})(bonf)$
$p_{FISHER}(bonf)$
exp. ERG11. 8 S<->R
no exp.. 20 S<->R
exp. FKS1. 6 S<->R
exp. ERG11,TAC1b. 7 S<->R
exp. ERG11,TAC1b. 7 S<->R
exp. FKS1. 10 S<->R
exp. ERG11,TAC1b. 9 S<->R
exp. ERG11,TAC1b. 5 S<->R
exp. PDR1. 21 S<->R
exp. FKS1,FKS2. 5 S<->R
exp. PDR1. 11 S<->R
exp. PDR1. 9 S<->R
sig. genes
100
80
60
40
20
0
C
GWAS datasets
GWAS filters
drug
FLC
ITR
POS
VRC
AMB
ANI
MIF
species
C. albicans
C. glabrata
C. auris
ASR method
MP
ML
ML/MP
min support
50
70
min ε
0.0
0.1
0.2
0.3
0.4
0.5
min $n_{Gt,Ph}$
2
3
p
yes
no
Bonferroni correction
No correction
p correction
$p(X^2)(maxT)$
$p(\varepsilon)(maxT)$
ASR method
min support
min ε
min $n_{Gt,Ph}$
$p(X^2)$
$p(n_{Gt,Ph})$
p_{FISHER}
exp. ERG11. 8 S<->R
no exp.. 20 S<->R
exp. FKS1. 6 S<->R
exp. ERG11,TAC1b. 7 S<->R
exp. ERG11,TAC1b. 7 S<->R
exp. FKS1. 10 S<->R
exp. ERG11,TAC1b. 9 S<->R
exp. ERG11,TAC1b. 5 S<->R
exp. PDR1. 21 S<->R
exp. FKS1,FKS2. 5 S<->R
exp. PDR1. 11 S<->R
exp. PDR1. 9 S<->R
sig. genes
100
80
60
40
20
0

Figure S6. Various GWAS parameters and filtering criteria largely alter the resulting hits. (A) Heatmap showing how different parameter/filter combinations (columns) yield various numbers of genes with significant GWAS hits (color scale) in each dataset (rows). As GWAS parameters, we varied both the Ancestral State Reconstruction (ASR) method (Maximum Parsimony (MP), Maximum Likelihood (ML) or a ML/MP consensus) and the minimum branch support to consider nodes (50 or 70). As filters, we considered filtering based on any combination of five p values (requiring them to be <0.05): $p(X^2)$ *(maxT)*, $p(\varepsilon)$ *(maxT)*, $p(X^2)$ *(bonferroni-corrected)*, $p(n_{Gt,Ph})$ *(bonferroni-corrected)* and/or p_{FISHER} *(bonferroni-corrected)*. In addition we applied various threshold on the minimum ε (0-0.5) and the minimum $n_{Gt,Ph}$ (2 or 3). The symbols represent the fraction of expected genes (*ERG11, FKS, PDR1* and *TAC1b*, shown in the y ticks) that have significant associations ('*' means 1/1 or 2/2 genes and '~' means 1/2 genes). The number in the ticks (i.e. 7 S<->R) represent the number of high-confidence phenotype transitions observed in each dataset. The yes/no boxes indicate which types of p values where required to be p<0.05 for significance. Note that any instance with >100 significant genes or without proper multiple testing correction (inferred with an empiric approach described in **Online Methods**) are set to 0 (blank cells) in this plot. **(B)** The same as in (A), but showing only the subset of filters that we chose to define high-confidence hits in each dataset (see **Online Methods**). The black boxes indicate these chosen filters. In the **Supplementary results** we discuss why we don't find the expected genes in some echinocandins and azoles. **(C)** The same as (A), but showing the effect of the non-corrected p values (right columns) in addition to the bonferroni-corrected ones (left columns). In addition, to generate this plot we did not consider whether the multiple testing burden is addressed (see **Online Methods**). Note that the *maxT* p values are readily corrected for multiple testing, and we reordered the column color boxes (as compared to (A)) to show the results of different combinations of correction methods (considering bonferroni-corrected and/or *maxT* p-values).

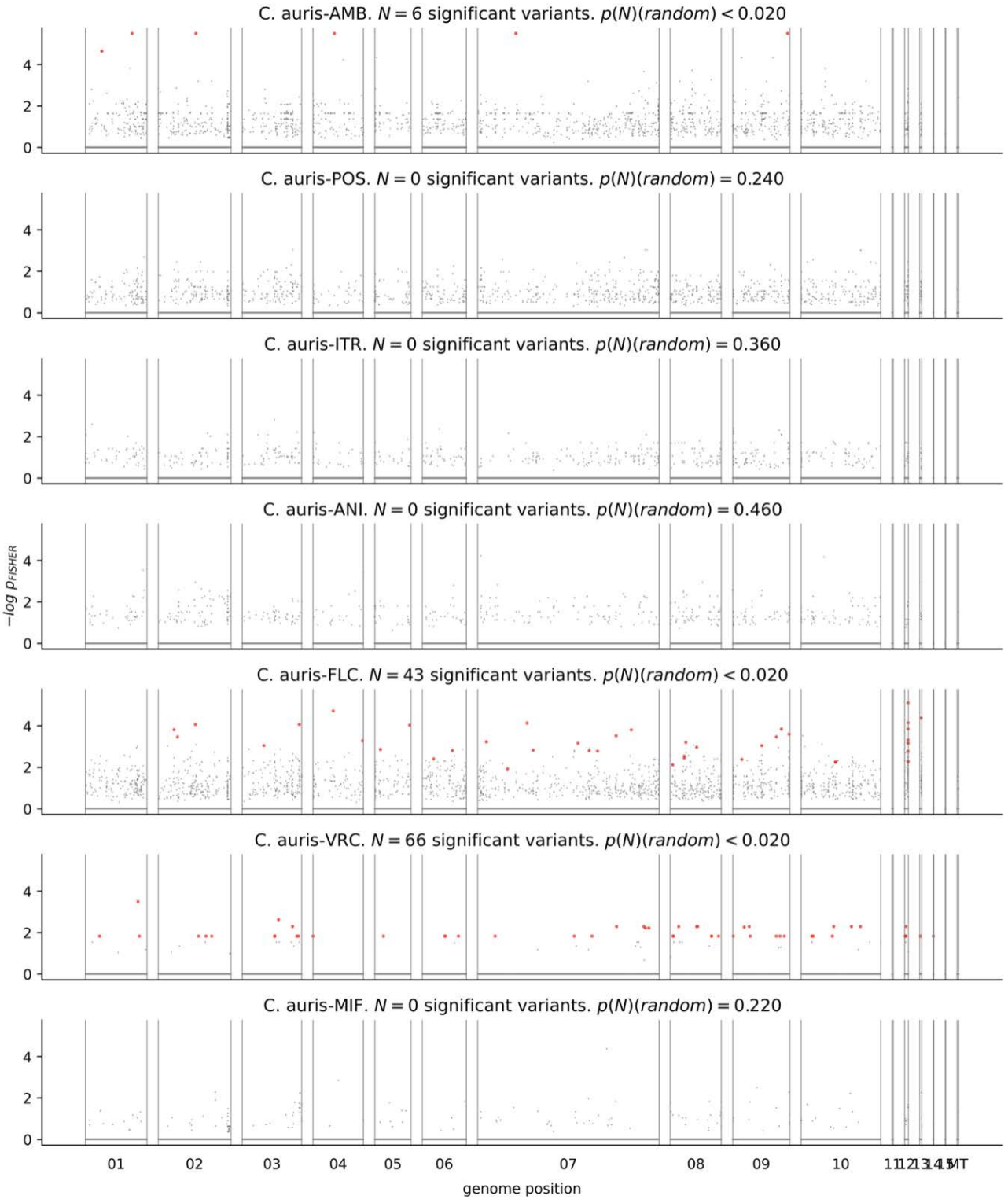
A


C. albicans-FLC. N = 1 significant variants. p(N)(random) = 0.040
−log pFISHER
2.5
0.0
1
2
3
4
5
6
7
R
MT

B


C. glabrata-POS. N = 0 significant variants. p(N)(random) = 0.300
C. glabrata-FLC. N = 65 significant variants. p(N)(random) = 0.040
C. glabrata-VRC. N = 0 significant variants. p(N)(random) = 0.300
C. glabrata-MIF. N = 52 significant variants. p(N)(random) = 0.020
−log pFISHER
6
4
2
0
A
B
C
D
E
F
G
H
I
J
K
L
M
MT

c


C. auris-AMB. N = 6 significant variants. p(N)(random) < 0.020
C. auris-POS. N = 0 significant variants. p(N)(random) = 0.240
C. auris-ITR. N = 0 significant variants. p(N)(random) = 0.360
C. auris-ANI. N = 0 significant variants. p(N)(random) = 0.460
C. auris-FLC. N = 43 significant variants. p(N)(random) < 0.020
C. auris-VRC. N = 66 significant variants. p(N)(random) < 0.020
C. auris-MIF. N = 0 significant variants. p(N)(random) = 0.220
−log pFISHER
0
2
4
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
MT
genome position

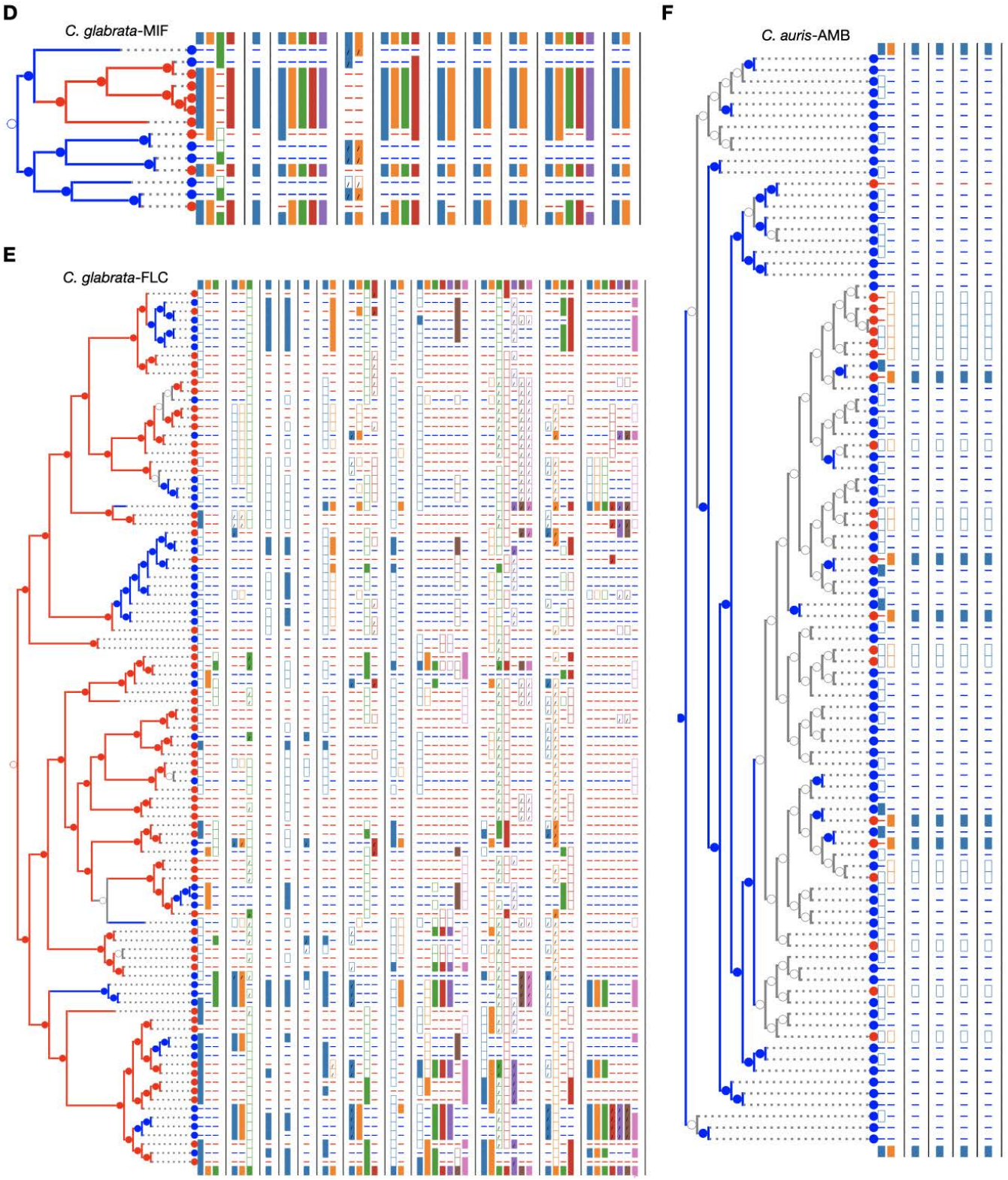

D
C. glabrata-MIF
E
C. glabrata-FLC
F
C. auris-AMB

Figure S7. Individual variants are associated with drug resistance in some GWAS datasets. (A,B,C) Manhattan plots showing the Fisher p value of the genotype-phenotype association for each variant (including SNPs, INDELs, CNVs and SVs) along the genomes of *C. albicans* (A), *C. glabrata* (B) and *C. auris* (C). Each subplot represents one drug for which we performed GWAS. Red points indicate variants that passed all the high-confidence filters (see **Online Methods** and **Figure S6B**). The *p(N)* represents the empirical probability of observing *N* or more variants under a null model of no association (see **Online Methods**). We use the fisher p value because it is appropriate for such a visualization, although it was not always used for the definition of significant hits. **(D,E,F)** These figures show the presence/absence pattern of all significant SNPs (red points in A-C) across strains for three example GWAS datasets (D,E,F). Each column represents a position (only biallelic positions) with one of these SNPs, and the gray vertical lines separate variants from different chromosomes. If a strain has a filled box for a given position/variant it means that the variant transition is correlated to a resistance transition in that strain or one of the ancestors. If a strain has an empty box it means that it has the variant, but it is not related to a resistance transition in that strain. The box colors are to ease visualization of where each SNP is in each strain. The '/' indicate SNPs that are heterozygous and found in duplicated regions. The circles of each node represent resistance (red), susceptibility (blue) or uncertain susceptibility (gray), according to ancestral state reconstruction. The nodes with a support < 70 have non-filled circles. In (C), we only show a clade of the tree with the significant SNPs.

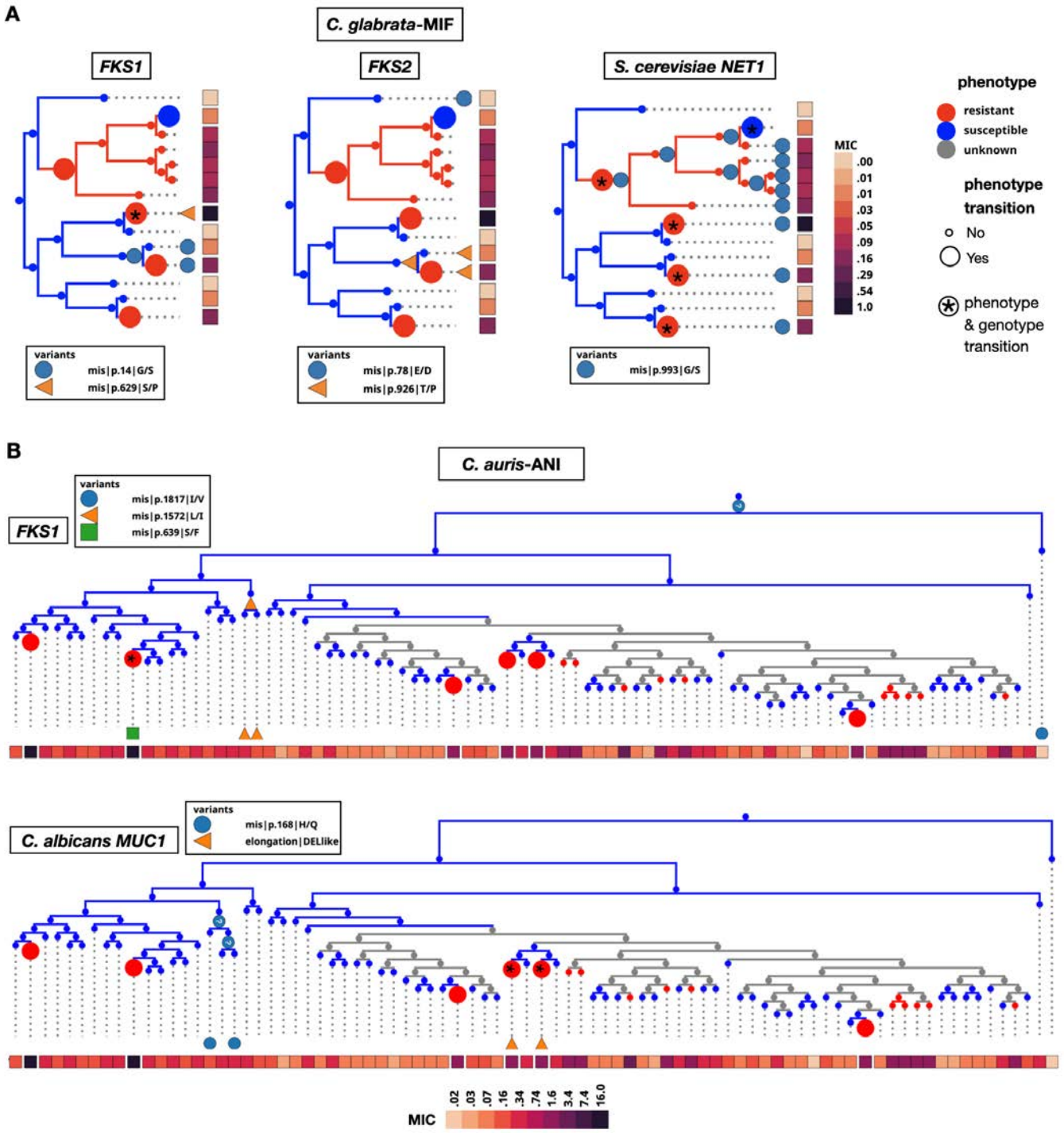


Figure S8. *FKS* mutations are not associated with resistance in some echinocandin datasets. (A) Representation of relevant variants and resistance phenotypes for the *C. glabrata*/micafungin (MIF) dataset, equivalent to what is shown in **Figure 5C**. The left and middle plots show *FKS1/FKS2* variants, which are not associated with resistance. The right plot refers to the top hit in this dataset: a G993S variant in *NET1*. The color scale shows the micafungin MIC for each of these strains. **(B)** The same as in (A), but for the *C. auris*/anidulafungin dataset. The upper plot shows *FKS1* variants, and the lower plot refers to the top hit in this dataset: small variants and an SV (rearrangement in the transcript that does not break the gene) in *B9J08_003526* (ortholog of *C. albicans' MUC1*).

SUPPLEMENTARY TABLES

Table S1. Strains used in this study. The tab **'All strains'** includes the metadata for all strains used in this study. We include the species, BioProject, Run, BioSample, numeric sample ID, clade, clonal cluster, type of strain, collection date, collection location, coverage, link of the paper from which metadata was obtained and susceptibility profiles. The antifungal drugs are amphotericin B (AMB), beauvericin (BVN), 5-flucytosine (5FC), terbinafine (TRB), miconazole (MIZ), ketoconazole (KET), isavuconazole (IVZ), voriconazole (VRC), posaconazole (POS), itraconazole (ITR), fluconazole (FLC), micafungin (MIF), caspofungin (CAS) and anidulafungin (ANI). We report the MIC_{50} for all of them except AMB, which includes MIC_{90}. The 'resistance', 'susceptibility' and 'intermediante_susceptibility' columns indicate the discrete susceptibility profile (for the tested drugs) either stated in the literature or derived from breakpoints on the MIC data (see **Online Methods**). The 'type' column is either 'clinical' (isolated from patients), 'environmental' (from soil or sea), 'genome_engineered' (strains that had some genetic engineering), 'genome_engineered/inmouse_evol_clone' (strains that had some genetic engineering and underwent *in-mouse* evolution), 'inmouse_evol_clone' (strains that underwent *in-mouse* evolution), 'invitro_evol_clone' (strains that underwent *in-vitro* evolution), 'invitro_evol_population' (strains that underwent *in-vitro* evolution and the whole population was sequenced) , 'one_homozygous_chromosome' (strains with one homozygous chromosome) or 'reference' (reference strains). The 'mean_coverage' is the mean read depth across windows of the genome. The 'pct_covered' is the mean percentage of the window covered across windows of the genome. Note that the strains here are all the ones for which we did variant calling except two *C. auris* samples that may be a mix of divergent lineages (see **Online Methods**). The column 'cladeID_systematic' indicates the systematically-defined clades from this work (see **Online Methods**). The column 'cladeID_previous' indicates the clades defined in previous studies (see **Online Methods**). Finally, the 'clonal_cluster' indicates the cluster of close strains (used in the selection analysis (**Figure 3**)) to which each strain belongs. The tab '**Strains overview**' includes the number of strains per species ('# strains'), the numbers of each type of strain ('# clinical', '# environmental' and '# other'), the number of clade ('# clades') and the average intraspecific diversity ('median pairwise SNPs/kb'). The tab '**GWAS drugs overview**' shows, for each species and drug where we performed a GWAS, the numbers of resistant (R) and susceptible (S) strains, the number of resistance phenotype transitions ('R>S or S>R transitions'), and the fraction of clades that have some resistant or susceptible strain. The tab '**Reference genomes**' tab has the information about the reference genomes and annotations used. We merged the gDNA and mtDNA if their source was not the same to get the reference genome. Note that the gff annotations were taken from the equivalent source for each gDNA and mtDNA. CGD stands for Candida Genome Database. The 'CGD gene features source' refers to the CGD table with chromosomal feature files from http://www.candidagenome.org/download/chromosomal_feature_files/. The 'CGD GO annotations

file' is the name of the file at http://www.candidagenome.org/download/go/archive from which we got Gene Ontology annotations. The tab '**Strain trees**' shows the trees generated for each strain in newick format, based on the numeric sample IDs. Note that in the github repository of this project (see **Data and Code Availability**) we provide the csv versions of most of these excel tabs, which are more suited for large-scale reanalysis.

Table S2. Recent selection data. The tab '**Genes under selection'** includes the selection scores and p values for all genes with a significantly high selection score (*S*), meaning that they have an excess of either recent non-synonymous SNPs, deletions (DEL), duplications (DUP) or in-frame INDELs (if_INDEL). This type of variant is indicated in the column 'type_var'. The selection score *S* (column 'selection_score_S') is calculated as the harmonic mean between the fraction of strains with signs of positive selection in the gene ($\pi_N > \pi_S$ for SNP and variant presence for DEL/DUP/if_INDEL) (column 'fraction_strains_w_selection') and the fraction of clonal strain clusters that have some strain with a sign of positive selection (column 'fraction_clusters_w_selection'). The columns 'total_number_strains' and 'total_number_clusters' indicate the absolute number of strains and clusters used to calculate *S*. In addition, 'significant_selection' is a TRUE/FALSE boolean showing whether the gene has significantly high *S*. For SNPs, the 'fdr_p_S' represents the FDR-corrected p value of observing a selection score $\geq S$ under a neutral model of evolution. Columns 'chromosome', 'start', 'end', 'gene_name', 'Scerevisiae_orthologs' and 'description' are gene features obtained from CGD. The column 'gff_upmost_parent' indicates the ID of the gene in the gff file, which is our proxy for gene ID. The column 'orthofinder_orthocluster' is the ID of the Orthologous Group (OG) in which the gene belongs, according to orthofinder. We used these to calculate the number of species that have genes of this OG under selection (column 'n_species_orthogroup') and the number of types of variants that yield a gene under selection in this OG in a given species (column 'n_types_vars_in_species_orthogroup'). Finally, the columns 'biological_process_GO', 'molecular_function_GO' and 'cellular_component_GO' indicate the annotated Gene Ontology terms. The tab '**Genes under selection >1 species'** is a subset of the 'Genes under selection**'** tab (with the same columns), only including genes where the OG has genes under selection in >1 species. The tab '**Selection scores all genes**' is equivalent to the 'Genes under selection**'** tab (with the same columns), but including all genes for which we could measure a selection score *S*. Most of these genes do not have significant signs of selection (as indicated in the column 'significant_selection'), but we provide them because they may be useful to validate the *S* values for specific genes. The tab '**Functional enrichments**' includes the Gene Ontology (GO) terms, Interproscan annotations, MetaCyc and Reactome pathways enriched in the sets of genes under selection by different species and types of variants. The column 'type_grouping' indicates the type of enrichment performed: GO Biological Process (GO_BP), GO Molecular Function (GO_MF), GO Cellular Component (GO_CC), Interposcan annotations (IP_domains), Metacyc or Reactome. 'ID' is the identifier of the enriched group. The columns 'ngenes_group_and_target', 'ngenes_no_group_target', 'ngenes_group_no_target', 'ngenes_no_group_no_target' indicate the numbers

of genes belonging to the target set (genes under selection) and/or to the group (i.e. genes with a given GO term), which were used for the association test. 'OR' and 'p_raw' are the odds-ratio and p value resulting from the Fisher test, and 'p_fdr' is the FDR-corrected p value. The 'group_name' indicates the annotation description. Finally, the column 'genes' indicates the set of genes that drive the enrichment (they belong to the group tested and are also under selection by a given species and type of variant). Note that we considered as 'significant enrichments' those that had a p_fdr<0.05 and an OR $\geq$ 2. The tab '**Gene features**' includes the gene name, *S. cerevisiae* ortholog, orthofinder OG and description for all genes (where the gene ID is 'gff_upmost_parent'). This tab is useful to do further analyses where linking between genes and OGs is required. Note that in the github repository of this project (see **Data and Code Availability**) we provide the csv versions of all these excel tabs, which are more suited for large-scale reanalysis.

Table S3. GWAS associations. The tab '**High-confidence GWAS hits**' includes all the GWAS hits that passed the strict, high-confidence filters (see **Online Methods**). Each row is one hit (it may be a variant, a domain, a gene or a pathway). This list is non-redundant, meaning that there is only one hit per gene or pathway in each GWAS dataset, and it includes hits for pathways that don't have any belonging gene as a hit. The columns 'type_vars', 'type_mutations' and 'type_collapsing' refer to the collapsing strategy used to find that hit (see **Online Methods** and **Figure 6**). The column 'group_name' is the unique identifier of the hit. For domain-level collapsing, it includes the ID of the domain, the gene name, and the range of the protein altered. The columns 'epsilon' and 'OR' indicate the strength of the association. The columns 'nodes_GenoAndPheno', 'nodes_noGenoAndNoPheno', 'nodes_GenoAndNoPheno' and 'nodes_noGenoAndPheno' indicate the numbers of nodes that have a genotype transition and/or a phenotype transition. In addition, 'orthogroups' refers to the Orthologous Groups (OGs) affected by that hit (only relevant for variants, domains and genes). The column 'n_spp_drug_worthogroups' refers to the number of GWAS datasets that have a hit affecting these OGs. Similarly, 'n_spp_drug_wpathway' indicates, for pathway hits, the number of GWAS datasets that have a hit involving a given pathway. On another line, the fields 'pval_chi_square_maxT', 'pval_epsilon_maxT', 'pval_chi_square_phenotypes', 'pval_GenoAndPheno_phenotypes' and 'pval_fisher' are the raw p values for each hit, while the 'pval_chi_square_phenotypes_bonferroni', 'pval_GenoAndPheno_phenotypes_bonferroni' and 'pval_fisher_bonferroni' include the corresponding bonferroni-corrected p values. Note that the 'maxT' p values are already corrected for multiple testing. Finally, 'description' is a text that indicates what is the hit, while 'biological_process_GO', 'cellular_component_GO', molecular_function_GO' are the GO terms annotated for each gene (only relevant for variants, genes and domains). The tab '**High-confidence GWAS hits >1 dataset**' includes a subset of high-confidence hits (from 'High-confidence GWAS hits') where the implicated OGs or pathways are significant in >1 dataset. The tab '**Low-confidence GWAS hits**' includes the GWAS hits obtained with relaxed filters. There may be some false positives here (so that this tab is not

useful for exploratory analysis), but these low-confidence hits may be suited to test hypotheses about specific drug resistance genes. The columns of this tab are the same as 'High-confidence GWAS hits', but also including 'ASR_method' and 'min_support', which refer to the GWAS parameters that yielded each hit. This table includes stacked low-confidence hits for six combinations of parameters, since 'ASR_method' can be 'DOWNPASS', 'MPPA' or 'MPPA,DOWNPASS' and 'min_support' can be '50' or '70'. This means that some hits could appear multiple times, since the different parameters often yield similar results. Note that in the github repository of this project (see **Data and Code Availability**) we provide the csv versions of all these excel tabs, which are more suited for large-scale reanalysis.

SUPPLEMENTARY REFERENCES

1. Heimark, L. *et al.* Mechanism of azole antifungal activity as determined by liquid chromatographic/mass spectrometric monitoring of ergosterol biosynthesis. *J. Mass Spectrom. JMS* **37**, 265–269 (2002).
2. Ni, Q. *et al.* CgPDR1 gain-of-function mutations lead to azole-resistance and increased adhesion in clinical Candida glabrata strains. *Mycoses* **61**, 430–440 (2018).
3. Rybak, J. M. *et al.* Mutations in TAC1B: a Novel Genetic Determinant of Clinical Fluconazole Resistance in Candida auris. *mBio* **11**, e00365-20 (2020).
4. Perlin, D. S. Resistance to echinocandin-class antifungal drugs. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **10**, 121–130 (2007).
5. Ksiezopolska, E. & Gabaldón, T. Evolutionary Emergence of Drug Resistance in Candida Opportunistic Pathogens. *Genes* **9**, 461 (2018).
6. Ksiezopolska, E. *et al.* Narrow mutational signatures drive acquisition of multidrug resistance in the fungal pathogen Candida glabrata. *Curr. Biol. CB* **31**, 5314-5326.e10 (2021).
7. Galocha, M. *et al.* Genomic evolution towards azole resistance in Candida glabrata clinical isolates unveils the importance of CgHxt4/6/7 in azole accumulation. *Commun. Biol.* **5**, 1–12 (2022).
8. Vu, B. G., Stamnes, M. A., Li, Y., Rogers, P. D. & Moye-Rowley, W. S. The Candida glabrata Upc2A transcription factor is a global regulator of antifungal drug resistance pathways. *PLoS Genet.* **17**, e1009582 (2021).
9. McTaggart, L. R., Cabrera, A., Cronin, K. & Kus, J. V. Antifungal Susceptibility of Clinical Yeast Isolates from a Large Canadian Reference Laboratory and Application of Whole-Genome Sequence Analysis To Elucidate Mechanisms of Acquired Resistance. *Antimicrob. Agents Chemother.* **64**, e00402-20 (2020).
10. Sanglard, D., Ischer, F., Calabrese, D., Majcherczyk, P. A. & Bille, J. The ATP binding cassette transporter gene CgCDR1 from Candida glabrata is involved in the resistance of clinical isolates to azole antifungal agents. *Antimicrob. Agents Chemother.* **43**, 2753–2765 (1999).
11. Chow, N. A. *et al.* Tracing the Evolutionary History and Global Expansion of Candida auris Using Population Genomic Analyses. *mBio* **11**, e03364-19 (2020).
12. O'Brien, C. E. *et al.* Population genomics of the pathogenic yeast Candida tropicalis identifies hybrid isolates in environmental samples. *PLoS Pathog.* **17**, e1009138 (2021).

4. GENERAL DISCUSSION

4. GENERAL DISCUSSION

Understanding recent adaptation in *Candida* pathogens is key to infer mechanisms of virulence, transmissibility and antifungal drug resistance. In the research articles that conform this PhD thesis (sections 3.2, 3.3 and 3.4) we addressed various open questions in this field by developing and using comparative genomics tools. In this summarizing discussion I integrate our findings, discussing broader implications and future directions.

On the one hand, we addressed various methodological gaps. Studying recent adaptation in *Candida* pathogens requires adequate bioinformatic tools for variant calling, filtering and functional annotation. Among other reasons, current tools are suboptimal due to i) limited accuracy to identify structural variants (SVs) and ii) the lack of easy-to-use, reproducible pipelines. To address these gaps we developed the "personalized Structural Variation detection" pipeline (perSVade), a Swiss-knife-like framework to call, filter and annotate several variant types, including SVs, directly from raw reads (see 3.2). In the sections below (4.1 and 4.2) I discuss the broad implications of this new tool on the study of recent adaptation in *Candida*.

On the other hand, we addressed open questions regarding recent adaptation in *Candida* pathogens. As reviewed in the introduction (see 1.2.5), our current knowledge is limited due to i) overlooking SVs, ii) exclusive focus on known driver genes, iii) statistically underpowered analyses, iv) understudied species and v) unknown mechanisms of adaptation to combination therapies. To address these gaps, we first analyzed the mechanisms and tradeoffs of resistance towards azoles and echinocandins in *C. glabrata*, using a large-scale *in vitro* evolution approach (see 3.3). Furthermore, we reanalyzed ~2,000 public genomes to understand the signs of recent selection and drug resistance in six major *Candida* species (see 3.4). In the sections below (4.3 and 4.4) I integrate our findings and discuss their broad implications for the *Candida* field.

4.1. PerSVade enables accurate detection of structural variants

Existing studies on *Candida* adaptation mostly explored small variants (SNPs and INDELs) and coverage-based CNVs (a subset of all SVs) (*139*, *156*, *157*). Since coverage-based CNV calling is inaccurate, and more complex SVs (i.e. translocations or inversions) are usually overlooked, the contribution of SVs to recent adaptation in *Candida* remains understudied. As discussed in our paper (see 3.2), this is mostly due to limitations in short read-based SV calling methods, whose applicability and accuracy on non-model organisms (i.e. *Candida*) remains unknown. Given the relevance of such variants (*161*, *236*, *278*) and the cost-effective nature of short read sequencing (see section 1.2.5), there is a need for improved SV-calling methods working with short reads. To solve this, perSVade automatically adapts a state-of-the-art SV calling pipeline to the input sample through simulations, which are used to choose optimal filtering parameters. This enables accurate SV identification for any species of interest, as shown with our benchmark analysis. In addition, the simulations inform about the expected calling accuracy, a key aspect to build trust in the called variants. Given that SV callers often yield low recall (*277*, *279*), this 'trust' is not trivial, and perSVade allows understanding whether the called SVs are meaningful in the sample of interest. Our testing on multiple eukaryotes suggested that this parameter optimization is necessary to achieve high accuracy, both in simulated and real SVs. Our results indicate that perSVade accurately adapts to variation in both i) species-specific genomic features (i.e. ploidy or repeat content) and ii) technical aspects of the library (i.e. coverage, read length or insert size). This underscores the importance of a tool like perSVade to automate the task of SV calling and filtering. From the perspective of computational resources, we find that perSVade is particularly suited for small genomes (i.e. <20 Mbp), although it works too on larger genomes (i.e. human) in a cost-effective manner. In summary, we developed a tool that enables and/or improves SV calling from short reads in *Candida* pathogens and beyond.

More specifically, various lines of evidence support the applicability of perSVade for SV calling in *Candida* species. First, they have small genomes, so that the computational burden of parameter optimization is minimal. Second, we validated that perSVade is accurate on *C. glabrata* and *C. albicans* in our simulation-based testing. For instance, we see that our tool can adapt to extremely high-coverage samples that are typical in these species (i.e. >300x). Third, we used perSVade to find SVs appearing during *in vitro* evolution under antifungal exposure in *C. glabrata*, and we could successfully validate all the high-confidence variants (8/8) through PCR (see 3.3). This illustrates the accuracy of our method. Accordingly, this tool was key to identify the precise rearrangements that lead to CNVs and/or truncations in *ERG3*, *CNE1 ERG11* and *FKS1* underlying *in vitro*-evolved drug resistance in *C. glabrata* (see 3.3). Similarly, perSVade enabled us to study the contribution of SVs to genetic diversity, recent selection and antifungal drug resistance across major *Candida* pathogens (see 3.4). These observations show the importance of

perSVade for SV calling in *Candida* pathogens. Given the growing availability of short read datasets for these pathogens, I envision that this pipeline will enable further studies that clarify the precise role of SVs in recent adaptation.

Admittedly, perSVade's reliance on simulations is a limitation of the method that deserves further discussion. Simulated SVs may not always resemble real SVs, possibly constraining the accuracy of our parameter optimization strategy (*277*). Real SVs may be biased towards certain genomic regions (i.e. those with repetitive elements) (*277*, *292*), so that some simulation types (i.e. randomly placed SVs) may not resemble reality. Accordingly, we find that SV calling accuracy is a bit lower on real human SVs and realistic simulations, as compared to random SVs. This may be due to the fact that some real SVs involve regions that are intractable by short read-based SV callers (i.e. regions with simple repeats or low complexity sequences), so that, even with parameter optimization, these variants are impossible to detect. This is consistent with the observation (done by us (see 3.2) and others (*277*, *279*)) that such algorithms can have limited recall, while precision is mostly high. These considerations suggest that, while the identified SVs are likely to be true, we may miss some variants. In *Candida* pathogens, this implies that the contribution of SVs to diversity, recent selection and drug resistance may be even higher than described here (see 3.4). To address this, an ideal solution would be to optimize parameters taking into account previously-defined sets of real SVs, which should be better for benchmarking parameters (*277*), as done in (*293*). However, this is only possible on species with such *bona fide* variants (i.e. humans), so that simulations may be useful in most species. In fact, our results empirically support the usage of simulations, since simulation-based parameter optimization greatly improves the calling of real SVs. Similarly, we find that, to identify real (or realistic) SVs, parameters optimized for random simulations (demanding minimal computational resources) are as accurate as parameters optimized for realistic simulations. This suggests that optimization based on random simulations is a cost-effective strategy for accurate SV calling, even if the simulations themselves are not entirely realistic. This is why we used random simulations for parameter optimization in our analysis of *Candida* genomes (see 3.3 and 3.4). In summary, while simulation-based optimization is not ideal, it remains a reasonable strategy for SV calling in species with no previously-defined sets of SVs, such as *Candida* species.

On another note, perSVade's SV calling procedure has some limitations that may be improved in future versions of the pipeline. First, memory usage and runtime may be a burden for certain (large) genomes. Beyond already discussed and already-implemented solutions (see 3.2), these may be optimized by i) benchmarking analyses recommending parameters for certain genomic features and technical properties and/or ii) re-writing bottleneck modules to more efficient programming languages (i.e. java). Second, perSVade's parameter optimization is focused on optimizing the balance between precision and recall (maximizing F-score), and it may be useful to provide options that allow for the maximization of either

precision and recall. For instance, in some datasets it may be essential to find the maximum number of true SVs, even if there are many false positives. This could be achieved with an option to choose parameters that have maximum recall, while preserving some precision (i.e. 50%). Third, given that various competitive SV callers exist (i.e. manta (*294*)), it may be interesting to implement parameter optimization modules for them, generating consensus SV calls across many programs. Fourth, the fact that we only simulate certain SV types (insertions, translocations, inversions, tandem duplications and deletions) may constrain real SV calling accuracy because unconsidered types of variants may exist, complicating the summarization of breakpoints into actual SVs. This is a common issue in SV callers, which can only deal with previously-defined known variant types, resulting in some unclassified breakpoints (*276*). Similarly, combinations of SVs around a certain region may complicate this breakpoint-to-SV summarization. For instance, in our *in vitro* study (see 3.3) we find that EF1620_7B_ANI has a combination of an unclassified breakpoint between chromosomes D and L, and a balanced translocation between these chromosomes (as compared to the reference genome). Through manual curation, our interpretation (also confirmed by PCR) is that there is a deletion (including the *CNE1* gene) in the left arm of this new D-L chromosome. Thus, manual examination of these variants (especially in regions with unclassified breakpoints) may be necessary to fully comprehend the functional impact of SVs in *Candida*. To solve this we may develop a module to visualize the results of perSVade across multiple samples, ideally including SVs, CNVs and small variants. In summary, despite their strengths, perSVade's SV calling modules may be further improved by i) optimization of computational efficiency, ii) enabling the maximization of diverse accuracy measurements, iii) consideration of multiple SV calling tools and iv) visualization tools that aid manual curation.

All in all, perSVade is a straightforward pipeline that enables accurate detection and filtering of SVs in a species of interest, which has been instrumental in this PhD thesis. In addition, based on current user feedback, I envision that this pipeline will be useful beyond this project, as it facilitates SV analysis in any species of interest.

4.2. PerSVade simplifies calling and annotation of several variant types

Beyond the SV calling capabilities, perSVade represents a solution to the lack of easy-to-use, reproducible variant calling pipelines for *Candida* genome analysis. Current analyses in these pathogens used custom approaches for read mapping, variant calling and annotation, hindering comparability and reproducibility of the results across studies and species (*155*, *158*). Typical pipelines require long development time and highly specialized knowledge because they rely on the integration of multiple specific tools. In addition, such workflows are usually not reproducible due to complex software dependencies or unavailable source code (*155*, *158*). To address such methodological issues, various efforts have been done in the broader genomics community. For instance, grenepipe (*295*) and MutantHuntWGS (*286*) are automatic pipelines for small variant identification, tested on plants and *Saccharomyces cerevisiae*, respectively. Similarly, YMAP is an online pipeline, suitable for *Candida* pathogens, to identify small variants, CNVs and LOH events (*244*). In addition, ALSgeneScanner is an automatic framework to find small and structural variants related to Amyotrophic Lateral Sclerosis (ALS) in humans (*296*, *297*). These examples show the potential of pipelines for variant analysis directly from the raw reads, specially for non-specialist users. However, to the best of our knowledge, there is a need for easy-to-use tools that can call, filter and annotate several variant types (small variants, CNVs and SVs) in *Candida* genomes.

This motivated us to develop perSVade as a flexible framework that can perform all these tasks from raw reads. First, our pipeline has modules for read quality control, trimming and mapping, which simplify the steps upstream of variant calling. Second, beyond optimized SV identification, perSVade has modules for small variant / CNV calling and filtering from the consensus of three different algorithms. Third, our tool enables the integration of CNV and SV calls into a single output vcf file, removing redundancy between SV and CNV calls. Fourth, perSVade has a module to infer coverage per genes, which can be useful for CNV and aneuploidy calling. Fifth, our pipeline has modules for SV, CNV and small variant functional annotation, which enables downstream analyses. In summary, perSVade is a Swiss-knife-like framework for straightforward variant analysis. Early versions of this tool were instrumental to understand recent mechanisms of adaptation in *Candida* species (see 3.3 and 3.4). Due to its reproducible installability and straightforward usability, I envision that perSVade will boost genomic analyses in *Candida* pathogens and beyond. Accordingly, as of 16/05/2023, the Docker image of perSVade had 479 pulls.

PerSVade has some pipeline design properties that make it broadly usable. First, it offers various options for reproducible installation, using either Docker, Singularity or Conda. This flexibility will likely maximize the usage of perSVade given i) the diverse computational environments in which the pipeline may be deployed and ii) the fact that users may prefer one mode of installation (i.e. Conda) over another (i.e. Docker). Such

simplified installation is significant in the field of *Candida* genomics, since current pipelines have complex software dependencies (*155*, *158*) that may hinder usage. Second, as compared to an one-liner workflow that generates all outputs directly from raw reads, perSVade's modular structure has many benefits. It allows to only use only the needed components, making it easily integrable into more complex pipelines. For instance, some users may be only interested in the small variant or repeat inference functions, and the current design allows this. In addition, although the modular structure requires executing a few more commands (as compared to a one-liner script), it ensures that users are aware of the underlying steps of the pipeline. I believe that this awareness is essential so that users understand the outputs of the pipeline, ensuring reasonable usage and downstream analyses. Similarly, this understanding is guaranteed by the fact that many arguments are mandatory (i.e. ploidy, genetic code or desired small variant calling algorithms), which require thinking about the pipeline before running it. Third, the intermediate steps of the pipeline are often parallelized (i.e. in small variant and CNV calling, variant annotation or repeat inference) to boost performance. Fourth, perSVade automatizes many intermediate tasks (i.e. genome indexing for read mapping or GFF sorting for variant annotation), simplifying variant analyses. In summary, perSVade's installation options, modular structure and simplicity make it a reproducible, highly flexible toolkit.

More specifically, the module 'call_small_variants' has novel, useful features that deserve further discussion. This is a pipeline to integrate the filtered SNPs and INDELs identified by three different callers: GATK HaplotypeCaller (*240*), freebayes (*241*) and bcftools (https://github.com/samtools/bcftools). Such consensus variant identification may maximize calling accuracy (*298*), and this is significant in this field because *Candida* genome analyses usually rely on a single caller algorithm (*155*, *156*).

In addition, we propose that having diverse algorithms for variant identification may be necessary to accurately compare the variants of closely-related samples, such as parent-daughter lineages from *in vitro* evolution experiments and/or studies of serial clinical isolates. To identify variants appearing during a short evolutionary time (i.e. during *in vitro* or within-patient evolution), it is necessary to compare the set of variants present in the strains before (t_0) and after (t) this time, respectively. For instance, in our *in vitro* study (see 3.3) t_0 variants are those in the wild-type (WT), while t variants are those in the evolved strains. To compare such variants, a typical strategy in *Candida* pathogens is to use a single variant caller to define the set of high-confidence variants in the t_0 and t samples (those passing some hard filters, such as a minimum quality score). Thus, variants that are in the high-confidence t variant set, but absent in the t_0 high-confidence variant set, are identified as new, potentially-causal variants (*139*, *182*). We consider that this strategy can yield false 'new' variants because there may be some missing true variants (false negatives) among the high-confidence t_0 set, due to i) usage of only one caller ii) limited accuracy of hard filters. As an illustrative example of this, a study in *C. auris* found that an isolate with ~1,000 SNPs (vs the reference) yielded up to 33 different SNPs (depending on the pipeline) between two independent sequencing runs

(*299*). Thus, this bias may be particularly relevant for strains that are highly divergent from the reference (with thousands of t_0 small variants), where even a small false negative calling rate at t_0 could result in various false positive 'new' variants. Such false positives may lead to significantly erroneous conclusions in such studies. Our solution to this issue (implemented in our *in vitro* study, see 3.3) was to define as t_0 variants those that were called by any of the three callers, even if they did not pass the filters. We propose this as a suitable strategy to minimize the number of false differential variants in such studies. This is possible with perSVade's 'call_small_variants' module because it outputs the unfiltered variants called by the three algorithms, likely maximizing recall at t_0.

Finally, the capacity to work with whole-population genome sequences (*300*) or the adaptability to multiple ploidies and read depths are further examples of the flexibility of the pipeline, enabling broad usage. All in all, perSVade simplifies and improves the calling of small variants due to the consideration of multiple callers and the flexibility to work with various genomic datasets. This explains why this module was instrumental in all our *Candida* genomic analyses (see 3.3 and 3.4).

Similarly, the module 'call_CNVs', has some key novel features. This is a pipeline to integrate the coverage-based variants of CONY (*301*), AneuFinder (*270*) and HMMcopy (*302*), generating consensus 'absolute' CNV calls. By 'absolute CNVs' we refer to variants reflecting the actual copy number (i.e. a strain has two copies of gene X), as opposed to 'relative CNVs' which refer to copy number changes between samples (i.e. for gene X, strain A has twice more copies than strain B). Although CNVs are technically a subset of SVs, I consider that coverage-based CNV calling is necessary because some CNVs may be missed by the SV-calling module ('call_SVs'). The reasons for this include i) variant types not inferred in 'call_SVs' (i.e. aneuploidies), ii) CNVs generated by combinations of rearrangements that difficult breakpoint summarization (see the EF1620_7B_ANI case, discussed above) and/or iii) low recall of 'call_SVs' in some regions. Thus, and despite the limitations of available methods (see 1.2.4), it is relevant to analyze CNVs inferred from coverage changes. Since most approaches for CNV calling in *Candida* rely on raw per-gene coverage calculations (*156*, *157*) or single callers (*139*, *155*, *207*), our multi-caller pipeline may represent an improvement in the field due to improved accuracy, as previously proposed (*303*). In addition, to maximize CNV calling accuracy, this module corrects the coverage biases derived from GC content, uneven mappability and distance to the telomere using non-parametric regression. Note that correction of the telomere-distance contribution is necessary to address the "smiley-pattern" bias, related to the artifactual presence of higher coverage in regions nearby telomere ends (discussed in 3.3). This is similar to YMAP's approach (*244*), although the mappability bias correction and the implementation as a stand-alone software (not online as YMAP) are relevant improvements. In summary, perSVade is a suitable tool to analyze absolute CNVs.

Furthermore, beyond providing 'absolute' CNV calls based on regression-based corrected coverage, this module has an option to perform CNV calling relative to a control sample. Such 'relative' calling is suitable to compare closely-related samples, such as parent-daughter strains from *in vitro* and/or within-patient evolution. While this is a common strategy in cancer genomics, where the coverage ratio between tumor and normal tissues is used to find somatic CNVs (*304*), it remains underused in *Candida* studies (*92*, *93*, *139*, *236*). As an example of the potential of such relative CNV calling in *Candida*, in our *in vitro* study (see 3.3) we used the coverage of the YPD-evolved strains as a control to find CNVs. This was key to address the "smiley-pattern" bias and get accurate CNVs, particularly in strains where balanced translocations complicate the calculation of real telomere distances (i.e. EF1620), which are necessary to faithfully correct this bias.In summary, perSVade's 'call_CNVs' module the calling of coverage-derived CNV calling due to i) the multi-caller approach, ii) various coverage corrections and iii) the option of relative CNV calling. This pipeline was instrumental to study CNVs in our analysis of public genomes (see 3.4).

However, despite the correction of coverage biases and the multi-caller framework, 'call_CNVs' may yield some false positive variants due to the intrinsic limitations of the underlying algorithms (*305*). To address this I recommend treating such variants with caution, verifying them through additional filters. For instance, in our analysis of public sequences (see 3.4) we only kept large CNVs meeting stringent coverage thresholds. Similarly, given the stochasticity of coverage data and the difficulty of defining precise CNV boundaries, one CNV may be split into various chunks in some samples, complicating variant comparison across strains. We speculate that this explains why the CNV and SNP diversity patterns are not correlated within *Candida* species (see 3.4). One solution to this is to only consider the functional effects of CNVs (i.e. gene loss or duplication) in downstream analyses and not the variants *per se,* as we did in our GWAS and selection measurements (see 3.4). Furthermore, the current solution to the "smiley pattern" bias is likely misleading in strains that have large balanced translocations, constraining the accuracy of absolute CNV calling. The main issue is that the (current) calculation of the distance to the telomere is based on coordinates of the reference genome, which is imprecise in such strains because translocations change the real distance to the telomere of affected regions. This has no trivial solution, especially in strains with multiple rearrangements. An alternative option to avoid this is to use relative CNV calling (discussed above and used in 3.3), but it is not suitable when the absolute variants are needed (i.e. in our population genomic study in 3.4). Thus, further improvements to this module may be necessary to maximize absolute CNV calling accuracy. For instance, building a strain-specific genome graph that considers the SVs (i.e. those inferred with the 'call_SVs' module) (*306*) may be useful to define the actual chromosomes in the sample of interest, enabling accurate calculation of the distance to telomeres. All in all, despite its capabilities, 'call_CNVs' has some limitations that may be addressed by i) careful downstream filtering and interpretation of these variants and ii) improved correction of the "smiley pattern" biases.

Beyond enabling the calling of SVs, CNVs, SNPs and INDELs, perSVade allows the integrated analysis of these variant types in two ways. On the one hand, there is an option in the 'call_small_variants' module to input the results of 'call_CNVs', generating an output that has the copy number of the region for each variant. From the perspective of small variant analysis, this is particularly useful to consider genotype calls in the context of gene copy number. For instance, in the haploid species of our GWAS and selection analyses (see 3.4) we considered diploid heterozygous small variants occurring in duplicated regions, which may have various alleles across duplications. This consideration of copy number changes for SNP/INDEL calling enables the identification of such heterozygous variants, which would be missed with typical haploid variant calling (*155*, *156*). Without perSVade, such integration would require specific, complex pipelines, underscoring the relevance of this option.

On the other hand, the module 'integrate_SV_CNV_calls' integrates all SVs and CNVs into a single output vcf file, simplifying downstream analyses of these variants due to various reasons. First, it removes CNVs that match an equivalent SV (i.e. a copy-paste insertion may be called by both SV and CNV calling pipelines). This ensures that the output variants have a unique identity, without redundancy. Second, this module adds various measurements to each variant (i.e. coverage and breakend metadata), which are relevant for downstream variant filtering. Third, the output vcf represents variants with regard to how they affect particular genomic regions, facilitating variant annotation with tools like VEP (*242*). Since most SV callers yield complex outputs, not directly integrable with functional annotators (*275*, *276*), this output represents an important improvement that enables downstream analyses such as GWAS or selection measurements in *Candida* pathogens (see 3.3 and 3.4). All in all, perSVade allows integrated analyses of multiple variant types, enabling more comprehensive variant interpretation and downstream analyses.

Despite the strengths of perSVade as a pipeline that simplifies several variant analysis tasks, there are various incremental improvements that may be relevant. First, it may be interesting to develop modules like 'optimize_parameters' to improve the filtering of small variants and CNVs, which currently relies on hard filters. The choice of filters and algorithms to call such variants is also key to get accurate results (*299*, *305*, *307*), suggesting that such parameter optimization could be useful. Second, a module to automatically integrate the SVs and CNVs of various samples may be suitable. This is not a trivial task because the same variant may be slightly different across samples (*279*). For example, in our studies (3.2 and 3.4) we considered as 'equal' variants those that reciprocally overlap by >75% of their total length, with breakpoints <50bp from each other. In its current state perSVade has some python functions to do this (mentioned in 3.4), but they are not generalizable to every multi-sample dataset. Thus, further additions to the pipeline may be needed to simplify this task. Third, the container-based usage of perSVade (with Docker or Singularity) may be simplified to be more friendly for users that are not comfortable with such tools. In the current version (at 16/05/2023) it is necessary to execute Docker/Singularity commands which sometimes

are confusing, although perSVade's tutorial is oriented to users with no experience on these tools. For instance, a recent user reported an issue with perSVade related to erroneous setting of Docker mounting volumes (see https://github.com/Gabaldonlab/perSVade/issues/14). A solution would be a wrapper script that takes care of the Docker/Singularity arguments, only requiring the input files and arguments from the user. Note that, although the Conda-based installation (available for perSVade) seems a more obvious alternative, it may not be reproducible in all environments due to the multiple dependencies of perSVade. Thus, a container-based installation that is as simple as possible is likely necessary for our tool. In summary, various enhancements of the pipeline could make it even more accurate, simple and user-friendly.

All in all, perSVade is a straightforward pipeline that boosts variant analysis in *Candida* species and beyond, which has been key in this project.

4.3. *In vitro* evolution reveals drug resistance mechanisms in *Candida glabrata*

Antifungal drug resistance dangerously limits therapeutic options for *Candida* infections, particularly in species like *C. glabrata* that frequently adapt to clinically-used azoles and echinocandins (*136*, *245*). The evolutionary mechanisms of this adaptive process remain poorly understood, although such knowledge could be key to improving current treatments and diagnostics (see 1.2). In *C. glabrata*, our limited understanding can be attributed to i) common focus on a limited set of expected resistance genes, ii) limited statistical power in most studies, iii) understudied adaptation to drug combinations, iv) unknown fitness tradeoffs of resistance and v) overlooked role of SVs and CNVs (see 1.2.5). Such missing knowledge is particularly relevant in *C. glabrata* because it may have unique resistance mechanisms due its relatively large evolutionary divergence to other *Candida* pathogens (see **Figure 1B**) (*49*). To address these gaps, we used a large-scale *in vitro* evolution approach to study adaptation to fluconazole and anidulafungin (see 3.3). Our results suggest widespread adaptation to both drugs, which was achieved with moderate fitness costs and through few mutations in nine recurrently-altered genes. In addition, we characterize a novel role of *ERG3* mutations in cross-resistance to fluconazole in anidulafungin-adapted strains. In summary, our study clarifies the mutational paths leading to resistance and cross-resistance to antifungal drugs in *C. glabrata*. Along with other studies in *Candida* pathogens, we (and others (*308*)) envision that our results will improve clinical management of such fungal infections.

Our large-scale experimental setup provided relevant novel insights into the drug adaptation process. As compared to studies of natural variation (see 1.2.1), our approach had many advantages to understand drug resistance. It allowed us to observe the adaptation in real-time, under controlled conditions and with parallel replicate populations, which facilitates pinpointing causal variants amongst those appearing during the experiment. In addition, the size of our collection, with 288 evolved populations and 77 sequenced clones from seven *C. glabrata* clades, likely ensures that the identified adaptive mechanisms are generalizable, and not due to clade-specific effects. This is significant given that previous studies analyzed a few strains covering a handful of clades, particularly for echinocandins (*92*, *159*, *234*, *247*). Studies about echinocandin resistance in this species were typically small-sized (*159*, *255*, *309*), and available genomic datasets with echinocandin susceptibility information remain scarce (see our manually-curated data from 3.4). This makes our anidulafungin resistance results particularly relevant. In addition, we analyzed the evolutionary paths of resistance to serial and combined exposure of both echinocandins and azoles, which gave insights about the suitability of such therapies (discussed below). In summary, various design features make our experimental setup suitable to understand antifungal resistance in *C. glabrata*. This represents an additional example of the suitability of *in vitro* evolution to understand antimicrobial adaptation.

By analyzing the phenotypes (fitness and drug susceptibility) of adapted clones we found that *C. glabrata* has a remarkable capacity to adapt to all the antifungal regimes assayed. We observed that multidrug resistance (MDR) appeared rapidly in all clades, through either serial (AinF, FinA), combined (ANIFLZ) or single (ANI) drug exposure. Such pervasive drug and multidrug resistance had mostly a small fitness cost, which is consistent with the observation that resistance mutations (i.e. in *PDR1*) may increase virulence (*310*), suggesting limited tradeoffs of reduced susceptibility. Together with the fact that only a few mutations are required for resistance, these observations may explain the (concerning) frequent clinical drug resistance in this species (*136*, *245*). There is one exception to this trend, related to the combined exposure to both anidulafungin and fluconazole (ANIFLZ samples). We found that almost half of populations (43.75%) did not survive this regime, suggesting reduced adaptive capability to combination therapies. In addition, among the strains that survived the ANIFLZ regime, several had largely reduced growth in YPD, suggesting that serious fitness tradeoffs may exist in this condition. These observations are consistent with previous reports indicating that combined drug exposure reduced the potential for drug resistance (*311*, *312*). Based on our findings, we speculate that such combination therapies for *C. glabrata* infections could be more effective and yield less resistance, although further clinical studies are needed. In summary, although this species has a high potential for MDR, combined therapies may provide a solution.

A key aspect of our genomic analysis is that it provides insights into the evolutionary mechanisms of azole resistance in *C. glabrata*. In almost all fluconazole-evolved strains, resistance can be attributed to a combination of both i) *PDR1* point mutations and ii) either *ERG11* point mutations or chromosome E (containing *ERG11*) duplications. The changes in *PDR1* likely represent GoF mutations that result in overexpression of drug efflux pumps, a broadly described azole resistance mechanism in this species (*251*).

Conversely, variants in *ERG11* (encoding the enzyme targeted by azoles) and chromosome E duplications (containing *ERG11*) are a rather novel, unexpected result. As reviewed in the Introduction (see 1.2), in several *Candida* species *ERG11* point mutations and duplications of *ERG11*-containing chromosomes may result in resistance due to reduced drug-target binding or drug titration, respectively. Despite this trend, similar observations are rare in *C. glabrata* (*313*), and *ERG11* changes were not considered to be a main resistance mechanism in this species (*181*, *245*). However, given the fact that most of our azole-adapted strains had these alterations, we conclude that changes in *ERG11* (also through chromosome E duplications) play a key role in azole resistance, at least in our experimental setup. This is further supported by the fact that *ERG11* point mutations affect mostly two residues (K152, Y141) near the azole binding pocket (perhaps 'hot spot' regions), potentially disrupting drug binding. In addition, contrary to the assumption that aneuploidies are only a transient adaptive step with high fitness costs (*261*), our results suggest that chromosome E duplications provide a stable selective advantage with minimal fitness tradeoffs. Evidence for this includes that i) almost half of the FLZ and AinF strains carried the duplication until the end of the

experiment, ii) the fitness of strains with this duplication was equivalent to that of strains with other variants iii) most of the aneuploidies were maintained upon removing the drug in FinA. In summary, we clarified the major mechanisms of azole resistance in this species.

Beyond these ubiquitously-altered genes, we find more sporadic recurrent alterations (in *CDR1* and chromosome I), which may represent additional azole resistance mechanisms with lower adaptive potential. For instance, given that *CDR1* is an azole efflux pump related to physiological tolerance (*314*), I speculate that the variants observed are GoF, leading to increased drug efflux that enables resistance. In addition, I propose that chromosome I duplications could be adaptive because they lead to overexpression of *TPO3*, a drug antiporter that has also been related to azole efflux (*315*). These two (mostly novel) resistance mechanisms, although less common, highlight the role of drug efflux for azole resistance. All in all, we gained various insights into novel mechanisms of azole resistance in *C. glabrata*.

Similarly, we pinpoint genetic mechanisms of adaptation to echinocandins in *C. glabrata*. All anidulafungin-evolved strains had point mutations in *FKS1* and/or *FKS2*, encoding the target of the echinocandins. As described for various *Candida* species, including *C. glabrata*, such variants likely impair drug binding, yielding reduced susceptibility (*245*). Consistent with such previous studies, most of the mutations are in 'hot spot' regions, but we also infer (unexpectedly) that mutations outside these regions sporadically contribute to resistance. This further confirms the ubiquitous relevance of *FKS* mutations for echinocandin resistance in this species, and underscores the need to analyze the whole gene (not only 'hot spots') in future research and diagnostic avenues.

Furthermore, our results shed light on the particular evolutionary relationship between *FKS* mutations and drug adaptation in *C. glabrata*. Such adaptation is unique in this species because it has two functionally-redundant *FKS* paralogs, while *Candida* species have mostly one (*316*). This likely generates adaptive features, further clarified by our results, that are relevant to understand the genotype-phenotype landscape of echinocandin resistance. First, having two genes may increase (perhaps double) the probability of acquiring adaptive mutations in at least one of them, which could explain the frequent adaptation to echinocandins in this species (*113*). Consistent with this hypothesis, we find that resistance can be attributed to variants in either *FKS1* or *FKS2*. However, since our analysis is only about *C. glabrata*, further studies about the adaptive potential of *FKS* genes across multiple species may be necessary to fully confirm this. Second, the fact that ~20% of strains had a combination of a non-synonymous mutation in one of the *FKS* genes (likely resistance-conferring) and a truncating variant (likely LoF) in the other gene suggests that interactions among these *FKS* paralogs shape echinocandin adaptation. This is consistent with previous reports in *C. glabrata* showing that *FKS* hot spot mutations confer higher resistance levels in a background

where the other *FKS* gene was deleted (*316*). All in all, our analyses revealed how having two redundant *FKS* paralogs makes echinocandin resistance unique in *C. glabrata*.

More specifically, I propose a model that may explain why a combination of a resistant *FKS* gene and a truncation of the other is adaptive during anidulafungin adaptation. To illustrate this model, let us examine the *FKS* mutational path where the WT strain (with $fks1_{WT}$ / $fks2_{WT}$ alleles, expressing both $FKS1_{WT}$ / $FKS2_{WT}$ proteins) turned to an anidulafungin-evolved strain (with $fks1_{R}$ / $fks2_{LoF}$ alleles, expressing only one functional *FKS* protein: the echinocandin-resistant $FKS1_{R}$). This $fks2_{LoF}$ allele may lead to reduced *FKS2* protein activity due to either i) gene deletion, ii) downregulation through nonsense-mediated decay, triggered by stop codons (*317*) and/or iii) translation of a non-functional truncated *FKS2* protein. I propose that a likely mutational path leading to such $fks1_{R}$ / $fks2_{LoF}$ is that the WT turns into a $fks1_{R}$ / $fks2_{WT}$ strain, with both $FKS1_{R}$ and $FKS2_{WT}$ proteins expressed. This strain likely has reduced susceptibility (since it has the $FKS1_{R}$), but it may have coexisting functional $FKS1_{R}$ and echinocandin-blocked $FKS2_{WT}$ in the cell wall. The presence of such blocked $FKS2_{WT}$ could be partially deleterious, so that additional $fks2_{LoF}$ allele, leading to the final $fks1_{R}$ / $fks2_{LoF}$ strain, may provide an adaptive advantage. In summary, having a single resistant *FKS* protein expressed may be optimal for echinocandin resistance, explaining why LoF mutations are adaptive.

Beyond this ubiquitous mechanism, based on *FKS* mutations, other less frequently-altered genes may be related to anidulafungin adaptation. A large fraction of anidulafungin-adapted strains acquired variants in *ERG3*, including, missense mutations, truncations and promoter alterations. The fitness competition experiments further indicate that such variants provide an adaptive benefit in anidulafungin. This is consistent with the role of *ERG3* LoF mutations in echinocandin adaptation in other *Candida* species (see 1.2.3). I speculate that this may be due to the linked nature of pathways coping with endoplasmic reticulum (ER) and cell wall stress. Studies in *Saccharomyces cerevisiae* have shown that cell wall stressors activate the ER unfolded protein response (UPR) through the cell wall integrity (CWI) MAP kinase cascade, so that UPR is necessary to ensure CWI. In addition, ER stress activates UPR and CWI signaling, suggesting that these pathways are coordinated (*318*). Given that ergosterol is synthesized in the ER (*319*), I propose that changes in *ERG3* alter sterol production and ER metabolism, in a way that reduces cell wall stress exerted by anidulafungin. This is consistent with the involvement of ER stress responses in echinocandin resistance, which is a cell wall stressor, in *A. fumigatus*, *C. neoformans*, *C. albicans* and *C. glabrata* (*320*). This speculative model explains why *ERG3* mutations appear during echinocandin exposure.

In addition, other less frequent mechanisms may be relevant. Two strains had variants in *ERG4*, which may have an effect that is similar to *ERG3*'s changes, given that both genes participate in the ergosterol biosynthesis pathway. However, since these strains also had *ERG3* variants, changes in *ERG4* may also represent compensatory mutations that could resolve biochemical imbalances in sterol production derived

from *ERG3* changes. Finally, three strains had variants in *CNE1* (including deletions), which is likely an ER protein involved in quality control of misfolded glycoproteins (*321*). This gene is also likely related to the ER-CWI link, as its disruption leads to ER stress and cell wall changes in *C. glabrata* (i.e. reduced β-1,6-glucan and increased chitin) (*322*). This direct influence on the cell wall through ER stress modulation may explain why *CNE1* variants are adaptive in anidulafungin treatment. Since one of these variants is a complete deletion, I propose that the adaptive effect of *CNE1* mutations is generally LoF, leading to lower expression or activity. In a way that is similar to *ERG3/4* changes, I speculate alterations on CWI caused by anidulafungin may be (partially) compensated by the changes in *CNE1* (sometimes deletions) through ER stress responses. In summary, we find that changes in *ERG3*, *ERG4* and *CNE1* sometimes influence echinocandin adaptation, maybe due to ER changes that result in beneficial cell wall alterations in the context of anidulafungin exposure.

Beyond single-drug resistance, our results clarify the mutational steps behind MDR derived from combinatorial or serial therapies. In strains evolved in both anidulafungin and fluconazole (AinF, FinA and ANIFLZ) the underlying genetic changes are the combination of those expected from single-drug exposure (ANI and FLZ). For instance, azole resistance could be attributed to changes in *PDR1*, *ERG11*, chromosome E, chromosome I and *CDR1*; while echinocandin resistance was likely related to variants in *FKS1/2*, *ERG3*, *ERG4* and *CNE1*. In addition, we found that, despite a few exceptions, these changes remained (and conferred resistance) after removal of the drug (in AinF, FinA), suggesting that resistance mutations are mostly stable and long-lasting, enabling MDR. Thus, overall, the mechanisms of resistance and MDR were mostly independent of the order of exposure or single vs combinatorial treatment. However, there are some exceptions to this trend suggesting that the treatment regime may shape how MDR is acquired. For instance, three AinF samples did not acquire new *PDR1/ERG11* alterations, likely because the parental ANI strains had levels of fluconazole resistance high enough through mutations in *ERG3*. In addition, *EPA13* adhesin deletions appeared only in two ANIFLZ samples, perhaps related to particular tradeoffs in the combined exposure. Although the role of *EPA13* in the adaptive process is unknown, we speculate that it may be important due to the resulting changes in adhesion, which may influence drug resistance through altering biofilm formation (*119*, *156*, *234*). In summary, despite some exceptions, we find that MDR after serial/combined exposure is acquired through genomic mechanisms that resemble single-drug adaptation.

In addition to the expected MDR derived from combined/serial adaptation, an unexpected result of our work is that MDR may appear in ANI samples, due to *ERG3* mutations conferring cross-resistance towards fluconazole. Importantly, we validated this mechanism through CRISPR-Cas9-mediated strain engineering, which further shows the suitability of this technique in such non-model species. While such cross-resistance has been described before in *C. glabrata* (*323*), the underlying mechanism involving *ERG3* is a novel finding. It is consistent with previous observations in other *Candida* species, where *ERG3* LoF variants may change

sterol composition in a way that is beneficial in azoles, which drive production of toxic intermediates through *ERG3* (see 1.2.3). A similar modulation of sterol profiles may explain this cross-resistance in *C. glabrata*. However, the precise mechanism may be unique in this species, as we find that mostly non-truncating mutations confer this cross-resistance (and not the truncating ones). We cannot fully discard that such non-truncating variants are LoF, but if complete *ERG3* disruption (LoF) was causing fluconazole resistance we would expect truncating mutations to be also correlated to resistance (which is not the case, see 3.3). This is also consistent with the observation that *ERG3* deletion does not yield fluconazole resistance in *C. glabrata* (*324*). Thus, I speculate that, contrary to other *Candida* species, resistance-conferring *ERG3* variants do not entirely disrupt the encoded enzyme, but rather modulate its biochemical activity resulting in sterol profiles that are adaptive in azoles. This species-specific effect, likely related to the unique metabolic properties of *C. glabrata*, underscores the need to study various *Candida* species to fully comprehend antifungal adaptation. In summary, we characterize a novel mechanism of cross-resistance in *C. glabrata*, which results in MDR as an unfortunate byproduct of anidulafungin adaptation.

The fact that *ERG3* variants may underlie adaptation towards both drugs provides further insights about the complexity of genotype-phenotype relationships in antifungal resistance. We find that truncating mutations are potentially adaptive in anidulafungin, but not in fluconazole, suggesting that the impact of *ERG3* variants on adaptation is fundamentally distinct in each drug. I propose that various variant types (including total gene disruption) may trigger the ER-CWI stress responses that are beneficial in anidulafungin (discussed above). Conversely, only a subset of these variants (i.e. those that modulate but do not eliminate enzymatic function) result, as a byproduct, in sterol composition changes that generate fluconazole resistance. In addition, our mutation re-introduction experiments showed that the cross-resistance effect of *ERG3* mutations is stronger in an anidulafungin-adapted background, as compared to a WT background. This suggests a role for epistatic interactions, where the combined effect of *FKS* and *ERG3* variants underlies fluconazole resistance, perhaps due to linked ER-CWI stress responses also influencing ER-produced sterols. Other studies suggested the importance of epistatic interactions for antifungal drug resistance. For instance, a typical azole resistance mechanism in *A. fumigatus* involves the combination of a 34-bp tandem repeat in the promoter and a missense mutation in the *CYP51A* gene (encoding the target enzyme lanosterol 14α-demethylase) (*325*). This suggests that epistatic interactions, despite being challenging to study, should be considered to fully understand drug resistance. All in all, we propose that *ERG3* variants influence drug resistance through distinct pathways, where epistatic interactions may play an important role.

These diverse paths towards MDR are relevant to further analyses of clinical drug resistance, such as our GWAS analysis, where treatment regimes are mostly unknown (see 3.4). Our findings that MDR after serial/combined exposure involves mechanisms analogous to single-drug adaptation suggests an additive

model, where MDR variants can be inferred from mutations underlying resistance towards each of the individual drugs. Such a model validates our approach of studying various drugs in isolation through GWAS. Similarly, testing different variants / genes independently with GWAS is more powerful under the assumption that resistance mechanisms do not rely on the epistatic effect of different variants. I consider that the main resistance mechanisms (through variants in *PDR1*, *ERG11* and *FKS*) likely meet this assumption, validating our GWAS strategy. However, the *ERG3*-mediated cross-resistance represents a notable exception to the additive model and the assumption of independence, as MDR may be acquired in a unique way in anidulafungin-evolved samples, with an importance of epistatic interactions. If such a mechanism exists in the clinics, it partially reduces the power of our GWAS, as fluconazole-resistant isolates may arise due to distinct mechanisms (i.e. through either *ERG3* or *PDR1* mutations). This may result in complex allelic heterogeneity for the resistance trait, perhaps reducing statistical power even if variant collapsing is considered (see 3.4 and 1.2.1). In my opinion, this does not mean that our GWAS strategy is completely powerless in the presence of such interactions / diverse mechanisms, but this is a limitation that should be mentioned. All in all, the fact that multiple evolutionary paths towards MDR exist impacts how clinical drug resistance should be studied, which is relevant for this PhD thesis and beyond.

Although several of the identified resistance mechanisms were expected, our results provide generally new insights into the genetic basis of resistance in *C. glabrata*. First, we find some novel, unexpected alterations underlying drug adaptation in *C. glabrata*. These include variants in *ERG11*, *CDR1*, *CNE1*, *ERG3*, *ERG4* and *EPA13*; and duplications of chromosomes I and E. This shows how a large-scale genome-wide study, not only focused on specific genes, provides a more complete picture about the resistance mechanisms. Second, the large-scale analysis of multiple clades spanning the intraspecific diversity ensures that the inferred adaptive processes are general to the species, and not restricted to single clades. Accordingly, the more common alterations in *ERG11* / chromosome E, *PDR1*, *FKS1/2* and *ERG3* were found in all assayed clades. Similarly, less frequent changes in *CDR1*, *ERG4*, *CNE1*, *EPA13* are found in multiple (at least >1) clades. The only exception to this trend are chromosome I duplications, which we always found in EB0911 backgrounds. My interpretation is that the smaller size of chromosome I (due to translocations, see 3.3) in this strain may result in a lower fitness cost of the duplication, making it a viable adaptive strategy. Despite this rare, clade-specific adaptive mechanism, the fact that our results are mostly clade-independent is significant, since most previous studies focused on one or few clades (*92*, *159*, *234*, *247*). Third, although the most important mechanisms inferred here (changes in *PDR1* and *FKS1/2*) were expected, they provide some relevant novel insights. These findings validate (to some extent) that our setting mimics previous (mostly clinical) studies, as we find similar mechanisms of adaptation. In addition, the fact that we found these amongst the most important genes using our large genome-wide approach (in contrast to previous targeted studies (*247*, *310*)) confirms their importance. Furthermore, our catalog of diverse mutations in these genes

provides a rich, unprecedented resource for further diagnostic applications or research avenues. In summary, our results validate the importance of known resistance mechanisms and suggest new ones.

Another relevant finding is related to our analyses of SVs, which contribute to drug resistance. On the one hand, we find the already-discussed aneuploidies and *EPA13* deletions, using standard coverage-based CNV calling. On the other hand, we find various SVs appearing during antifungal adaptation that are only detectable with a breakpoint-calling tool, like perSVade. These include i) small deletions around *ERG3* and *FKS1*, ii) an unbalanced translocation driving *ERG11* duplication, iii) a combination of a translocation and a deletion leading to loss of *CNE1*, iv) balanced translocations that may enable chromosome I aneuploidies and v) a translocation disrupting *FKS1*. Some of these variants would have been missed by typical coverage-based approaches due to either i) the translocation variants not changing read depth or ii) changes in coverage spanning very small regions (in the case of *ERG3* and *FKS1* deletions). In addition, although some of these SVs (*ERG11* duplications the *CNE1* deletion) are also detectable from coverage changes, perSVade's SV calling approach allowed us to dissect the precise underlying rearrangements, which enables a more detailed interpretation. This further shows the need to consider such complex SVs in further studies. In addition, this supports the suitability of perSVade to precisely understand SVs in *Candida* pathogens.

Beyond the genetic analyses, our unique antifungal susceptibility measurements provide new insights, deserving some discussion. We used both the (standard) Minimum Inhibitory Concentration (MIC) and the resistance Area Under the concentration-vs-growth Curve (rAUC), a novel metric measuring growth over a range of concentrations. We find that MIC and rAUC are generally correlated, although each of them provides some unique insights. MIC is more suitable (as compared to rAUC) to classify strains into discrete susceptibility categories (i.e. resistant vs susceptible). Such a discretization was useful for some of our biological interpretations (i.e. to analyze 'loss-of-resistance' in AinF, FinA samples), and it is practical to define clinical resistance breakpoints. However, despite the general similarity of both measurements, there are some exceptions where MIC may provide misleading susceptibility measurements. As shown in our study and others (*326*), two strains may have the same MIC but very different concentration-vs-growth curves, which represent the actual susceptibility phenotype. A particularly striking example is found in strains growing slightly above 50% (relative to no-drug) even in high concentrations. Despite having some degree of drug susceptibility, such strains would have a MIC that is as high as that of other strains that are not affected by the drug (with ~100% growth relative to no-drug across all concentrations). This bias in MIC, known as the trailing effect (*326*), justifies the need for alternative measurements like rAUC, which better summarize concentration-vs-growth curves. In summary, while MIC was useful for discretization, rAUC allowed us to measure more accurately the quantitative level of susceptibility. These observations support the need to also consider resistance as a continuous quantitative trait in future studies.

Such susceptibility measurements also provide relevant insights into the genetic mechanisms of antifungal resistance. Most notably, we find that *ERG3*-mediated fluconazole cross-resistance leads to such a trailing effect. By inspecting the curves and using rAUC, we found that the fluconazole resistance levels are higher in fluconazole-evolved strains (FLZ, AinF, FinA and ANIFLZ) than in anidulafungin-evolved cross-resistant samples. This indicates that *ERG3*-mediated fluconazole resistance is not as strong as *PDR1*/*ERG11*-mediated resistance, which leads us to two relevant conclusions. First, changes in drug efflux (via *PDR1*) and/or reduced drug binding (to *ERG11*) may be more efficient adaptations than changes in *ERG3* (perhaps modulating sterol composition). Second, this explains why most AinF samples acquired new *PDR1*/*ERG11* variants, as these are likely necessary to achieve maximum resistance. Note that such considerations cannot be made based on MIC, as these strains have equally-high MICs. In summary, by analyzing susceptibility with rAUC, we could better dissect the fundamental differences across various azole resistance mechanisms.

Our findings about the mechanisms of antifungal adaptation, combined with previous evidence, may improve the clinical management of *C. glabrata* infections in the short term. First, our results on ANIFLZ populations suggest that combining azoles and echinocandins is an interesting approach to minimize resistance and improve therapeutic outcomes. This is consistent with the observation that synergistic combined therapies may be suitable in *C. albicans*, as they may require lower doses and yield reduced toxicity (*327*). However, such therapies may also yield unexpected drug interactions and/or adverse reactions, so that further studies are needed to assess their clinical viability in *C. glabrata*. Second, the cross-resistance effect suggests that, if echinocandin resistance appears during treatment, the choice of a secondary azole antifungal should be based on prior susceptibility profiling (as azole resistance may be frequent). Third, the comprehensive list of mutations in *PDR1*/*ERG11*/*FKS*/*ERG3* may be useful to design molecular tools that identify resistance variants, enabling fast assessment of the susceptibility profile, essential to optimize the therapeutic choices (*68*). Combined with i) improved diagnosis of the species and ii) better antifungal stewardship, I consider that these guidelines may improve the clinical outcomes of *C. glabrata* infections.

In addition, our work opens future long-term avenues to prevent resistance from appearing by exploiting the resistance mechanisms inferred here. For instance, the fact that loss of fluconazole resistance in FinA can be attributed to *PDR1* truncation further supports the notion that *PDR1* inhibitors may be a key coadjuvant for azole therapy, likely due to downregulation of efflux pumps. Accordingly, the iKIX1 *PDR1* inhibitor resensitizes azole-resistant strains of *C. glabrata*, improving survival in a murine model (*153*). More broadly, this is supported by the suitability of combining drug efflux inhibitors (i.e. azoffluxin) and azoles in various *Candida* species (*154*). Conversely, such a strategy of inhibiting resistance mechanisms that do not involve the drug target (i.e. *PDR1* in azoles) may not be as feasible in echinocandins, where resistance

mutations mostly affect the drug target. In summary, while azole resistance may be targetable, this may be more complicated for echinocandins.

As an alternative to preventing the emergence of resistance, our analyses reveal a way to generate loss of resistance in echinocandin-resistant clinical infections. The two sequenced AinF strains that lost anidulafungin resistance have unique mutational paths, providing an unexpected clue about how treatment regime may be used to modulate susceptibility. Indeed, the parentals of these strains are the only ANI samples harboring resistance mutations in both *FKS2* ($fks2_R$ allele) and *CNE1* (potentially LoF, discussed above), maintaining unaltered *FKS1* ($fks1_{WT}$ allele) and *ERG3*. I speculate that this pre-existing combination of mutations increased the likelihood of acquiring further *FKS2* truncations ($fks2_{LoF}$ allele) that led to loss of anidulafungin resistance in AinF. I propose that, although the $fks1_{WT}$/ $fks2_R$ may be good enough for survival in the presence of anidulafungin, it may generate some incompatibilities that make it deleterious in the absence of anidulafungin (AinF), where there is no pressure to keep $fks2_R$. For instance, in the same way that a blocked WT *FKS* could be damaging in ANI (discussed above), having a *FKS* with resistance mutations may be less efficient in the absence of the drug. This may increase the chances to turn into $fks1_{WT}$/ $fks2_{LoF}$ in AinF. The fact that these strains have mutated *CNE1* and unaltered *ERG3* may also be relevant, as ANI strains with $fks1_{WT}$/ $fks2_R$ but other variant combinations did not lose resistance in AinF. This may be explained by a (yet to be understood) influence of the ER-CWI stress responses, modulated by specific *CNE1*/*ERG3* sequences, on the $fks1_{WT}$/ $fks2_R$ incompatibilities.

Such findings could be relevant to find ways to modulate resistance in patients where echinocandin therapy failed. By using a combination of *CNE1* inhibitors (mimicking the effect of the variants) and frequent monitoring of *FKS*/*ERG3* mutations, I propose a diagnostic pipeline that may 'force' the loss of echinocandin resistance. In patients infected with anidulafungin-resistant strains with mutations in only one *FKS* gene (ideally *FKS2*) and no *ERG3* alterations, switching the therapy to a combination of azoles and *CNE1* inhibitors may generate loss of echinocandin resistance (due to evolutionary constraints that are similar to the AinF strains discussed above). This strategy, while dependent on strain mutational background, may yield echinocandins viable again for these patients. Such a clinical pipeline is highly-speculative and futuristic, since it will require i) validation that the such *CNE1-FKS-ERG3* interactions are not only a coincidence of two strains, ii) development of viable nucleic acid or drug-based *CNE1* inhibitors, iii) testing of the proposed mutational steps and iv) molecular diagnostic tools that can accurately pinpoint resistance mutations. All in all, I propose a strategy, requiring further validation, to modulate resistance in *C. glabrata.* These new ideas illustrate how the study of fundamental resistance mechanisms may aid the design of new clinical guidelines in the context of precision medicine.

In addition to the resistance mechanisms described, this study is relevant because we generated a valuable resource: our collection with hundreds of drug-evolved populations spanning *C. glabrata*'s intraspecific diversity, including intermediate timepoints. These samples may open future research avenues. For instance, a recent study used our anidulafungin-adapted strains to validate a cellular assay that predicts echinocandin susceptibility (*328*). Similarly, our collection was used to validate that chemically-modified echinocandins can be effective against anidulafungin-resistant strains (*329*). In addition, further studies restarting the evolutionary process from intermediate timepoints may illuminate the predictability of some evolutionary outcomes observed. For example, we observe that, contrary to the norm, a few anidulafungin-adapted (ANI) strains with *ERG3* mutations (and fluconazole cross-resistance) do not acquire new *PDR1*/*ERG11* variants when evolved in fluconazole (AinF). This could be because i) *ERG3* mutations in ANI highly reduce fluconazole susceptibility so that new *PDR1* variants are not adaptive in AinF or ii) random genetic drift effects yielding an AinF clone that lacks these *PDR1* mutations. By restarting the evolutionary process from these ANI strains we may distinguish between the two, further illuminating the adaptive path. Finally, our collection is currently being used to screen for drugs effective on multidrug resistant (MDR) strains. This is a promising avenue to find compounds that have novel mechanisms of action or synergistic effects with current antifungals. In summary, our collection is a valuable resource to i) pursue follow-up questions about the adaptive process and ii) improve current diagnostics and therapies for *C. glabrata*.

One of the main limitations of our approach is that we only analyzed clones at the end of the evolution experiment, which leaves many open questions regarding i) the evolutionary dynamics of the process and ii) the mutational steps leading to antifungal adaptation. To address these, a follow up project may be to sequence the intermediate populations (as in (*300*)) of our collection to understand these dynamics, and dissect the order of appearance of resistance mutations. Such an analysis may shed light on some of the speculative mechanisms discussed above. For instance, if *FKS* truncations are only adaptive with prior resistance-conferring *FKS* variants, we'd expect the truncations to appear later in the experiment. In addition, understanding whether *ERG3* mutations predate *ERG4* changes may clarify the compensatory nature of these variants. Similarly, the relationships between changes in *ERG11*/chromosome E and *PDR1* remain elusive, which could be clarified by analyzing the order of appearance. Furthermore, such a longitudinal, whole-population analysis may elucidate whether the three AinF clones with no new *PDR1/ERG11* mutations are exceptional cases or representatives of the population. Finally, such analyses may clarify if the evolutionary dynamics of antifungal adaptation are simple (i.e. involving sequential fixation of adaptive alleles) or complex (i.e. following patterns of clonal interference (*330*)). In fact, we are currently analyzing the populations of ANI and AinF samples precisely to answer such questions. In summary, studying the intermediate populations of our experiments may further clarify some of the questions raised in our study.

In addition to the open questions that can be directly inferred through such population sequencing, various speculative scenarios discussed above, related to the fundamental properties of antifungal adaptation, may be investigated through further experiments. This will be necessary to fully dissect how such evolutionary mechanisms impact cellular functions. First, the proposed selective advantages of mutations generating novel mechanisms (*FKS* truncations and variants in *ERG11*, *CDR1*, *CNE1*, *ERG4*) may be validated by re-introducing these variants in isolation and testing the resulting fitness in each drug, as we did for the *ERG3* mutations. Second, the precise molecular mechanisms explaining the (sometimes epistatic) effects in *FKS1/2*, *CNE1* and *ERG3*, in relation to the changes in anidulafungin and fluconazole susceptibility, remain poorly understood. This may be further clarified by analyzing, in strains that have various combinations of these mutations, changes in i) sterol composition, ii) gene expression, iii) drug susceptibility, iv) cell wall composition and/or v) CWI and ER stress response signaling. These strains may be amongst our collection and/or artificially generated (such as our *ERG3* CRISPR-Cas9 constructs). Third, a limitation of our study is that the set of analyzed clades is only a subset of all currently-described clades in *C. glabrata* (see 3.4), indicating that future studies may be needed to fully ensure that the mechanisms found here are fully generalizable. I consider this as a minor limitation, since the actual intraspecific SNP diversity covered by this subset of clades (measured in (*156*)) is comparable to the diversity found in the broader collection of *C. glabrata* strains (see 3.4). In summary, further studies are needed to clarify some of the mechanisms of resistance proposed here.

Similarly, to fully comprehend antifungal drug adaptation, further studies should analyze the impact of varying experimental designs. On the one hand, it is unclear whether the mechanisms inferred here, based on fluconazole and anidulafungin, are applicable to other azoles and echinocandins. Various lines of evidence suggest that there can be differences in adaptation mechanisms across drugs of the same type, particularly for azoles. For instance, a recent *in vitro* study in *C. glabrata* found that variants in *CgHxt4/6/7* hexose transporters are a main driver of posaconazole resistance (*92*), which contrasts with our results in fluconazole. In addition, we see that fluconazole and voriconazole susceptibility profiles are not always equivalent in a subset of our strains. Finally, our GWAS study suggests that some of the canonical genes (i.e. *ERG11* in *C. auris* or *CDR1* in *C. glabrata*) are unrelated to resistance towards some of the tested azoles (see 3.4). These observations underscore the need to study different compounds, without always assuming that one drug is representative of its class.

On the other hand, the drug concentration may occasionally affect the type of adaptive mutations acquired, as shown in a recent *in vitro* evolution study in *C. albicans* (*331*). Thus, some of the resistance mechanisms described here may only reflect our particular experimental design, and it could be relevant to study the effects of changing the concentrations. I consider this a minor concern, since i) our evolution experiment was actually carried on multiple concentrations and ii) the mechanisms inferred mostly overlap previous

findings in *Candida* species and our GWAS results (see 3.4). Still, additional sequencing of intermediate populations, each evolved at different concentrations, may be relevant to fully understand whether such concentration-dependent mechanisms exist. Conversely, it may be also interesting to study whether there is a difference between using a single fixed concentration vs increasing sequentially the concentrations (as we did). For instance, the fact that *CgHxt4/6/7* variants play a role in posaconazole adaptation (*92*), but not in our fluconazole experiments, could also be because the posaconazole evolution was performed with a different regime (a single fixed concentration). All in all, to fully comprehend antifungal adaptation, it may be interesting to study adaptation to different concentration regimes.

In addition, to translate our findings into the clinics various aspects of the drug adaptation process require further research. On the one hand, although the fitness tradeoffs of resistance appear to be minimal, we studied them in rich media, which may not represent (more harsh) host conditions. To clarify these tradeoffs, we may need further experiments assessing i) virulence in *G. mellonella* and mice and/or ii) fitness under different stressors (i.e. osmotic, oxidative or reductive). Such experiments were performed for a small fraction of our strains (two FLZ strains and the *ERG3* mutants), and suggest that the fitness/virulence tradeoffs are actually limited, but further large-scale analyses are required. For instance, we find that some *ERG3* mutants have lower tolerance to membrane and oxidative stressors, indicating that such tradeoffs may be relevant, deserving further attention. A comprehensive understanding of these tradeoffs may enable future coadjuvant-based therapies that target specific fitness / virulence vulnerabilities of resistant strains.

On the other hand, the stability of the antifungal resistance in the absence of the selecting drug remains understudied. This is a relevant question to understand the effects of sequential therapies, which may allow modulating resistance as proposed above for echinocandin-adapted strains. The FinA and AinF analyses suggest that resistance is remarkably stable, but further experiments with other therapies or no drug treatment are needed to fully understand stability. For instance, the loss of resistance resulting from *PDR1* truncation in one FinA strain may be an 'evolutionary trap' (*332*) in terms of azole adaptation, as this lineage may need a reversion of this truncation (likely less probable than typical *PDR1* GoF mutations) to regain resistance. By studying such mechanisms of stability we may be able to find these evolutionary traps, which could be exploited to artificially modulate resistance in the clinics. All in all, further studies of fitness tradeoffs and resistance stability may enable better therapeutic guidelines and/or guide the development of novel drugs.

Our results may also contribute to developing further diagnostic applications for antifungal susceptibility profiling, but various challenges remain. Although the catalog of mutations presented here is potentially useful, it is likely still insufficient to trivially predict resistance from sequence, which is essential for

diagnostic applications. We infer that the set of mutations that may confer resistance is highly diverse, only partially covered by our study and previous surveys. This means that a trivial strategy of identifying particular mutations in a sample to infer resistance may not be yet possible, particularly if the goal is to predict the emergence of new resistant strains through *de novo* mutations. Even machine learning-based applications predicting resistance based on the presence/absence of certain variants, as done for *C. auris* (*333*), may not be powerful enough to do such predictions in *C. glabrata*, as there may be several unaccounted variants with a potential to drive resistance.

An alternative would be to predict resistance based on the functional effects of these variants, and not only their presence/absence patterns. For instance, a putative solution would be a machine learning classifier trained on variant functional annotations such as the affected regions, the type of variant (truncating vs non-truncating) and/or the predicted effect on protein stability. This would be conceptually similar to previous work in anticancer drug resistance, where the effects of resistance mutations on protein structure have been used to make sequence-based predictions (*334*). Such a model may deal better with new, unaccounted mutations. Another option would be to predict the structure of the drug target enzyme based on sequence (with any variants found), and then infer computationally the drug-enzyme binding affinity as a proxy for susceptibility. This may be particularly useful for echinocandins, where mutations in the target are likely essential for resistance. Finally, gathering additional data, by sequencing more strains or with techniques like deep mutational scans (*335*), may enrich our understanding about the genotype-phenotype landscape of antifungal resistance, perhaps increasing the power of such prediction tools. However, such applications will require extensive testing, and its implementation may be challenging due to i) the quantitative nature of antifungal susceptibility limiting the power of discrete classification and ii) epistatic interactions across genes (i.e. between *FKS/ERG3* or *FKS1/FKS2* variants) complicating such predictions. All in all, although our work provides key data for future diagnostic tools, further efforts are needed to accurately predict antifungal drug resistance from sequence.

Admittedly, due to its *in vitro* nature, our study has some theoretical limitations that deserve further discussion. As reviewed above (see 1.2.2), experimental evolution may not recapitulate entirely the drug adaptation process in the host, leaving open the question of how applicable is the knowledge gained here to directly infer clinical resistance mutations. For instance, the fitness and virulence trade offs influencing drug adaptation in the clinics may not be equal in our experiment, potentially affecting some of the resistance mechanisms inferred. In addition, given the proposed importance of drug concentration on adaptive mechanisms (*331*), the fact that our drug regime may be different between our experiments and clinical treatment could lead to differences in the genes affected. These limitations imply that some of the resistance mechanisms inferred *in vitro* may not operate in the same exact way in the clinics. For instance, our results about *ERG11*, which contrast with previous clinical studies (and also our GWAS results, see 3.4),

may be explained by such differences. These considerations show why, if enough data is available, it is important to also study clinical resistance directly as we did in our GWAS analysis (see 3.4). Most notably, this may be essential to validate whether *ERG3* mutations also mediate cross-resistance between anidulafungin and fluconazole in the clinics. Alternatively, further studies of virulence of the adapted strains, as we did for some strains in *G. mellonella*, may be useful to predict whether all mechanisms inferred here are also implicated in clinical drug adaptation. I consider that this is likely the case, as our main findings are overly consistent with described mechanisms of clinical resistance (from other studies and our GWAS, see 3.4). However, this limitation may still be relevant for some of our inferred mechanisms.

Despite these considerations that deserve further attention, I consider that our *in vitro* setting is overly a valid model to understand clinical resistance due to various reasons. On the one hand, given the limitations of currently available clinical data (see 3.4), such a model is likely the only way to study some aspects of drug adaptation. First, it allowed us to understand the effects of specific drug regimes, which is not possible with current clinical data where treatment information is mostly unavailable (see 3.4). Second, we could explore echinocandin adaptation in *C. glabrata*, which would not be feasible on clinical isolates due to insufficient sampling (see 3.4). Third, our *in vitro* model enables a much cleaner dissection of causal mutations as compared to clinical studies, as the conditions are more controlled and the evolutionary distances between strains are much lower (reviewed in 1.2.2).

On the other hand, even as compared to clinical studies that do not have such limitations (i.e. analyses of echinocandin adaptation in close serial isolates considering treatment regimes), *in vitro* evolution has many advantages. First, it enables the study of larger, less noisy collections of strains, as clinical studies face constraints like i) bioethical considerations, ii) the need to pool patient data from multiple hospitals and/or iii) non-uniform treatments across patients. Second, a good understanding of drug adaptation (also in the clinics) requires to not only study the resistance mutations, but also their fitness tradeoffs. This may be essential to actually target such tradeoffs and improve current therapies (discussed above), and I consider that our *in vitro* setting is better to understand such compromises than studies of clinical resistance. I consider that variants appearing during clinical drug adaptation likely have reduced fitness tradeoffs, enabling resistant strains to be viable in harsh host conditions. Thus, studying only clinical resistance mechanisms may be insufficient to understand such constraints. In the next section (see 4.4) I discuss how this may be relevant to modulate azole susceptibility, by targeting such vulnerabilities in strains with resistance attributable to *ERG11* changes. In summary, I consider that the limitations of our *in vitro* setting are mostly unavoidable consequences of its benefits as a model to study clinical antifungal resistance.

All in all, despite some limitations and open questions, our study improves our understanding about the mechanisms of resistance and MDR in *C. glabrata*, providing lessons that are also relevant for other *Candida* pathogens.

4.4. Public sequences illuminate the genomic signs of recent selection and antifungal drug resistance

Beyond the knowledge gained from artificial techniques like *in vitro* evolution, the study of genomic variation in *Candida* populations is essential to fully understand recent adaptation in its natural niche (reviewed in 1.2.1). For instance, various studies have analyzed how genetic variants across clinical isolates contribute to antifungal resistance, sometimes using techniques like GWAS (*155*, *207*). In addition, the genomic signs of selection have been used to infer processes of recent adaptation, potentially changing drug susceptibility, virulence or human transmissibility (*156*, *212*). However, as reviewed in 1.2.5, our current knowledge is limited due to i) exclusive focus on a subset of genes, ii) incomplete characterization of SVs, iii) small-sized studies, iv) lack of multi-species analyses, v) inadequate GWAS tools and/or vi) focus on ancient selective pressures, sometimes unrelated to clinical adaptation. To address these gaps we generated and analyzed a dataset of variants and phenotypes for ~2,000 public genomes of (mostly) clinical isolates, including *C. glabrata*, *C. auris*, *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. orthopsilosis* (see 3.4). Our results provide novel insights and resources that clarify processes of recent selection and mechanisms of drug resistance across major *Candida* pathogens.

Our population genomic dataset is significant in the field of *Candida* pathogens due to various reasons. First, from the perspective of the individual species it is particularly comprehensive because we combined all available data from the SRA, in contrast with most previous studies that analyzed only one (typically smaller) collection (*156*, *158*). By using such a comprehensive dataset, we likely ensure that the inferred mechanisms of adaptation are as general as possible, and not only related to a particular collection that could be biased towards certain geographic locations and/or years. In addition, by maximizing the sample size we likely increased our power to detect genome-wide signatures of selection and drug resistance variants. These considerations are particularly relevant in *C. parapsilosis*, *C. auris*, *C. glabrata* and *C. albicans*, where we analyzed more than twice more strains than previous similar studies (*155*, *158*, *225*, *336*).

Second, the multispecies nature of our dataset / analysis makes it relevant in the field of *Candida* genomics, as most studies focused on a single species with particular methods that do not allow for rigorous comparisons. By analyzing different species with the same methods we could infer similarities in genetic diversity, recent selection and drug resistance mechanisms across species. Such an understanding was key to infer shared cellular functions involved in recent adaptation, such as the changes in adhesion and *NRG1/2* transcriptional regulators, that could be useful to design future broad-spectrum antifungals. Third, our mining of phenotypic metadata (isolation country, date, type of strain, source study and antifungal

susceptibility) enabled our comprehensive selection and drug resistance analyses. This is relevant because there is still no automatic way to obtain such information, particularly the type of strain and drug susceptibility profiles, without the manual curation of tens of studies that we performed. Although further efforts should enable the automatic linking of sequences to phenotypes, our approach underscores the importance of careful curation of the literature to perform population genomic analyses and GWAS. In summary, our dataset is relevant due to i) improved sample size, ii) consideration of multiple species and iii) availability of carefully curated phenotypes. This illustrates the value of depositing data in common repositories that can be used for further integrative research.

Our analysis provides new insights into the intraspecific diversification of *Candida* pathogens. As expected, most species (except *C. parapsilosis*) have a high SNP diversity which is consistent with previous studies (*95*, *156–158*, *158*), likely reflecting the hybrid nature of some species and/or the ancient pre-human diversification of *Candida* pathogens. In *C. glabrata* and *C. albicans* we find several new clades as compared to previous similar studies, suggesting that our collection not only has more strains, but also better represents the population structure of these pathogens. Nevertheless, all newly identified clades have closely-related known clades, suggesting that either i) the current sampling covers the breadth of genetic diversity of the species or ii) more deeply divergent clades may remain undiscovered. In contrast, the additional *C. auris* isolates analyzed here mostly belong to previously-defined clades (*155*), indicating that we do not uncover new diversity for this species. In addition, our systematic clade-definition method is relevant due to various reasons. First, it enables automatic clade inference based on rational criteria that are the same across different species, in contrast with typical arbitrary definitions based on manual tree inspection (*95*, *156–158*, *158*), which are not suited for comparisons across taxa. Second, the automatically-inferred clades are mostly consistent with previous accepted clade definitions, suggesting that our method yields meaningful results. In conclusion, our study provides a higher resolution about the intraspecific diversification of these pathogens, and also shows the suitability of systematic clade definition.

This study also highlights the quantitative importance of SVs and coverage-derived CNVs on intraspecific diversification, recent selection and drug resistance. Our results show that all of the identifiable complex SVs (deletions, tandem duplications, inversions, insertions and translocations) are present in the studied populations. Many such variants would be missed by more common coverage-based CNV calling (as discussed in 4.3), which justifies the need for a tool like perSVade (see 3.2) to study these complex variants in *Candida*. The importance of such SVs was expected given that previous work found such complex rearrangements in various *Candida* species (*161*, *162*). However, our SV analysis is novel because it can detect nine types of SVs, including small ones (50-100 bp), in contrast to previous studies that focused on larger, less varied variant types (i.e. >5,000 bp (*161*)). Such a detailed characterization enabled a better understanding of the SVs present in a population. For example, we found that some types of SVs (cut-paste

insertions and inverted translocations) are only present in *C. albicans* and *C. tropicalis*, suggesting some species-specific mechanisms of SV formation.

On another line, we find that unclassified breakpoints and coverage-derived CNVs (not overlapping equivalent SVs, see 3.2) are frequent, suggesting that our SV survey is still incomplete. As discussed above (see 4.1), this could be due to i) unaccounted types of SVs or ii) combinations of rearrangements in a region, which complicates breakpoint summarization. This shows the importance of analyzing coverage-derived CNVs in addition to SVs, which is facilitated by perSVade's multi-variant calling capacities. In summary, our results further confirm the importance of SVs and CNVs for the intraspecific diversification of these pathogens, suggesting that they should not be overlooked. Furthermore, from a functional perspective, we found that such variants i) contribute to a significant fraction of the intraspecific diversity, ii) are likely affected by recent selection and iii) contribute to antifungal drug resistance. These findings further illustrate the importance of perSVade (see 3.2) to analyze *Candida* genomes.

Despite the relevance of our SV/CNV analysis, it has some limitations in the context of this study that deserve further attention. A first limitation is that unclassified breakpoints are common, which means that we are likely missing important variants. Given the large diversity of these populations, I speculate that a major driver for this are the multiple rearrangements present in many strains, as compared to the reference (i.e. see EF1620_7B_ANI discussed in 4.1 and 4.3), which complicate breakpoint summarization. This is consistent with the observation that reference genome choice impacts the results of population genomic studies (*337*). We partially addressed this limitation by considering coverage-based CNVs and also the transcript-breaking effects of such unclassified breakpoints, but this solution may be improved. For instance, one way to improve our analysis (without the need to manually validate each variant as proposed in 4.1) would be to use multiple reference genomes (i.e. one per clade), reducing the divergence between each strain and its reference. This is not yet possible, as it will require the generation of high-quality clade-specific genome assemblies, as done in (*161*).

In addition, the fact that variant zygosity is difficult to infer in SVs (as discussed in https://github.com/PapenfussLab/gridss/issues/234) represents a potential limitation. As an example of an SV with complex zygosity, in our *in vitro* study (see 3.3) we find a strain (CBS138_9F_FLZ) that acquired an unbalanced translocation where the right arm of chromosome E ($chrE_R$) got duplicated and fused to the right arm of chromosome J ($chrJ_R$). Since this strain has both chrE and the fused $chrE_R$-$chrJ_R$, the breakend in chrE is heterozygous, while the breakend in chrJ is homozygous. Such complexity for genotyping SVs, where a given variant may have both homozygous and heterozygous breakends, shows how our understanding of these rearrangements, and possibly the functional inferences made on them, remain limited. Further studies, perhaps using long reads or other SV callers that perform such genotyping, may address this.

Finally, although the small variants can be found in CandidaMine (https://candidamine.org), the generated SV/CNV datasets are currently not available due to the non-trivial integration of these variants into the database structure. We are currently working to include them in future releases, which will improve the reproducibility of our results and enable further re-analyses of our dataset. In conclusion our results related to SVs/CNVs have some limitations derived from the inherent complexity of these variants.

Similarly, our coverage-based CNV analyses have some limitations that deserve further discussion. The fact that the CNV diversity patterns are different from the SNP/INDEL/SV patterns is a puzzling observation. Although we propose that this is mostly due to imprecise boundary definitions, implying that considering the CNV's functional effects (i.e. deleted/duplicated genes) is a good solution (see 3.4 and 4.2), there may be other more concerning factors. For instance, I propose that an additional cause could be that different samples have varying levels of the 'smiley-pattern' bias, perhaps due to differences in DNA preparation as previously suggested (*244*). PerSVade's coverage correction should be able to solve this because it is tailored to each sample, but it may not be 100% accurate, particularly in samples with multiple rearrangements that complicate calculation of the distance to the telomere (discussed in 4.2). This is why we applied additional coverage threshold filters to the called CNVs, and we gave these variants the lowest priority in the redundancy reduction of GWAS hits. However, these represent partial solutions, and future studies may improve CNV analyses by either i) improving the automatic 'smiley-pattern' correction in perSVade by using genome graphs (discussed above), ii) working with various reference genomes that minimize the number of rearrangements or iii) improving DNA extraction and library preparation protocols to avoid the 'smiley-pattern' bias. All in all, despite the relevant insights provided by our CNV analyses, it has some limitations derived from the noisy nature of coverage data.

Beyond these insights into the intraspecific diversity and SV relevance in *Candida* pathogens, our study illuminates the genomic signatures of recent selection, which hint to important human-related adaptive processes. By combining Ancestral State Reconstruction (ASR) of each variant with statistical modeling, we developed novel methods to identify genes that recently acquired an excess of functional variants, a hallmark of selection (*212*). On the one hand, we detected genes with an excess of nonsynonymous SNPs (nsyn_SNPs) as those with $\pi_N > \pi_S$ recurrently in multiple strain clusters, a method inspired by previous standard analyses of π_N/π_S-like metrics (*156*, *212*). Such genes have the classical hallmarks of positive selection. Our approach is novel because it detects selection from recently-appeared variants in a given gene, by using an empirical model of neutral evolution specifically tailored for such recent variants. This was essential to understand recent, SNP-related selection because standard π_N/π_S-based analyses cannot work with such low variant counts (further discussed below).

On the other hand, for other types of variants (in-frame INDELs (if_INDELs), gene duplications (DUPs) or truncations (DELs)) we pinpointed genes that frequently acquired such recent mutations across multiple strain clusters. Although this is a less standard sign of selection because frequent variants could reflect higher intrinsic mutation rates (*338*), we consider it relevant because there is no trivial way to define 'neutral' variants for such if_INDELs, DUPs and DELs. In addition, frequent CNVs and SVs in some genes have been interpreted as a sign of adaptation in *Candida* (*161*, *207*) and rice plants (*339*) (among others), justifying our approach. The novelty of our method in *Candida* pathogens relies on the systematic, integrated and genome-wide analysis, not only focusing on a few cherry-picked genes. In summary, we identify genes potentially affected by recent selection on different variant types. This represents a major improvement vs previous studies due to our consideration of i) only recent variants, which ensures that the selection signs are potentially clinically-relevant, ii) various variant types and not only SNPs and iii) genome-wide signs of selection, not only focused on certain genes. Beyond our analyses, these methods may be applicable to future population genomic studies in fungi and beyond.

More specifically, our novel statistical method to detect genes with a significant excess of nsyn_SNPs deserves additional justification (see the Online Methods of 3.4). Given our focus on the few recent variants that appeared within clusters of clonal strains, we considered that we had insufficient mutations to infer selection based only on raw π_N/π_S or dN/dS values, as commonly done (*156*, *225*). As synonymous SNPs are the least common, strains with some adaptive nonsynonymous variants ($\pi_N > 0$) may have a $\pi_S = 0$, which does not allow for π_N/π_S calculations. In addition, even in strains with some synonymous SNP, the low variant counts would likely result in inaccurate π_N/π_S calculations due to single variants dramatically changing the ratio. Thus, we reasoned that we lacked resolution to detect selection for a given gene in each strain, as previously done when considering all (not only recent) variants (*225*). In addition, given the inaccurate nature of such π_N/π_S values, we also considered that measuring the average π_N/π_S for a given gene across all strains (as done in (*156*)) may not be appropriate for our purposes. These constraints justified the need for a better method to detect recent selection.

To solve this we used alternative metrics and an empirical statistical method to pinpoint genes with an excess of recurrent nsyn_SNPs. To avoid problems with solely relying on π_N/π_S calculations, but still capture average selective pressures, we calculated, for each gene, a selection score (*S*) that takes into account the fraction of clusters and strains with $\pi_N > \pi_S$. In addition, as *S* may also be inflated by low (or 0) π_S values, we only defined as 'genes under selection' those that had an *S* that is higher than expected under an empirical model of neutral evolution. Such neutral models were tailored to each gene based on synonymous mutation rates inferred from all variants, which likely ensured that genes with high *S* due to

increased mutation rates (not due to selection) were filtered out. We validated this approach by showing that real π_S values are likely under this neutral model, except in a few outlier genes that were discarded. This illustrates the suitability of empirical strategies to infer recent selection, as these model neutral evolution in a way that is tailored to the gene and population of interest, without relying on the assumptions of similar parametric techniques. In summary, we developed a novel algorithm to detect genes with an excess of recent nonsynonymous SNPs, which was useful to detect signs of recent selection and has also potential beyond this project.

In addition, the way in which we measured the selection score per gene (*S*), based on a harmonic mean that considers both the number of strains and clonal clusters with signs of selection, provides various advantages. *S* can only be high if a relevant number of strains from multiple clusters acquired an excess of variants, so that it reflects selective pressures that are convergently present in multiple parts of the phylogeny. This focus on convergent signatures is analogous to i) our approach of considering recurrently-mutated genes in the *in vitro* study (see 3.3) and ii) the assumptions behind convergence-based GWAS methods (*192*, *201*). It ensures that inferred selection signatures are (likely) related to general adaptive processes, and not the result of cluster-specific effects.

This consideration of only convergent selective pressures has its own limitations because we cannot detect evolutionarily-relevant cluster-specific processes, such as the fact that chromosome I aneuploidies may only appear in EB0911 due to strain-specific trade offs (discussed in 4.3). However, I consider that focusing on general convergent adaptive processes is appropriate because cluster-specific signals may also come from confounding factors, unrelated to selection. For instance, particular historical events within each cluster, such as population bottlenecks or founder effects, may generate patterns of variants that inflate the functional variant counts in some genes in a way that is unrelated to selection. By considering the integrated signatures from various clusters we likely reduce such confounding effects, as it is unlikely that genes with inflated variant counts due to such random effects are shared across multiple clusters. In addition, taking into account convergent signals is suitable given the (likely) non-uniform nature of the analyzed collection. Despite our efforts to generate a comprehensive collection covering multiple continents, our isolates are likely biased towards countries that perform more sequencing, and not necessarily a random sample representing natural *Candida* populations. For instance, there is a clade in *C. albicans* that includes almost half of the strains (see 3.4), which may represent such a bias. This is an inevitable limitation of population genomic datasets from such globally-distributed pathogens (*155*, *158*, *340*), but it is relevant to such selection inferences. A simple approach of measuring average selection signals across strains (as done in (*156*)) could lead to confounding effects in our analysis because strains within overrepresented clades would contribute more to these average metrics. However, by using a harmonic mean measurement (*S*) that weights the fraction of clusters affected we likely minimize these

confounding effects of overrepresented clades. All in all, we used selection measurements that likely capture general adaptive processes, while limiting the misleading effects of i) complex population histories and ii) biases in sampling.

By using these methods we could define, for each species, a catalog including hundreds of genes with signs of recent selection or excess of recurrent variants (for non-SNP variants). The fact that so many processes are under recent selection suggests that clinical isolates are not fully adapted to human-related environments. This is consistent with the hypothesis that *Candida* pathogens often switch between the human host and other environments (*73*, *132*, *341*). Further sequencing of environmental isolates, which are currently limited, may allow a better exploration of the selective constraints of this switch. In addition, the large number of functions that are under selection illustrates the biologically-relevant diversity of *Candida* populations, underscoring the need to study various strains of each species to draw general conclusions (as we did in our *in vitro* study, see 3.3). From a functional perspective, we find that most of the related orthologous groups (OGs) and affected pathways are not shared across taxa, suggesting mostly species-specific mechanisms of adaptation. Together with the observation that the transcriptomic responses after host interaction of different *Candida* species are mostly non-overlapping (*342*), our findings further support the idea that each of these pathogens has unique virulence and drug resistance mechanisms. This likely reflects the fact that host adaptation in various *Candida* pathogens has evolved independently multiple times across Saccharomycotina (*21*). Thus, it may be important to apply species-specific therapies and diagnostics for *Candida* infections, as previously proposed (*47*, *104*). This also highlights the relevance of a precision medicine approach for infectious diseases. In summary, we provide an overview about the recent selective pressures suffered by *Candida* pathogens, which hints to mechanisms of host adaptation and drug resistance.

These species-specific processes may be used to infer mechanisms of host adaptation and/or drug resistance that are particularly important in each of the *Candida* pathogens. Thus, in the paragraphs below I propose speculative scenarios explaining why changes in these species-specific functions could be adaptive (described in 3.4). For instance, in *C. glabrata* the changes in filamentous growth could reflect a modulation of pseudohyphae formation, which may promote virulence by increasing macrophage escape and host invasion, as previously proposed (*147*, *343*). In addition, the duplications in tRNA methylation genes may be adaptive because they provide general stress tolerance, given that upregulation of such genes has been observed as a response to various stresses in this species (i.e. cell wall, heat and oxidative stresses, faced in the host) (*344*). Conversely, the changes in actin nucleation in this species could reflect adaptations to azoles, as these drugs induce actin cytoskeleton remodeling (*345*). Similarly, the alterations in bud site selection, which is also related to the actin cytoskeleton (*346*), could also reflect such antifungal adaptation. In addition, the changes in (cAMP-related) G protein-coupled receptor signaling may also represent

adaptations to azoles, given that cAMP-related pathways articulate the physiological response to this antifungal (*347*). In summary, I propose various host and drug adaptation mechanisms that are unique in *C. glabrata*, including the changes in i) pseudohyphae formation, ii) general stress responses, iii) actin cytoskeleton and iv) cAMP signaling. Such a large number of distinct functions are consistent with the large distance between this species and the other CTG-clade *Candida* species (**Figure 1B**).

However, we also find relevant species-specific functions under selection within the CTG taxa, suggesting that some of the important host adaptation / drug resistance mechanisms also differ between these closer species. For instance, in *C. auris* the changes in rRNA binding proteins may be an adaptation to echinocandins, as the transcriptomic response to caspofungin involves upregulation of ribosomal genes (*348*). In addition, the changes in the TTT complex may reflect the importance of phosphatidylinositol 3-kinase signaling for adhesion and filamentation in this species, as previously proposed in *C. albicans* (*349*). Furthermore, the selective processes affecting genes from the host cellular component in *C. auris* may be related to the lower ability of phagocytic cells to engulf and kill this pathogen, as compared to *C. albicans* (*350*). Conversely, in *C. albicans*, the changes in temperature responses may enhance host survival and virulence during invasive infections that trigger fever (*351*). Conversely, the changes in inositol phosphate dephosphorylation may be related to changes in virulence through altered filamentation, as inositol phosphate changes are related to this process (*352*). Finally, the changes in the cell wall could be related to adaptation to cell wall stressors, such as echinocandins (discussed above).

On another line, in *C. tropicalis* the changes in histone deacetylases (HDACs) are consistent with the role of these proteins in modulating virulence, biofilm formation and host dissemination in *Candida* species (*353*). In addition, changes in oligosaccharide synthesis, glucan degradation and protein mannosylation may reflect the importance of protein glycosylation for virulence, cell wall integrity, interactions with the immune system and/or hyphal growth (*354*). Finally, the changes in carbohydrate metabolism and transport in *C. orthopsilosis* may reflect the importance of energy obtention processes and/or sugar-related signaling (*355*) in this species. Taken together, these speculative scenarios suggest multiple host and drug adaptation mechanisms that are unique in *C. auris*, *C. albicans*, *C. tropicalis* or *C.orthopsilosis*. These involve changes in i) ribosomal regulation, ii) inositol signaling, iii) interaction with phagocytes, iv) temperature responses, v) cell wall integrity, vi) HDAC-mediated regulation, vii) protein glycosylation and ix) carbohydrate metabolism. Such differences in recent adaptation reveal the unique selective constraints of each *Candida* pathogen, which hints to relevant species-specific cell functions and/or therapeutic targets. These highly speculative ideas open many interesting research questions that may be pursued through further experimentation.

Beyond these species-specific adaptations, we also find some convergently-affected functions in multiple species. On the one hand, we analyzed the overlaps in OGs altered in various pathogens. Our analysis is

relevant from a methodological perspective, as we used a resampling method to ensure that the number of overlapping OGs is higher than expected by chance. This further shows the suitability of empirical modeling to identify such significant biological signals. More importantly, the analysis of shared OGs reveals interesting convergently-altered processes. As expected, these involve functions related to clinically-relevant adaptation, such as adhesion, pseudohyphal growth, drug efflux, fluconazole resistance, yeast-to-hyphae transitions and/or white-opaque phenotypic switching (related to parasexual cycles (*165*)). However, I consider that this list of shared OGs is novel and relevant because it provides a comprehensive overview, not tailored towards pre-defined hypotheses, about the specific genes involved in convergent adaptive processes in the clinics.

On the other hand, we investigated the functions that are enriched in genes under selection across multiple species. Our analysis revealed that changes in adhesion are at the core of such convergent adaptations, maybe modulating both i) host tissue adherence and ii) biofilm formation through altered *Candida-Candida* interactions. This is consistent with the importance of these functions for colonization, virulence and drug resistance (reviewed in 1.2). Since these processes are mostly enriched in genes with recurrent deletions, I speculate that it is the loss of specific adhesion proteins that often drives such adaptation. This could be due to i) loss of some adhesins increasing adherence and biofilm formation (as previously proposed (*156*)) and/or ii) deletion of adhesins leading to decreased adherence, which may promote more efficient systemic dissemination. Further research should clarify this role of adhesion, as it appears to be a central factor influenced by recent clinical adaptation. Regardless of the precise underlying mechanism, this example further illustrates the importance of i) reductive evolution in budding yeasts (*20*) and ii) considering multiple variants (also SVs), and a tool like perSVade, to study recent adaptation in *Candida*. In summary, our analyses about convergently-altered processes revealed common selective constraints across *Candida* pathogens, hinting to the most important aspects of host-pathogen interactions.

In addition to these novel insights into the mechanisms of recent adaptation, our collection of selection signatures constitutes a valuable resource. It provides information about the implication of different genes in clinical adaptation, which may be used as an empirical hint to understand gene function. This may be significant for *Candida* pathogens because functional inferences often come from (perhaps inaccurate) homology predictions (except maybe in *C. albicans*). Such an unbiased source of information likely opens several research avenues. First, our list of genes may guide further experimental efforts towards the genes and pathways most likely related to clinically-relevant adaptation, such as those with higher selection scores in our collection. For instance, let us imagine a research project trying to dissect the molecular mechanisms governing virulence in *C. tropicalis*. Instead of relying on hypotheses based on better-studied species (i.e. *C. albicans*) or costly genome-wide screens, one option would be to study the effects of disrupting the genes under selection (provided here), some of them likely related to virulence such as the HDACs discussed

above. Second, this functional information may be used to pinpoint clinically-relevant genes in genome-wide studies. For instance, we are currently involved in a project about finding variants that explain why various *C. parapsilosis* isolates induce different cytokine responses *in vitro* (manuscript in preparation). We found various potential genes involved, and the selection scores inferred here were key to narrow down the list of candidate genes, which will be validated experimentally. Third, the information about convergently-affected adaptive processes may aid the design of new pan-*Candida* drugs and/or coadjuvant therapies. In summary, our dataset of selection signatures may aid further research and drug design efforts. Given the highly-speculative nature of our specific biological insights, as we cannot know the phenotypes related to the selective processes, I consider that this dataset actually constitutes one of the most relevant contributions of our work.

Despite these novel insights, our selection inferences have several limitations that deserve some discussion. First, the fact that *Candida* species can sometimes undergo sexual reproduction (*156*, *165*) may bias our results, as it could generate a false impression of convergent, independent signals, which actually come from sexual admixture. For instance, our GWAS results (see 3.4) suggest that drug adaptation (likely driven by selection) may occasionally be related to such sexual recombination, so that this is not a purely theoretical concern. However, the impact of sexual reproduction is likely limited because asexual propagation is the major source of diversification in these species (*160*), supporting the applicability of our approach. Still, further studies may use sweep-detection methods (see 1.2.1) on our data to better understand the impact of sexual reproduction on selection inferences.

Second, our definition of 'recent' variants is arbitrary to some extent, as it is based on a threshold of 1 SNP/kb to define clonal clusters which may leave out of the analysis some variants that are related to diversification in clinical settings. This has no trivial solution, as there is no easy way to establish a threshold that certainly captures all the clinically-associated evolution of these pathogens. This is especially hard due to our ignorance of relevant aspects such as the temporal history of association of these species with humans or their mutation rates. However, to further understand clinically-related selective pressures it could be interesting to study the trajectories of π_N/π_S metrics across the evolutionary history of each species, as done before in *M. tuberculosis* (*214*). Such a study would not completely solve the issue of separating clinical from non-clinical diversification, but it may complement our analyses.

Third, we cannot be sure that the genes with an excess of non-SNP variants (if_INDELs, DELs, DUPs) reflect selection, which opens relevant questions. Such genes may have higher intrinsic mutation rates, which is something that deserves further investigation with techniques like mutation accumulation experiments (*356*). However, even if these signals come solely from higher mutation rates, they likely provide relevant insights as these genes appear to be involved in functions for host and drug adaptation (i.e. adhesion). In

fact, such (potential) differences in mutation rates could also be the result of selection, as keeping genes that are important for host interaction in regions with higher mutation rates (i.e. subtelomeric) may be a mechanism to optimize the adaptive potential of these pathogens. Further studies about such mutation rates combined with multi-species genome comparisons may illuminate these adaptive constraints. In summary, despite the novel insights provided, further research is needed to fully understand recent selection in *Candida* pathogens.

As a specific case of recent adaptation, our study provides novel insights into the mechanisms of clinical antifungal adaptation in major *Candida* species. By using convergence-based GWAS, we detected the variants, genes and pathways associated with drug resistance in *C. auris*, *C. glabrata* and *C. albicans*, which are the species where we could gather sufficient data to perform powerful inferences. In *C. auris* we analyzed these associations for four azoles (fluconazole, itraconazole, posaconazole and voriconazole), two echinocandins (anidulafungin and micafungin) and amphotericin B. Furthermore, in *C. glabrata* we investigated resistance towards three azoles (fluconazole, posaconazole and voriconazole) and micafungin. Finally, in *C. albicans* we analyzed fluconazole resistance.

This represents a large-scale study that is unprecedented in the field of *Candida* genomics due to various reasons. First, we performed genome-wide testing of genotype-phenotype associations in a large collection, which contrasts with typical studies that are either i) based on small sample sizes, ii) only focused on a narrow set of genes and/or iii) missing rigorous statistical tests (reviewed in 1.2.1). This likely ensures increased power to obtain a more complete picture about the mechanisms of clinical resistance in *Candida* pathogens. Second, we analyzed multiple species and drugs in a consistent way, as opposed to most current studies focused on one species/drug (*155*, *207*), enabling systematic comparisons between datasets. Third, in contrast to previous GWAS analyses of single SNPs based on allele counting methods (see 1.2.1), we used a method that considers convergence, variant grouping and multiple mutation types (including SVs and CNVs). I consider these major improvements, necessary to take into account i) the asexual nature of *Candida* pathogens, ii) the potential allelic heterogeneity of the resistance trait (i.e. as seen in our *in vitro* study, see 3.3) and iii) the importance of SVs and CNVs (discussed above). Fourth, we analyzed resistance in clinical isolates, which ensures that the mechanisms inferred are clinically-relevant. Fifth, we provide evidence about resistance mechanisms towards drugs that are understudied, such as amphotericin B in *C. auris*, echinocandins and the non-fluconazole azoles. All in all, we present a comprehensive, unprecedented study of clinical resistance towards compounds of all main antifungal classes (azoles, echinocandins and polyenes) in three major *Candida* pathogens.

More specifically, to illustrate why this large-scale meta-analysis improves our understanding of drug resistance, let us compare it with the approach of reviewing the findings of individual (smaller) studies for a

given antifungal, which is a more common and seemingly straightforward strategy (*104*, *357*). Given their integrative nature, such reviews apparently provide a complete picture about the resistance mechanisms, but they are likely constrained by the statistical power and limited focus of individual studies. This is relevant for *Candida* pathogens because most studies of drug resistance had reduced sample sizes and/or a narrow focus on previously-expected genes. I hypothesize that less frequent (but important) mechanisms, such as *CDR1* variants in *C. glabrata* (see 3.3), may 'fly under the radar' in such smaller studies, and thus they would also not be captured in integrative review studies. A similar limitation may apply to literature reviews discussing the similarities in drug adaptation mechanisms across different species and/or drugs (*104*). Thus, our re-analysis strategy based on pooling together various studies has the potential to increase our detection power, ultimately providing a more complete picture about the resistance mechanisms and the similarities across species and/or drugs. In fact, I consider this to be the primary reason explaining why we find novel, unexpected resistance mechanisms. In addition, this increased power also allowed us to make more confident inferences about the differences in known resistance mechanisms across various drugs and species. In summary, our integrative, multi-species GWAS strategy provided a comprehensive view about antifungal resistance mechanisms, confirming known drivers of drug resistance alongside potentially novel players.

To enable the analysis of our *Candida* datasets we developed a novel convergence GWAS method, which presents unique advantages as compared to similar tools. Our approach is fundamentally equivalent to hogwash's synchronous algorithm, which is similar to the simultaneous score method from treeWAS (*201*), but enabling variant grouping (*192*). In brief, we find variants whose transition (appearance or loss) is correlated with the transition in the resistance phenotype. This ensures that significant hits (variants or groups of variants) represent stringent genotype-phenotype associations, as the underlying variants change with the phenotype. In addition, this method is suitable to address complex allelic heterogeneity because it allows opposing directions of the phenotype and genotype transitions. This may capture instances where different variants in a gene or pathway (analyzed in a grouped manner) have opposite effects in the phenotype, which could be relevant to fully understand antifungal resistance. For example, in our *in vitro* study (see 3.4) we find that point mutations and truncating variants in *FKS* or *PDR1* have opposite effects for echinocandin or azole susceptibility, respectively.

In addition, despite the similarities with hogwash, our approach presents many advantages. First, we improve the versatility of the analysis by considering several ASR methods, association statistics, p value types and p value correction methods. For instance, the calculation of maxT p values is significant as it yields p values that are already corrected for multiple testing in a way that is tailored to each dataset, which explains its suitability for GWAS (*190*). This consideration of multiple statistical approaches allowed us to carefully filter our hits to ensure accurate results. Second, instead of using genotype reshuffling as in

hogwash, we considered phenotype reshuffling to calculate empirical p values, which may yield less biased results with our trees that have highly variable branch length. Third, by parallelizing the code and reducing redundant steps we developed a highly-optimized pipeline, which was necessary to work with our large amount of datasets, grouping techniques and varying parameters. In summary, we developed and used a GWAS tool that was essential to analyze drug resistance in *Candida* pathogens. In addition, we provide the underlying code, so that it may be useful to analyze genotype-phenotype associations beyond this project.

Given that such genome-wide techniques have been underused in *Candida* pathogens, our study provides important novel lessons about how to do GWAS to understand drug resistance in these species. On the one hand, we learn which variant collapsing strategies are most effective. For instance, we find that the consideration of multiple variant types (including SVs and CNVs, not only SNPs) could be key to uncover all relevant associations. In addition, we see that grouping variants at domain and gene levels is essential to capture important signals due to the heterogeneity of mutations affecting the same gene. Conversely, although pathway-level collapsing does not usually add relevant information, it could be useful in smaller datasets, where detecting associations from more narrow gene collapsing may not be possible. Finally, a large fraction of associations are based on specifically collapsing truncating variants, reinforcing the importance of considering gene loss to understand drug resistance. On the other hand, we gain insights into the most suitable GWAS parameters. For instance, using solely maxT p values and filtering based on the convergence level is a suitable strategy in almost all datasets. Conversely, the choice of ASR methods and minimum branch support may be less universal, indicating that these parameters may need to be tailored to each dataset. However, the most common useful strategy is based on using i) a consensus between different ASR methods based on either maximum likelihood or maximum parsimony and ii) requiring a minimum support of 70.

Beyond these methodological considerations, our GWAS analysis improves our understanding about the mechanisms of resistance. Notably, our results allowed us to evaluate the clinical relevance of previously-established mechanisms, which mostly confirm the role of variants in *ERG11*, *TAC1b*, *PDR1* and *FKS* genes (reviewed in 1.2.3). However, our analyses of azole resistance in *C. auris* showed some interesting, less established mechanisms. Although we find that the mechanisms are overly similar across different azoles, there was one relevant unexpected result: *ERG11* variants are related to resistance towards fluconazole and voriconazole in *C. auris*, but they appear to be unrelated to itraconazole or posaconazole resistance. Accordingly, previous studies re-introducing *C. auris ERG11* variants in susceptible strains showed a strong impact on fluconazole / voriconazole resistance, but a minimal or inexistent effect on itraconazole / posaconazole susceptibility (*358*, *359*). From a chemical structure point of view, this could be due to the fact that fluconazole and voriconazole are short-tailed azoles, while itraconazole and posaconazole are long-tailed compounds (*358*, *360*). Thus, our findings confirm that the role of *ERG11*

variants in azole resistance is limited to some, perhaps only short-tailed, azoles in *C. auris*. These results underscore the need of studying resistance towards different drugs, even if they have a presumably similar mechanism of action.

Conversely, we find that *TAC1b* variants are more universally-related to azole resistance in *C. auris*, suggesting that changes in drug efflux (derived from mutations in this transcription factor) are the main azole adaptation mechanism in this species. This is consistent with previous evidence suggesting that, even in fluconazole/voriconazole, the combined effect of *ERG11* and *TAC1b* variants is necessary to reach strong clinically-relevant resistance (*359*). All in all, our findings confirm the role of these known driver genes in clinical antifungal resistance. Although none of these findings are entirely novel, they represent a suitable confirmation given the size of our dataset, the rigorous statistical analysis employed and our focus on naturally-occuring resistance.

As a specific example of expected mechanisms that is relevant to this PhD thesis, we used our GWAS hits to evaluate the clinical relevance of the mechanisms of azole and echinocandin resistance in *C. glabrata* inferred in our *in vitro* study (see 3.3). We find overlapping results for most of the azole-resistance mechanisms, including the changes in *CDR1* and *PDR1*, further indicating that major mechanisms inferred *in vitro* are clinically-relevant. In addition, this analysis clarified the relationship between *EPA13* changes and drug resistance, something that was not fully clear in our *in vitro* study because these mutations only appeared during combined therapy (of both anidulafungin and fluconazole). We find a correlation between *EPA13* variants and resistance to various azoles, suggesting that such mutations were adaptive towards fluconazole in our *in vitro* setting. Still, further research is needed to validate this, as we likely have insufficient data to ensure that such variants are not correlated also to echinocandin resistance.

Conversely, although we could infer the involvement of *FKS* mutations, we could not validate some of the anidulafungin adaptation mechanisms found *in vitro* (related to *ERG3*, *ERG4* and *CNE1*) due to insufficient sampling of echinocandin-resistant strains. Thus, further sequencing of such strains is necessary to fully comprehend clinical adaptation to this drug category. Most importantly, the clinical relevance of our *ERG3*-mediated cross-resistance mechanism should be further investigated. In summary, we could validate most of the *in vitro*-inferred azole resistance mechanisms, supporting the relevance of our GWAS dataset as a resource to explore pre-specified hypotheses (as also discussed for the selection signatures). In addition, this analysis validates the suitability of our *in vitro* study (see 3.3) as a model for clinical resistance due to i) the high overlap with clinical azole resistance mechanisms and ii) the insufficient sampling of clinical echinocandin-resistant strains, which makes *in vitro* studies more necessary for such drugs.

Despite this high overlap regarding azole resistance mechanisms, there is one important exception that deserves discussion. In contrast to our *in vitro* results, and consistently with previous clinical reports (*245*), we do not find an association between *ERG11* variants or chromosome E duplications and azole resistance in our GWAS, even after considering different filterings and/ or manually inspecting such mutations. Together with our *in vitro* results, this suggests that *ERG11* changes are secondary resistance mechanisms in this species, perhaps only providing resistance in certain conditions. One explanation for this could be that *ERG11* changes carry a clinically-relevant fitness trade off, so that they do not appear in our GWAS or previous clinical studies. This partial discrepancy between our studies highlights the importance of combining *in vitro* and clinical analyses to fully understand resistance.

In addition, I consider that this discrepancy does not necessarily represent a limitation of *in vitro* studies, but could actually open exciting novel research and clinical avenues. Investigating this complex relationship between *ERG11* mutations and fitness, virulence and resistance could be relevant to identify vulnerabilities in the antifungal adaptation process, which could perhaps guide the development of new drugs or improved therapeutic guidelines. For instance, let us assume a speculative future scenario where we i) understand the source of this (potential) tradeoff associated with *ERG11* mutations and ii) possess a drug to compensate for this vulnerability and make *ERG11* mutants viable, called CD (for compensatory drug). If the (potential) fitness trade off of *ERG11* mutations can be replicated experimentally (i.e. by growing *ERG11* mutants in a certain stressor), CD could be developed by screening drugs that restore this tradeoff in growth assays. This may allow for the design of an evolutionary trap-based therapy using azoles and CD. First, a combination of azoles and CD may be administered to promote the acquisition of resistance through *ERG11*. Once azole resistance appears, and after validation of the presence of *ERG11* changes, interrupting the therapy may result in cell death of the pathogens and clearance of the infection, due to the reduced fitness of the strains with *ERG11* mutations in the absence of CD. This speculative therapy would need extensive validation, and it could have its own limitations as it may promote the emergence of resistance. However, this idea shows how a good understanding about the full landscape of resistance mutations, and not only those that confer resistance in the clinics, could be relevant. Similarly, it illustrates how an understanding of the genomic mechanisms underlying such fitness tradeoffs could be key to modulate resistance. More broadly, it further supports the suitability of studying both *in vitro* and clinical resistance, as done in this PhD thesis.

In addition to these confirmatory analyses, our results suggest additional, novel mechanisms playing a role in the emergence of drug resistance. On the one hand, we identified novel gene families related to resistance towards multiple drugs in the same species, which allowed us to predict mechanisms of cross-resistance and multi-drug resistance. In *C. glabrata*, the most notable observations are the multiple hits affecting adhesin genes in relation to azole and/or echinocandin resistance. This resembles previous studies showing how changes in the expression of certain adhesins modulate biofilm formation and thus

azole resistance in *C. glabrata* (*234*). In addition, it is consistent with the observation that genetic changes in adhesins are correlated to the acquisition of resistance during *in vitro* evolution, as shown by us (see 3.3) and others (*92*). Taken together, our findings likely reflect the importance of biofilm formation for clinical resistance in *C. glabrata*. Conversely, in *C. auris* the gene families related to resistance towards multiple drugs (azoles and polyenes) appear to be involved in the regulation of gene expression, either at the level of transcription or RNA degradation. This is consistent with the known role of transcription factors for drug resistance in this species (i.e. *MRR1* and *TAC1b*, discussed in 1.2.3), and it suggests that we unearth potential novel regulators of antifungal susceptibility.

On the other hand, we unexpectedly found that variants in *NRG1* and *NRG2* are correlated to fluconazole resistance in *C. glabrata* and *C. auris*, respectively. This is consistent with the fact that the orthologs of these genes in *S. cerevisiae* have been related to azole susceptibility (*361*, *362*). Both are (predicted) transcriptional regulators, putatively controlling pseudohyphal growth, adhesion, biofilm formation and the responses to osmotic stress and glucose sensing. I speculate that its relationship to fluconazole resistance comes from its role in adhesion and biofilm formation, a key process in antifungal resistance (*119*). Accordingly, *S. cerevisiae NRG1/2* deletion mutants had changes in both miconazole susceptibility and biofilm formation (*361*). More broadly, this represents yet another example of how changes in adherence are essential for recent adaptation in *Candida* pathogens. In summary, our GWAS hits reveal various novel drug resistance mechanisms, often related to adhesion and biofilm formation.

Another unexpected relevant result of our GWAS analysis is the observation that (para)sexual reproduction could sometimes underlie drug resistance, at least in *C. auris* and *C. glabrata*. This is a surprising result contrasting with a typical focus on *de novo* mutations as the primary source of drug adaptation in these pathogens (*92*, *139*). Our results suggest that, once a resistance variant appears in the population, it can spread through recombination resulting in patterns that involve the whole genome, which is why we propose a parasexual-like underlying mechanism. Such recombination-based acquisition of resistance may yield a faster spread of resistance mechanisms, as compared to the emergence based on *de novo* mutations, which could explain frequent resistance in these species. This represents a specific example of the selective advantages of parasexual propagation in *Candida* pathogens (*165*, *363*). However, it is a significant finding as it shows the relevance of recombination for a clinically-relevant phenotype, also illustrating why it is necessary to understand sex in *Candida* species (*102*, *156*, *158*) which were traditionally considered asexual organisms.

Conversely, various questions remain open. The biological mechanisms underlying this mating remain elusive in these species, and our results suggest that they should be further clarified to fully understand the emergence of resistance. In addition, our genomic evidence is indirect and only based on a few linked

variants, so further validation may be necessary to clarify the process. This will not be a trivial task because recombination studies in *Candida* pathogens are typically powerful to detect signatures that involve thousands of variants (i.e. mating between strains from different clades (*156*, *158*)), so that more tailored approaches could be needed to better understand our data. In summary we provide novel evidence suggesting an occasional involvement of (para)sexual recombination in the spread of resistance mechanisms of *Candida* pathogens. However, further genomic and experimental efforts should be made to validate these findings, further clarifying i) the underlying biological mechanism and ii) the genomic events leading to these patterns of recombination.

Furthermore, our findings on drug resistance have novel implications for the sequence-based prediction of antifungal resistance. Even with all our new list of associations, the current catalog of resistance mutations is still likely incomplete, so that it may be insufficient to make predictions solely through the trivial strategy of identifying particular mutations (also discussed in 4.3). However, our results provide some insights that could allow (partially) accurate predictions based on more complex strategies. Our findings suggest that the resistance phenotype in a given strain can be attributed to i) *de novo* mutations, ii) recombination with other resistant strains and/or iii) clonal spread of a previously-resistant strain. Given these different scenarios, I propose various future directions that could be used to predict resistance, using a combination of our dataset (with variants and phenotypes) and previous catalogs (i.e. from our *in vitro* study) for training. For instance, a machine learning-based classifier trained on this data could be interesting to capture resistance that comes from recombination and/or clonal propagation. This has been done previously in *C. auris* (*333*), but this study had some limitations. It used randomized cross-validation to define training/testing sets, likely ignoring the (biasing) fact that the phylogenetic structure of *C. auris* is strongly correlated to resistance. Thus, it is likely that the classifiers used captured only the clonal spread of resistance, and not necessarily causal resistance variants. This is a valid strategy, as clonal spread of resistance is likely relevant, but further efforts should be made to also capture resistance derived from recombination. For instance, training these models on some clades and testing them on others could be an interesting approach. Conversely, such predictions would likely miss resistance coming from *de novo* variants, so that they are likely not a universal solution.

To also predict the emergence of resistance based *de novo* mutations we may need alternative strategies, and we speculate about two of them. On the one hand, we may build predictors working with functional variant effects and/or predicted changes in drug-target binding (discussed in 4.3). Such strategies, although challenging to implement, could be useful to profile strains that have resistance due to any of the three scenarios, as they may capture the true determinants of drug resistance. On the other hand, instead of predicting susceptibility directly, we could combine our knowledge about drug resistance genes (DRGs) with phylogenetic reconstruction to predict changes in resistance. To illustrate our pipeline, let us imagine a

newly sequenced strain where we want to infer resistance towards a certain drug. We could use tree reconstruction to find its closest isolate in the training dataset, and determine the genes with variants as compared to this closest strain. If these genes do not include any DRG, we could predict the susceptibility based on the phenotypes of this closest strain. Such a strategy has the potential to capture mostly clonal and recombination-based spread of resistance, but it could also be useful for *de novo* mutations affecting previously-defined DRGs. This would be powerful for drugs in which most of the resistance-related genes are known, which seems a fair assumption for various azoles in *C. auris* and *C. glabrata*. However, it has the downside that, given that not all variants in DRGs necessarily affect susceptibility, the prediction of a change in the resistance will often be uncertain. Thus, this approach may a yield predictions that are either i) accurate and confident (i.e. strain A is susceptible to fluconazole because it has no DRG variants vs strain B, which is also susceptible) or ii) uncertain (i.e. strain A could be resistant to fluconazole, because it has a DRG variant vs strain B). Despite this limitation, such predictions could aid therapeutic choices for *Candida* infections. For instance, if a strain has a prediction of 'no change in fluconazole susceptibility', then fluconazole is likely a suitable option. Conversely, strains with uncertain predictions of a 'change in resistance' may be further profiled with standard growth assays. Note that this contrasts with machine-learning learning tools, which may yield confident but inaccurate results (*333*), potentially leading to therapeutic failure if used in the clinics. In addition, given the availability of fast tree reconstruction methods (*364*), such a strategy may be more cost-effective than machine learning tools, and perhaps it could better capture the evolutionary nature of drug adaptation. In summary, based on our findings and generated datasets, we propose future directions to perform sequence-based prediction of antifungal resistance.

Despite all these valuable insights, our analysis of drug resistance had some general limitations that open future directions. On the one hand, our phenotypic data has some shortcomings. First, the fact that our resistance data comes from multiple experimental setups could introduce some noise in analysis. This is an inevitable consequence of integrating multiple studies, and we addressed this by only considering sharp dichotomous resistance transitions, which likely ensures that we capture relevant susceptibility changes. However, such a discretization of resistance has its own tradeoffs, as different drug susceptibility is likely a quantitative trait. This is relevant because different resistance mechanisms may lead to varying levels of susceptibility, as shown by i) our echinocandin GWAS (3.4) and ii) the differences in fluconazole resistance between ANI and FLZ strains in our *in vitro* study (3.3). Such subtle differences may be missed by our discretized analysis, so further GWAS considering resistance quantitatively could be interesting. Second, our drug susceptibility data is still sparse and incomplete, since not all drugs were assayed on all strains, which limited the power of our GWAS on certain drugs. We addressed this by carefully selecting filters and analyzing high-confidence and low-confidence hits, but these are just partial solutions tailored to our dataset. Further efforts to obtain more phenotypic data of these strains, ideally in a consistent way, may

further increase our power to detect genotype-phenotype associations. This is particularly relevant for echinocandins, polyenes and flucytosine, with insufficient available data in most species. In fact, for these undersampled compounds, artificial techniques like *in vitro* evolution (see 3.3 and 4.3) may currently be better to study drug resistance. In summary, our phenotypic dataset has some limitations that could impact the GWAS results and the proposed diagnostic applications (see previous paragraphs).

On the other hand, various aspects of the analysis could be improved. First, as in our selection analyses (discussed above), the partially-sexual nature of *Candida* species represents a limitation to our convergence-based analyses, as the phylogenetic tree is not a 100% accurate representation of the population structure. This is also consistent with the occasional role of recombination in the spread of resistance. However, it is likely an acceptable limitation because asexual propagation is likely the major source of diversification, overly supporting our approach. Still, further studies may combine both convergence-based and allele-counting GWAS to fully understand the emergence and spread of resistance. Second, to ensure a focus on potentially relevant mechanisms, in the paper we only discussed novel genes if they were found in multiple independent GWAS datasets. I consider this to be a sensible solution due to i) the limited space in the paper and ii) potential biases (multiple testing, noisy resistance data, lack of power and impact of recombination) reducing the confidence in results that are found only in one GWAS dataset. However, this means that our collection of GWAS hits likely includes many interesting, novel results that should be further explored and experimentally validated. Third, the role of epistatic interactions, which could be relevant according to our *in vitro* study (3.3) and previous evidence (*365*), remains underexplored in our analysis. I consider that we had insufficient strains to explore such interactions, but it is an interesting question that may be pursued with further sequencing efforts. In summary, despite all the novel insights, our analyses had some limitations that open future research directions.

Beyond the results and datasets related to recent selection and drug resistance, our study provides various processed datasets and pipelines that constitute valuable resources. On the one hand, we provide the small variants (available in CandidaMine), strain trees, clade annotations (of our work and also previous hallmark studies) and phenotypic metadata. These data enable the reproducibility of our results, and also facilitate further epidemiological or population genomic studies on these pathogens. To mention a few examples, our dataset could be directly re-analyzed to i) characterize the phylogeographic dynamics of clades, ii) understand the phylogenetic position of newly-sequenced isolates or iii) study recombination between divergent strains. The availability of the variants and trees is particularly relevant, as such data is often not provided. In addition, the integrative nature of our dataset (with drug susceptibility data and clade assignments of various studies) makes it useful for such future research / clinical avenues. On the other hand, we provide the code and software environments used to generate our results. This is not a common practice in similar studies, but I consider it essential to facilitate i) reproducibility of the results and ii) use of

the methods (i.e. for variant filtering, GWAS, selection analyses or tree generation) in further studies. In particular, we developed our novel GWAS pipeline to be installable as a standalone package, which may be used in future drug resistance studies in fungi and beyond. In addition, by providing the necessary Conda environments we facilitate deploying this code on new machines. This is not a trivial aspect, as there are hundreds of dependencies that would be tedious to handle without such Conda-based installation. All in all, we provided various datasets and software that not only enable reproducibility of our results, but also pave the way for future interesting research avenues.

Despite the relevance of our collection, it has various limitations that generally apply to our analyses of diversity, selection and drug resistance. On the one hand, the fact that our dataset comes from aggregating data from multiple studies may lead to batch effects influencing the variant calling results, as proposed before (*366*), due to multiple experimental protocols for sequencing. This might be particularly relevant for CNV calling due to sample-specific 'smiley patterns' (discussed above), but batch effects could also have an impact on the results drawn from other variant types (*366*). We addressed this by i) filtering out low quality samples, ii) performing sample-tailored SV/CNV calling with perSVade, iii) considering only positions with high coverage in all strains for tree reconstruction, iv) focusing on results that are consistent across multiple independent datasets (i.e. different species or drugs). However, we cannot fully discard some impact of batch effects. Thus, future studies may clarify the influence of varying sequencing parameters (i.e. sequencing machine, DNA extraction kit, library preparation protocol, coverage, read length and/or insert size) on such selection inferences or GWAS results in *Candida* pathogens. Such analyses may be necessary to validate whether our filtering strategies are appropriate to correct these biases. In summary, although aggregating data from multiple studies is a novel, insightful strategy, it can theoretically lead to some batch effects that may be further explored.

On the other hand, despite our effort to make our collection as comprehensive as possible, it is still incomplete due to various factors. First, the fact that our isolates are likely biased towards certain countries (discussed above) implies that our results may not be entirely applicable to strains from lesser studied regions. Thus, further sampling of isolates from underrepresented countries may be necessary to ensure the general relevance of our results. Second, the fact that the collection of strains for each species and/or each drug (in GWAS) is different, both in terms of sample size and representativeness of the natural *Candida* populations, limits our ability to make comparisons between species/drugs. Although the similarities between datasets (i.e. role of *NRG1/2* for fluconazole adaptation in both *C. glabrata* and *C. auris*) are insightful, we cannot discard that differences are due to varying levels of power. For instance, the fact that 'response to temperature stimulus' is only enriched in genes under selection in *C. albicans*, but not in the other close species (i.e. *C. tropicalis*), may be due to the higher sample size in *C. albicans*. This is why we focused most of our conclusions on the overlaps across datasets, rather than their differences. Such

limitation applies to our diversity analysis, the types of SVs inferred (i.e. the fact that some SV types are only present in some species, as discussed above), the selection analyses and the resistance mechanisms inferred. Thus, it is possible that we underestimated the fraction of genes related to selective processes and/or drug resistance in multiple species/drugs. Further sequencing of underrepresented species (*C. orthopsilosis*, *C. parapsilosis* and *C. tropicalis*) may be necessary to improve this. Third, to enable more comprehensive analyses, it could be relevant to enrich our collection with further metadata or experimental measurements. Further efforts to integrate various available datasets (i.e. genomic, phenotypic, transcriptomic and/or epigenomic) for each strain may enable a precise understanding of recent adaptation in *Candida* pathogens. This may allow mechanistically insightful studies that better explain how such genomic changes contribute to resistance. All in all, despite the comprehensive nature of our dataset, further data collection efforts are needed to fully understand selection and drug resistance in natural populations of *Candida* pathogens.

In summary, despite some limitations and open questions, our study improves our understanding about the signs of selection and mechanisms of drug resistance of major *Candida* pathogens, unearthing candidate genes that deserve future attention.

5. CONCLUSIONS

5. CONCLUSIONS

Based on the results shown here we conclude that:

1) The perSVade pipeline developed here simplifies and improves the analysis of structural variants (SVs) from short reads. This enables the study of SVs in non-model organisms, such as *Candida* species.

2) PerSVade can predict the SV calling accuracy on simulated genomes, which informs about the reliability of the calling process in an automatic manner.

3) PerSVade's parameter optimization is essential to maximize SV calling accuracy on simulated variants for six eukaryotic organisms, and on a reference dataset of validated human variants. There is no universal set of "optimal" parameters, which underscores the need for species-specific parameter optimization for SV calling.

4) PerSVade can be used to analyze SNPs, INDELs, SVs and CNVs, so that it facilitates multi-variant reproducible genomic studies.

5) *Candida glabrata* has a high ability to evolve resistance towards fluconazole and anidulafungin, with minor fitness costs and requiring few mutations *in vitro*.

6) There is a large set of mutations, affecting a narrow set of genes, driving resistance *in vitro* in *Candida glabrata*. *ERG11* and *PDR1* alterations underlie fluconazole resistance, while several (not only hotspot) mutations in *FKS* genes drive anidulafungin resistance.

7) *ERG3* mutations often drive cross-resistance from anidulafungin to fluconazole *in vitro* in *Candida glabrata*.

8) Recent, clinically-relevant, selection has shaped hundreds of gene families and pathways in six major *Candida* pathogens. Adaptations are mostly species-specific, suggesting highly variable, multifactorial adaptive mechanisms. In addition, there are various conserved adaptive processes involving drug resistance, adhesion and filamentous growth.

9) Convergence-based GWAS is a suitable approach to study the mechanisms of drug resistance in clinical isolates of *Candida auris, glabrata* and *albicans*. Hundreds of genes and pathways potentially affect resistance towards all major antifungal drugs. These include known drivers of resistance (*ERG11*, *PDR1*, *TAC1b* and *FKS* genes) and also novel players related to adhesion, biofilm formation and transcriptional regulation.

10) Structural and copy-number variants are likely drivers of genetic diversity, recent selection and antifungal drug resistance in *Candida* pathogens. They should be considered more often in genomic analyses.

11) The dataset of variants, phenotypes, selection signatures and genotype-drug resistance associations generated here constitute a valuable resource to validate specific hypotheses in future studies and to develop clinical applications.

6. BIBLIOGRAPHY

6. BIBLIOGRAPHY

1. D. L. Hawksworth, R. Lücking, Fungal Diversity Revisited: 2.2 to 3.8 Million Species. *Microbiol. Spectr.* **5** (2017), doi:10.1128/microbiolspec.FUNK-0052-2016.
2. G. M. Mueller, J. P. Schmit, Fungal biodiversity: what do we know? What can we predict? *Biodivers. Conserv.* **16**, 1–5 (2007).
3. J. P. Latgé, Aspergillus fumigatus and aspergillosis. *Clin. Microbiol. Rev.* **12**, 310–350 (1999).
4. G. A. Kowalchuk, S. E. Jones, L. L. Blackall, Microbes orchestrate life on Earth. *ISME J.* **2**, 795–796 (2008).
5. M. C. Fisher, D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, S. J. Gurr, Emerging fungal threats to animal, plant and ecosystem health. *Nature*. **484**, 186–194 (2012).
6. M. Wainwright, N. C. Wickramasinghe, J. V. Narlikar, P. Rajaratnam, Microorganisms cultured from stratospheric air samples obtained at 41 km. *FEMS Microbiol. Lett.* **218**, 161–165 (2003).
7. K. R. Freeman, A. P. Martin, D. Karki, R. C. Lynch, M. S. Mitter, A. F. Meyer, J. E. Longcore, D. R. Simmons, S. K. Schmidt, Evidence that chytrids dominate fungal communities in high-elevation soils. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18315–18320 (2009).
8. M. Blackwell, Made for Each Other: Ascomycete Yeasts and Insects. *Microbiol. Spectr.* **5** (2017), doi:10.1128/microbiolspec.FUNK-0081-2016.
9. M. J. McCullough, B. C. Ross, P. C. Reade, Candida albicans: a review of its history, taxonomy, epidemiology, virulence attributes, and methods of strain differentiation. *Int. J. Oral Maxillofac. Surg.* **25**, 136–144 (1996).
10. M. A. Naranjo-Ortiz, T. Gabaldón, Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol. Rev.* **94**, 2101–2137 (2019).
11. L. J. Galindo, G. Torruella, P. López-García, M. Ciobanu, A. Gutiérrez-Preciado, S. A. Karpov, D. Moreira, Phylogenomics Supports the Monophyly of Aphelids and Fungi and Identifies New Molecular Synapomorphies. *Syst. Biol.*, syac054 (2022).
12. T. Y. James, D. Porter, C. A. Leander, R. Vilgalys, J. E. Longcore, Molecular phylogenetics of the Chytridiomycota supports the utility of ultrastructural data in chytrid systematics. *Can. J. Bot.* **78**, 336–350 (2000).
13. S. Sekimoto, D. Rochon, J. E. Long, J. M. Dee, M. L. Berbee, A multigene phylogeny of Olpidium and its implications for early fungal evolution. *BMC Evol. Biol.* **11**, 331 (2011).
14. R. A. Hanafy, V. B. Lanjekar, P. K. Dhakephalkar, T. M. Callaghan, S. S. Dagar, G. W. Griffith, M. S. Elshahed, N. H. Youssef, Seven new Neocallimastigomycota genera from wild, zoo-housed, and domesticated herbivores greatly expand the taxonomic diversity of the phylum. *Mycologia*. **112**, 1212–1239 (2020).

15. T. Y. James, P. M. Letcher, J. E. Longcore, S. E. Mozley-Standridge, D. Porter, M. J. Powell, G. W. Griffith, R. Vilgalys, A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). *Mycologia*. **98**, 860–871 (2006).
16. J. W. Spatafora, Y. Chang, G. L. Benny, K. Lazarus, M. E. Smith, M. L. Berbee, G. Bonito, N. Corradi, I. Grigoriev, A. Gryganskyi, T. Y. James, K. O'Donnell, R. W. Roberson, T. N. Taylor, J. Uehling, R. Vilgalys, M. M. White, J. E. Stajich, A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia*. **108**, 1028–1046 (2016).
17. A. Schüβler, D. Schwarzott, C. Walker, A new fungal phylum, the Glomeromycota: phylogeny and evolution* *Dedicated to Manfred Kluge (Technische Universität Darmstadt) on the occasion of his retirement. *Mycol. Res.* **105**, 1413–1421 (2001).
18. R.-L. Zhao, G.-J. Li, S. Sánchez-Ramírez, M. Stata, Z.-L. Yang, G. Wu, Y.-C. Dai, S.-H. He, B.-K. Cui, J.-L. Zhou, F. Wu, M.-Q. He, J.-M. Moncalvo, K. D. Hyde, A six-gene phylogenetic overview of Basidiomycota and allied phyla with estimated divergence times of higher taxa and a phyloproteomics perspective. *Fungal Divers.* **84**, 43–74 (2017).
19. J. W. Spatafora, M. C. Aime, I. V. Grigoriev, F. Martin, J. E. Stajich, M. Blackwell, The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies. *Microbiol. Spectr.* **5** (2017), doi:10.1128/microbiolspec.FUNK-0053-2016.
20. X.-X. Shen, D. A. Opulente, J. Kominek, X. Zhou, J. L. Steenwyk, K. V. Buh, M. A. B. Haase, J. H. Wisecaver, M. Wang, D. T. Doering, J. T. Boudouris, R. M. Schneider, Q. K. Langdon, M. Ohkuma, R. Endoh, M. Takashima, R.-I. Manabe, N. Čadež, D. Libkind, C. A. Rosa, J. DeVirgilio, A. B. Hulfachor, M. Groenewald, C. P. Kurtzman, C. T. Hittinger, A. Rokas, Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*. **175**, 1533-1545.e20 (2018).
21. T. Gabaldón, M. A. Naranjo-Ortíz, M. Marcet-Houben, Evolutionary genomics of yeast pathogens in the Saccharomycotina. *FEMS Yeast Res.* **16**, fow064 (2016).
22. M. Figueroa, K. E. Hammond-Kosack, P. S. Solomon, A review of wheat diseases-a field perspective. *Mol. Plant Pathol.* **19**, 1523–1536 (2018).
23. S. V. Avery, I. Singleton, N. Magan, G. H. Goldman, The fungal threat to global food security. *Fungal Biol.* **123**, 555–557 (2019).
24. H. C. J. Godfray, D. Mason-D'Croz, S. Robinson, Food system consequences of a fungal disease epidemic in a major crop. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150467 (2016).
25. P. Tripathi, N. K. Dubey, Exploitation of natural products as an alternative strategy to control postharvest fungal rotting of fruit and vegetables. *Postharvest Biol. Technol.* **32**, 235–245 (2004).
26. D. P. Bebber, M. A. T. Ramotowski, S. J. Gurr, Crop pests and pathogens move polewards in a warming world. *Nat. Clim. Change*. **3**, 985–988 (2013).
27. S. Sun, M. J. Hoy, J. Heitman, Fungal pathogens. *Curr. Biol.* **30**, R1163–R1169 (2020).
28. S. Federhen, The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).

29. J. Pleadin, J. Frece, K. Markov, Mycotoxins in food and feed. *Adv. Food Nutr. Res.* **89**, 297–345 (2019).

30. A. Medina, A. Akbar, A. Baazeem, A. Rodriguez, N. Magan, Climate change, food security and mycotoxins: Do we know enough? *Fungal Biol. Rev.* **31**, 143–154 (2017).

31. N. Benkerroum, Aflatoxins: Producing-Molds, Structure, Health Issues and Incidence in Southeast Asian and Sub-Saharan African Countries. *Int. J. Environ. Res. Public. Health*. **17**, 1215 (2020).

32. F. Malir, V. Ostry, A. Pfohl-Leszkowicz, J. Malir, J. Toman, Ochratoxin A: 50 Years of Research. *Toxins*. **8**, 191 (2016).

33. M. M. Moake, O. I. Padilla-Zakour, R. W. Worobo, Comprehensive Review of Patulin Control Methods in Foods. *Compr. Rev. Food Sci. Food Saf.* **4**, 8–21 (2005).

34. L. Qu, L. Wang, H. Ji, Y. Fang, P. Lei, X. Zhang, L. Jin, D. Sun, H. Dong, Toxic Mechanism and Biological Detoxification of Fumonisins. *Toxins*. **14**, 182 (2022).

35. I. Laraba, S. P. McCormick, M. M. Vaughan, D. M. Geiser, K. O'Donnell, Phylogenetic diversity, trichothecene potential, and pathogenicity within Fusarium sambucinum species complex. *PLoS ONE*. **16**, e0245037 (2021).

36. V. Zingales, M. Taroncher, P. A. Martino, M.-J. Ruiz, F. Caloni, Climate Change and Effects on Molds and Mycotoxins. *Toxins*. **14**, 445 (2022).

37. G. S. Shephard, Impact of mycotoxins on human health in developing countries. *Food Addit. Contam. Part Chem. Anal. Control Expo. Risk Assess.* **25**, 146–151 (2008).

38. S. Seyedmousavi, S. de M. G. Bosco, S. de Hoog, F. Ebel, D. Elad, R. R. Gomes, I. D. Jacobsen, H. E. Jensen, A. Martel, B. Mignon, F. Pasmans, E. Piecková, A. M. Rodrigues, K. Singh, V. A. Vicente, G. Wibbelt, N. P. Wiederhold, J. Guillot, Fungal infections in animals: a patchwork of different situations. *Med. Mycol.* **56**, 165–187 (2018).

39. B. C. Scheele, F. Pasmans, L. F. Skerratt, L. Berger, A. Martel, W. Beukema, A. A. Acevedo, P. A. Burrowes, T. Carvalho, A. Catenazzi, I. De la Riva, M. C. Fisher, S. V. Flechas, C. N. Foster, P. Frías-Álvarez, T. W. J. Garner, B. Gratwicke, J. M. Guayasamin, M. Hirschfeld, J. E. Kolby, T. A. Kosch, E. La Marca, D. B. Lindenmayer, K. R. Lips, A. V. Longo, R. Maneyro, C. A. McDonald, J. Mendelson, P. Palacios-Rodriguez, G. Parra-Olea, C. L. Richards-Zawacki, M.-O. Rödel, S. M. Rovito, C. Soto-Azat, L. F. Toledo, J. Voyles, C. Weldon, S. M. Whitfield, M. Wilkinson, K. R. Zamudio, S. Canessa, Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*. **363**, 1459–1463 (2019).

40. N. T. Case, J. Berman, D. S. Blehert, R. A. Cramer, C. Cuomo, C. R. Currie, I. V. Ene, M. C. Fisher, L. K. Fritz-Laylin, A. C. Gerstein, N. L. Glass, N. A. R. Gow, S. J. Gurr, C. T. Hittinger, T. M. Hohl, I. D. Iliev, T. Y. James, H. Jin, B. S. Klein, J. W. Kronstad, J. M. Lorch, V. McGovern, A. P. Mitchell, J. A. Segre, R. S. Shapiro, D. C. Sheppard, A. Sil, J. E. Stajich, E. E. Stukenbrock, J. W. Taylor, D. Thompson, G. D. Wright, J. Heitman, L. E. Cowen, The future of fungi: threats and opportunities. *G3 GenesGenomesGenetics*. **12**, jkac224 (2022).

41. S. J. O'Hanlon, A. Rieux, R. A. Farrer, G. M. Rosa, B. Waldman, A. Bataille, T. A. Kosch, K. A. Murray, B. Brankovics, M. Fumagalli, M. D. Martin, N. Wales, M. Alvarado-Rybak, K. A. Bates, L. Berger, S. Böll, L. Brookes, F. Clare, E. A. Courtois, A. A. Cunningham, T. M. Doherty-Bone, P. Ghosh, D. J. Gower, W. E. Hintz, J. Höglund, T. S. Jenkinson, C.-F. Lin, A. Laurila, A. Loyau, A. Martel, S. Meurling, C. Miaud, P. Minting, F. Pasmans, D. S. Schmeller, B. R. Schmidt, J. M. G. Shelton, L. F. Skerratt, F. Smith, C. Soto-Azat, M. Spagnoletti, G. Tessa, L. F. Toledo, A. Valenzuela-Sánchez, R. Verster, J. Vörös, R. J. Webb, C. Wierzbicki, E. Wombwell, K. R. Zamudio, D. M. Aanensen, T. Y. James, M. T. P. Gilbert, C. Weldon, J. Bosch, F. Balloux, T. W. J. Garner, M. C. Fisher, Recent Asian origin of chytrid fungi causing global amphibian declines. *Science*. **360**, 621–627 (2018).

42. D. S. Blehert, A. C. Hicks, M. Behr, C. U. Meteyer, B. M. Berlowski-Zier, E. L. Buckles, J. T. H. Coleman, S. R. Darling, A. Gargas, R. Niver, J. C. Okoniewski, R. J. Rudd, W. B. Stone, Bat white-nose syndrome: an emerging fungal pathogen? *Science*. **323**, 227 (2009).

43. M. C. Fisher, S. J. Gurr, C. A. Cuomo, D. S. Blehert, H. Jin, E. H. Stukenbrock, J. E. Stajich, R. Kahmann, C. Boone, D. W. Denning, N. A. R. Gow, B. S. Klein, J. W. Kronstad, D. C. Sheppard, J. W. Taylor, G. D. Wright, J. Heitman, A. Casadevall, L. E. Cowen, Threats Posed by the Fungal Kingdom to Humans, Wildlife, and Agriculture. *mBio*. **11**, e00449-20 (2020).

44. D. Rigling, S. Prospero, Cryphonectria parasitica, the causal agent of chestnut blight: invasion history, population biology and disease control. *Mol. Plant Pathol.* **19**, 7–20 (2018).

45. G. M. Lovett, M. A. Arthur, K. C. Weathers, J. M. Griffin, Long-term Changes in Forest Carbon and Nitrogen Cycling Caused by an Introduced Pest/Pathogen Complex. *Ecosystems*. **13**, 1188–1200 (2010).

46. F. Bongomin, S. Gago, R. O. Oladele, D. W. Denning, Global and Multi-National Prevalence of Fungal Diseases-Estimate Precision. *J. Fungi Basel Switz.* **3**, 57 (2017).

47. W. H. O. A. R. Division, W. H. O. C. of N. T. Diseases, W. H. O. G. C. and Partnership, A. Alastruey-Izquierdo, "WHO fungal priority pathogens list to guide research, development and public health action" (VoR, Organización Mundial de la Salud (OMS), 2022), (available at https://repisalud.isciii.es/handle/20.500.12105/15113).

48. G. D. Brown, D. W. Denning, N. A. R. Gow, S. M. Levitz, M. G. Netea, T. C. White, Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **4**, 165rv13-165rv13 (2012).

49. T. Gabaldón, L. Carreté, The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in Candida glabrata. *FEMS Yeast Res.* **16**, fov110 (2016).

50. A. B. Arsenault, J. M. Bliss, Neonatal Candidiasis: New Insights into an Old Problem at a Unique Host-Pathogen Interface. *Curr. Fungal Infect. Rep.* **9**, 246–252 (2015).

51. U. Hofer, How antibiotics predispose to candidiasis. *Nat. Rev. Microbiol.* **20**, 382–382 (2022).

52. M. Shahi, S. Ayatollahi Mousavi, M. Nabili, M. Aliyali, S. Khodavaisy, H. Badali, Aspergillus colonization in patients with chronic obstructive pulmonary disease. *Curr. Med. Mycol.* **1**, 45–51

(2015).

53. A. Raut, N. T. Huy, Rising incidence of mucormycosis in patients with COVID-19: another challenge for India amidst the second wave? *Lancet Respir. Med.* **9**, e77 (2021).

54. W. MacNee, Is Chronic Obstructive Pulmonary Disease an Accelerated Aging Disease? *Ann. Am. Thorac. Soc.* **13**, S429–S437 (2016).

55. A. Vermes, H.-J. Guchelaar, J. Dankert, Flucytosine: a review of its pharmacology, clinical indications, pharmacokinetics, toxicity and drug interactions. *J. Antimicrob. Chemother.* **46**, 171–179 (2000).

56. J. J. Gallagher, N. Williams-Bouyer, C. Villarreal, J. P. Heggers, D. N. Herndon, "Chapter 12 - Treatment of infection in burns" in *Total Burn Care (Third Edition)*, D. N. Herndon, Ed. (W.B. Saunders, Edinburgh, 2007; https://www.sciencedirect.com/science/article/pii/B9781416032748500155), pp. 136–176.

57. D. S. Perlin, Resistance to echinocandin-class antifungal drugs. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **10**, 121–130 (2007).

58. L. Heimark, P. Shipkova, J. Greene, H. Munayyer, T. Yarosh-Tomaine, B. DiDomenico, R. Hare, B. N. Pramanik, Mechanism of azole antifungal activity as determined by liquid chromatographic/mass spectrometric monitoring of ergosterol biosynthesis. *J. Mass Spectrom. JMS*. **37**, 265–269 (2002).

59. H. Carolus, S. Pierson, K. Lagrou, P. Van Dijck, Amphotericin B and Other Polyenes—Discovery, Clinical Use, Mode of Action and Drug Resistance. *J. Fungi*. **6**, 321 (2020).

60. D. Farmakiotis, D. P. Kontoyiannis, Epidemiology of antifungal resistance in human pathogenic yeasts: current viewpoint and practical recommendations for management. *Int. J. Antimicrob. Agents*. **50**, 318–324 (2017).

61. A. Arastehfar, T. Gabaldón, R. Garcia-Rubio, J. D. Jenks, M. Hoenigl, H. J. F. Salzer, M. Ilkit, C. Lass-Flörl, D. S. Perlin, Drug-Resistant Fungi: An Emerging Challenge Threatening Our Limited Antifungal Armamentarium. *Antibiot. Basel Switz.* **9**, 877 (2020).

62. D. W. Denning, Antifungal drug resistance: an update. *Eur. J. Hosp. Pharm. Sci. Pract.* **29**, 109–112 (2022).

63. J. R. Perfect, The antifungal pipeline: a reality check. *Nat. Rev. Drug Discov.* **16**, 603–616 (2017).

64. M. A. Pfaller, Antifungal drug resistance: mechanisms, epidemiology, and consequences for treatment. *Am. J. Med.* **125**, S3-13 (2012).

65. J. W. M. van der Linden, E. Snelders, G. A. Kampinga, B. J. A. Rijnders, E. Mattsson, Y. J. Debets-Ossenkopp, E. J. Kuijper, F. H. Van Tiel, W. J. G. Melchers, P. E. Verweij, Clinical implications of azole resistance in Aspergillus fumigatus, The Netherlands, 2007-2009. *Emerg. Infect. Dis.* **17**, 1846–1854 (2011).

66. M. W. McCarthy, T. J. Walsh, Drugs currently under investigation for the treatment of invasive candidiasis. *Expert Opin. Investig. Drugs*. **26**, 825–831 (2017).

67. M. C. Arendrup, T. Boekhout, M. Akova, J. F. Meis, O. A. Cornely, O. Lortholary, European Society of

Clinical Microbiology and Infectious Diseases Fungal Infection Study Group, European Confederation of Medical Mycology, ESCMID and ECMM joint clinical guidelines for the diagnosis and management of rare invasive yeast infections. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **20 Suppl 3**, 76–98 (2014).

68. Consortium OPATHY, T. Gabaldón, Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS Microbiol. Rev.* **43**, 517–547 (2019).
69. V. A. Robert, A. Casadevall, Vertebrate Endothermy Restricts Most Fungi as Potential Pathogens. *J. Infect. Dis.* **200**, 1623–1626 (2009).
70. M. A. Garcia-Solache, A. Casadevall, Global warming will bring new fungal diseases for mammals. *mBio*. **1**, e00061-10 (2010).
71. N. E. Nnadi, D. A. Carter, Climate change and the emergence of fungal pathogens. *PLOS Pathog.* **17**, e1009503 (2021).
72. X. Wu, Y. Lu, S. Zhou, L. Chen, B. Xu, Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environ. Int.* **86**, 14–23 (2016).
73. A. Casadevall, D. P. Kontoyiannis, V. Robert, On the Emergence of Candida auris: Climate Change, Azoles, Swamps, and Birds. *mBio*. **10**, e01397-19 (2019).
74. *One Health: Fungal Pathogens of Humans, Animals, and Plants: Report on an American Academy of Microbiology Colloquium held in Washington, DC, on October 18, 2017* (American Society for Microbiology, Washington (DC), 2019; http://www.ncbi.nlm.nih.gov/books/NBK549988/), *American Academy of Microbiology Colloquia Reports*.
75. A. B. Araúz, P. Papineni, Histoplasmosis. *Infect. Dis. Clin. North Am.* **35**, 471–491 (2021).
76. L.-J. Ma, D. M. Geiser, R. H. Proctor, A. P. Rooney, K. O'Donnell, F. Trail, D. M. Gardiner, J. M. Manners, K. Kazan, Fusarium pathogenomics. *Annu. Rev. Microbiol.* **67**, 399–416 (2013).
77. S. Yadlapati, S. P. Chaudhari, "Eumycetoma" in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2022; http://www.ncbi.nlm.nih.gov/books/NBK574511/).
78. A. Flevari, M. Theodorakopoulou, A. Velegraki, A. Armaganidis, G. Dimopoulos, Treatment of invasive candidiasis in the elderly: a review. *Clin. Interv. Aging*. **8**, 1199–1208 (2013).
79. J. D. Sobel, Vulvovaginal candidosis. *Lancet Lond. Engl.* **369**, 1961–1971 (2007).
80. J. Morgan, Global trends in candidemia: Review of reports from 1995–2005. *Curr. Infect. Dis. Rep.* **7**, 429–439 (2005).
81. N. Papon, V. Courdavault, M. Clastre, R. J. Bennett, Emerging and Emerged Pathogenic Candida Species: Beyond the Candida albicans Paradigm. *PLoS Pathog.* **9**, e1003550 (2013).
82. G. Bravo Ruiz, A. Lorenz, What do we know about the biology of the emerging fungal pathogen of humans Candida auris? *Microbiol. Res.* **242**, 126621 (2021).
83. C. W. J. Africa, P. M. dos S. Abrantes, Candida antifungal drug resistance in sub-Saharan African populations: A systematic review. *F1000Research*. **5**, 2832 (2017).

84. M. J. Biagi, N. P. Wiederhold, C. Gibas, B. L. Wickes, V. Lozano, S. C. Bleasdale, L. Danziger, Development of High-Level Echinocandin Resistance in a Patient With Recurrent Candida auris Candidemia Secondary to Chronic Candiduria. *Open Forum Infect. Dis.* **6**, ofz262 (2019).

85. S. S. Rathore, J. Sathiyamoorthy, C. Lalitha, J. Ramakrishnan, A holistic review on Cryptococcus neoformans. *Microb. Pathog.* **166**, 105521 (2022).

86. J. M. Steinbrink, M. H. Miceli, Mucormycosis. *Infect. Dis. Clin. North Am.* **35**, 435–452 (2021).

87. A.-L. Bidaud, P. Schwarz, G. Herbreteau, E. Dannaoui, Techniques for the Assessment of In Vitro and In Vivo Antifungal Combinations. *J. Fungi*. **7**, 113 (2021).

88. F. Almeida, M. L. Rodrigues, C. Coelho, The Still Underestimated Problem of Fungal Diseases Worldwide. *Front. Microbiol.* **10** (2019) (available at https://www.frontiersin.org/articles/10.3389/fmicb.2019.00214).

89. L. V. N. Oliveira, R. Wang, C. A. Specht, S. M. Levitz, Vaccines for human fungal diseases: close but still a long way to go. *Npj Vaccines*. **6**, 1–8 (2021).

90. J. D. Jenks, C. I. Aneke, M. M. Al-Obaidi, M. Egger, L. Garcia, T. Gaines, M. Hoenigl, G. R. T. Iii, Race and ethnicity: Risk factors for fungal infections? *PLOS Pathog.* **19**, e1011025 (2023).

91. G. Singh, C. W. Pitoyo, D. Aditianingsih, C. M. Rumende, Risk factors for early invasive fungal disease in critically ill patients. *Indian J. Crit. Care Med. Peer-Rev. Off. Publ. Indian Soc. Crit. Care Med.* **20**, 633–639 (2016).

92. M. Galocha, R. Viana, P. Pais, A. Silva-Dias, M. Cavalheiro, I. M. Miranda, M. Van Ende, C. S. Souza, C. Costa, J. Branco, C. M. Soares, P. Van Dijck, A. G. Rodrigues, M. C. Teixeira, Genomic evolution towards azole resistance in Candida glabrata clinical isolates unveils the importance of CgHxt4/6/7 in azole accumulation. *Commun. Biol.* **5**, 1–12 (2022).

93. J. Bing, T. Hu, Q. Zheng, J. F. Muñoz, C. A. Cuomo, G. Huang, Experimental Evolution Identifies Adaptive Aneuploidy as a Mechanism of Fluconazole Resistance in Candida auris. *Antimicrob. Agents Chemother.* **65**, e01466-20 (2020).

94. R. A. C. Dos Santos, J. L. Steenwyk, O. Rivero-Menendez, M. E. Mead, L. P. Silva, R. W. Bastos, A. Alastruey-Izquierdo, G. H. Goldman, A. Rokas, Genomic and Phenotypic Heterogeneity of Clinical Isolates of the Human Pathogens Aspergillus fumigatus, Aspergillus lentulus, and Aspergillus fumigatiaffinis. *Front. Genet.* **11**, 459 (2020).

95. M. S. Schröder, K. Martinez de San Vicente, T. H. R. Prandini, S. Hammel, D. G. Higgins, E. Bagagli, K. H. Wolfe, G. Butler, Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species. *PLoS Genet.* **12**, e1006404 (2016).

96. L. Mukaremera, K. K. Lee, H. M. Mora-Montes, N. A. R. Gow, Candida albicans Yeast, Pseudohyphal, and Hyphal Morphogenesis Differentially Affects Immune Recognition. *Front. Immunol.* **8**, 629 (2017).

97. E. Sasani, S. Khodavaisy, S. Agha Kuchak Afshari, S. Darabian, F. Aala, S. Rezaie, Pseudohyphae

formation in Candida glabrata due to CO2 exposure. *Curr. Med. Mycol.* **2**, 49–52 (2016).

98. T. Chakraborty, Z. Tóth, R. Tóth, C. Vágvölgyi, A. Gácser, Iron Metabolism, Pseudohypha Production, and Biofilm Formation through a Multicopper Oxidase in the Human-Pathogenic Fungus Candida parapsilosis. *mSphere*. **5**, e00227-20 (2020).

99. H. Du, Q. Zheng, R. J. Bennett, G. Huang, Ploidy changes in human fungal pathogens: Going beyond sexual reproduction. *PLOS Pathog.* **18**, e1010954 (2022).

100. J. Heitman, J. W. Kronstad, J. W. Taylor, L. A. Casselton, Sex in fungi: molecular determination and evolutionary implications. *Sex Fungi Mol. Determ. Evol. Implic.* (2007) (available at https://www.cabdirect.org/cabdirect/abstract/20073244184).

101. M. A. S. Santos, A. C. Gomes, M. C. Santos, L. C. Carreto, G. R. Moura, The genetic code of the fungal CTG clade. *C. R. Biol.* **334**, 607–611 (2011).

102. L. P. Pryszcz, T. Németh, A. Gácser, T. Gabaldón, Genome Comparison of Candida orthopsilosis Clinical Strains Reveals the Existence of Hybrids between Two Distinct Subspecies. *Genome Biol. Evol.* **6**, 1069–1078 (2014).

103. I. Stefanini, E. Stoakes, H. H. T. Wu, L. Xu-McCrae, A. Hussain, J. Moat, C. G. Dowson, M. D. David, C. Constantinidou, Genomic Assembly of Clinical Candida glabrata (Nakaseomyces glabrata) Isolates Reveals within-Species Structural Plasticity and Association with *In Vitro* Antifungal Susceptibility. *Microbiol. Spectr.*, e01827-22 (2022).

104. P. Pais, M. Galocha, R. Viana, M. Cavalheiro, D. Pereira, M. C. Teixeira, Microevolution of the pathogenic yeasts Candida albicans and Candida glabrata during antifungal therapy and host infection. *Microb. Cell Graz Austria*. **6**, 142–159 (2019).

105. J. Talapko, M. Juzbašić, T. Matijević, E. Pustijanac, S. Bekić, I. Kotris, I. Škrlec, Candida albicans—The Virulence Factors and Clinical Manifestations of Infection. *J. Fungi*. **7**, 79 (2021).

106. P. G. Pappas, C. A. Kauffman, D. R. Andes, C. J. Clancy, K. A. Marr, L. Ostrosky-Zeichner, A. C. Reboli, M. G. Schuster, J. A. Vazquez, T. J. Walsh, T. E. Zaoutis, J. D. Sobel, Clinical Practice Guideline for the Management of Candidiasis: 2016 Update by the Infectious Diseases Society of America. *Clin. Infect. Dis.* **62**, e1–e50 (2016).

107. M. Toda, Population-Based Active Surveillance for Culture-Confirmed Candidemia — Four Sites, United States, 2012–2016. *MMWR Surveill. Summ.* **68** (2019), doi:10.15585/mmwr.ss6808a1.

108. P. Montravers, O. Leroy, C. Eckmann, Intra-abdominal candidiasis: it's still a long way to get unquestionable data. *Intensive Care Med.* **41**, 1682–1684 (2015).

109. E. Roilides, E. Farmaki, J. Evdoridou, A. Francesconi, M. Kasai, J. Filioti, M. Tsivitanidou, D. Sofianou, G. Kremenopoulos, T. J. Walsh, Candida tropicalis in a Neonatal Intensive Care Unit: Epidemiologic and Molecular Analysis of an Outbreak of Infection with an Uncommon Neonatal Pathogen. *J. Clin. Microbiol.* **41**, 735–741 (2003).

110. D. L. Zuza-Alves, W. P. Silva-Rocha, G. M. Chaves, An Update on Candida tropicalis Based on Basic and

Clinical Approaches. *Front. Microbiol.* **8** (2017) (available at https://www.frontiersin.org/articles/10.3389/fmicb.2017.01927).

111. L. Wang, A. Xu, P. Zhou, M. Zhao, C. Xu, Y. Wang, K. Wang, F. Wang, Y. Miao, W. Zhao, X. Gao, Rapid Detection of Candida tropicalis in Clinical Samples From Different Sources Using RPA-LFS. *Front. Cell. Infect. Microbiol.* **12** (2022) (available at https://www.frontiersin.org/articles/10.3389/fcimb.2022.898186).
112. X. Fan, M. Xiao, K. Liao, T. Kudinha, H. Wang, L. Zhang, X. Hou, F. Kong, Y.-C. Xu, Notable Increasing Trend in Azole Non-susceptible Candida tropicalis Causing Invasive Candidiasis in China (August 2009 to July 2014): Molecular Epidemiology and Clinical Azole Consumption. *Front. Microbiol.* **8**, 464 (2017).
113. M. A. Pfaller, D. J. Diekema, J. D. Turnidge, M. Castanheira, R. N. Jones, Twenty Years of the SENTRY Antifungal Surveillance Program: Results for Candida Species From 1997-2016. *Open Forum Infect. Dis.* **6**, S79–S94 (2019).
114. D. Trofa, A. Gácser, J. D. Nosanchuk, Candida parapsilosis, an Emerging Fungal Pathogen. *Clin. Microbiol. Rev.* **21**, 606–625 (2008).
115. V. Krcmery, A. J. Barnes, Non-albicans Candida spp. causing fungaemia: pathogenicity and antifungal resistance. *J. Hosp. Infect.* **50**, 243–260 (2002).
116. G. Garcia-Effron, S. K. Katiyar, S. Park, T. D. Edlind, D. S. Perlin, 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 (2008).
117. M. A. Pfaller, L. Boyken, R. J. Hollis, S. A. Messer, S. Tendolkar, D. J. Diekema, Global Surveillance of In Vitro Activity of Micafungin against Candida: a Comparison with Caspofungin by CLSI-Recommended Methods. *J. Clin. Microbiol.* **44**, 3533–3538 (2006).
118. M. Cavalheiro, M. C. Teixeira, Candida Biofilms: Threats, Challenges, and Promising Strategies. *Front. Med.* **5** (2018) (available at https://www.frontiersin.org/articles/10.3389/fmed.2018.00028).
119. S. Silva, C. F. Rodrigues, D. Araújo, M. E. Rodrigues, M. Henriques, Candida Species Biofilms' Antifungal Resistance. *J. Fungi Basel Switz.* **3**, 8 (2017).
120. A. Tavanti, A. D. Davidson, N. A. R. Gow, M. C. J. Maiden, F. C. Odds, Candida orthopsilosis and Candida metapsilosis spp. nov. To Replace Candida parapsilosis Groups II and III. *J. Clin. Microbiol.* **43**, 284–292 (2005).
121. M. T. Blanco-Blanco, A. C. Gómez-García, C. Hurtado, M. A. Galán-Ladero, M. del C. Lozano, A. García-Tapias, M. T. Blanco, Candida orthopsilosis fungemias in a Spanish tertiary care hospital: Incidence, epidemiology and antifungal susceptibility. *Rev. Iberoam. Micol.* **31**, 145–148 (2014).
122. S. R. Lockhart, S. A. Messer, M. A. Pfaller, D. J. Diekema, Geographic distribution and antifungal susceptibility of the newly described species Candida orthopsilosis and Candida metapsilosis in

comparison to the closely related species Candida parapsilosis. *J. Clin. Microbiol.* **46**, 2659–2664 (2008).

123. E. C. van Asbeck, K. V. Clemons, D. A. Stevens, Candida parapsilosis: a review of its epidemiology, pathogenesis, clinical aspects, typing and antimicrobial susceptibility. *Crit. Rev. Microbiol.* **35**, 283–309 (2009).

124. A. P. Silva, I. M. Miranda, C. Lisboa, C. Pina-Vaz, A. G. Rodrigues, Prevalence, Distribution, and Antifungal Susceptibility Profiles of Candida parapsilosis, C. orthopsilosis, and C. metapsilosis in a Tertiary Care Hospital. *J. Clin. Microbiol.* **47**, 2392–2397 (2009).

125. F. Lamoth, D. P. Kontoyiannis, The Candida auris Alert: Facts and Perspectives. *J. Infect. Dis.* **217**, 516–520 (2018).

126. T. S. N. Ku, C. J. Walraven, S. A. Lee, Candida auris: Disinfectants and Implications for Infection Control. *Front. Microbiol.* **9**, 726 (2018).

127. K. Satoh, K. Makimura, Y. Hasumi, Y. Nishiyama, K. Uchida, H. Yamaguchi, Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiol. Immunol.* **53**, 41–44 (2009).

128. C.-S. Tsai, S. S.-J. Lee, W.-C. Chen, C.-H. Tseng, N.-Y. Lee, P.-L. Chen, M.-C. Li, L.-S. Syue, C.-L. Lo, W.-C. Ko, Y.-P. Hung, COVID-19-associated candidiasis and the emerging concern of Candida auris infections. *J. Microbiol. Immunol. Infect. Wei Mian Yu Gan Ran Za Zhi*, S1684-1182(22)00283–3 (2022).

129. C. Keighley, K. Garnham, S. A. J. Harch, M. Robertson, K. Chaw, J. C. Teng, S. C.-A. Chen, Candida auris: Diagnostic Challenges and Emerging Opportunities for the Clinical Microbiology Laboratory. *Curr. Fungal Infect. Rep.* **15**, 116–126 (2021).

130. M. G. Frías-De-León, R. Hernández-Castro, T. Vite-Garín, R. Arenas, A. Bonifaz, L. Castañón-Olivares, G. Acosta-Altamirano, E. Martínez-Herrera, Antifungal Resistance in Candida auris: Molecular Determinants. *Antibiotics*. **9**, 568 (2020).

131. M. Lyman, Notes from the Field: Transmission of Pan-Resistant and Echinocandin-Resistant Candida auris in Health Care Facilities — Texas and the District of Columbia, January–April 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70** (2021), doi:10.15585/mmwr.mm7029a2.

132. T. Gabaldón, C. Fairhead, Genomes shed light on the secret life of Candida glabrata: not so asexual, not so commensal. *Curr. Genet.* **65**, 93–98 (2019).

133. K. Kumar, F. Askari, M. S. Sahu, R. Kaur, Candida glabrata: A Lot More Than Meets the Eye. *Microorganisms*. **7**, 39 (2019).

134. P. G. Pappas, M. S. Lionakis, M. C. Arendrup, L. Ostrosky-Zeichner, B. J. Kullberg, Invasive candidiasis. *Nat. Rev. Dis. Primer*. **4**, 1–20 (2018).

135. S. Vallabhaneni, A. A. Cleveland, M. M. Farley, L. H. Harrison, W. Schaffner, Z. G. Beldavs, G. Derado,

C. D. Pham, S. R. Lockhart, R. M. Smith, Epidemiology and Risk Factors for Echinocandin Nonsusceptible Candida glabrata Bloodstream Infections: Data From a Large Multisite Population-Based Candidemia Surveillance Program, 2008-2014. *Open Forum Infect. Dis.* **2**, ofv163 (2015).

136. K. E. Pristov, M. A. Ghannoum, Resistance of Candida to azoles and echinocandins worldwide. *Clin. Microbiol. Infect.* **25**, 792–798 (2019).
137. D. Kadosh, J. L. Lopez-Ribot, Candida albicans: Adapting to Succeed. *Cell Host Microbe*. **14**, 483–485 (2013).
138. G. H. W. Tso, J. A. Reales-Calderon, A. S. M. Tan, X. Sem, G. T. T. Le, T. G. Tan, G. C. Lai, K. G. Srinivasan, M. Yurieva, W. Liao, M. Poidinger, F. Zolezzi, G. Rancati, N. Pavelka, Experimental evolution of a fungal pathogen into a gut symbiont. *Science*. **362**, 589–595 (2018).
139. H. Carolus, S. Pierson, J. F. Muñoz, A. Subotić, R. B. Cruz, C. A. Cuomo, P. Van Dijck, Genome-Wide Analysis of Experimentally Evolved Candida auris Reveals Multiple Novel Mechanisms of Multidrug Resistance. *mBio*. **12**, e03333-20 (2021).
140. M. L. Branchini, M. A. Pfaller, J. Rhine-Chalberg, T. Frempong, H. D. Isenberg, Genotypic variation and slime production among blood and catheter isolates of Candida parapsilosis. *J. Clin. Microbiol.* **32**, 452–456 (1994).
141. E. Bojang, H. Ghuman, P. Kumwenda, R. A. Hall, Immune Sensing of Candida albicans. *J. Fungi*. **7**, 119 (2021).
142. S. Silva, M. Negri, M. Henriques, R. Oliveira, D. W. Williams, J. Azeredo, Candida glabrata, Candida parapsilosis and Candida tropicalis: biology, epidemiology, pathogenicity and antifungal resistance. *FEMS Microbiol. Rev.* **36**, 288–305 (2012).
143. J. P. Lopes, M. S. Lionakis, Pathogenesis and virulence of Candida albicans. *Virulence*. **13**, 89–121 (2021).
144. D. L. Moyes, D. Wilson, J. P. Richardson, S. Mogavero, S. X. Tang, J. Wernecke, S. Höfs, R. L. Gratacap, J. Robbins, M. Runglall, C. Murciano, M. Blagojevic, S. Thavaraj, T. M. Förster, B. Hebecker, L. Kasper, G. Vizcay, S. I. Iancu, N. Kichik, A. Häder, O. Kurzai, T. Luo, T. Krüger, O. Kniemeyer, E. Cota, O. Bader, R. T. Wheeler, T. Gutsmann, B. Hube, J. R. Naglik, Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature*. **532**, 64–68 (2016).
145. L. Böhm, S. Torsin, S. H. Tint, M. T. Eckstein, T. Ludwig, J. C. Pérez, The yeast form of the fungus Candida albicans promotes persistence in the gut of gnotobiotic mice. *PLoS Pathog.* **13**, e1006699 (2017).
146. S. Brunke, B. Hube, Two unlike cousins: Candida albicans and C. glabrata infection strategies. *Cell. Microbiol.* **15**, 701–708 (2013).
147. M. Galocha, P. Pais, M. Cavalheiro, D. Pereira, R. Viana, M. C. Teixeira, Divergent Approaches to Virulence in C. albicans and C. glabrata: Two Sides of the Same Coin. *Int. J. Mol. Sci.* **20**, 2345 (2019).

148. M. V. Horton, J. E. Nett, Candida auris infection and biofilm formation: going beyond the surface. *Curr. Clin. Microbiol. Rep.* **7**, 51–56 (2020).

149. E. R. Ballou, G. M. Avelar, D. S. Childers, J. Mackie, J. M. Bain, J. Wagener, S. L. Kastora, M. D. Panea, S. E. Hardison, L. A. Walker, L. P. Erwig, C. A. Munro, N. A. R. Gow, G. D. Brown, D. M. MacCallum, A. J. P. Brown, Lactate signalling regulates fungal β-glucan masking and immune evasion. *Nat. Microbiol.* **2**, 16238 (2016).

150. K. Seider, S. Brunke, L. Schild, N. Jablonowski, D. Wilson, O. Majer, D. Barz, A. Haas, K. Kuchler, M. Schaller, B. Hube, The facultative intracellular pathogen Candida glabrata subverts macrophage cytokine production and phagolysosome maturation. *J. Immunol. Baltim. Md 1950*. **187**, 3072–3086 (2011).

151. M. S. Ahmad Khan, F. Alshehrei, S. B. Al-Ghamdi, M. A. Bamaga, A. S. Al-Thubiani, M. Z. Alam, Virulence and biofilms as promising targets in developing antipathogenic drugs against candidiasis. *Future Sci. OA*. **6**, FSO440 (2020).

152. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The New Tree of Eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).

153. J. L. Nishikawa, A. Boeszoermenyi, L. A. Vale-Silva, R. Torelli, B. Posteraro, Y.-J. Sohn, F. Ji, V. Gelev, D. Sanglard, M. Sanguinetti, R. I. Sadreyev, G. Mukherjee, J. Bhyravabhotla, S. J. Buhrlage, N. S. Gray, G. Wagner, A. M. Näär, H. Arthanari, Inhibiting Fungal Multidrug Resistance by Disrupting an Activator-Mediator Interaction. *Nature*. **530**, 485–489 (2016).

154. K. R. Iyer, K. Camara, M. Daniel-Ivad, R. Trilles, S. M. Pimentel-Elardo, J. L. Fossen, K. Marchillo, Z. Liu, S. Singh, J. F. Muñoz, S. H. Kim, J. A. Porco, C. A. Cuomo, N. S. Williams, A. S. Ibrahim, J. E. Edwards, D. R. Andes, J. R. Nodwell, L. E. Brown, L. Whitesell, N. Robbins, L. E. Cowen, An oxindole efflux inhibitor potentiates azoles and impairs virulence in the fungal pathogen Candida auris. *Nat. Commun.* **11**, 6429 (2020).

155. N. A. Chow, J. F. Muñoz, L. Gade, E. L. Berkow, X. Li, R. M. Welsh, K. Forsberg, S. R. Lockhart, R. Adam, A. Alanio, A. Alastruey-Izquierdo, S. Althawadi, A. B. Araúz, R. Ben-Ami, A. Bharat, B. Calvo, M. Desnos-Ollivier, P. Escandón, D. Gardam, R. Gunturu, C. H. Heath, O. Kurzai, R. Martin, A. P. Litvintseva, C. A. Cuomo, Tracing the Evolutionary History and Global Expansion of Candida auris Using Population Genomic Analyses. *mBio*. **11**, e03364-19 (2020).

156. L. Carreté, E. Ksiezopolska, C. Pegueroles, E. Gómez-Molero, E. Saus, S. Iraola-Guzmán, D. Loska, O. Bader, C. Fairhead, T. Gabaldón, Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans. *Curr. Biol. CB*. **28**, 15-27.e7 (2018).

157. C. E. O'Brien, J. Oliveira-Pacheco, E. Ó Cinnéide, M. A. B. Haase, C. T. Hittinger, T. R. Rogers, O. Zaragoza, U. Bond, G. Butler, Population genomics of the pathogenic yeast Candida tropicalis identifies hybrid isolates in environmental samples. *PLoS Pathog.* **17**, e1009138 (2021).

158. J. Ropars, C. Maufrais, D. Diogo, M. Marcet-Houben, A. Perin, N. Sertour, K. Mosca, E. Permal, G. Laval, C. Bouchier, L. Ma, K. Schwartz, K. Voelz, R. C. May, J. Poulain, C. Battail, P. Wincker, A. M. Borman, A. Chowdhary, S. Fan, S. H. Kim, P. Le Pape, O. Romeo, J. H. Shin, T. Gabaldon, G. Sherlock, M.-E. Bougnoux, C. d'Enfert, Gene flow contributes to diversification of the major fungal pathogen Candida albicans. *Nat. Commun.* **9**, 2253 (2018).

159. A. E. Barber, M. Weber, K. Kaerger, J. Linde, H. Gölz, D. Duerschmied, A. Markert, R. Guthke, G. Walther, O. Kurzai, Comparative Genomics of Serial Candida glabrata Isolates and the Rapid Acquisition of Echinocandin Resistance during Therapy. *Antimicrob. Agents Chemother.* **63**, e01628-18 (2019).

160. K. Alby, R. J. Bennett, Sexual reproduction in the Candida clade: cryptic cycles, diverse mechanisms, and alternative functions. *Cell. Mol. Life Sci. CMLS*. **67**, 3275–3285 (2010).

161. M. Marcet-Houben, M. Alvarado, E. Ksiezopolska, E. Saus, P. W. J. de Groot, T. Gabaldón, Chromosome-level assemblies from diverse clades reveal limited structural and gene content variation in the genome of Candida glabrata. *BMC Biol.* **20**, 226 (2022).

162. T. Sekizuka, S. Iguchi, T. Umeyama, Y. Inamine, K. Makimura, M. Kuroda, Y. Miyazaki, K. Kikuchi, Clade II Candida auris possess genomic structural variations related to an ancestral strain. *PloS One*. **14**, e0223433 (2019).

163. M. P. Hirakawa, D. A. Martinez, S. Sakthikumar, M. Z. Anderson, A. Berlin, S. Gujja, Q. Zeng, E. Zisson, J. M. Wang, J. M. Greenberg, J. Berman, R. J. Bennett, C. A. Cuomo, Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res.* **25**, 413–425 (2015).

164. G. Janbon, F. Sherman, E. Rustchenko, Monosomy of a specific chromosome determines L-sorbose utilization: a novel regulatory mechanism in Candida albicans. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5150–5155 (1998).

165. R. J. Bennett, The Parasexual Lifestyle of Candida albicans. *Curr. Opin. Microbiol.* **28**, 10–17 (2015).

166. R. N. H. Seervai, S. K. Jones, M. P. Hirakawa, A. M. Porman, R. J. Bennett, Parasexuality and ploidy change in Candida tropicalis. *Eukaryot. Cell*. **12**, 1629–1640 (2013).

167. Y. Wang, J. Xu, Population genomic analyses reveal evidence for limited recombination in the superbug Candida auris in nature. *Comput. Struct. Biotechnol. J.* **20**, 3030–3040 (2022).

168. Z. K. Ross, A. Lorenz, Is Candida auris sexual? *PLoS Pathog.* **16**, e1009094 (2020).

169. R. E. Fundyga, R. J. Kuykendall, W. Lee-Yang, T. J. Lott, Evidence for aneuploidy and recombination in the human commensal yeast Candida parapsilosis. *Infect. Genet. Evol.* **4**, 37–43 (2004).

170. J. Xu, Is Natural Population of Candida tropicalis Sexual, Parasexual, and/or Asexual? *Front. Cell. Infect. Microbiol.* **11** (2021) (available at https://www.frontiersin.org/articles/10.3389/fcimb.2021.751676).

171. V. Mixão, T. Gabaldón, Genomic evidence for a hybrid origin of the yeast opportunistic pathogen Candida albicans. *BMC Biol.* **18**, 48 (2020).

172. V. Mixão, T. Gabaldón, Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast*. **35**, 5–20 (2018).

173. D. S. Perlin, Echinocandin Resistance in Candida. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **61 Suppl 6**, S612-617 (2015).

174. A. Lupetti, R. Danesi, M. Campa, M. D. Tacca, S. Kelly, Molecular basis of resistance to azole antifungals. *Trends Mol. Med.* **8**, 76–81 (2002).

175. L. Vale-Silva, F. Ischer, S. Leibundgut-Landmann, D. Sanglard, Gain-of-function mutations in PDR1, a regulator of antifungal drug resistance in Candida glabrata, control adherence to host cells. *Infect. Immun.* **81**, 1709–1720 (2013).

176. D. Cavalieri, M. Di Paola, L. Rizzetto, N. Tocci, C. De Filippo, P. Lionetti, A. Ardizzoni, B. Colombari, S. Paulone, I. G. Gut, L. Berná, M. Gut, J. Blanc, M. Kapushesky, E. Pericolini, E. Blasi, S. Peppoloni, Genomic and Phenotypic Variation in Morphogenetic Networks of Two Candida albicans Isolates Subtends Their Different Pathogenic Potential. *Front. Immunol.* **8** (2018) (available at https://www.frontiersin.org/articles/10.3389/fimmu.2017.01997).

177. J. F. Muñoz, L. Gade, N. A. Chow, V. N. Loparev, P. Juieng, E. L. Berkow, R. A. Farrer, A. P. Litvintseva, C. A. Cuomo, Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. *Nat. Commun.* **9**, 5346 (2018).

178. J. Gao, E. W. L. Chow, H. Wang, X. Xu, C. Cai, Y. Song, J. Wang, Y. Wang, LncRNA DINOR is a virulence factor and global regulator of stress responses in Candida auris. *Nat. Microbiol.* **6**, 842–851 (2021).

179. H.-H. Chou, H.-J. Lo, K.-W. Chen, M.-H. Liao, S.-Y. Li, Multilocus sequence typing of Candida tropicalis shows clonal cluster enriched in isolates with resistance or trailing growth of fluconazole. *Diagn. Microbiol. Infect. Dis.* **58**, 427–433 (2007).

180. M. Desnos-Ollivier, S. Bretagne, C. Bernède, V. Robert, D. Raoux, E. Chachaty, E. Forget, C. Lacroix, F. Dromer, Yeasts Group, Clonal population of flucytosine-resistant Candida tropicalis from blood cultures, Paris, France. *Emerg. Infect. Dis.* **14**, 557–565 (2008).

181. C. Biswas, V. R. Marcelino, S. Van Hal, C. Halliday, E. Martinez, Q. Wang, S. Kidd, K. Kennedy, D. Marriott, C. O. Morrissey, I. Arthur, K. Weeks, M. A. Slavin, T. C. Sorrell, V. Sintchenko, W. Meyer, S. C.-A. Chen, Whole Genome Sequencing of Australian Candida glabrata Isolates Reveals Genetic Diversity and Novel Sequence Types. *Front. Microbiol.* **9**, 2946 (2018).

182. L. Carreté, E. Ksiezopolska, E. Gómez-Molero, A. Angoulvant, O. Bader, C. Fairhead, T. Gabaldón, Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations. *Front. Microbiol.* **10**, 112 (2019).

183. E. Garbe, S. Vylkova, Role of Amino Acid Metabolism in the Virulence of Human Pathogenic Fungi. *Curr. Clin. Microbiol. Rep.* **6**, 108–119 (2019).

184. E. Sitterlé, A. T. Coste, T. Obadia, C. Maufrais, M. Chauvel, N. Sertour, D. Sanglard, A. Puel, C. D'Enfert, M.-E. Bougnoux, Large-scale genome mining allows identification of neutral polymorphisms and

novel resistance mutations in genes involved in Candida albicans resistance to azoles and echinocandins. *J. Antimicrob. Chemother.* **75**, 835–848 (2020).

185. O. Leiva-Peláez, G. Gutiérrez-Escobedo, E. López-Fuentes, J. Cruz-Mora, A. De Las Peñas, I. Castaño, Molecular characterization of the silencing complex SIR in Candida glabrata hyperadherent clinical isolates. *Fungal Genet. Biol.* **118**, 21–31 (2018).
186. J. E. San, S. Baichoo, A. Kanzi, Y. Moosa, R. Lessells, V. Fonseca, J. Mogaka, R. Power, T. de Oliveira, Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front. Microbiol.* **10**, 3119 (2019).
187. D. Sanglard, Finding the needle in a haystack: Mapping antifungal drug resistance in fungal pathogen by genomic approaches. *PLoS Pathog.* **15**, e1007478 (2019).
188. G. Povysil, S. Petrovski, J. Hostyk, V. Aggarwal, A. S. Allen, D. B. Goldstein, Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
189. R. A. Power, J. Parkhill, T. de Oliveira, Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2017).
190. S. Gundlach, J. C. Kässens, L. Wienbrandt, Genome-wide Association Interaction Studies with MB-MDR and maxT Multiple Testing Correction on FPGAs. *Procedia Comput. Sci.* **80**, 639–649 (2016).
191. Y. Swiel, J.-T. Brandenburg, M. Hayat, W. C. Chen, M. Cox, S. Hazelhurst, FPGA Acceleration of GWAS Permutation Testin (2022), , doi:10.1101/2022.03.11.483235.
192. K. Saund, E. S. Snitkin, Hogwash: three methods for genome-wide association studies in bacteria. *Microb. Genomics*. **6**, mgen000469 (2020).
193. P. E. Chen, B. J. Shapiro, The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25**, 17–24 (2015).
194. P. Chen, B. Shapiro, *Classic genome-wide association methods are unlikely to identify causal variants in strongly clonal microbial populations.* (2021).
195. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
196. C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, D. Heckerman, FaST linear mixed models for genome-wide association studies. *Nat. Methods*. **8**, 833–835 (2011).
197. A. Sánchez-Vallet, F. E. Hartmann, T. C. Marcel, D. Croll, Nature's genetic screens: using genome-wide association studies for effector discovery. *Mol. Plant Pathol.* **19**, 3–6 (2018).
198. M.-J. Xiang, J.-Y. Liu, P.-H. Ni, S. Wang, C. Shi, B. Wei, Y.-X. Ni, H.-L. Ge, Erg11 mutations associated with azole resistance in clinical isolates of Candida albicans. *FEMS Yeast Res.* **13**, 386–393 (2013).
199. J. Bergelson, F. Roux, Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat. Rev. Genet.* **11**, 867–879 (2010).
200. M. R. Farhat, B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, E. M.

Streicher, A. Calver, A. Sloutsky, D. Kaur, J. E. Posey, B. Plikaytis, M. R. Oggioni, J. L. Gardy, J. C. Johnston, M. Rodrigues, P. K. C. Tang, M. Kato-Maeda, M. L. Borowsky, B. Muddukrishna, B. N. Kreiswirth, N. Kurepina, J. Galagan, S. Gagneux, B. Birren, E. J. Rubin, E. S. Lander, P. C. Sabeti, M. Murray, Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45**, 1183–1189 (2013).

201. C. Collins, X. Didelot, A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* **14**, e1005958 (2018).

202. J. Corander, N. J. Croucher, S. R. Harris, J. A. Lees, G. Tonkin-Hill, "Bacterial Population Genomics" in *Handbook of Statistical Genomics* (John Wiley & Sons, Ltd, 2019; https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119487845.ch36), pp. 997–1020.

203. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

204. M. H. Guo, A. Dauber, M. F. Lippincott, Y.-M. Chan, R. M. Salem, J. N. Hirschhorn, Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *Am. J. Hum. Genet.* **99**, 527–539 (2016).

205. M. J. White, B. L. Yaspan, O. J. Veatch, P. Goddard, O. S. Risse-Adams, M. G. Contreras, Strategies for Pathway Analysis using GWAS and WGS Data. *Curr. Protoc. Hum. Genet.* **100**, e79 (2019).

206. K. Saund, Z. Lapp, S. N. Thiede, A. Pirani, E. S. Snitkin, prewas: data pre-processing for more informative bacterial GWAS. *Microb. Genomics*. **6** (2020), doi:10.1099/mgen.0.000368.

207. X. Guo, R. Zhang, Y. Li, Z. Wang, O. P. Ishchuk, K. M. Ahmad, J. Wee, J. Piskur, J. A. Shapiro, Z. Gu, Understand the genomic diversity and evolution of fungal pathogen Candida glabrata by genome-wide analysis of genetic variations. *Methods*. **176**, 82–90 (2020).

208. K. M. de San Vicente, M. S. Schröder, L. Lombardi, E. Iracane, G. Butler, Correlating Genotype and Phenotype in the Asexual Yeast Candida orthopsilosis Implicates ZCF29 in Sensitivity to Caffeine. *G3 Bethesda Md*. **9**, 3035–3043 (2019).

209. J. Rhodes, A. Abdolrasouli, K. Dunne, T. R. Sewell, Y. Zhang, E. Ballard, A. P. Brackin, N. van Rhijn, H. Chown, A. Tsitsopoulou, R. B. Posso, S. H. Chotirmall, N. G. McElvaney, P. G. Murphy, A. F. Talento, J. Renwick, P. S. Dyer, A. Szekely, P. Bowyer, M. J. Bromley, E. M. Johnson, P. Lewis White, A. Warris, R. C. Barton, S. Schelenz, T. R. Rogers, D. Armstrong-James, M. C. Fisher, Population genomics confirms acquisition of drug-resistant Aspergillus fumigatus infection by humans from the environment. *Nat. Microbiol.* **7**, 663–674 (2022).

210. J. F. Guerra Maldonado, A. T. Vincent, M. Chenal, F. J. Veyrier, CAPRIB: a user-friendly tool to study amino acid changes and selection for the exploration of intra-genus evolution. *BMC Genomics*. **21**, 832 (2020).

211. S. Mohammadi, A. Leduc, S. J. Charette, J. Barbeau, A. T. Vincent, Amino acid substitutions in specific

proteins correlate with farnesol unresponsiveness in Candida albicans. *BMC Genomics*. **24**, 93 (2023).

212. M. C. Derbyshire, Bioinformatic Detection of Positive Selection Pressure in Plant Pathogens: The Neutral Theory of Molecular Sequence Evolution in Action. *Front. Microbiol.* **11**, 644 (2020).
213. G. Aguileta, G. Refrégier, R. Yockteng, E. Fournier, T. Giraud, Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect. Genet. Evol.* **9**, 656–670 (2009).
214. Á. Chiner-Oms, M. G. López, M. Moreno-Molina, V. Furió, I. Comas, Gene evolutionary trajectories in Mycobacterium tuberculosis reveal temporal signs of selection. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2113600119 (2022).
215. J. M. Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
216. B. H, G. P, G. J, S. S, A. G, S. A, L. P. S, J. C, B. A, G. T, Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Mol. Ecol.* **26** (2017), doi:10.1111/mec.13976.
217. M. L. Bendall, S. L. Stevens, L.-K. Chan, S. Malfatti, P. Schwientek, J. Tremblay, W. Schackwitz, J. Martin, A. Pati, B. Bushnell, J. Froula, D. Kang, S. G. Tringe, S. Bertilsson, M. A. Moran, A. Shade, R. J. Newton, K. D. McMahon, R. R. Malmstrom, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
218. N. Takeuchi, O. X. Cordero, E. V. Koonin, K. Kaneko, Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* **13**, 20 (2015).
219. C. Kosiol, T. Vinař, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, A. Siepel, Patterns of Positive Selection in Six Mammalian Genomes. *PLOS Genet.* **4**, e1000144 (2008).
220. T. Miyata, T. Yasunaga, Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36 (1980).
221. Z. Yang, J. P. Bielawski, Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
222. S. Kryazhimskiy, J. B. Plotkin, The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
223. D. J. Wilson, CRyPTIC Consortium, GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes. *Mol. Biol. Evol.* **37**, 2450–2460 (2020).
224. P. C. Brunner, B. A. McDonald, Evolutionary analyses of the avirulence effector AvrStb6 in global populations of Zymoseptoria tritici identify candidate amino acids involved in recognition. *Mol. Plant Pathol.* **19**, 1836–1846 (2018).
225. N. Helmstetter, A. D. Chybowska, C. Delaney, A. Da Silva Dantas, H. Gifford, T. Wacker, C. Munro, A. Warris, B. Jones, C. A. Cuomo, D. Wilson, G. Ramage, R. A. Farrer, Population genetics and microevolution of clinical Candida glabrata reveals recombinant sequence types and hyper-variation within mitochondrial genomes, virulence genes, and drug targets. *Genetics*. **221**, iyac031 (2022).

226. M. Muñoz, L. M. Wintaco, S. A. Muñoz, J. D. Ramírez, Dissecting the Heterogeneous Population Genetic Structure of Candida albicans: Limitations and Constraints of the Multilocus Sequence Typing Scheme. *Front. Microbiol.* **10** (2019) (available at https://www.frontiersin.org/articles/10.3389/fmicb.2019.01052).

227. M. Norling, R. P. Bishop, R. Pelle, W. Qi, S. Henson, E. F. Drábek, K. Tretina, D. Odongo, S. Mwaura, T. Njoroge, E. Bongcam-Rudloff, C. A. Daubenberger, J. C. Silva, The genomes of three stocks comprising the most widely utilized live sporozoite Theileria parva vaccine exhibit very different degrees and patterns of sequence divergence. *BMC Genomics*. **16**, 729 (2015).

228. J. Mu, P. Awadalla, J. Duan, K. M. McGee, J. Keebler, K. Seydel, G. A. T. McVean, X. Su, Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nat. Genet.* **39**, 126–130 (2007).

229. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. **351**, 652–654 (1991).

230. S. L. Simmons, G. DiBartolo, V. J. Denef, D. S. A. Goltsman, M. P. Thelen, J. F. Banfield, Population Genomic Analysis of Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLOS Biol.* **6**, e177 (2008).

231. C. W. Nelson, A. L. Hughes, Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing. *Infect. Genet. Evol.* **30**, 1–7 (2015).

232. Q. Wu, J. Patocka, E. Nepovimova, K. Kuca, A Review on the Synthesis and Bioactivity Aspects of Beauvericin, a Fusarium Mycotoxin. *Front. Pharmacol.* **9** (2018) (available at https://www.frontiersin.org/articles/10.3389/fphar.2018.01338).

233. S. F. Elena, R. E. Lenski, Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).

234. M. Cavalheiro, C. Costa, A. Silva-Dias, I. M. Miranda, C. Wang, P. Pais, S. N. Pinto, D. Mil-Homens, M. Sato-Okamoto, A. Takahashi-Nakaguchi, R. M. Silva, N. P. Mira, A. M. Fialho, H. Chibana, A. G. Rodrigues, G. Butler, M. C. Teixeira, A Transcriptomics Approach To Unveiling the Mechanisms of In Vitro Evolution towards Fluconazole Resistance of a Candida glabrata Clinical Isolate. *Antimicrob. Agents Chemother.* **63**, e00995-18 (2019).

235. J. Branco, M. Ola, R. M. Silva, E. Fonseca, N. C. Gomes, C. Martins-Cruz, A. P. Silva, A. Silva-Dias, C. Pina-Vaz, C. Erraught, L. Brennan, A. G. Rodrigues, G. Butler, I. M. Miranda, *Clin. Microbiol. Infect.*, in press, doi:10.1016/j.cmi.2017.02.002.

236. R. T. Todd, T. D. Wikoff, A. Forche, A. Selmecki, Genome plasticity in Candida albicans is driven by long repeat sequences. *eLife*. **8**, e45954 (2019).

237. C. Rafaluk, G. Jansen, H. Schulenburg, G. Joop, When experimental selection for virulence leads to loss of virulence. *Trends Parasitol.* **31**, 426–434 (2015).

238. A. Forche, P. T. Magee, A. Selmecki, J. Berman, G. May, Evolution in Candida albicans Populations

During a Single Passage Through a Mouse Host. *Genetics*. **182**, 799–811 (2009).

239. S. F. Bailey, F. Blanquart, T. Bataillon, R. Kassen, What drives parallel evolution? *BioEssays*. **39**, e201600176 (2017).

240. R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, E. Banks, Scaling accurate genetic variant discovery to tens of thousands of samples (2018), p. 201178.

241. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing (2012), , doi:10.48550/arXiv.1207.3907.

242. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

243. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. **6**, 80–92 (2012).

244. D. A. Abbey, J. Funt, M. N. Lurie-Weinberger, D. A. Thompson, A. Regev, C. L. Myers, J. Berman, YMAP: a pipeline for visualization of copy number variation and loss of heterozygosity in eukaryotic pathogens. *Genome Med.* **6**, 100 (2014).

245. E. Ksiezopolska, T. Gabaldón, Evolutionary Emergence of Drug Resistance in Candida Opportunistic Pathogens. *Genes*. **9**, 461 (2018).

246. S. Costa-de-Oliveira, A. G. Rodrigues, Candida albicans Antifungal Resistance and Tolerance in Bloodstream Infections: The Triad Yeast-Host-Antifungal. *Microorganisms*. **8**, 154 (2020).

247. K. Spettel, W. Barousch, A. Makristathis, I. Zeller, M. Nehr, B. Selitsch, M. Lackner, P.-M. Rath, J. Steinmann, B. Willinger, Analysis of antifungal resistance genes in Candida albicans and Candida glabrata using next generation sequencing. *PloS One*. **14**, e0210397 (2019).

248. B. G. Vu, M. A. Stamnes, Y. Li, P. D. Rogers, W. S. Moye-Rowley, The Candida glabrata Upc2A transcription factor is a global regulator of antifungal drug resistance pathways. *PLoS Genet.* **17**, e1009582 (2021).

249. L. S. Burrack, R. T. Todd, N. Soisangwan, N. P. Wiederhold, A. Selmecki, Genomic Diversity across Candida auris Clinical Isolates Shapes Rapid Development of Antifungal Resistance In Vitro and In Vivo. *mBio*. **13**, e00842-22 (2022).

250. G. F. Ribeiro, E. Denes, H. Heaney, D. S. Childers, What 'Omics can tell us about antifungal adaptation. *FEMS Yeast Res.* **21**, foab070 (2021).

251. Q. Ni, C. Wang, Y. Tian, D. Dong, C. Jiang, E. Mao, Y. Peng, CgPDR1 gain-of-function mutations lead to azole-resistance and increased adhesion in clinical Candida glabrata strains. *Mycoses*. **61**, 430–440 (2018).

252. F. Daneshnia, S. Hilmioğlu Polat, M. Ilkit, E. Shor, J. N. de Almeida Júnior, L. M. Favarello, A. L.

Colombo, A. Arastehfar, D. S. Perlin, Determinants of fluconazole resistance and the efficacy of fluconazole and milbemycin oxim combination against Candida parapsilosis clinical isolates from Brazil and Turkey. *Front. Fungal Biol.* **3** (2022) (available at https://www.frontiersin.org/articles/10.3389/ffunb.2022.906681).

253. F. Barchiesi, D. Calabrese, D. Sanglard, L. Falconi Di Francesco, F. Caselli, D. Giannini, A. Giacometti, S. Gavaudan, G. Scalise, Experimental Induction of Fluconazole Resistance in Candida tropicalis ATCC 750. *Antimicrob. Agents Chemother.* **44**, 1578–1584 (2000).

254. J. Li, A. T. Coste, D. Bachmann, D. Sanglard, F. Lamoth, Deciphering the Mrr1/Mdr1 Pathway in Azole Resistance of Candida auris. *Antimicrob. Agents Chemother.* **66**, e0006722 (2022).

255. S. D. Singh-Babak, T. Babak, S. Diezmann, J. A. Hill, J. L. Xie, Y.-L. Chen, S. M. Poutanen, R. P. Rennie, J. Heitman, L. E. Cowen, Global analysis of the evolution and mechanism of echinocandin resistance in Candida glabrata. *PLoS Pathog.* **8**, e1002718 (2012).

256. J. M. Rybak, C. M. Dickens, J. E. Parker, K. E. Caudle, K. Manigaba, S. G. Whaley, A. T. Nishimoto, A. Luna-Tapia, S. Roy, Q. Zhang, K. S. Barker, G. E. Palmer, T. R. Sutter, R. Homayouni, N. P. Wiederhold, S. L. Kelly, P. D. Rogers, Loss of C-5 Sterol Desaturase Activity Results in Increased Resistance to Azole and Echinocandin Antifungals in a Clinical Isolate of Candida parapsilosis. *Antimicrob. Agents Chemother.* **61**, e00651-17 (2017).

257. F. Chaabane, A. Graf, L. Jequier, A. T. Coste, Review on Antifungal Resistance Mechanisms in the Emerging Pathogen Candida auris. *Front. Microbiol.* **10**, 2788 (2019).

258. P. Escandón, N. A. Chow, D. H. Caceres, L. Gade, E. L. Berkow, P. Armstrong, S. Rivera, E. Misas, C. Duarte, H. Moulton-Meissner, R. M. Welsh, C. Parra, L. A. Pescador, N. Villalobos, S. Salcedo, I. Berrio, C. Varón, A. Espinosa-Bode, S. R. Lockhart, B. R. Jackson, A. P. Litvintseva, M. Beltran, T. M. Chiller, Molecular Epidemiology of Candida auris in Colombia Reveals a Highly Related, Countrywide Colonization With Regional Patterns in Amphotericin B Resistance. *Clin. Infect. Dis.* **68**, 15–21 (2019).

259. A. Coste, V. Turner, F. Ischer, J. Morschhäuser, A. Forche, A. Selmecki, J. Berman, J. Bille, D. Sanglard, A Mutation in Tac1p, a Transcription Factor Regulating CDR1 and CDR2, Is Coupled With Loss of Heterozygosity at Chromosome 5 to Mediate Antifungal Resistance in Candida albicans. *Genetics*. **172**, 2139 (2006).

260. A. Selmecki, A. Forche, J. Berman, Aneuploidy and isochromosome formation in drug-resistant Candida albicans. *Science*. **313**, 367–370 (2006).

261. J. Berman, Ploidy plasticity: a rapid and reversible strategy for adaptation to stress. *FEMS Yeast Res.* **16**, fow020 (2016).

262. P. Marichal, H. Vanden Bossche, F. C. Odds, G. Nobels, D. W. Warnock, V. Timmerman, C. Van Broeckhoven, S. Fay, P. Mose-Larsen, Molecular biological characterization of an azole-resistant Candida glabrata isolate. *Antimicrob. Agents Chemother.* **41**, 2229–2237 (1997).

263. S. Poláková, C. Blume, J. A. Zárate, M. Mentel, D. Jørck-Ramberg, J. Stenderup, J. Piskur, Formation of

new chromosomes as a virulence mechanism in yeast Candida glabrata. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2688–2693 (2009).

264. F. Yang, F. Teoh, A. S. M. Tan, Y. Cao, N. Pavelka, J. Berman, Aneuploidy Enables Cross-Adaptation to Unrelated Drugs. *Mol. Biol. Evol.* **36**, 1768–1782 (2019).

265. F. Yang, R. T. Todd, A. Selmecki, Y.-Y. Jiang, Y.-B. Cao, J. Berman, The fitness costs and benefits of trisomy of each Candida albicans chromosome. *Genetics*. **218**, iyab056 (2021).

266. P. Kakade, S. Sircaik, C. Maufrais, I. V. Ene, R. J. Bennett, Aneuploidy and gene dosage regulate filamentation and host colonization by Candida albicans. *Proc. Natl. Acad. Sci.* **120**, e2218163120 (2023).

267. E. L. Berkow, S. R. Lockhart, Fluconazole resistance in Candida species: a current perspective. *Infect. Drug Resist.* **10**, 237–245 (2017).

268. R. T. Todd, A. Selmecki, Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. *eLife*. **9**, e58349 (2020).

269. J. Bing, Z. Guan, T. Zheng, Z. Zhang, S. Fan, C. L. Ennis, C. J. Nobile, G. Huang, Clinical isolates of Candida auris with enhanced adherence and biofilm formation due to genomic amplification of ALS4. *PLOS Pathog.* **19**, e1011239 (2023).

270. B. Bakker, A. Taudt, M. E. Belderbos, D. Porubsky, D. C. J. Spierings, T. V. de Jong, N. Halsema, H. G. Kazemier, K. Hoekstra-Wakker, A. Bradley, E. S. J. M. de Bont, A. van den Berg, V. Guryev, P. M. Lansdorp, M. Colomé-Tatché, F. Foijer, Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* **17**, 115 (2016).

271. J. M. Kebschull, A. M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143 (2015).

272. J. P. Szatkiewicz, W. Wang, P. F. Sullivan, W. Wang, W. Sun, Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic Acids Res.* **41**, 1519–1532 (2013).

273. B. Gallone, J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, C. Teiling, B. Steffy, M. Taylor, A. Schwartz, T. Richardson, C. White, G. Baele, S. Maere, K. J. Verstrepen, Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell*. **166**, 1397-1410.e16 (2016).

274. M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, F. J. Sedlazeck, Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).

275. D. L. Cameron, J. Baber, C. Shale, J. E. Valle-Inclan, N. Besselink, A. van Hoeck, R. Janssen, E. Cuppen, P. Priestley, A. T. Papenfuss, GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 202 (2021).

276. J. Schröder, A. Wirawan, B. Schmidt, A. T. Papenfuss, CLOVE: classification of genomic fusions into structural variation events. *BMC Bioinformatics*. **18**, 346 (2017).

277. D. L. Cameron, L. Di Stefano, A. T. Papenfuss, Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.* **10**, 3240 (2019).

278. J. F. Muñoz, R. M. Welsh, T. Shea, D. Batra, L. Gade, D. Howard, L. A. Rowe, J. F. Meis, A. P. Litvintseva, C. A. Cuomo, Clade-specific chromosomal rearrangements and loss of subtelomeric adhesins in Candida auris. *Genetics*. **218**, iyab029 (2021).

279. S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, Y. Kamatani, Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).

280. A. Ameur, W. P. Kloosterman, M. S. Hestand, Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnol.* **37**, 72–85 (2019).

281. M. U. Ahsan, Q. Liu, L. Fang, K. Wang, NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biol.* **22**, 261 (2021).

282. S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, Q. Gouil, Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

283. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).

284. D. Merkel, *Linux J.*, in press.

285. G. M. Kurtzer, V. Sochat, M. W. Bauer, Singularity: Scientific containers for mobility of compute. *PLOS ONE*. **12**, e0177459 (2017).

286. M. A. Ellison, J. L. Walker, P. J. Ropp, J. D. Durrant, K. M. Arndt, MutantHuntWGS: A Pipeline for Identifying Saccharomyces cerevisiae Mutations. *G3 GenesGenomesGenetics*. **10**, 3009–3014 (2020).

287. S. J. N. Duxbury, S. Bates, R. E. Beardmore, I. Gudelj, Evolution of drug-resistant and virulent small colonies in phenotypically diverse populations of the human fungal pathogen Candida glabrata. *Proc. Biol. Sci.* **287**, 20200761 (2020).

288. O. Avramovska, A. C. Smith, E. Rego, M. A. Hickman, Tetraploidy accelerates adaptation under drug selection in a fungal pathogen. *Front. Fungal Biol.* **3** (2022) (available at https://www.frontiersin.org/articles/10.3389/ffunb.2022.984377).

289. R. Leinonen, H. Sugawara, M. Shumway, International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).

290. N. Zhang, T. Huang, Y.-D. Cai, Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol. Genet. Genomics*. **290**, 343–352 (2015).

291. M. Sautour, J.-P. Lemaître, L. Ranjard, C. Truntzer, L. Basmaciyan, G. Depret, A. Hartmann, F. Dalle, Detection and survival of Candida albicans in soils. *Environ. DNA*. **3**, 1093–1101 (2021).

292. A. W. C. Pang, O. Migita, J. R. MacDonald, L. Feuk, S. W. Scherer, Mechanisms of Formation of

Structural Variation in a Fully Sequenced Human Genome. *Hum. Mutat.* **34**, 345–354 (2013).

293. J. Valls-Margarit, I. Galván-Femenía, D. Matías-Sánchez, N. Blay, M. Puiggròs, A. Carreras, C. Salvoro, B. Cortés, R. Amela, X. Farre, J. Lerga-Jaso, M. Puig, J. F. Sánchez-Herrero, V. Moreno, M. Perucho, L. Sumoy, L. Armengol, O. Delaneau, M. Cáceres, R. de Cid, D. Torrents, GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res.* **50**, 2464–2479 (2022).

294. X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, C. T. Saunders, Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma. Oxf. Engl.* **32**, 1220–1222 (2016).

295. L. Czech, M. Exposito-Alonso, grenepipe: a flexible, scalable and reproducible pipeline to automate variant calling from sequence reads. *Bioinformatics*. **38**, 4809–4811 (2022).

296. A. Iacoangeli, A. Al Khleifat, W. Sproviero, A. Shatunov, A. R. Jones, S. Opie-Martin, E. Naselli, S. D. Topp, I. Fogh, A. Hodges, R. J. Dobson, S. J. Newhouse, A. Al-Chalabi, ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients. *Amyotroph. Lateral Scler. Front. Degener.* **20**, 207–215 (2019).

297. A. Al Khleifat, A. Iacoangeli, J. J. F. A. van Vugt, H. Bowles, M. Moisse, R. A. J. Zwamborn, R. A. A. van der Spek, A. Shatunov, J. Cooper-Knock, S. Topp, R. Byrne, C. Gellera, V. López, A. R. Jones, S. Opie-Martin, A. Vural, Y. Campos, W. van Rheenen, B. Kenna, K. R. Van Eijk, K. Kenna, M. Weber, B. Smith, I. Fogh, V. Silani, K. E. Morrison, R. Dobson, M. A. van Es, R. L. McLaughlin, P. Vourc'h, A. Chio, P. Corcia, M. de Carvalho, M. Gotkine, M. P. Panades, J. S. Mora, P. J. Shaw, J. E. Landers, J. D. Glass, C. E. Shaw, N. Basak, O. Hardiman, W. Robberecht, P. Van Damme, L. H. van den Berg, J. H. Veldink, A. Al-Chalabi, Structural variation analysis of 6,500 whole genome sequences in amyotrophic lateral sclerosis. *Npj Genomic Med.* **7**, 1–8 (2022).

298. D. C. Koboldt, Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).

299. X. Li, J. F. Muñoz, L. Gade, S. Argimon, M.-E. Bougnoux, J. R. Bowers, N. A. Chow, I. Cuesta, R. A. Farrer, C. Maufrais, J. Monroy-Nieto, D. Pradhan, J. Uehling, D. Vu, C. A. Yeats, D. M. Aanensen, C. d'Enfert, D. M. Engelthaler, D. W. Eyre, M. C. Fisher, F. Hagen, W. Meyer, G. Singh, A. Alastruey-Izquierdo, A. P. Litvintseva, C. A. Cuomo, Comparing genomic variant identification protocols for Candida auris. *Microb. Genomics*. **9**, 000979 (2023).

300. D. J. Kvitek, G. Sherlock, Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9**, e1003972 (2013).

301. Y.-C. Wei, G.-H. Huang, CONY: A Bayesian procedure for detecting copy number variations from sequencing read depths. *Sci. Rep.* **10**, 10493 (2020).

302. S. P. Shah, X. Xuan, R. J. DeLeeuw, M. Khojasteh, W. L. Lam, R. Ng, K. P. Murphy, Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinforma. Oxf. Engl.* **22**, e431-439 (2006).

303. M. Coutelier, M. Holtgrewe, M. Jäger, R. Flöttman, M. A. Mensah, M. Spielmann, P. Krawitz, D. Horn, D. Beule, S. Mundlos, Combining callers improves the detection of copy number variants from whole-genome sequencing. *Eur. J. Hum. Genet.* **30**, 178–186 (2022).

304. G. F. Gao, C. Oh, G. Saksena, D. Deng, L. C. Westlake, B. A. Hill, M. Reich, S. E. Schumacher, A. C. Berger, S. L. Carter, A. D. Cherniack, M. Meyerson, B. Tabak, R. Beroukhim, G. Getz, Tangent normalization for somatic copy-number inference in cancer genome analysis. *Bioinformatics*. **38**, 4677–4686 (2022).

305. L. Zhang, W. Bai, N. Yuan, Z. Du, Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput. Biol.* **15**, e1007069 (2019).

306. S. Chen, P. Krusche, E. Dolzhenko, R. M. Sherman, R. Petrovski, F. Schlesinger, M. Kirsche, D. R. Bentley, M. C. Schatz, F. J. Sedlazeck, M. A. Eberle, Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).

307. M. Suvakov, A. Panda, C. Diesh, I. Holmes, A. Abyzov, CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *GigaScience*. **10**, giab074 (2021).

308. N. Robbins, L. E. Cowen, Antifungal drug resistance: Deciphering the mechanisms governing multidrug resistance in the fungal pathogen Candida glabrata. *Curr. Biol.* **31**, R1520–R1523 (2021).

309. X. Hou, K. R. Healey, E. Shor, M. Kordalewska, C. J. Ortigosa, P. Paderu, M. Xiao, H. Wang, Y. Zhao, L.-Y. Lin, Y.-H. Zhang, Y.-Z. Li, Y.-C. Xu, D. S. Perlin, Y. Zhao, Novel FKS1 and FKS2 modifications in a high-level echinocandin resistant clinical isolate of Candida glabrata. *Emerg. Microbes Infect.* **8**, 1619–1625 (2019).

310. S. Ferrari, F. Ischer, D. Calabrese, B. Posteraro, M. Sanguinetti, G. Fadda, B. Rohde, C. Bauser, O. Bader, D. Sanglard, Gain of function mutations in CgPDR1 of Candida glabrata not only mediate antifungal resistance but also enhance virulence. *PLoS Pathog.* **5**, e1000268 (2009).

311. C. Munck, H. K. Gumpert, A. I. Nilsson Wallin, H. H. Wang, M. O. A. Sommer, Prediction of resistance development against drug combinations by collateral responses to component drugs. *Sci. Transl. Med.* **6**, 262ra156 (2014).

312. M. Spitzer, N. Robbins, G. D. Wright, Combinatorial strategies for combating invasive fungal infections. *Virulence*. **8**, 169–185 (2016).

313. C. M. Hull, J. E. Parker, O. Bader, M. Weig, U. Gross, A. G. S. Warrilow, D. E. Kelly, S. L. Kelly, Facultative sterol uptake in an ergosterol-deficient clinical isolate of Candida glabrata harboring a missense mutation in ERG11 and exhibiting cross-resistance to azoles and amphotericin B. *Antimicrob. Agents Chemother.* **56**, 4223–4232 (2012).

314. H.-F. Tsai, A. A. Krol, K. E. Sarti, J. E. Bennett, Candida glabrata PDR1, a transcriptional regulator of a pleiotropic drug resistance network, mediates azole resistance in clinical isolates and petite mutants. *Antimicrob. Agents Chemother.* **50**, 1384–1392 (2006).

315. C. Costa, J. Ribeiro, I. M. Miranda, A. Silva-Dias, M. Cavalheiro, S. Costa-de-Oliveira, A. G. Rodrigues, M. C. Teixeira, Clotrimazole Drug Resistance in Candida glabrata Clinical Isolates Correlates with Increased Expression of the Drug:H(+) Antiporters CgAqr1, CgTpo1_1, CgTpo3, and CgQdr2. *Front. Microbiol.* **7**, 526 (2016).

316. S. K. Katiyar, A. Alastruey-Izquierdo, K. R. Healey, M. E. Johnson, D. S. Perlin, T. D. Edlind, Fks1 and Fks2 are functionally redundant but differentially regulated in Candida glabrata: implications for echinocandin resistance. *Antimicrob. Agents Chemother.* **56**, 6304–6309 (2012).

317. K. E. Baker, R. Parker, Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.* **16**, 293–299 (2004).

318. D. J. Krysan, The cell wall and endoplasmic reticulum stress responses are coordinately regulated in Saccharomyces cerevisiae. *Commun. Integr. Biol.* **2**, 233–235 (2009).

319. Q. Lv, L. Yan, Y. Jiang, The synthesis, regulation, and functions of sterols in Candida albicans: Well-known but still lots to learn. *Virulence*. **7**, 649–659 (2016).

320. T. Miyazaki, S. Kohno, ER stress response mechanisms in the pathogenic yeast Candida glabrata and their roles in virulence. *Virulence*. **5**, 365–370 (2014).

321. F. Parlati, M. Dominguez, J. J. Bergeron, D. Y. Thomas, Saccharomyces cerevisiae CNE1 encodes an endoplasmic reticulum (ER) membrane protein with sequence similarity to calnexin and calreticulin and functions as a constituent of the ER quality control apparatus. *J. Biol. Chem.* **270**, 244–253 (1995).

322. Y. Tanaka, M. Sasaki, F. Ito, T. Aoyama, M. Sato-Okamoto, A. Takahashi-Nakaguchi, H. Chibana, N. Shibata, Cooperation between ER stress and calcineurin signaling contributes to the maintenance of cell wall integrity in Candida glabrata. *Fungal Biol.* **122**, 19–33 (2018).

323. C. Hatwig, E. A. Balbueno, V. Z. Bergamo, B. Pippi, A. M. Fuentefria, G. P. Silveira, Multidrug-resistant *Candida glabrata* strains obtained by induction of anidulafungin resistance in planktonic and biofilm cells. *Braz. J. Pharm. Sci.* **55**, e18025 (2019).

324. A. Geber, C. A. Hitchcock, J. E. Swartz, F. S. Pullen, K. E. Marsden, K. J. Kwon-Chung, J. E. Bennett, Deletion of the Candida glabrata ERG3 and ERG11 genes: effect on cell viability, cell growth, sterol composition, and antifungal susceptibility. *Antimicrob. Agents Chemother.* **39**, 2708–2717 (1995).

325. L. Trovato, G. Scalia, M. Domina, S. Oliveri, Environmental Isolates of Multi-Azole-Resistant Aspergillus spp. in Southern Italy. *J. Fungi*. **4**, 131 (2018).

326. K. Zomorodian, A. Bandegani, H. Mirhendi, K. Pakshir, N. Alinejhad, A. Poostforoush Fard, In Vitro Susceptibility and Trailing Growth Effect of Clinical Isolates of Candida Species to Azole Drugs. *Jundishapur J. Microbiol.* **9**, e28666 (2016).

327. J. Cui, B. Ren, Y. Tong, H. Dai, L. Zhang, Synergistic combinations of antifungals and anti-virulence agents to fight against Candida albicans. *Virulence*. **6**, 362–371 (2015).

328. Q. Z. Jaber, M. Bibi, E. Ksiezopolska, T. Gabaldon, J. Berman, M. Fridman, Elevated Vacuolar Uptake of

Fluorescently Labeled Antifungal Drug Caspofungin Predicts Echinocandin Resistance in Pathogenic Yeast. *ACS Cent. Sci.* **6**, 1698–1712 (2020).

329. D. Logviniuk, Q. Z. Jaber, R. Dobrovetsky, N. Kozer, E. Ksiezopolska, T. Gabaldón, S. Carmeli, M. Fridman, Benzylic Dehydroxylation of Echinocandin Antifungal Drugs Restores Efficacy against Resistance Conferred by Mutated Glucan Synthase. *J. Am. Chem. Soc.* **144**, 5965–5975 (2022).
330. A. N. Nguyen Ba, I. Cvijović, J. I. Rojas Echenique, K. R. Lawrence, A. Rego-Costa, X. Liu, S. F. Levy, M. M. Desai, High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*. **575**, 494–499 (2019).
331. R. T. Todd, N. Soisangwan, S. Peters, B. Kemp, T. Crooks, A. Gerstein, A. Selmecki, Antifungal Drug Concentration Impacts the Spectrum of Adaptive Mutations in Candida albicans. *Mol. Biol. Evol.* **40**, msad009 (2023).
332. R. Hale, J. R. Morrongiello, S. E. Swearer, Evolutionary traps and range shifts in a rapidly changing world. *Biol. Lett.* **12**, 20160003 (2016).
333. D. Li, Y. Wang, W. Hu, F. Chen, J. Zhao, X. Chen, L. Han, Application of Machine Learning Classifier to Candida auris Drug Resistance Analysis. *Front. Cell. Infect. Microbiol.* **11**, 742062 (2021).
334. Y.-F. Lin, J.-J. Liu, Y.-J. Chang, C.-S. Yu, W. Yi, H.-Y. Lane, C.-H. Lu, Predicting Anticancer Drug Resistance Mediated by Mutations. *Pharmaceuticals*. **15**, 136 (2022).
335. G. Pines, R. G. Fankhauser, C. A. Eckert, Predicting Drug Resistance Using Deep Mutational Scanning. *Molecules*. **25**, 2265 (2020).
336. P. T. West, S. L. Peters, M. R. Olm, F. B. Yu, H. Gause, Y. C. Lou, B. A. Firek, R. Baker, A. D. Johnson, M. J. Morowitz, R. L. Hettich, J. F. Banfield, Genetic and behavioral adaptation of Candida parapsilosis to the microbiome of hospitalized infants revealed by in situ genomics, transcriptomics, and proteomics. *Microbiome*. **9**, 142 (2021).
337. C. Valiente-Mullor, B. Beamud, I. Ansari, C. Francés-Cuesta, N. García-González, L. Mejía, P. Ruiz-Hueso, F. González-Candelas, One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol.* **17**, e1008678 (2021).
338. X. Ma, M. V. Rogacheva, K. T. Nishant, S. Zanders, C. D. Bustamante, E. Alani, Mutation hotspots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Rep.* **1**, 36–42 (2012).
339. R. R. Fuentes, D. Chebotarov, J. Duitama, S. Smith, J. F. De la Hoz, M. Mohiyuddin, R. A. Wing, K. L. McNally, T. Tatarinova, A. Grigoriev, R. Mauleon, N. Alexandrov, Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880 (2019).
340. V. Hill, C. Ruis, S. Bajaj, O. G. Pybus, M. U. G. Kraemer, Progress and challenges in virus genomic epidemiology. *Trends Parasitol.* **37**, 1038–1049 (2021).
341. D. Bensasson, J. Dicks, J. M. Ludwig, C. J. Bond, A. Elliston, I. N. Roberts, S. A. James, Diverse Lineages of Candida albicans Live on Old Oaks. *Genetics*. **211**, 277–288 (2019).
342. M. Pekmezovic, H. Hovhannisyan, M. S. Gresnigt, E. Iracane, J. Oliveira-Pacheco, S. Siscar-Lewin, E.

Seemann, B. Qualmann, T. Kalkreuter, S. Müller, T. Kamradt, S. Mogavero, S. Brunke, G. Butler, T. Gabaldón, B. Hube, Candida pathogens induce protective mitochondria-associated type I interferon signalling and a damage-driven response in vaginal epithelial cells. *Nat. Microbiol.* **6**, 643–657 (2021).

343. S. Brunke, K. Seider, D. Fischer, I. D. Jacobsen, L. Kasper, N. Jablonowski, A. Wartenberg, O. Bader, A. Enache-Angoulvant, M. Schaller, C. d'Enfert, B. Hube, One small step for a yeast--microevolution within macrophages renders Candida glabrata hypervirulent due to a single point mutation. *PLoS Pathog.* **10**, e1004478 (2014).

344. M. N. Rai, R. Rai, P. Sethiya, C. Parsania, Transcriptome analysis reveals a common adaptive transcriptional response of Candida glabrata to diverse environmental stresses. *Res. Microbiol.*, 104073 (2023).

345. P. Bhakt, R. Shivarathri, D. K. Choudhary, S. Borah, R. Kaur, Fluconazole-induced actin cytoskeleton remodeling requires phosphatidylinositol 3-phosphate 5-kinase in the pathogenic yeast Candida glabrata. *Mol. Microbiol.* **110**, 425–443 (2018).

346. J. B. Moseley, B. L. Goode, The Yeast Actin Cytoskeleton: from Cellular Function to Biochemical Mechanism. *Microbiol. Mol. Biol. Rev.* **70**, 605–645 (2006).

347. P. Jain, I. Akula, T. Edlind, Cyclic AMP Signaling Pathway Modulates Susceptibility of Candida Species and Saccharomyces cerevisiae to Antifungal Azoles and Other Sterol Biosynthesis Inhibitors. *Antimicrob. Agents Chemother.* **47**, 3195–3201 (2003).

348. D. Zamith-Miranda, R. F. Amatuzzi, I. F. Munhoz da Rocha, S. T. Martins, A. C. R. Lucena, A. Z. Vieira, G. Trentin, F. Almeida, M. L. Rodrigues, E. S. Nakayasu, J. D. Nosanchuk, L. R. Alves, Transcriptional and translational landscape of Candida auris in response to caspofungin. *Comput. Struct. Biotechnol. J.* **19**, 5264–5277 (2021).

349. A. Bruckmann, W. Künkel, A. Härtl, R. Wetzker, R. Eck, A phosphatidylinositol 3-kinase of Candida albicans influences adhesion, filamentous growth and virulence. *Microbiol. Read. Engl.* **146 (Pt 11)**, 2755–2764 (2000).

350. Y. Wang, Y. Zou, X. Chen, H. Li, Z. Yin, B. Zhang, Y. Xu, Y. Zhang, R. Zhang, X. Huang, W. Yang, C. Xu, T. Jiang, Q. Tang, Z. Zhou, Y. Ji, Y. Liu, L. Hu, J. Zhou, Y. Zhou, J. Zhao, N. Liu, G. Huang, H. Chang, W. Fang, C. Chen, D. Zhou, Innate immune responses against the fungal pathogen Candida auris. *Nat. Commun.* **13**, 3553 (2022).

351. D. Casagrande Pierantoni, L. Corte, A. Casadevall, V. Robert, G. Cardinali, C. Tascini, How does temperature trigger biofilm adhesion and growth in Candida albicans and two non-Candida albicans Candida species? *Mycoses*. **64**, 1412–1421 (2021).

352. T. Ma, Q. Yu, C. Ma, X. Mao, Y. Liu, X. Peng, M. Li, Role of the inositol polyphosphate kinase Vip1 in autophagy and pathogenesis in Candida albicans. *Future Microbiol.* **15**, 1363–1377 (2020).

353. C. Garnaud, M. Champleboux, D. Maubon, M. Cornet, J. Govin, Histone Deacetylases and Their Inhibition in Candida Species. *Front. Microbiol.* **7**, 1238 (2016).

354. H. M. Mora-Montes, P. Ponce-Noyola, J. C. Villagómez-Castro, N. A. Gow, A. Flores-Carreón, E. López-Romero, Protein glycosylation in Candida. *Future Microbiol.* **4**, 1167–1183 (2009).

355. M. Van Ende, S. Wijnants, P. Van Dijck, Sugar Sensing and Signaling in Candida albicans and Candida glabrata. *Front. Microbiol.* **10** (2019) (available at https://www.frontiersin.org/articles/10.3389/fmicb.2019.00099).

356. H. Liu, J. Zhang, The rate and molecular spectrum of mutation are selectively maintained in yeast. *Nat. Commun.* **12**, 4044 (2021).

357. F. Daneshnia, J. N. de Almeida Júnior, M. Ilkit, L. Lombardi, A. M. Perry, M. Gao, C. J. Nobile, M. Egger, D. S. Perlin, B. Zhai, T. M. Hohl, T. Gabaldón, A. L. Colombo, M. Hoenigl, A. Arastehfar, Worldwide emergence of fluconazole-resistant Candida parapsilosis: current framework and future research roadmap. *Lancet Microbe* (2023), doi:10.1016/S2666-5247(23)00067-8.

358. B. Williamson, A. Wilk, K. D. Guerrero, T. D. Mikulski, T. N. Elias, I. Sawh, G. Cancino-Prado, D. Gardam, C. H. Heath, N. P. Govender, D. S. Perlin, M. Kordalewska, K. R. Healey, Impact of Erg11 Amino Acid Substitutions Identified in Candida auris Clade III Isolates on Triazole Drug Susceptibility. *Antimicrob. Agents Chemother.* **66**, e0162421 (2022).

359. J. M. Rybak, C. Sharma, L. A. Doorley, K. S. Barker, G. E. Palmer, P. D. Rogers, Delineation of the Direct Contribution of Candida auris ERG11 Mutations to Clinical Triazole Resistance. *Microbiol. Spectr.* **9**, e01585-21.

360. M. M. Matin, P. Matin, Md. R. Rahman, T. Ben Hadda, F. A. Almalki, S. Mahmud, M. M. Ghoneim, M. Alruwaily, S. Alshehri, Triazoles and Their Derivatives: Chemistry, Synthesis, and Therapeutic Applications. *Front. Mol. Biosci.* **9**, 864286 (2022).

361. D. Vandenbosch, E. De Canck, I. Dhondt, P. Rigole, H. J. Nelis, T. Coenye, Genomewide screening for genes involved in biofilm formation and miconazole susceptibility in Saccharomyces cerevisiae. *FEMS Yeast Res.* **13**, 720–730 (2013).

362. M. Guan, P. Xia, M. Tian, D. Chen, X. Zhang, Molecular fingerprints of conazoles via functional genomic profiling of Saccharomyces cerevisiae. *Toxicol. In Vitro*. **69**, 104998 (2020).

363. N. Zhang, B. B. Magee, P. T. Magee, B. R. Holland, E. Rodrigues, A. R. Holmes, R. D. Cannon, J. Schmid, Selective Advantages of a Parasexual Cycle for the Yeast Candida albicans. *Genetics*. **200**, 1117–1132 (2015).

364. D. Dylus, A. Altenhoff, S. Majidian, F. J. Sedlazeck, C. Dessimoz, Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.*, 1–9 (2023).

365. S. Sun, C. Roth, A. Floyd Averette, P. M. Magwene, J. Heitman, Epistatic genetic interactions govern morphogenesis during sexual reproduction and infection in a global human fungal pathogen. *Proc. Natl. Acad. Sci.* **119**, e2122293119 (2022).

366. J. A. Tom, J. Reeder, W. F. Forrest, R. R. Graham, J. Hunkapiller, T. W. Behrens, T. R. Bhangale, Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics*. **18**, 351 (2017).