Title: Population genetics and microevolution of clinical *Candida glabrata* reveals recombinant sequence types and hyper-variation within mitochondrial genomes, virulence genes and drug-targets

Authors: Nicolas Helmstetter[*,††], Aleksandra D. Chybowska[†,††], Christopher Delaney[‡], Alessandra Da Silva Dantas[*], Hugh Gifford[*], Theresa Wacker[*], Carol Munro[†], Adilia Warris[*], Brian Jones[§], Christina A. Cuomo[**], Duncan Wilson[*], Gordon Ramage[‡] and Rhys A. Farrer[*,**]

Authors institutional affiliations:

[*]Medical Research Council Centre for Medical Mycology at the University of Exeter, Exeter, UK, EX4 4QD.

[†]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK, AB25 2ZD.

[‡]School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. G12 8QQ.

[§]Institute of Infection, Immunity & Inflammation, University of Glasgow, UK. G12 8TA.

[**]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 02142.

[††]These authors contributed equally.

Raw sequences for all isolates of *C. glabrata* from this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA669061.

Short Running Title: Population genetics of *C. glabrata*

Keywords: *Candida glabrata*, genome sequencing, epidemiology, candidiasis, microevolution, mitochondria, drug-resistance, evolution.

Corresponding author: Rhys A. Farrer. MRC Centre for Medical Mycology University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK. T: +44(0)1392 727594; E: r.farrer@exeter.ac.uk

Abstract

Candida glabrata is the second most common etiological cause of worldwide systemic candidiasis in adult patients. Genome analysis of 68 isolates from 8 hospitals across Scotland, together with 83 global isolates, revealed insights into the population genetics and evolution of *C. glabrata*. Clinical isolates of *C. glabrata* from across Scotland are highly-genetically diverse, including at least 19 separate sequence types (STs) that have been recovered previously in globally diverse locations, and one newly discovered ST. Several STs had evidence for ancestral recombination, suggesting transmission between distinct geographical regions has coincided with genetic exchange arising in new clades. Three isolates were missing MATα1, potentially representing a second mating type. Signatures of positive selection were identified in every ST including enrichment for Epithelial Adhesins (EPA) thought to facilitate fungal adhesion to human epithelial cells. In patent microevolution was identified from seven sets of recurrent cases of candidiasis, revealing an enrichment for non-synonymous and frameshift indels in cell surface proteins. Microevolution within patients also affected EPA genes, and several genes involved in drug resistance including the ergosterol synthesis gene *ERG4* and the echinocandin target *FKS1*/2, the latter coinciding with a marked drop in fluconazole MIC. In addition to nuclear genome diversity, the *C. glabrata* mitochondrial genome was particularly diverse, appearing reduced in size and with fewer conserved protein encoding genes in all non-reference ST15 isolates. Together, this study highlights the genetic diversity present within the *C. glabrata* population that may impact virulence and drug resistance, and two major mechanisms generating this diversity: microevolution and genetic exchange/recombination.

Article Summary (80 words)

Candida glabrata is a leading human fungal pathogen worldwide, which is increasing in prevalence and evolving antifungal resistance. Here, we report the largest whole-genome sequencing project for *C. glabrata* to date based on clinically derived candidemia isolates from hospitals across Scotland in the United Kingdom. We discover evidence for a second mating type, evidence for recombination between sequence types, hyper-diverse mitochondrial genomes, signatures of positive selection in pathogenicity genes, and in patient microevolution of drug-resistance genes.

Introduction

Candida is the most prominent genus of the Debaryomycetaceae family, with over 400 genetically and phenotypically diverse species currently described [1]. Many of these species are harmless commensals of the mucous membranes and digestive tracts of healthy individuals. Approximately 30 *Candida* species are of clinical importance in humans. Most of these species that are capable of causing disease in humans belong to the CTG-Serine clade, including *C. albicans*, *C. dubliniensis*, *C. tropicalis*, *C. parapsilosis*, *C. lusitaniae*, *C. guilliermondii and C. auris*, while others such as *C. glabrata* and *C. bracarensis* belong to the genetically distant Nakaseomyces clade [1,2]. In adult patients, *C. glabrata* is the second most commonly isolated species after *C. albicans*, which together cause approximately three quarters of all systemic candidiasis [3,4]. Infections caused by these species range from mild vulvovaginal candidiasis (VVC or thrush) to severe, drug resistant and difficult to treat invasive infections affecting single organs or the blood stream

(candidemia) with or without dissemination to the heart, brain, kidneys and other parts of the body [5]. Bloodstream infections caused by *Candida spp.* are associated with mortality rates of 30-60% [6,7]. Candidemia is associated with diverse risk factors including neutropenia, chemotherapy, diabetes, old age, compromised immune function, prolonged antibiotic and steroid treatment, and intravenous catheters that can harbour fungal biofilms [8]. Pathogenic *Candida* species including *C. glabrata* have also exhibited alarming increases in resistance against all major classes of antifungal drugs, hindering effective treatments and resulting in increasing mortality rates [4,9–12].

C. glabrata typically grows in the yeast form and is considered to have evolved an infection strategy based on stealth and evasion without causing severe damage in murine models [13]. This ability of *C. glabrata* and some of its relatives in the Nakaseomyces clade to infect humans is thought to have evolved recently [14], as several of its closest relatives have to date been exclusively isolated environmentally (*C. castellii*, *N. baccilisporus*, *N. delphensis*) [14]. Pathogenicity in the Nakaseomyces correlates with the number of Epithelial Adhesins (*EPA*) encoded in their genomes, which facilitate adherence and colonisation of human epithelial cells [15]. In contrast to *C. albicans*, the pathogenicity of Nakaseomyces species does not coincide with number or presence of Phospholipase-B and Superoxide Dismutase genes [16–19]. Many *Candida* genes involved in virulence are therefore likely to have diverse functions, some of which may not be conserved between distant clades.

Like many fungal pathogens, *C. glabrata*'s niche(s) and life cycle are poorly understood. *C. glabrata* is increasingly identified among clinical samples where it is responsible for an increasing proportion of cases of candidemia [4,20]. *C. glabrata* has also been identified environmentally, including as a component of the mycobiota of yellow-legged gulls [21], in droppings and cloaca swabs of birds of prey, migratory birds and passeriformes [16], and other potentially transitory sources including spontaneously fermenting coffee beans [22]. Concerted efforts for sampling are required to determine the true ecological distribution of *C. glabrata*, as they are for several other important *Candida* species such as *C. auris* [23]. Furthermore, the relatedness of global isolates and their routes of transmission (either patient to patient, or between patient and the environment) requires further studies comparing genotypes to collected metadata including location of isolation.

High levels of genetic heterogeneity have been identified in the *C. glabrata* population, as molecular methods have identified diverse strains, clades and Sequence Types (STs) both inter- and intra-nationally [24,25]. As of January 2022, the Multi Locus Sequence Type (MLST) database for *C. glabrata* included 233 STs from 1,414 isolates from 29 countries, based on the sequence identity for 6 genetic loci [26]. All isolates of *C. glabrata* reported to date have been haploid, with occasional aneuploids e.g. transitory disomies of chromosomes E and G [24].

Genetic heterogeneity within many fungal populations is shaped by their ability to switch between clonal and sexual recombination [27]. The ability for *C. glabrata* to undergo a sexual cycle remains unknown, with all reported attempts in the laboratory to encourage mating thus far unsuccessful. *C. glabrata* has therefore

been regarded as an asexual species, despite the presence of well conserved mating loci [28,29] and 14 examples of phylogenetic incompatibilities from multi locus sequencing [30]. More recent genomic analysis from 34 globally isolated *C. glabrata* strains revealed evidence of population admixture, suggesting a thus far undiscovered sexual cycle [24]. Greater sampling efforts and genomic analyses are therefore required to fully explore signatures of adaptation, virulence and recombination.

In this study, we explore the population genetics and microevolution of *C. glabrata* using comparative genomic analysis of 68 clinical *C. glabrata* isolates from 8 hospitals across Scotland, combined with 83 publicly available and globally isolated genomes, finding evidence of recombinant STs, hypervariable mitochondrial genomes, as well as variation in virulence genes and drug-targets between STs and between serial isolates from prolonged or recurrent infection.

Materials and Methods

Library preparation, sequencing and antifungal tests

Candida glabrata was collected from blood in 2012 from eight hospitals in Scotland (Table S1). These isolates were collected as part of a retrospective study of all cases of *Candida* blood stream infections carried out within Scotland under NHS Caldicott Guardian approval from March 2012 to February 2013, as described previously [4,31].

Genomic and mitochondrial DNA was extracted from 68 isolates using the QIAamp® DNA mini extraction kit (Qiagen) according to the manufacturer's

instructions. A small modification was made prior to extraction which was to mechanical disrupt the yeast. This was achieved by bead-beating the cells with sterile acid-washed 0.5 mm diameter glass beads (Thistle Scientific) for 3 x 30s. Following isolation and extraction using the QIAamp columns the DNA was eluted into elution buffer before storage at -20°C and transport to the sequencing facility.

Library preparation was performed by the Centre for Genome-Enabled Biology and Medicine at the University of Aberdeen. Briefly, gDNA quality was assessed on a Tapestation 4200 with a high sensitivity genomic DNA tape (Agilent) and quantified by fluorimetry using Quant-IT dsDNA High-sensitivity (HS) assay (Thermo Fisher). Dual indexed Illumina libraries were prepared from 1 ng purified gDNA using an Illumina Nextera XT DNA library preparation kit and Nextera XT v2 indices, which were purified from free adapters using AMPure XP beads (Beckman Coulter). Libraries were quantified using Quant-IT dsDNA High-sensitivity (HS) assay (Thermo Fisher) and average fragment size was calculated on Tapestation 4200 (Agilent), then equimolar pooled at 10nM. Concentration of the pool was verified by qPCR (Kapa library quantification kit, Roche) on QuantStudio 6 using SYBR green, and 1.8 pM of the library pool was sequenced on an Illumina NextSeq500 with 150bp paired end reads and 8bp index reads to average alignment depths of 41.9X (Table S2). This data was supplemented with paired end Illumina reads from Carreté *et al.* [24], and isolate CBS138 from Xu *et al*. [32].

Minimum inhibitory concentration (MIC) tests for fluconazole were performed at the Mycology Reference Laboratory, Public Health England, Bristol, according to

199 standard Clinical and Laboratory Standards Institute (CLSI) broth microdilution M27
200 guidelines [33].
201
202 **Variant calling**
203 The Genome Analysis Toolkit (GATK) v.4.1.2.0 [34] was used to call variants.
204 Our Workflow Description Language (WDL) scripts were executed by Cromwell
205 workflow execution engine v.48 [35]. Briefly, raw sequences were pre-processed by
206 mapping reads to the reference genome *C. glabrata* CBS138 using BWA-MEM
207 v.0.7.17 [36]. Next, duplicates were marked, and the resulting file was sorted by
208 coordinate order. Intervals were created using a custom bash script allowing parallel
209 analysis of large batches of genomics data. Using the scatter-gather approach,
210 HaplotypeCaller was executed in GVCF mode with the haploid ploidy flag. Variants
211 were imported to GATK 4 GenomicsDB and hard filtered if QualByDepth (QD) < 2.0,
212 FisherStrand (FS) > 60.0, root mean square mapping quality (MQ) < 40.0, Genotype
213 Quality (GQ) $\geq$ 50, Allele Depth (AD) $\geq$ 0.8, or Coverage (DP) $\geq$ 10.
214
215 To identify aneuploid chromosomes, depth of coverage was calculated for
216 each of 206 fungal samples. Sorted BAM files prepared in the pre-processing phase
217 of SNP calling were passed to Samtools v.1.2 [37] and mpileup files were generated.
218 Read depth was normalised by total alignment depth and plotted against the location
219 in the genome using 10 kb non-overlapping sliding windows. To identify structural
220 variation, assembly *de novo* was achieved using Spades v3.12 [38] using default
221 parameters.
222
223 **Phylogenetic and population genetic analysis**

To construct species-specific phylogenetic trees, all sites that were either a homozygous reference or SNP in every isolate were identified using ECATools (https://github.com/rhysf/ECATools) and concatenated into a FASTA file. Our rooted tree *C. bracarensis* included 1,198 phylogenetically informative sites, while the unrooted C. glabrata tree included 34,980 phylogenetically informative sites. Phylogenetic trees were constructed with RAxML PThreads v.7.7.8 [39] using the general-time-reversible model and CAT rate approximation with 100 bootstrap support, both with rooting to *C. bracarensis* AGP [40] or midpoint rooting without *C. bracarensis*. We constructed a tree using the same models with 1000 bootstrap support for all Saccharomycetaceae species that had a genome assembly in NCBI or JGI Mycocosm. We also constructed neighbour-joining trees using PAUP v4.0b10 [41] and a NeighborNet Network with SplitsTree v.4.15.1 [42]. Trees were visualised using FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

A multisample variant call format (VCF) corresponding to all 151 genomes was made with VCFTools v0.1.12 vcf-merge [43] and converted to ped and map file formats for use in PLINK v1.90 [44]. VCFTools was used to calculate genetic diversity metric pi, using the --site-pi parameter. Unsupervised ADMIXTURE [45] (settings: --haploid="*" -s time) was run on a moderately linkage disequilibrium (LD)-pruned alignment (PLINK --indep-pairwise 60 10 0.1) for values of K between 1 and 35. A value of K = 20 provided the lowest cross-validation error. Principle component analysis was performed using SmartPCA v4 [46]. Consensus gene sequences for each isolate were generated, and genes *FKS2* (CAGL0K04037g) positions 240-828, *LEU2* (CAGL0H03795g), *NMT1* (CAGL0A04059g), *TRP1* (CAGL0C04092g), *UGP1* (CAGL0L01925g), and *URA3* (CAGL0I03080g) were used to identify known MLST

for each of the isolates, and confirm that isolate CG57 was an unknown MLST and registered as the new ST204 on PubMLST (https://pubmlst.org/organisms/candida-glabrata) [47].

We applied Weir's estimator [48] of Wright's Fixation Index (F_{ST}) according to the equations given in Multilocus 1.3 [49] using non-overlapping sliding windows. The scripts have been made available online (https://github.com/rhysf/FSTwindows).

Selection and microevolutionary analysis

The direction and magnitude of natural selection for each ST were assessed by measuring the rates of non-synonymous substitution (*dN*), synonymous substitution (*dS*) and omega (ω = *dN*/*dS*) using the yn00 program of PAML [50], which implements the Yang and Nielsen method, taking into account codon bias [51]. Further GC corrections were not applied. The program was run on every gene in each isolate using the standard nuclear code translation table. To examine the functional significance of genes with ω > 1, we evaluated their Pfam domains and Gene Ontology (GO) terms for statistical enrichment (genes with ω > 1 vs, the remaining genes) using the two-tailed Fisher exact test with Storey false discovery rate (FDR)-corrected P values (q) of < 0.05. GO Terms were acquired using Blast2Go v6.0.1 [52] using Blastp-fast to the NCBI BLAST nr-database (E-value < 1E-5).

Genes of interest were defined including both *FKS* and 12 *ERG* pathway genes, as well as all genes listed in Table 1 of [53], which included adhesions including *EPA* genes, aspartic proteases, phospholipases, cell wall biogenesis,

structural wall proteins, regulatory, efflux pumps. This gene list was then screened for genes with either signatures of positive selection, or those undergoing microevolutionary changes (non-synonymous and frameshift indels).

Results

Recombinant sequence types in Scottish clinical samples

Clinical isolates of *Candida glabrata* from across Scotland are highly-genetically diverse. Using whole-genome sequencing, we analysed the genomes for 68 isolates of *C. glabrata* from 47 separate patients across eight Scottish hospitals, generating the largest panel of *C. glabrata* genome sequences to date. These 68 isolates belonged to twenty separate sequence-types (ST) of *C. glabrata*, which represent genetically related sub-populations based on alleles from six loci/genes. One isolate (CG57 from a single patient in Forth Valley Royal Hospital) belonged to a new ST that has not been previously identified anywhere else (ST204) (**Fig. 1, Table S1, Table S2**). Variant calling using the diploid model of GATK found few examples of heterozygosity (< 0.41 per kb for every isolate) suggesting all isolates were haploid (**Table S3**). Our panel of *C. glabrata* isolates was supplemented with a further 83 genomes from three recent studies of global *C. glabrata* isolates [24,25,32], as well as sequences from the outgroup *C. bracarensis* [40], which has also been identified from clinical settings and is the closest known relative of *C. glabrata* [14].

Phylogenetic analysis of our Scottish collection along with the worldwide *C. glabrata* isolates revealed high genetic diversity among the 29 separate STs represented by our combined panel (**Fig. 2, Fig. S1**). Allelic diversity among *C.*

glabrata isolates (mean nucleotide diversity (π) = 0.00665, σ = 0.047) was higher than previously reported (π = 0.00485 based on the Internal Transcribed Spacer (ITS) of 29 strains [54]). Our WGS-based calculation of *C. glabrata* π was the highest of any species in the Saccharomycetaceae that had both an available genome assembly and a calculation of π (albeit those are based on ITS sequences and fewer strains than we had) [54] (**Fig. 3a**). However, *C. glabrata* genetic diversity was typical among the Saccharomycotina (mean/$\bar{x}$ = 0.0055, median = 0.0039, standard deviaton/σ = 0.0055) (**Fig. 3b**). Nucleotide diversity within the population was present across the nuclear genome (**Fig. 3c**), with window length having some impact on the result (smaller window lengths (5 kb) resulted in higher average π in approximately half of the genome: chromosomes A-F, M and H). Most of the allelic diversity across the 151 *C. glabrata* isolates came from the nuclear genome (min. = 0.09 SNPs / kb, max. = 6.54 SNPs / kb, $\bar{x}$ = 5.55 SNPs / kb) compared with the mitochondrial genome (min. = 0.05 SNPs / kb, max. = 3.64 SNPs / kb, $\bar{x}$ = 1.21 SNPs / kb). Indeed, a significant difference between nuclear SNPs / kb and mitochondrial SNPs / kb was found using a two-tailed *t*-test for all 151 genomes (*p* = 5.6147E-111).

Seven clade (C) delineations for *C. glabrata* were recently proposed [24], which were equivalent to pre-existing STs including C1 (ST19), C2 (ST7), C3 (ST8), C4 (ST22), C6 (ST136) and C7 (ST3). We found that C5 was polyphyletic, encompassing isolates belonging to the genetically divergent ST6, ST10 and ST15 (**Fig. S1**). Therefore, we recommend the use of ST delineations rather than those clade delineations.

Several *C. glabrata* STs had evidence of genetic recombination. Our neighbour-net network tree of all isolates suggested historic gene-flow between several STs including for example ST7, ST55 and ST162 (**Fig. 2**). Genomic regions with low Wright's fixation index (F_{ST}) values, consistent with genetic exchange, were also identified from pairwise comparisons (n = 435) across 5 kb and 10 kb non-overlapping windows of all STs (**Fig. S2**). xF_{ST} values calculated from 5 kb windows were slightly lower than those calculated from 10 kb windows (averaging -0.046 for each ST pairwise comparison), indicating that window length impacts this measure of genetic variation. Twelve pairwise comparisons from 10 kb windows had $F_{ST} < 0.9$ across the genome (**Fig. S3**), with the lowest for ST18 and ST26 (F_{ST} = 0.64). Additionally, ST7, ST55 and ST162 had lower F_{ST} values across the genome (F_{ST} = 0.65 - 0.83) than other pairwise comparisons demonstrating incompatible phylogenetic signals between these STs (**Fig. S2**).

Principal-component analysis (PCA) of whole-genome SNPs revealed little evidence of clustering of several *C. glabrata* STs, which is consistent with gene flow between them (**Fig. 4A**). For unsupervised model-based clustering with ADMIXTURE, we first identified that K = 20 had the lowest cross-validation error (**Fig. 4B**), and was therefore used for subsequent analysis. Two isolates were consistently (6 independent Admixture runs) found to have evidence for mixed ancestry: ST177 CG1 and our newly discovered ST204 CG57 (**Fig. 4C, Fig. S4**). Other isolates were found to have evidence of mixed ancestry in the majority of runs including ST124 WM18.66, ST126 WM05.155 and ST8 M17.

Only one of the Scottish isolates (CG46) had evidence for Chromosome Copy Number Variation (CCNV)/aneuploidy, found in Chromosome C (**Fig. S5**). Distributions of normalised chromosome read depths of chromosome C (average depth per 10kb window = 0.68) differ significantly from the rest of the genome of CG46 (average depth per 10kb window = 1.05; Kolmogorov-Smirnov Test: p = 2.09E-25), with coverages of chromosome C significantly lower than in the rest of the genome (Wilcoxon rank-sum test: p = 1.353E-25). No other CCNVs were found, despite many isolates having been treated with antifungals that have previously been correlated with CCNV [55]. Together, these results suggest occasional genetic recombination within the *C. glabrata* population, without an association with aneuploidy.

Mating types and mating type switching are poorly understood in *C. glabrata*, although it is thought that Mating-type regulatory protein α2 is expressed in all MTLα strains and not in MTLa strains [56,57]. MATα2 (CAGL0B01265g) was present in all Scottish isolates (breadth of coverage; BOC > 87%). However, MATα1 appeared to be absent or partially absent in 3/9 ST6 isolates (CG12 = 16% BOC, CG121 = 18% BOC, CG42 = 12.5% BOC), while present in the remaining six ST6 isolates, and all the other STs (BOC 100%). The functional relevance of MATα1 deletion or truncation is unclear but may be a hallmark of the rarer of the two mating types.

Hyper-variable mitochondrial genomes among sequence types

F_{ST} analysis highlighted the mitochondrial genome of *C. glabrata* as hyper-variable (**Table S2, Fig. S6**). Forty-three genes were identified in ≥10 pairwise F_{ST} comparisons, including all eleven mitochondrial protein encoding genes. To explain

this enrichment of low F_{ST} mitochondrial genes, we studied the 151 genome alignments. While the nuclear genome had 97.3 - 99.4% BOC, the mitochondrial genome had 20.4 - 99.9% BOC, with 42% of isolates (*n* = 63/151) containing >10% ambiguous mitochondrial bases (2 kb) (Table S2) (here, we define ambiguous as sites with too few reads aligning to be called by GATK, or reads that cannot be called by GATK due to not passing variant filtration). Surprisingly, a pattern of low and/or patchy read coverage was identified in every isolate including the ST15 reference isolate CBS138 (**Fig. S6**), indicating that the reference mitochondrial sequence assembly [58] may have a high error rate, and given additional differences identified in non-reference isolates, that *C. glabrata* mitochondrial genomes are highly heterogenous.

The mitochondrial genome for some *C. glabrata* isolates appears reduced in size and encodes fewer protein encoding genes (**Fig. 5**). As many as 22/37 (59%) mitochondrial encoded genes were entirely absent in at least one isolate, including Cg1, Cg1II, and Cg1III (putative endonucleases of exons and introns in the mitochondrial COX1 gene), *ATP8* and *ATP9* (subunits 8 and 9 of the enzyme complex required for ATP synthesis), *RPM1*/*RPR1* (RNA component of mitochondrial RNAse P), *VAR1* (putative mitochondrial ribosomal protein of the small subunit,) and most of the tRNA genes (15/23). Nine separate STs had absent mitochondrial genes. Normalised depth of coverage was variable across the genes, with < 1 average normalised depth across all isolates for Cg1, Cg1II, and Cg1III, *ATP8, RPM1, VAR1* and tRNA-Met1. While non-uniform coverage in terms of depth and breadth was found across all mitochondrial genomes belonging to all datasets, our newly sequenced isolates have the lowest mean breadth across mitochondrial

genes ($\bar{x}$ = 92.18, σ = 20.07) compared with Biswas *et al.* [25] ($\bar{x}$ = 98.48, σ = 11.47) and Carreté *et al.* [55] ($\bar{x}$ = 97.03, σ = 14.11), suggesting there are some discrepancies between library preparation impacting mitochondrial read sequencing. Notably, only 1/50 Biswas *et al.* [25] isolates (WM03.450) had entirely absent mitochondrial genes compared with 8/32 Carreté *et al.* [55] isolates and 15/68 of our newly sequence isolates. Total sequencing depth can be ruled out as the main cause for low mitochondrial coverage, given Carreté *et al.* [55] had the highest sequencing depth ($\bar{x}$ = 360X) and had many isolates with absent mitochondrial genes, compared with Biswas *et al.* [25] ($\bar{x}$ = 74X) and ours ($\bar{x}$ = 42X).

We used assembly *de novo* to further explore the mitochondrial sequence for isolate WM03.450 (ST83), which had the greatest number of ambiguous bases across its mitochondrial genome (16 kb / 80%). Our WM03.450 Illumina-based assembly (12.9 Mb; N.contigs = ~3 thousand; N_{50} = 85 kb) is 562 kb longer than the CBS138 reference sequence, indicating substantial genomic differences between these isolates and STs. Aligning our assembly to the reference CBS138 mitochondrial genome using Blastn identified 10 contig matches with a combined alignment length of only 1.9 kb (mean 157 nt per contig), suggesting the low alignment is not due to conserved nucleotide sequences that have undergone large rearrangements. Aligning the assembly to the eleven mitochondrial protein sequences using Blastx identified only 6/11 genes across six separate contigs, four of which were < 364 nt length, and two that are 10.4 kb and 81.5 kb. Conversely, assembly *de novo* and Blastx of our Illumina reads for the reference isolate CBS138 against the published CBS138 genome identified all eleven mitochondrial genes present on four contigs, with 18.9 kb total sequence length, of which Blastn aligned

9.3 kb to the published mitochondrial assembly. Together, these analyses suggest that whole gene deletions in the *C. glabrata* mitochondria are common.

Signatures of selection identified among sequence types

In contrast to the *C. glabrata* mitochondrial genome, we found that gene deletions in the nuclear genome are rare. Indeed, fewer than six presence/absence (P/A) polymorphisms (strictly defined as zero reads aligning) were identified per isolate (~0.1% of 5,210 protein-encoding genes) (**Table S4**). Of these, two consecutive nuclear-encoded genes (CAGL0A02255g and CAGL0A02277g) on Chromosome A were entirely absent of read coverage in 25 out of the 68 Scottish isolates (37%), which included all representatives of eleven separate STs (ST4, 7, 8, 24, 25, 55, 67, 83, 177, and our newly described 204). These STs do not cluster phylogenetically, ruling out a single evolutionary event causing this deletion. The two genes have identical nucleotide sequences and encode the same amino acid sequence, which is conserved across a range of other Ascomycota species, as well as having sequence similarity to the K62 Killer Preprotoxin protein encoded by the *Saccharomyces paradoxus* L-A virus M62 satellite (BLASTp E-value = 1e-36), suggesting a possible viral origin. CAGL0F09273g is a separate, putative adhesin-like protein (adhesin cluster V) with a “hyphally regulated cell wall protein N-terminal” PFAM that is lost in eleven isolates including all ST4 (CG68A, CG77), four ST7 (CG157, CG48A, CG48F, CG78), three ST8 (CG127, CG52, CG82), ST19 CG119, ST24 CG166 and ST147 CG133. Again, this gene must have been lost multiple times, given its presence in several ST7 and ST8 isolates. This gene is the last gene on Chromosome F, has an $\pi = 0.00244$, which is lower than the overall average across the genome, and has previously been reported to undergo CCNV within serial

clinical isolates [55], suggesting it is able to undergo variation within microevolutionary timescales, which may impact the adhesive ability of these *C. glabrata* isolates.

Between 61 and 85 genes with a signature of positive selection ($dN/dS = \omega$, and $\omega > 1$ [59]) were found in each ST apart from the reference ST (ST15 CG151), for which only a single gene with $\omega > 1$ ($\omega = 1.0019$) was identified (**Table S5**). Apart from the reference ST, STs had between 4 and 14 genes with $\omega > 2$, showing stronger signatures of diversifying or positive selection. Of the 2,083 total genes with $\omega > 1$ across all clades, 608 were unique genes (11.6% of all genes) i.e., they had this signature in multiple clades, owing to either ancestry or selection acting on the same gene families. To explore selection, we took an unbiased approach using PFAM and GO-term enrichment comparing the numbers of each term in those 608 genes compared with the remaining non-selected genes, as well as a targeted approach for genes of interest (see Methods Selection and microevolutionary analysis) including adhesins, proteases, efflux pumps, FKS, and ERG pathway genes.

Genes with signatures of positive selection within the *C. glabrata* population targets diverse genes and gene functions. Our unbiased approach for enrichment of functional domains in 608 gene products with signatures of positive selection identified only three significantly enriched (two-tailed Fisher exact test with false discovery rate (FDR)-corrected *p*-values (q) of < 0.05) PFAM domains and 16 GO terms (Table 1). The enriched PFAM domains were 1) Flocculin repeat (PF00624.20; $q = 7.42E-17$), 2) GLEYA domain (PF10528.11; $q = 0.01$) and 3) Armadillo repeat

(PF00514.25; *q* = 0.02). Flocculin is a sub-telomeric gene family involved in flocculation or cell aggregation in *S. cerevisiae* [60], while GLEYA domains are present in *C. glabrata EPA* proteins. Thirty Flocculin PFAM domains were assigned to only six genes in *C. glabrata*, two of which have $\omega > 1$: CAGL0I07293g and CAGL0I00220g, and together account for 23/30 Flocculin repeat PFAMs. Enriched GO-terms covered a range of possible biological functions including ribosomal/RNA-binding and mitochondrial structural proteins.

Our targeted approach highlighted 21/129 genes of interest that have $\omega > 1$, with at least one found in every ST apart from the reference ST15 and ST46 (**Table S6**). Notably, none of the aspartic proteases, phospholipases, cell wall biogenesis, efflux pumps, ergosterol biosynthesis pathway genes or *FKS* genes were found to have hallmarks of positive selection, implying these are conserved within the population. Several genes with $\omega > 1$ were found in multiple STs, including adhesive protein CAGL0J01727g (adhesin cluster VI) that is under positive selection in seven STs (18, 26, 36, 45, 147, 177, 204) and adhesive protein CAGL0I07293g (adhesin cluster V) under positive selection in seven mostly distinct STs (3, 8, 25, 83, 123, 136, 177). *C. glabrata* encodes 17 putative adhesive proteins without N-terminal signal peptides, casting doubt on their role in adhesion. One of these is a pseudogene (CAGL0E00110g) with $\omega > 1$ in 13/29 STs. The structural cell wall protein *AWP7* belonging to the Srp1p/Tip1p family was under selection in 7 STs.

C. glabrata **nosocomial in-patient microevolution targets pathogenicity factors and drug targets**

Our Scottish *C. glabrata* panel included seven sets of between 2 and 9 isolates from recurrent cases of candidiasis. To explore the microevolution of *C. glabrata* within a human host, and the effects of antifungal treatment (fluconazole, nystatin, and posaconazole) on fungal genetics, we documented all genetic changes between serial isolates (**Table 2, Table S7**). Although exact dates of isolation were not documented, phylogenetic analysis (**Fig. S1, Fig. 6**) confirmed these serial isolates were highly related, with between 64 and 140 mutations (1.13468×10^{-5} per base pair) identified between pairs of serial isolates (**Fig. 6**). While the mutation rate or generation time for *C. glabrata* is not known [55], this small number of mutations likely suggests recent clonal origins appropriate for microevolutionary analysis. Serial isolates had an estimated time between isolation (based on blood culture dates) between 0 and 239 days (mean 15 days). Five serial isolates from 4 separate patients/cases showed MIC changes from the earlier sampled isolate (**Table 3**), including 2 increases (CG107A->B +8 ug/mL, CG97B->C +4 ug/mL), 1 decrease (CG84G->H -4 ug/mL,) and 1 large transient increase (CG93A, B, C, D, E = 4 ug/mL; CG93H, I, K >64 ug/mL, CG93K = 4 ug/mL).

Mutations identified between serial isolates were mostly in protein coding sequence (CDS) regions (between 53 and 127 mutations per pairs of serial isolates, collectively adding up to 1,741/1,995 total mutations = 87%), despite protein-coding regions taking up only 7.9 / 12.3 Mb (64%) (**Fig. 6B, 6C**). The remaining serial mutations were within intergenic regions (236 mutations; 12%) and intronic regions (18 mutations; 1%). Intronic regions had the highest count of serial mutations after accounting for the total sequence in introns (**Fig. 6C**), albeit with ≤ 3 found per pair of serial isolates. Hypergeometric tests revealed that the number of mutations in coding

sequence compared with non-coding sequence was higher than expected by chance, suggesting a highly significant enrichment of mutations in protein coding genes (*p* = 3e-120).

To explore the 1,741 microevolutionary changes within coding regions, we categorised them into five groups of newly arising mutations (regardless of prior state): 1) insertions/deletions (indels) (*n* = 362; 21%), 2) synonymous mutations (*n* = 264; 15%), 3) non-synonymous mutations (*n* = 303, 17%), 4) nonsense mutations (*n* = 2), and 5) reversion back to reference base (*n* = 810; 47%). Of the indels, 147/362 (41%) were frameshifts that disrupted 54 genes. Non-synonymous mutations were detected in 139 genes (**Fig. 6D, 6E**).

Enrichment for PFAM/GO-terms of these genes with frameshift and non-synonymous mutations (two-tailed Fisher exact test with false discovery rate (FDR)-corrected *p* values (*q*) of < 0.05) revealed three enriched GO-terms and eight enriched terms (**Table 1**). Both categories (frameshifts and non-synonymous mutations) were enriched for GO:0009986 Cell Surface (*q* = 3.21E-08 and 1.09E-06, respectively), suggesting that *C. glabrata* undergoes rapid mutations in several of its cell surface proteins during prolonged/serial blood stream infections. Enriched PFAM terms included the “RNA polymerase *RPB1* C-terminal repeat” for genes with frameshift indels (*q* = 1.25E-36), GLEYA domains for genes with either frameshift (*q* = 3.15E-12) or non-synonymous mutations (*q* = 3.58E-08). Several repeat associated PFAMs and the “Hyphally regulated cell wall protein N-terminal” domain were enriched for non-synonymous mutations (*q* = 5.87E-04).

Several genes of interest (see Methods: Selection and microevolutionary analysis) had microevolutionary changes (*n* = 29/129) (**Table S8**). Notably, one of the two newly acquired nonsense mutations was identified in *FKS2* (Case 6 J->K), coinciding with a substantial drop in fluconazole MIC (**Table 3**). The other was in the uncharacterised CAGL0K04631g at an earlier time point in the same patient (Case 6 D->E).

Twenty adhesins including *EPA* genes were mutated between serial isolates, including in all seven sets/cases of isolates and at every time point. For example, *EPA3* had 5 indels in Case 1 (A->B), a synonymous mutation in Case 2 (A->B; nucleotide position (pos.) 2304), Case 3 (D->E; pos. 1539), Case 4 (A->F; pos. 1119), Case 5 (F->G; pos. 2259), a non-synonymous mutation (pos. 2224) and large (30nt) insertion in Case 5 (G->H), two large deletions (42nt and 16nt), and two synonymous and one non-synonymous mutations in Case 6 (A->B; pos. 1002, 2319 and 2276 respectively).

The longer 42nt deletion from Case 6 (A->B) reverts back to reference in Case 6 (B->C), suggesting either a) a non-descendent isolate (intra-host variation), b) a false negative reference in 6C or c) a false positive deletion in 6A. The same 42nt deletion, along with a new insertion at the site of the previous synonymous mutation appears in Case 6 (C->D), thereby suggesting the variant is real and option c less likely. That 42nt deletion reverts back to reference in Case 6 (D->E), and appears again in Case 6 (E->H). By Case 6 (H->I), the gene has a new synonymous mutation, and in Case 6 (I->J) it has accumulated a new 15nt deletion. *EPA3* is therefore a hot-spot of variation. Another EPA gene that accumulated a large

number of mutations was *AWP12*, which accumulated five non-synonymous mutations and one synonymous mutation (Case 6H->I).

Other genes that had accumulated mutations between serial isolates included those encoding an aspartic protease *YPS5*, several structural wall proteins belonging to the Srp1/Tip family, regulatory protein *PDR1*, the ergosterol synthesis gene *ERG4* (a non-synonymous mutation in Case 3A->B), and both *FKS1* and *FKS2*. Therefore, *C. glabrata* genes that are antifungal targets and gene families involved in drug-resistance and pathogenicity can therefore undergo rapid mutation within a human host.

Discussion

In this study, we sequenced and analysed the largest panel of *C. glabrata* genomes to date. These isolates were collected from blood-stream infections of patients at several Scottish hospitals in 2012. Our 68 genomes were analysed alongside 83 further publicly available and globally isolated genomes [25,32,55], revealing greater genetic diversity than previously recognised, including a nucleotide diversity of 0.00665, which is much higher than has been calculated for the distantly related *C. albicans* at 0.00298 [54]. Surprisingly, we found that only one of our Scottish isolates had evidence of aneuploidy, despite many having been treated with antifungals, which has previously been correlated [55]. Chromosome C in CG46 had lower depth of coverage compared with the rest of the genome, perhaps due to chromosome loss in a subset of cells. The patient that CG46 was isolated from was initially treated with Fluconazole. Following resistance to Fluconazole being

596 detected, the patient was subsequently treated with Caspofungin, suggesting a
597 potential link between those antifungal treatments and the observed aneuploidy.
598
599 We found that the mitochondrial genome of *C. glabrata* was hyper-diverse
600 compared with its nuclear genome for many isolates, including several long deletions
601 spanning one or more genes, with the potential to impact many important biological
602 functions including drug resistance and persistence [61]. High levels of variation in
603 mitochondrial genomes within the major fungal phyla have previously been noted in
604 terms of gene order, genome size, composition of intergenic regions, presence of
605 repeats, introns, associated ORFs, and evidence for mitochondrial recombination in
606 all fungal phyla [62]. This variation is lacking in Metazoa [62]. Our results suggest
607 some of these types of mitochondrial genetic diversity are likely present within the *C.*
608 *glabrata* population.
609
610 Isolates in this study belonged to twenty-nine separate sequence types (STs)
611 of *C. glabrata,* each of which was separated by large number of variants. However,
612 as many as 193 MLST STs have been documented [47]. Therefore, it is likely that
613 the true genetic diversity of *C. glabrata* is much higher than we have been able to
614 calculate with whole-genome sequences (albeit the largest panel studied to date).
615 Indeed, several further STs may yield further evidence of recombination or lack of,
616 and may ultimately require a new effort to group STs into lineages (also dependent
617 on the frequency of recombination that erode these divisions). The genetic diversity
618 of *C. glabrata* in hospitals around Scotland is extremely high, with representatives
619 from 20 STs. Such high genetic diversity (and many of the same STs) have also
620 been found from genome sequencing and phylogenetic analysis of isolates collected

621 in other countries such as Australia [25], suggesting they must have been
622 transported across or between continents, perhaps by anthropogenic or even natural
623 means (for example its association with birds [16,21] and food [22]). Greater
624 sampling and genotyping of clinical and environmental isolates will be required for
625 understanding ancestry or endemicity.
626
627 *C. glabrata* has long been regarded as a haploid asexual yeast, although
628 evidence has recently emerged of a cryptic sexual cycle [28–30]. Our genome
629 sequencing and population genetics supports this work, revealing compelling
630 evidence that at least 12 sequence types (STs) stem from recombination between
631 other STs. However, further work remains to document and describe individual
632 recombinant isolates. Providing genetic recombination between isolates is naturally
633 occurring, the mechanisms of genetic exchange are also unknown, although likely
634 relate to the conserved mating type locus, which play a central role in the sexual
635 cycle of diverse fungi [63]. Here, we show that the MATα1 gene was absent or
636 partially absent in three isolates belonging to ST6, which could potentially impact or
637 even be a hallmark of a rarer second mating type of *C. glabrata*. Together, genetic
638 recombination among *C. glabrata* isolates appears much more common than
639 previously recognised, and likely contributes to increased genetic diversity.
640
641 The nuclear genome for isolates belonging to every ST (apart from the
642 reference ST15 that was included as a control) included evidence of positive or
643 diversifying selection. Signatures of positive selection were found enriched in genes
644 with diverse functions, including several with repeat domains, as well as EPA genes
645 with GLEYA domains. EPA genes are a large sub-telomeric family of virulence-

related surface glycoprotein-encoding genes encoded by several other pathogens including *Plasmodium*, *Trypanosoma*, and *Pneumocystis* [64]. Such gene differences between STs of *C. glabrata* may result in clinically-relevant phenotypic differences.

In host microevolutionary changes between serial isolates were enriched within coding-sequences, which is a surprise, given the expectation for intergenic regions to be more permissive to mutations due to relaxed selection within intergenic regions and purifying selection within coding sequence. The reason for this abundance of serial mutations in coding sequence is unclear, although it could potentially be technical (e.g. false negative variants within repetitive sequences) or biological (e.g. drug exposure and host immune pressure). Alternatively, enriched mutations in genes could potentially be driven by processes such as DNA polymerase induced mutations, or differences in chromatin states (e.g. heterochromatin could lead to increased exposure to DNA damaging agents resulting in higher mutation rates, or conversely, greater surveillance and correction of mutations in euchromatin regions by cellular DNA repair enzymes [65]).

Selection may explain why we identified similar numbers of non-synonymous mutations to synonymous mutations, given random mutations are expected to be non-synonymous in ~2/3 nucleotides of each codon. Furthermore, accumulations of deleterious mutations could be occurring in the serial isolates due to small population sizes, although population size estimates could not be calculated accurately from the metadata.

Genes with GLEYA domains including EPA genes were significantly enriched for frameshift and non-synonymous mutations in the coding sequence between serial isolates. Combined with our finding of positive selection in EPA genes across STs, suggests that EPA genes are undergoing variation at both longer time frames and microevolutionary time-scales.

Genes encoding several important drug-targets also underwent mutations between serial isolates, including a non-synonymous mutation in the ergosterol biosynthesis pathway gene *ERG4*, and a nonsense mutation in the 1,3-β-glucan synthase component *FKS2* (mutations in these genes can confer resistance to azoles [66] and echinocandins [67] respectively). Indeed, the nonsense mutation in FKS2 coincided with a marked drop in fluconazole MIC for isolate CG93K, suggesting a possible link.

Our study highlights the need for further sampling and genomic analysis of *C. glabrata* in order to better inform the population structure and mechanisms underlying its increasing emergence, pathogenicity and multi-drug resistance. While we have largely focused on differences among the conserved regions of the *C. glabrata* ST15 CBS138 genome using an alignment-based strategy, our discoveries of a hyper-diverse mitochondrial sequence highlight the value for future long-read sequencing and assemblies to characterise the pan-genomes of *C. glabrata* and structural genomic diversity that exists among and perhaps within STs, and to explore the mechanisms driving those changes. Furthermore, given the genetic diversity between STs that we document, it would likely be valuable to sequence and assemble additional high-quality reference sequences for the purposes of increasing

variant-calling accuracy and quantifying gene content between different STs. Given the low and patchy alignment depth across the ST15 CBS138 mitochondrial sequence for that same isolate, a review and update for the published CBS138 mitochondrial genome is likely required as well. Indeed, high (~0.5%–1%) frequencies of structural variation in the nuclear genomes of *C. glabrata* isolates was recently found using *de novo* assemblies from long single-molecule real-time reads [68].

The rapidity that *C. glabrata* can mutate important genes and gene families, both *via* microevolution and putative recombination highlights an obstacle for future drug-development, given that individual gene targets are able to mutate within short time spans, and substantial diversity already present between STs. In addition, the epidemiology of *C. glabrata* is poorly understood. Future sampling and genomic comparison studies are necessary to identify the routes and mechanisms of its spread and evolution.

Data availability

Raw sequences for all haploid isolates of *C. glabrata* from this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA669061.

Acknowledgments

R.A.F, D.W., T.W., and A.W. are supported by the Medical Research Council Centre for Medical Mycology MR/N006364/2. R.A.F. is supported by a Wellcome Trust Seed Award (215239/Z/19/Z). D.W. is supported by a Wellcome Trust Senior

721 Research Fellowship (214317/Z/18/Z). C.A.M is funded by the European Union's
722 Horizon 2020, Innovative Training Network: FunHoMic (grant N° 812969) and
723 consortium ‘Host-Directed Medicine in invasive FUNgal infections’—HDM-FUN
724 (Grant Agreement 847507). The authors thank Zeynab Heidari and Elaina Collie-
725 Duguid of the Centre for Genome-Enabled Biology and Medicine , University of
726 Aberdeen for their support and assistance in this work. We would also like to thank
727 Dr. Lucy van Dorp for valuable suggestions for running ADMIXTURE, and Prof. Ian
728 Stansfield for useful discussions.

729

730 **Competing Interest Statement**

731 The authors declare no competing interests.

732

733 **Figures and Tables**

734

735 Figure 1. *Candida glabrata* isolates were collected across eight health boards across
736 Scotland in 2012, belonging to 20 separate sequence types, including the newly
737 described ST204. Duplicate isolates stemming from the same patient at different
738 time points have been excluded.

739

740 Figure 2. A NeighborNet network using SplitsTree, with sequence types (ST) labels
741 replacing isolate names at the nodes. Green = found in Scotland, purple = not found
742 in Scotland. The scale bar represents nucleotide substitutions per site.

743

744 Figure 3. a) A RAxML phylogenetic tree with 1000 bootstrap support of all
745 Saccharomycetaceae species that had a genome assembly in NCBI or JGI

Mycocosm and nucleotide diversity (π). Note: C. glabrata is calculated from whole genome sequence data presented in this study, while the other species are based on ITS sequences only [54]. b) π (based on ITS sequences only) for all Saccharomycotina and non-Saccharomycotina that are listed in the ISHAM ITS reference DNA barcoding database [54]. c) non-overlapping 5 kb, 10 kb and 20 kb windows of $\bar{x}\pi$ (π for all sites in the genome divided by window length).

Figure 4. Population genetics of *C. glabrata* sequence types (ST). a) Principal-component analysis (PCA) of whole-genome SNPs using SmartPCA revealed little evidence of sub-clustering among STs (isolates are calculated and plotted ndividually, but labelled by their ST alone for clarity). SmartPCA failed to calculate the eigenvalues for some isolates including those belonging to ST4. b) The cross-validation (CV) error from running unsupervised ADMIXTURE for variant-sites across the *C. glabrata* population, testing K values between 1 and 35. K = 20 provided the lowest CV error. c) ADMIXTURE plot for all isolates using K = 20, revealing several isolates with evidence of mixed ancestry. Isolates are ordered according to the neighbor-joining tree constructed with PAUP in Figure S1.

Figure 5. Breadth of coverage and depth of coverage across each of the 37 mitochondrial encoded genes for all 151 isolates compared in this study (Each point represents an isolate). A) Breadth of coverage as a % across each gene. B) The Normalised depth of coverage for each gene (total read depth for each gene / total read depth across both nuclear and mitochondrial genomes). C) Breadth of coverage as a % across each gene, categorised by sequence types (ST)s. D) Normalised depth of coverage for each gene, categorised by ST.

Figure 6. Microevolutionary changes across seven sets of *C. glabrata* isolates. A) a RAxML phylogenetic tree of the serial isolates using the general-time-reversible model and CAT rate approximation with 100 bootstrap support. Branch lengths indicate the mean number of changes per site. B) The number of serial mutations total (All), those within protein-coding sequence (CDS), intergenic and intronic regions. C) Those same serial mutations per kb (calculated as the count of serial mutations divided by the total length of the feature (where All = whole genome) and multiplied by 1000). D) Serial mutations within CDS categorized by their effect on the sequencing: Insertion/Deletion (Indel), synonymous mutation (Syn.), non-synoynmous mutation (Non.Syn.) and nonsense mutation. E) Those same serial mutations within CDS per kb.

Figure S1. Phylogenetic trees of *C. glabrata*. All genomic sites that were either a homozygous reference or SNP in every isolate of *C. glabrata* and *C. bracarensis* AGP [40] for rooting were identified using ECATools and concatenated into a FASTA file. A) A neighbor-joining tree constructed with PAUP. Scale bar indicates the distance based on substitutions per site. B) A maximum likelihood tree constructed using RAxML PThreads v.7.7.8 [39] using the general-time-reversible model and CAT rate approximation with 100 bootstrap support. Branch lengths indicate the mean number of changes per site. The clade according to Carreté L *et al.* 2018 [24], as well as sequence type (ST), country code (AU = Australia, BE = Belgium, DE = Germany, FR = France, GB = Great Britain, IT = Italy, TW = Taiwan, US = United States), MAT and reference are also shown.

Figure S2. Mean F_{ST} values from pairwise comparisons of each sequence type (ST) calculated from a) 10 kb non-overlapping windows and b) 5 kb windows. C) Mean F_{ST} values from 10 kb windows were similar to values calculated from 5 kb windows, with a mean difference of -0.046 per pairwise comparison.

Figure S3. Non-overlapping 10 kb windows showing F_{ST} values for 12 pairwise comparisons that had long genomic regions with lower values.

Figure S4. Five independent runs of ADMIXTURE using K = 20 and time-based seed values, revealing several isolates with evidence of mixed ancestry. Isolates are ordered according to the neighbor-joining tree constructed with PAUP in Figure S1.

Figure S5. Non-overlapping 10 kb windows showing normalized depth of coverage (including GC normalization by percentiles; GC, and excluding ambiguous sites (effective window length)).

Figure S6. Integrated Genome Viewer (IGV) screenshots for the reference isolate CBS138, and all isolates compared in this study, indicating substantial differences between our Illumina sequences and the mitochondrial assembly. Gene features are shown as a track (directionality indicated by arrows), and the read alignment from the BAM files are shown for each isolate, where peaks indicate higher depth, and colors on the peaks indicate discrepancies to the reference base: green = A, blue = C, red = T, brown = G, purple = insertion)

Table 1. GO-term and PFAM enrichment (two-tailed Fisher exact test with false discovery rate (FDR)-corrected P values (q) of <0.05) for genes with *dN*/*dS* (ω) > 1, and genes with either microevolutionary frameshifts or non-synonymous mutations across the seven sets of serial isolates. The relative proportion (Rel. prop) was calculated as (number of terms in set 1 / number of terms in set 2) * (genes with any terms in set 2 / genes with any terms in set 1).

Table 2. Summary of microevolution across seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis. We documented 1,995 mutations between all serial isolates, which were either in protein-coding regions (Coding) or Intergenic and intron regions (Non-coding). Coding mutations were further characterised into Coding Indels, some of which caused frameshifts (Coding Indel (frameshift)), non-synonymous mutations (Coding Non.Syn.), nonsense mutations, (Coding Nonsense), synonymous mutations (Coding Syn.), and bases that reverted back to the ST15 CBS138 reference base (either from a previous microevolutionary change or a pre-existing variant between the initial isolate and the reference ST15 CBS138).

Table 3. Minimum inhibitory concentration (MIC) values of fluconazole for each of the serial isolates.

Table S1. Metadata for clinical Scottish *C. glabrata* sequenced and analysed in this study.

Table S2. Haploid variant call summary based on alignments to the published nuclear and mitochondrial assembly of ST15 CBS138. These variants form the basis for the population genetic and comparative genomics tests.

Table S3. Diploid variant call summary. Diploid variant calls were used to check for any evidence of heterozygosity suggestive of diploidy. All heterozygous sites (single nucleotide heterozygous positions + heterozygous insertions + heterozygous deletions) amounted to <0.0404% (0.4 per Kb) of total positions called per isolate, suggesting these were errors and not evidence of diploidy.

Table S4. Counts of presence/absence (P/A) polymorphisms in each isolate, based on zero reads aligning to the ST15 CBS138 nuclear and mitochondrial genomic regions encoding gene sequences.

Table S5. A summary of *dN*/*dS* (ω) and nonsense mutations found across every gene in isolates representing each of the sequence types (ST).

Table S6. Details of 21 genes (found 67 times across all STs) with *dN*/*dS* (ω) > 1, which belonged to our set of 129 “genes of interest” including adhesions (e.g. EPA genes), aspartic proteases, phospholipases, cell wall biogenesis, structural wall proteins, regulatory, efflux pumps (all genes in Table 1 of [53]), as well as both FKS and 12 ERG pathway genes.

Table S7. Counts of all microevolutionary changes documented between seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis.

Table S8. Details of 29 genes that had microevolutionary changes documented between seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis, and belonged to our set of 129 “genes of interest” including adhesions (e.g. EPA genes), aspartic proteases, phospholipases, cell wall biogenesis, structural wall proteins, regulatory, efflux pumps (all genes in Table 1 of [53]), as well as both FKS and 12 ERG pathway genes. For each mutation, the information is encoded in a string with details separated by a semi colon. The first detail in variant type (e.g. ref_to_snp, where ref=reference), the second is location in CDS by nucleotide count, the third is the codon position (1, 2 or 3), the fourth is codon found along with amino acid position and type. Finally, a short description is given e.g. INSERTION and DELETION along with the number of nucleotides, or SYN=Synonymous, NSY=Non-synonymous, and NON=Nonsense.

References

1. Daniel H-M, Lachance M-A, Kurtzman CP. 2014 On the reclassification of species assigned to *Candida* and other anamorphic ascomycetous yeast genera based on phylogenetic circumscription. *Antonie van Leeuwenhoek* **106**, 67–84. (doi:10.1007/s10482-014-0170-z)

2. Turner SA, Butler G. 2014 The *Candida* pathogenic species complex. *Cold Spring Harb Perspect Med* **4**. (doi:10.1101/cshperspect.a019778)

3. Perlroth J, Choi B, Spellberg B. 2007 Nosocomial fungal infections: epidemiology, diagnosis, and treatment. *Med Mycol* **45**, 321–346. (doi:10.1080/13693780701218689)

4. Rajendran R, Sherry L, Deshpande A, Johnson EM, Hanson MF, Williams C, Munro CA, Jones BL, Ramage G. 2016 A prospective surveillance study of candidaemia: epidemiology, risk factors, antifungal treatment and outcome in hospitalized patients. *Front Microbiol* **7**, 915. (doi:10.3389/fmicb.2016.00915)

5. Mullick A, Leon Z, Min-Oo G, Berghout J, Lo R, Daniels E, Gros P. 2006 Cardiac failure in C5-deficient A/J mice after *Candida albicans* infection. *Infect Immun* **74**, 4439–4451. (doi:10.1128/IAI.00159-06)

6. Hirano R, Sakamoto Y, Kudo K, Ohnishi M. 2015 Retrospective analysis of mortality and *Candida* isolates of 75 patients with candidemia: a single hospital experience. *Infect Drug Resist* **8**, 199–205. (doi:10.2147/IDR.S80677)

7. Lamoth F, Lockhart SR, Berkow EL, Calandra T. 2018 Changes in the epidemiological landscape of invasive candidiasis. *J Antimicrob Chemother* **73**, i4–i13. (doi:10.1093/jac/dkx444)

8. Yapar N. 2014 Epidemiology and risk factors for invasive candidiasis. *Ther Clin Risk Manag* **10**, 95–105. (doi:10.2147/TCRM.S40160)

9. Odds FC, Hanson MF, Davidson AD, Jacobsen MD, Wright P, Whyte JA, Gow NAR, Jones BL. 2007 One year prospective survey of *Candida* bloodstream infections in Scotland. *J Med Microbiol* **56**, 1066–1075. (doi:10.1099/jmm.0.47239-0)

10. Cowen LE, Anderson JB, Kohn LM. 2002 Evolution of drug resistance in *Candida albicans*. *Annu. Rev. Microbiol.* **56**, 139–165. (doi:10.1146/annurev.micro.56.012302.160907)

11. Pfaller MA, Castanheira M, Lockhart SR, Ahlquist AM, Messer SA, Jones RN. 2012 Frequency of Decreased Susceptibility and Resistance to Echinocandins among Fluconazole-Resistant Bloodstream Isolates of Candida glabrata. *J Clin Microbiol* **50**, 1199–1203. (doi:10.1128/JCM.06112-11)

12. Barber AE *et al.* 2019 Comparative genomics of serial *Candida glabrata* isolates and the rapid acquisition of echinocandin resistance during therapy. *Antimicrob Agents Chemother* **63**. (doi:10.1128/AAC.01628-18)

13. Brunke S, Hube B. 2013 Two unlike cousins: *Candida albicans* and *C. glabrata* infection strategies. *Cell Microbiol* **15**, 701–708. (doi:10.1111/cmi.12091)

14. Gabaldón T *et al.* 2013 Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics* **14**, 623. (doi:10.1186/1471-2164-14-623)

15. Cormack BP, Ghori N, Falkow S. 1999 An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. *Science* **285**, 578–582. (doi:10.1126/science.285.5427.578)

16. Cafarchia C, Romito D, Coccioli C, Camarda A, Otranto D. 2008 Phospholipase activity of yeasts from wild birds and possible implications for human disease. *Medical Mycology* **46**, 429–434. (doi:10.1080/13693780701885636)

17. Bink A, Vandenbosch D, Coenye T, Nelis H, Cammue BPA, Thevissen K. 2011 Superoxide Dismutases Are Involved in Candida albicans Biofilm Persistence against Miconazole ▿ . *Antimicrob Agents Chemother* **55**, 4033–4037. (doi:10.1128/AAC.00280-11)

18. Ibrahim AS, Mirbod F, Filler SG, Banno Y, Cole GT, Kitajima Y, Edwards JE, Nozawa Y, Ghannoum MA. 1995 Evidence implicating phospholipase as a virulence factor of *Candida albicans*. *Infection and Immunity* **63**, 1993–1998. (doi:10.1128/iai.63.5.1993-1998.1995)

19. Martchenko M, Alarco A-M, Harcus D, Whiteway M. 2004 Superoxide Dismutases in *Candida albicans*: transcriptional regulation and functional characterization of the hyphal-induced SOD5 gene. *Mol Biol Cell* **15**, 456–467. (doi:10.1091/mbc.E03-03-0179)

20. Beck-Sagué C, Jarvis WR. 1993 Secular trends in the epidemiology of nosocomial fungal infections in the United States, 1980-1990. National Nosocomial Infections Surveillance System. *J Infect Dis* **167**, 1247–1251. (doi:10.1093/infdis/167.5.1247)

21. Al-Yasiri MH, Normand A-C, L'Ollivier C, Lachaud L, Bourgeois N, Rebaudet S, Piarroux R, Mauffrey J-F, Ranque S. 2016 Opportunistic fungal pathogen *Candida glabrata* circulates between humans and yellow-legged gulls. *Sci Rep* **6**, 36157. (doi:10.1038/srep36157)

22. de Melo Pereira GV, Soccol VT, Pandey A, Medeiros ABP, Andrade Lara JMR, Gollo AL, Soccol CR. 2014 Isolation, selection and evaluation of yeasts for use in fermentation of coffee beans by the wet process. *International Journal of Food Microbiology* **188**, 60–66. (doi:10.1016/j.ijfoodmicro.2014.07.008)

23. Arora P, Singh P, Wang Y, Yadav A, Pawar K, Singh A, Padmavati G, Xu J, Chowdhary A. 2021 Environmental isolation of *Candida auris* from the coastal wetlands of Andaman Islands, India. *mBio* **12**. (doi:10.1128/mBio.03181-20)

24. Carreté L *et al.* 2018 Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr Biol* **28**, 15-27.e7. (doi:10.1016/j.cub.2017.11.027)

25. Biswas C *et al.* 2018 Whole genome sequencing of Australian *Candida glabrata* isolates reveals genetic diversity and novel sequence types. *Front Microbiol* **9**. (doi:10.3389/fmicb.2018.02946)

26. Gabaldón T, Gómez-Molero E, Bader O. 2020 Molecular Typing of *Candida glabrata*. *Mycopathologia* **185**, 755–764. (doi:10.1007/s11046-019-00388-x)

27. Drenth A, McTaggart AR, Wingfield BD. 2019 Fungal clones win the battle, but recombination wins the war. *IMA Fungus* **10**, 18. (doi:10.1186/s43008-019-0020-8)

28. Wong S, Fares MA, Zimmermann W, Butler G, Wolfe KH. 2003 Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata*. *Genome Biol* **4**, R10. (doi:10.1186/gb-2003-4-2-r10)

29. Gabaldón T, Fairhead C. 2019 Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. *Curr Genet* **65**, 93–98. (doi:10.1007/s00294-018-0867-z)

30. Dodgson AR, Pujol C, Pfaller MA, Denning DW, Soll DR. 2005 Evidence for recombination in *Candida glabrata*. *Fungal Genet Biol* **42**, 233–243. (doi:10.1016/j.fgb.2004.11.010)

31. Rajendran R *et al.* 2016 Biofilm formation is a risk factor for mortality in patients with *Candida albicans* bloodstream infection-Scotland, 2012-2013. *Clin Microbiol Infect* **22**, 87–93. (doi:10.1016/j.cmi.2015.09.018)

32. Xu Z, Green B, Benoit N, Schatz M, Wheelan S, Cormack B. 2020 *De novo* genome assembly of *Candida glabrata* reveals cell wall protein complement and structure of dispersed tandem repeat arrays. *Mol Microbiol* **113**, 1209–1224. (doi:10.1111/mmi.14488)

33. Alexander BD, Clinical and Laboratory Standards Institute. 2017 *Reference method for broth dilution antifungal susceptibility testing of yeasts*.

34. McKenna A *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110)

35. Voss K, Auwera GV der, Gentry J. 2017 Full-stack genomics pipelining with GATK4 + WDL + Cromwell. In *18th Annual Bioinformatics Open Source Conference (BOSC 2017)*, (doi:10.7490/f1000research.1114634.1)

36. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*

37. Li H *et al.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

38. Bankevich A *et al.* 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. (doi:10.1089/cmb.2012.0021)

39. Stamatakis A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)

40. Correia A, Sampaio P, James S, Pais C. 2006 *Candida bracarensis sp. nov.*, a novel anamorphic yeast species phenotypically similar to *Candida glabrata*. *Int J Syst Evol Microbiol* **56**, 313–317. (doi:10.1099/ijs.0.64076-0)

41. Swofford DL. In press. *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) Version 4.0 beta (2001).*

42. Huson DH. 1998 SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73.

43. Danecek P *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

44. Purcell S *et al.* 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575. (doi:10.1086/519795)

45. Alexander DH, Novembre J, Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. (doi:10.1101/gr.094052.109)

46. Abraham G, Inouye M. 2014 Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**. (doi:10.1371/journal.pone.0093766)

47. Jolley KA, Bray JE, Maiden MCJ. 2018 Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124. (doi:10.12688/wellcomeopenres.14826.1)

48. Hudson RR, Kaplan NL. 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164. (doi:10.1093/genetics/111.1.147)

49. Agapow P-M, Burt A. 2001 Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes* **1**, 101–102. (doi:10.1046/j.1471-8278.2000.00014.x)

50. Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

51. Yang Z, Nielsen R. 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.

52. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676. (doi:10.1093/bioinformatics/bti610)

53. Weig M, Jänsch L, Gross U, De Koster CG, Klis FM, De Groot PWJ. 2004 Systematic identification *in silico* of covalently bound cell wall proteins and analysis of protein-polysaccharide linkages of the human pathogen *Candida glabrata*. *Microbiology (Reading)* **150**, 3129–3144. (doi:10.1099/mic.0.27256-0)

54. Irinyi L *et al.* 2015 International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med Mycol* **53**, 313–337. (doi:10.1093/mmy/myv008)

55. Carreté L, Ksiezopolska E, Gómez-Molero E, Angoulvant A, Bader O, Fairhead C, Gabaldón T. 2019 Genome comparisons of *Candida glabrata* serial clinical isolates reveal patterns of genetic variation in infecting clonal populations. *Front. Microbiol.* **10**. (doi:10.3389/fmicb.2019.00112)

56. Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH. 2004 Evolution of the MAT locus and its Ho endonuclease in yeast species. *PNAS* **101**, 1632–1637.

57. Srikantha T, Lachke SA, Soll DR. 2003 Three Mating Type-Like Loci in Candida glabrata. *Eukaryot Cell* **2**, 328–340. (doi:10.1128/EC.2.2.328-340.2003)

58. Koszul R *et al.* 2003 The complete mitochondrial genome sequence of the pathogenic yeast *Candida (Torulopsis) glabrata*. *FEBS Lett* **534**, 39–48. (doi:10.1016/s0014-5793(02)03749-3)

59. Kryazhimskiy S, Plotkin JB. 2008 The population genetics of dN/dS. *PLoS Genet* **4**. (doi:10.1371/journal.pgen.1000304)

60. Teunissen AW, Steensma HY. 1995 Review: the dominant flocculation genes of Saccharomyces cerevisiae constitute a new subtelomeric gene family. *Yeast* **11**, 1001–1013. (doi:10.1002/yea.320111102)

61. Garcia-Rubio R, Jimenez-Ortigosa C, DeGregorio L, Quinteros C, Shor E, Perlin DS. In press. Multifactorial Role of Mitochondria in Echinocandin Tolerance Revealed by Transcriptome Analysis of Drug-Tolerant Cells. *mBio* **0**, e01959-21. (doi:10.1128/mBio.01959-21)

62. Aguileta G, de Vienne DM, Ross ON, Hood ME, Giraud T, Petit E, Gabaldón T. 2014 High variability of mitochondrial gene order among fungi. *Genome Biol Evol* **6**, 451–465. (doi:10.1093/gbe/evu028)

63. Fraser JA, Heitman J. 2003 Fungal mating-type loci. *Curr Biol* **13**, R792-795. (doi:10.1016/j.cub.2003.09.046)

64. De Las Peñas A, Pan S-J, Castaño I, Alder J, Cregg R, Cormack BP. 2003 Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev* **17**, 2245–2258. (doi:10.1101/gad.1121003)

65. Makova KD, Hardison RC. 2015 The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213–223. (doi:10.1038/nrg3890)

66. Feng W, Yang J, Xi Z, Qiao Z, Lv Y, Wang Y, Ma Y, Wang Y, Cen W. 2017 Mutations and/or overexpressions of ERG4 and ERG11 genes in clinical azoles-resistant isolates of *Candida albicans*. *Microb Drug Resist* **23**, 563–570. (doi:10.1089/mdr.2016.0095)

67. Katiyar SK, Alastruey-Izquierdo A, Healey KR, Johnson ME, Perlin DS, Edlind TD. 2012 Fks1 and Fks2 are functionally redundant but differentially regulated in Candida glabrata: implications for echinocandin resistance. *Antimicrob Agents Chemother* **56**, 6304–6309. (doi:10.1128/AAC.00813-12)

68. Xu Z, Green B, Benoit N, Sobel JD, Schatz MC, Wheelan S, Cormack BP. 2021 Cell wall protein variation, break-induced replication, and subtelomere dynamics in *Candida glabrata*. *Molecular Microbiology* **116**, 260–276. (doi:10.1111/mmi.14707)

Tables

Table 1

Category	GO/PFAM term	Genes ω< 1	Genes ω> 1	fisher p	q value	Rel. prop	GO/PFAM description
dN/dS>1	GO:0003723	428	25	2.26E-04	3.04E-02	1.95	RNA binding
	GO:0003735	152	3	7.38E-05	1.49E-02	5.78	structural constituent of ribosome
	GO:0003824	1745	155	8.61E-05	1.64E-02	1.28	catalytic activity
	GO:0005515	1100	81	4.48E-06	1.41E-03	1.55	protein binding
	GO:0005740	278	11	5.08E-05	1.06E-02	2.88	mitochondrial envelope
	GO:0005759	160	4	1.90E-04	2.65E-02	4.56	mitochondrial matrix
	GO:0006412	287	11	2.68E-05	6.05E-03	2.98	translation
	GO:0019693	100	1	3.63E-04	4.35E-02	11.4	ribose phosphate metabolic process
	GO:0022626	87	0	1.16E-04	1.99E-02	N/A	cytosolic ribosome
	GO:0022857	277	13	4.08E-04	4.70E-02	2.43	transmembrane transporter activity
	GO:0036094	724	52	2.58E-04	3.31E-02	1.59	small molecule binding
	GO:0043168	689	48	1.89E-04	2.65E-02	1.64	anion binding
	GO:0044281	498	27	1.57E-05	3.86E-03	2.1	small molecule metabolic process
	GO:0044391	143	3	2.32E-04	3.05E-02	5.44	ribosomal subunit
	GO:0071840	1325	113	2.76E-04	3.47E-02	1.34	cellular component organization or biogenesis
	GO:1901362	364	19	1.92E-04	2.65E-02	2.18	organic cyclic compound biosynthetic process
	PF00624.20	4	23	6.53E-20	7.42E-17	0.02	Flocculin repeat
	PF10528.11	12	10	1.87E-05	1.07E-02	0.13	GLEYA domain
	PF00514.25	8	8	5.72E-05	2.17E-02	0.11	Armadillo repeat

Microevolution (Frameshift)	GO/PFAM term	Genes without frameshift	Genes with frameshift	fisher p	q value	Rel. prop	GO/PFAM description
	GO:0009986	30	9	7.53E-12	4.26E-08	0.03	cell surface
	GO:0009987	3616	17	1.20E-05	3.38E-02	1.7	cellular process
	PF05001.15	0	17	1.10E-39	1.25E-36	0	RNA polymerase Rpb1 C-terminal repeat
	PF10528.11	13	9	5.54E-15	3.15E-12	0.01	GLEYA domain
	PF00399.21	24	9	4.09E-13	1.55E-10	0.02	Yeast PIR protein repeat
	PF08238.14	14	6	2.64E-09	7.49E-07	0.02	Sel1 repeat
	PF11765.10	6	3	2.58E-05	5.86E-03	0.01	Hyphally regulated cell wall protein N-terminal
	PF09770.11	0	2	4.79E-05	9.08E-03	0	Topoisomerase II-associated protein PAT1

Microevolution (Non-synonymous)	GO/PFAM term	Genes without non-synonymous	Genes with non-synonymous	fisher p	q value	Rel. prop	GO/PFAM description
	GO:0009986	28	11	7.88E-10	4.45E-06	0.06	cell surface
	PF10528.11	12	10	3.14E-11	3.58E-08	0.03	GLEYA domain
	PF11765.10	4	5	1.03E-06	5.87E-04	0.02	Hyphally regulated cell wall protein N-terminal

Table 2

Case ID	Initial	Relapse	ST	All Mutations	Coding	Non-coding	Coding Indel	Coding Indel (frameshift)	Coding (Non.Syn.)	Coding Nonsense	Coding (Syn.)	Coding (revert to ref.)
1	CG107A	CG107B	36	97	83	14	18	9	16	0	17	32
2	CG18A	CG18B	10	140	127	13	20	7	24	0	20	63
3	CG191A	CG191B	10	64	53	11	13	4	11	0	8	21
3	CG191B	CG191C	10	78	66	12	14	9	10	0	8	34
3	CG191C	CG191D	10	83	72	11	17	12	14	0	10	31
3	CG191D	CG191E	10	96	85	11	14	4	12	0	9	50
3	CG191E	CG191F	10	87	77	10	23	16	19	0	17	18
4	CG48A	CG48F	7	92	79	13	10	3	11	0	13	45
5	CG84F	CG84G	67	76	64	12	17	5	8	0	5	34
5	CG84G	CG84H	67	71	59	12	21	5	9	0	3	26
6	CG93A	CG93B	162	125	114	11	20	4	15	0	16	63
6	CG93B	CG93C	162	119	105	14	23	10	23	0	18	41
6	CG93C	CG93D	162	124	110	14	19	7	16	0	9	66
6	CG93D	CG93E	162	112	97	15	24	10	20	1	12	40
6	CG93E	CG93H	162	119	105	14	21	5	20	0	18	46
6	CG93H	CG93I	162	116	102	14	20	6	13	0	14	55
6	CG93I	CG93J	162	96	82	14	22	13	13	0	11	36
6	CG93J	CG93K	162	135	120	15	21	11	22	1	28	48
7	CG97A	CG97B	25	79	66	13	14	3	15	0	10	27
7	CG97B	CG97C	25	86	75	11	11	4	12	0	18	34

Title: Population genetics and microevolution of clinical *Candida glabrata* reveals recombinant sequence types and hyper-variation within mitochondrial genomes, virulence genes and drug-targets

Authors: Nicolas Helmstetter[*,††], Aleksandra D. Chybowska[†,††], Christopher Delaney[‡], Alessandra Da Silva Dantas[*], Hugh Gifford[*], Theresa Wacker[*], Carol Munro[†], Adilia Warris[*], Brian Jones[§], Christina A. Cuomo[**], Duncan Wilson[*], Gordon Ramage[‡] and Rhys A. Farrer[*,**]

Authors institutional affiliations:

[*]Medical Research Council Centre for Medical Mycology at the University of Exeter, Exeter, UK, EX4 4QD.

[†]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK, AB25 2ZD.

[‡]School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. G12 8QQ.

[§]Institute of Infection, Immunity & Inflammation, University of Glasgow, UK. G12 8TA.

[**]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 02142.

[††]These authors contributed equally.

Raw sequences for all isolates of *C. glabrata* from this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA669061.

Short Running Title: Population genetics of *C. glabrata*

Keywords: *Candida glabrata*, genome sequencing, epidemiology, candidiasis, microevolution, mitochondria, drug-resistance, evolution.

Corresponding author: Rhys A. Farrer. MRC Centre for Medical Mycology University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK. T: +44(0)1392 727594; E: r.farrer@exeter.ac.uk

Abstract

Candida glabrata is the second most common etiological cause of worldwide systemic candidiasis in adult patients. Genome analysis of 68 isolates from 8 hospitals across Scotland, together with 83 global isolates, revealed insights into the population genetics and evolution of *C. glabrata*. Clinical isolates of *C. glabrata* from across Scotland are highly-genetically diverse, including at least 19 separate sequence types (STs) that have been recovered previously in globally diverse locations, and one newly discovered ST. Several STs had evidence for ancestral recombination, suggesting transmission between distinct geographical regions has coincided with genetic exchange arising in new clades. Three isolates were missing MATα1, potentially representing a second mating type. Signatures of positive selection were identified in every ST including enrichment for Epithelial Adhesins (EPA) thought to facilitate fungal adhesion to human epithelial cells. In patent microevolution was identified from seven sets of recurrent cases of candidiasis, revealing an enrichment for non-synonymous and frameshift indels in cell surface proteins. Microevolution within patients also affected EPA genes, and several genes involved in drug resistance including the ergosterol synthesis gene *ERG4* and the echinocandin target *FKS1*/2, the latter coinciding with a marked drop in fluconazole MIC. In addition to nuclear genome diversity, the *C. glabrata* mitochondrial genome was particularly diverse, appearing reduced in size and with fewer conserved protein encoding genes in all non-reference ST15 isolates. Together, this study highlights the genetic diversity present within the *C. glabrata* population that may impact virulence and drug resistance, and two major mechanisms generating this diversity: microevolution and genetic exchange/recombination.

Article Summary (80 words)

Candida glabrata is a leading human fungal pathogen worldwide, which is increasing in prevalence and evolving antifungal resistance. Here, we report the largest whole-genome sequencing project for *C. glabrata* to date based on clinically derived candidemia isolates from hospitals across Scotland in the United Kingdom. We discover evidence for a second mating type, evidence for recombination between sequence types, hyper-diverse mitochondrial genomes, signatures of positive selection in pathogenicity genes, and in patient microevolution of drug-resistance genes.

Introduction

Candida is the most prominent genus of the Debaryomycetaceae family, with over 400 genetically and phenotypically diverse species currently described [1]. Many of these species are harmless commensals of the mucous membranes and digestive tracts of healthy individuals. Approximately 30 *Candida* species are of clinical importance in humans. Most of these species that are capable of causing disease in humans belong to the CTG-Serine clade, including *C. albicans*, *C. dubliniensis*, *C. tropicalis*, *C. parapsilosis*, *C. lusitaniae*, *C. guilliermondii and C. auris*, while others such as *C. glabrata* and *C. bracarensis* belong to the genetically distant Nakaseomyces clade [1,2]. In adult patients, *C. glabrata* is the second most commonly isolated species after *C. albicans*, which together cause approximately three quarters of all systemic candidiasis [3,4]. Infections caused by these species range from mild vulvovaginal candidiasis (VVC or thrush) to severe, drug resistant and difficult to treat invasive infections affecting single organs or the blood stream

(candidemia) with or without dissemination to the heart, brain, kidneys and other parts of the body [5]. Bloodstream infections caused by *Candida spp.* are associated with mortality rates of 30-60% [6,7]. Candidemia is associated with diverse risk factors including neutropenia, chemotherapy, diabetes, old age, compromised immune function, prolonged antibiotic and steroid treatment, and intravenous catheters that can harbour fungal biofilms [8]. Pathogenic *Candida* species including *C. glabrata* have also exhibited alarming increases in resistance against all major classes of antifungal drugs, hindering effective treatments and resulting in increasing mortality rates [4,9–12].

C. glabrata typically grows in the yeast form and is considered to have evolved an infection strategy based on stealth and evasion without causing severe damage in murine models [13]. This ability of *C. glabrata* and some of its relatives in the Nakaseomyces clade to infect humans is thought to have evolved recently [14], as several of its closest relatives have to date been exclusively isolated environmentally (*C. castellii*, *N. baccilisporus*, *N. delphensis*) [14]. Pathogenicity in the Nakaseomyces correlates with the number of Epithelial Adhesins (*EPA*) encoded in their genomes, which facilitate adherence and colonisation of human epithelial cells [15]. In contrast to *C. albicans*, the pathogenicity of Nakaseomyces species does not coincide with number or presence of Phospholipase-B and Superoxide Dismutase genes [16–19]. Many *Candida* genes involved in virulence are therefore likely to have diverse functions, some of which may not be conserved between distant clades.

Like many fungal pathogens, *C. glabrata*'s niche(s) and life cycle are poorly understood. *C. glabrata* is increasingly identified among clinical samples where it is responsible for an increasing proportion of cases of candidemia [4,20]. *C. glabrata* has also been identified environmentally, including as a component of the mycobiota of yellow-legged gulls [21], in droppings and cloaca swabs of birds of prey, migratory birds and passeriformes [16], and other potentially transitory sources including spontaneously fermenting coffee beans [22]. Concerted efforts for sampling are required to determine the true ecological distribution of *C. glabrata*, as they are for several other important *Candida* species such as *C. auris* [23]. Furthermore, the relatedness of global isolates and their routes of transmission (either patient to patient, or between patient and the environment) requires further studies comparing genotypes to collected metadata including location of isolation.

High levels of genetic heterogeneity have been identified in the *C. glabrata* population, as molecular methods have identified diverse strains, clades and Sequence Types (STs) both inter- and intra-nationally [24,25]. As of January 2022, the Multi Locus Sequence Type (MLST) database for *C. glabrata* included 233 STs from 1,414 isolates from 29 countries, based on the sequence identity for 6 genetic loci [26]. All isolates of *C. glabrata* reported to date have been haploid, with occasional aneuploids e.g. transitory disomies of chromosomes E and G [24].

Genetic heterogeneity within many fungal populations is shaped by their ability to switch between clonal and sexual recombination [27]. The ability for *C. glabrata* to undergo a sexual cycle remains unknown, with all reported attempts in the laboratory to encourage mating thus far unsuccessful. *C. glabrata* has therefore

been regarded as an asexual species, despite the presence of well conserved mating loci [28,29] and 14 examples of phylogenetic incompatibilities from multi locus sequencing [30]. More recent genomic analysis from 34 globally isolated *C. glabrata* strains revealed evidence of population admixture, suggesting a thus far undiscovered sexual cycle [24]. Greater sampling efforts and genomic analyses are therefore required to fully explore signatures of adaptation, virulence and recombination.

In this study, we explore the population genetics and microevolution of *C. glabrata* using comparative genomic analysis of 68 clinical *C. glabrata* isolates from 8 hospitals across Scotland, combined with 83 publicly available and globally isolated genomes, finding evidence of recombinant STs, hypervariable mitochondrial genomes, as well as variation in virulence genes and drug-targets between STs and between serial isolates from prolonged or recurrent infection.

Materials and Methods

Library preparation, sequencing and antifungal tests

Candida glabrata was collected from blood in 2012 from eight hospitals in Scotland (Table S1). These isolates were collected as part of a retrospective study of all cases of *Candida* blood stream infections carried out within Scotland under NHS Caldicott Guardian approval from March 2012 to February 2013, as described previously [4,31].

Genomic and mitochondrial DNA was extracted from 68 isolates using the QIAamp® DNA mini extraction kit (Qiagen) according to the manufacturer's

instructions. A small modification was made prior to extraction which was to mechanical disrupt the yeast. This was achieved by bead-beating the cells with sterile acid-washed 0.5 mm diameter glass beads (Thistle Scientific) for 3 x 30s. Following isolation and extraction using the QIAamp columns the DNA was eluted into elution buffer before storage at -20°C and transport to the sequencing facility.

Library preparation was performed by the Centre for Genome-Enabled Biology and Medicine at the University of Aberdeen. Briefly, gDNA quality was assessed on a Tapestation 4200 with a high sensitivity genomic DNA tape (Agilent) and quantified by fluorimetry using Quant-IT dsDNA High-sensitivity (HS) assay (Thermo Fisher). Dual indexed Illumina libraries were prepared from 1 ng purified gDNA using an Illumina Nextera XT DNA library preparation kit and Nextera XT v2 indices, which were purified from free adapters using AMPure XP beads (Beckman Coulter). Libraries were quantified using Quant-IT dsDNA High-sensitivity (HS) assay (Thermo Fisher) and average fragment size was calculated on Tapestation 4200 (Agilent), then equimolar pooled at 10nM. Concentration of the pool was verified by qPCR (Kapa library quantification kit, Roche) on QuantStudio 6 using SYBR green, and 1.8 pM of the library pool was sequenced on an Illumina NextSeq500 with 150bp paired end reads and 8bp index reads to average alignment depths of 41.9X (Table S2). This data was supplemented with paired end Illumina reads from Carreté *et al.* [24], and isolate CBS138 from Xu *et al*. [32].

Minimum inhibitory concentration (MIC) tests for fluconazole were performed at the Mycology Reference Laboratory, Public Health England, Bristol, according to

199 standard Clinical and Laboratory Standards Institute (CLSI) broth microdilution M27
200 guidelines [33].
201
202 **Variant calling**
203 The Genome Analysis Toolkit (GATK) v.4.1.2.0 [34] was used to call variants.
204 Our Workflow Description Language (WDL) scripts were executed by Cromwell
205 workflow execution engine v.48 [35]. Briefly, raw sequences were pre-processed by
206 mapping reads to the reference genome *C. glabrata* CBS138 using BWA-MEM
207 v.0.7.17 [36]. Next, duplicates were marked, and the resulting file was sorted by
208 coordinate order. Intervals were created using a custom bash script allowing parallel
209 analysis of large batches of genomics data. Using the scatter-gather approach,
210 HaplotypeCaller was executed in GVCF mode with the haploid ploidy flag. Variants
211 were imported to GATK 4 GenomicsDB and hard filtered if QualByDepth (QD) < 2.0,
212 FisherStrand (FS) > 60.0, root mean square mapping quality (MQ) < 40.0, Genotype
213 Quality (GQ) $\geq$ 50, Allele Depth (AD) $\geq$ 0.8, or Coverage (DP) $\geq$ 10.
214
215 To identify aneuploid chromosomes, depth of coverage was calculated for
216 each of 206 fungal samples. Sorted BAM files prepared in the pre-processing phase
217 of SNP calling were passed to Samtools v.1.2 [37] and mpileup files were generated.
218 Read depth was normalised by total alignment depth and plotted against the location
219 in the genome using 10 kb non-overlapping sliding windows. To identify structural
220 variation, assembly *de novo* was achieved using Spades v3.12 [38] using default
221 parameters.
222
223 **Phylogenetic and population genetic analysis**

To construct species-specific phylogenetic trees, all sites that were either a homozygous reference or SNP in every isolate were identified using ECATools (https://github.com/rhysf/ECATools) and concatenated into a FASTA file. Our rooted tree *C. bracarensis* included 1,198 phylogenetically informative sites, while the unrooted C. glabrata tree included 34,980 phylogenetically informative sites. Phylogenetic trees were constructed with RAxML PThreads v.7.7.8 [39] using the general-time-reversible model and CAT rate approximation with 100 bootstrap support, both with rooting to *C. bracarensis* AGP [40] or midpoint rooting without *C. bracarensis*. We constructed a tree using the same models with 1000 bootstrap support for all Saccharomycetaceae species that had a genome assembly in NCBI or JGI Mycocosm. We also constructed neighbour-joining trees using PAUP v4.0b10 [41] and a NeighborNet Network with SplitsTree v.4.15.1 [42]. Trees were visualised using FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

A multisample variant call format (VCF) corresponding to all 151 genomes was made with VCFTools v0.1.12 vcf-merge [43] and converted to ped and map file formats for use in PLINK v1.90 [44]. VCFTools was used to calculate genetic diversity metric pi, using the --site-pi parameter. Unsupervised ADMIXTURE [45] (settings: --haploid="*" -s time) was run on a moderately linkage disequilibrium (LD)-pruned alignment (PLINK --indep-pairwise 60 10 0.1) for values of K between 1 and 35. A value of K = 20 provided the lowest cross-validation error. Principle component analysis was performed using SmartPCA v4 [46]. Consensus gene sequences for each isolate were generated, and genes *FKS2* (CAGL0K04037g) positions 240-828, *LEU2* (CAGL0H03795g), *NMT1* (CAGL0A04059g), *TRP1* (CAGL0C04092g), *UGP1* (CAGL0L01925g), and *URA3* (CAGL0I03080g) were used to identify known MLST

for each of the isolates, and confirm that isolate CG57 was an unknown MLST and registered as the new ST204 on PubMLST (https://pubmlst.org/organisms/candida-glabrata) [47].

We applied Weir's estimator [48] of Wright's Fixation Index (F_{ST}) according to the equations given in Multilocus 1.3 [49] using non-overlapping sliding windows. The scripts have been made available online (https://github.com/rhysf/FSTwindows).

Selection and microevolutionary analysis

The direction and magnitude of natural selection for each ST were assessed by measuring the rates of non-synonymous substitution (*dN*), synonymous substitution (*dS*) and omega (ω = *dN*/*dS*) using the yn00 program of PAML [50], which implements the Yang and Nielsen method, taking into account codon bias [51]. Further GC corrections were not applied. The program was run on every gene in each isolate using the standard nuclear code translation table. To examine the functional significance of genes with ω > 1, we evaluated their Pfam domains and Gene Ontology (GO) terms for statistical enrichment (genes with ω > 1 vs, the remaining genes) using the two-tailed Fisher exact test with Storey false discovery rate (FDR)-corrected P values (q) of < 0.05. GO Terms were acquired using Blast2Go v6.0.1 [52] using Blastp-fast to the NCBI BLAST nr-database (E-value < 1E-5).

Genes of interest were defined including both *FKS* and 12 *ERG* pathway genes, as well as all genes listed in Table 1 of [53], which included adhesions including *EPA* genes, aspartic proteases, phospholipases, cell wall biogenesis,

structural wall proteins, regulatory, efflux pumps. This gene list was then screened for genes with either signatures of positive selection, or those undergoing microevolutionary changes (non-synonymous and frameshift indels).

Results

Recombinant sequence types in Scottish clinical samples

Clinical isolates of *Candida glabrata* from across Scotland are highly-genetically diverse. Using whole-genome sequencing, we analysed the genomes for 68 isolates of *C. glabrata* from 47 separate patients across eight Scottish hospitals, generating the largest panel of *C. glabrata* genome sequences to date. These 68 isolates belonged to twenty separate sequence-types (ST) of *C. glabrata*, which represent genetically related sub-populations based on alleles from six loci/genes. One isolate (CG57 from a single patient in Forth Valley Royal Hospital) belonged to a new ST that has not been previously identified anywhere else (ST204) (**Fig. 1, Table S1, Table S2**). Variant calling using the diploid model of GATK found few examples of heterozygosity (< 0.41 per kb for every isolate) suggesting all isolates were haploid (**Table S3**). Our panel of *C. glabrata* isolates was supplemented with a further 83 genomes from three recent studies of global *C. glabrata* isolates [24,25,32], as well as sequences from the outgroup *C. bracarensis* [40], which has also been identified from clinical settings and is the closest known relative of *C. glabrata* [14].

Phylogenetic analysis of our Scottish collection along with the worldwide *C. glabrata* isolates revealed high genetic diversity among the 29 separate STs represented by our combined panel (**Fig. 2, Fig. S1**). Allelic diversity among *C.*

glabrata isolates (mean nucleotide diversity (π) = 0.00665, σ = 0.047) was higher than previously reported (π = 0.00485 based on the Internal Transcribed Spacer (ITS) of 29 strains [54]). Our WGS-based calculation of *C. glabrata* π was the highest of any species in the Saccharomycetaceae that had both an available genome assembly and a calculation of π (albeit those are based on ITS sequences and fewer strains than we had) [54] (**Fig. 3a**). However, *C. glabrata* genetic diversity was typical among the Saccharomycotina (mean/$\bar{x}$ = 0.0055, median = 0.0039, standard deviaton/σ = 0.0055) (**Fig. 3b**). Nucleotide diversity within the population was present across the nuclear genome (**Fig. 3c**), with window length having some impact on the result (smaller window lengths (5 kb) resulted in higher average π in approximately half of the genome: chromosomes A-F, M and H). Most of the allelic diversity across the 151 *C. glabrata* isolates came from the nuclear genome (min. = 0.09 SNPs / kb, max. = 6.54 SNPs / kb, $\bar{x}$ = 5.55 SNPs / kb) compared with the mitochondrial genome (min. = 0.05 SNPs / kb, max. = 3.64 SNPs / kb, $\bar{x}$ = 1.21 SNPs / kb). Indeed, a significant difference between nuclear SNPs / kb and mitochondrial SNPs / kb was found using a two-tailed *t*-test for all 151 genomes (*p* = 5.6147E-111).

Seven clade (C) delineations for *C. glabrata* were recently proposed [24], which were equivalent to pre-existing STs including C1 (ST19), C2 (ST7), C3 (ST8), C4 (ST22), C6 (ST136) and C7 (ST3). We found that C5 was polyphyletic, encompassing isolates belonging to the genetically divergent ST6, ST10 and ST15 (**Fig. S1**). Therefore, we recommend the use of ST delineations rather than those clade delineations.

Several *C. glabrata* STs had evidence of genetic recombination. Our neighbour-net network tree of all isolates suggested historic gene-flow between several STs including for example ST7, ST55 and ST162 (**Fig. 2**). Genomic regions with low Wright's fixation index (F_{ST}) values, consistent with genetic exchange, were also identified from pairwise comparisons (n = 435) across 5 kb and 10 kb non-overlapping windows of all STs (**Fig. S2**). xF_{ST} values calculated from 5 kb windows were slightly lower than those calculated from 10 kb windows (averaging -0.046 for each ST pairwise comparison), indicating that window length impacts this measure of genetic variation. Twelve pairwise comparisons from 10 kb windows had $F_{ST} < 0.9$ across the genome (**Fig. S3**), with the lowest for ST18 and ST26 (F_{ST} = 0.64). Additionally, ST7, ST55 and ST162 had lower F_{ST} values across the genome (F_{ST} = 0.65 - 0.83) than other pairwise comparisons demonstrating incompatible phylogenetic signals between these STs (**Fig. S2**).

Principal-component analysis (PCA) of whole-genome SNPs revealed little evidence of clustering of several *C. glabrata* STs, which is consistent with gene flow between them (**Fig. 4A**). For unsupervised model-based clustering with ADMIXTURE, we first identified that K = 20 had the lowest cross-validation error (**Fig. 4B**), and was therefore used for subsequent analysis. Two isolates were consistently (6 independent Admixture runs) found to have evidence for mixed ancestry: ST177 CG1 and our newly discovered ST204 CG57 (**Fig. 4C, Fig. S4**). Other isolates were found to have evidence of mixed ancestry in the majority of runs including ST124 WM18.66, ST126 WM05.155 and ST8 M17.

Only one of the Scottish isolates (CG46) had evidence for Chromosome Copy Number Variation (CCNV)/aneuploidy, found in Chromosome C (**Fig. S5**). Distributions of normalised chromosome read depths of chromosome C (average depth per 10kb window = 0.68) differ significantly from the rest of the genome of CG46 (average depth per 10kb window = 1.05; Kolmogorov-Smirnov Test: p = 2.09E-25), with coverages of chromosome C significantly lower than in the rest of the genome (Wilcoxon rank-sum test: p = 1.353E-25). No other CCNVs were found, despite many isolates having been treated with antifungals that have previously been correlated with CCNV [55]. Together, these results suggest occasional genetic recombination within the *C. glabrata* population, without an association with aneuploidy.

Mating types and mating type switching are poorly understood in *C. glabrata*, although it is thought that Mating-type regulatory protein α2 is expressed in all MTLα strains and not in MTLa strains [56,57]. MATα2 (CAGL0B01265g) was present in all Scottish isolates (breadth of coverage; BOC > 87%). However, MATα1 appeared to be absent or partially absent in 3/9 ST6 isolates (CG12 = 16% BOC, CG121 = 18% BOC, CG42 = 12.5% BOC), while present in the remaining six ST6 isolates, and all the other STs (BOC 100%). The functional relevance of MATα1 deletion or truncation is unclear but may be a hallmark of the rarer of the two mating types.

Hyper-variable mitochondrial genomes among sequence types

F_{ST} analysis highlighted the mitochondrial genome of *C. glabrata* as hyper-variable (**Table S2, Fig. S6**). Forty-three genes were identified in ≥10 pairwise F_{ST} comparisons, including all eleven mitochondrial protein encoding genes. To explain

this enrichment of low F_{ST} mitochondrial genes, we studied the 151 genome alignments. While the nuclear genome had 97.3 - 99.4% BOC, the mitochondrial genome had 20.4 - 99.9% BOC, with 42% of isolates (*n* = 63/151) containing >10% ambiguous mitochondrial bases (2 kb) (Table S2) (here, we define ambiguous as sites with too few reads aligning to be called by GATK, or reads that cannot be called by GATK due to not passing variant filtration). Surprisingly, a pattern of low and/or patchy read coverage was identified in every isolate including the ST15 reference isolate CBS138 (**Fig. S6**), indicating that the reference mitochondrial sequence assembly [58] may have a high error rate, and given additional differences identified in non-reference isolates, that *C. glabrata* mitochondrial genomes are highly heterogenous.

The mitochondrial genome for some *C. glabrata* isolates appears reduced in size and encodes fewer protein encoding genes (**Fig. 5**). As many as 22/37 (59%) mitochondrial encoded genes were entirely absent in at least one isolate, including Cg1, Cg1II, and Cg1III (putative endonucleases of exons and introns in the mitochondrial COX1 gene), *ATP8* and *ATP9* (subunits 8 and 9 of the enzyme complex required for ATP synthesis), *RPM1*/*RPR1* (RNA component of mitochondrial RNAse P), *VAR1* (putative mitochondrial ribosomal protein of the small subunit,) and most of the tRNA genes (15/23). Nine separate STs had absent mitochondrial genes. Normalised depth of coverage was variable across the genes, with < 1 average normalised depth across all isolates for Cg1, Cg1II, and Cg1III, *ATP8, RPM1, VAR1* and tRNA-Met1. While non-uniform coverage in terms of depth and breadth was found across all mitochondrial genomes belonging to all datasets, our newly sequenced isolates have the lowest mean breadth across mitochondrial

genes ($\bar{x}$ = 92.18, σ = 20.07) compared with Biswas *et al.* [25] ($\bar{x}$ = 98.48, σ = 11.47) and Carreté *et al.* [55] ($\bar{x}$ = 97.03, σ = 14.11), suggesting there are some discrepancies between library preparation impacting mitochondrial read sequencing. Notably, only 1/50 Biswas *et al.* [25] isolates (WM03.450) had entirely absent mitochondrial genes compared with 8/32 Carreté *et al.* [55] isolates and 15/68 of our newly sequence isolates. Total sequencing depth can be ruled out as the main cause for low mitochondrial coverage, given Carreté *et al.* [55] had the highest sequencing depth ($\bar{x}$ = 360X) and had many isolates with absent mitochondrial genes, compared with Biswas *et al.* [25] ($\bar{x}$ = 74X) and ours ($\bar{x}$ = 42X).

We used assembly *de novo* to further explore the mitochondrial sequence for isolate WM03.450 (ST83), which had the greatest number of ambiguous bases across its mitochondrial genome (16 kb / 80%). Our WM03.450 Illumina-based assembly (12.9 Mb; N.contigs = ~3 thousand; N_{50} = 85 kb) is 562 kb longer than the CBS138 reference sequence, indicating substantial genomic differences between these isolates and STs. Aligning our assembly to the reference CBS138 mitochondrial genome using Blastn identified 10 contig matches with a combined alignment length of only 1.9 kb (mean 157 nt per contig), suggesting the low alignment is not due to conserved nucleotide sequences that have undergone large rearrangements. Aligning the assembly to the eleven mitochondrial protein sequences using Blastx identified only 6/11 genes across six separate contigs, four of which were < 364 nt length, and two that are 10.4 kb and 81.5 kb. Conversely, assembly *de novo* and Blastx of our Illumina reads for the reference isolate CBS138 against the published CBS138 genome identified all eleven mitochondrial genes present on four contigs, with 18.9 kb total sequence length, of which Blastn aligned

9.3 kb to the published mitochondrial assembly. Together, these analyses suggest that whole gene deletions in the *C. glabrata* mitochondria are common.

Signatures of selection identified among sequence types

In contrast to the *C. glabrata* mitochondrial genome, we found that gene deletions in the nuclear genome are rare. Indeed, fewer than six presence/absence (P/A) polymorphisms (strictly defined as zero reads aligning) were identified per isolate (~0.1% of 5,210 protein-encoding genes) (**Table S4**). Of these, two consecutive nuclear-encoded genes (CAGL0A02255g and CAGL0A02277g) on Chromosome A were entirely absent of read coverage in 25 out of the 68 Scottish isolates (37%), which included all representatives of eleven separate STs (ST4, 7, 8, 24, 25, 55, 67, 83, 177, and our newly described 204). These STs do not cluster phylogenetically, ruling out a single evolutionary event causing this deletion. The two genes have identical nucleotide sequences and encode the same amino acid sequence, which is conserved across a range of other Ascomycota species, as well as having sequence similarity to the K62 Killer Preprotoxin protein encoded by the *Saccharomyces paradoxus* L-A virus M62 satellite (BLASTp E-value = 1e-36), suggesting a possible viral origin. CAGL0F09273g is a separate, putative adhesin-like protein (adhesin cluster V) with a “hyphally regulated cell wall protein N-terminal” PFAM that is lost in eleven isolates including all ST4 (CG68A, CG77), four ST7 (CG157, CG48A, CG48F, CG78), three ST8 (CG127, CG52, CG82), ST19 CG119, ST24 CG166 and ST147 CG133. Again, this gene must have been lost multiple times, given its presence in several ST7 and ST8 isolates. This gene is the last gene on Chromosome F, has an $\pi = 0.00244$, which is lower than the overall average across the genome, and has previously been reported to undergo CCNV within serial

clinical isolates [55], suggesting it is able to undergo variation within microevolutionary timescales, which may impact the adhesive ability of these *C. glabrata* isolates.

Between 61 and 85 genes with a signature of positive selection ($dN/dS = \omega$, and $\omega > 1$ [59]) were found in each ST apart from the reference ST (ST15 CG151), for which only a single gene with $\omega > 1$ ($\omega = 1.0019$) was identified (**Table S5**). Apart from the reference ST, STs had between 4 and 14 genes with $\omega > 2$, showing stronger signatures of diversifying or positive selection. Of the 2,083 total genes with $\omega > 1$ across all clades, 608 were unique genes (11.6% of all genes) i.e., they had this signature in multiple clades, owing to either ancestry or selection acting on the same gene families. To explore selection, we took an unbiased approach using PFAM and GO-term enrichment comparing the numbers of each term in those 608 genes compared with the remaining non-selected genes, as well as a targeted approach for genes of interest (see Methods Selection and microevolutionary analysis) including adhesins, proteases, efflux pumps, FKS, and ERG pathway genes.

Genes with signatures of positive selection within the *C. glabrata* population targets diverse genes and gene functions. Our unbiased approach for enrichment of functional domains in 608 gene products with signatures of positive selection identified only three significantly enriched (two-tailed Fisher exact test with false discovery rate (FDR)-corrected *p*-values (q) of < 0.05) PFAM domains and 16 GO terms (Table 1). The enriched PFAM domains were 1) Flocculin repeat (PF00624.20; $q = 7.42E-17$), 2) GLEYA domain (PF10528.11; $q = 0.01$) and 3) Armadillo repeat

(PF00514.25; q = 0.02). Flocculin is a sub-telomeric gene family involved in flocculation or cell aggregation in *S. cerevisiae* [60], while GLEYA domains are present in *C. glabrata EPA* proteins. Thirty Flocculin PFAM domains were assigned to only six genes in *C. glabrata*, two of which have $\omega > 1$: CAGL0I07293g and CAGL0I00220g, and together account for 23/30 Flocculin repeat PFAMs. Enriched GO-terms covered a range of possible biological functions including ribosomal/RNA-binding and mitochondrial structural proteins.

Our targeted approach highlighted 21/129 genes of interest that have $\omega > 1$, with at least one found in every ST apart from the reference ST15 and ST46 (**Table S6**). Notably, none of the aspartic proteases, phospholipases, cell wall biogenesis, efflux pumps, ergosterol biosynthesis pathway genes or *FKS* genes were found to have hallmarks of positive selection, implying these are conserved within the population. Several genes with $\omega > 1$ were found in multiple STs, including adhesive protein CAGL0J01727g (adhesin cluster VI) that is under positive selection in seven STs (18, 26, 36, 45, 147, 177, 204) and adhesive protein CAGL0I07293g (adhesin cluster V) under positive selection in seven mostly distinct STs (3, 8, 25, 83, 123, 136, 177). *C. glabrata* encodes 17 putative adhesive proteins without N-terminal signal peptides, casting doubt on their role in adhesion. One of these is a pseudogene (CAGL0E00110g) with $\omega > 1$ in 13/29 STs. The structural cell wall protein *AWP7* belonging to the Srp1p/Tip1p family was under selection in 7 STs.

C. glabrata nosocomial in-patient microevolution targets pathogenicity factors and drug targets

Our Scottish *C. glabrata* panel included seven sets of between 2 and 9 isolates from recurrent cases of candidiasis. To explore the microevolution of *C. glabrata* within a human host, and the effects of antifungal treatment (fluconazole, nystatin, and posaconazole) on fungal genetics, we documented all genetic changes between serial isolates (**Table 2, Table S7**). Although exact dates of isolation were not documented, phylogenetic analysis (**Fig. S1, Fig. 6**) confirmed these serial isolates were highly related, with between 64 and 140 mutations (1.13468×10^{-5} per base pair) identified between pairs of serial isolates (**Fig. 6**). While the mutation rate or generation time for *C. glabrata* is not known [55], this small number of mutations likely suggests recent clonal origins appropriate for microevolutionary analysis. Serial isolates had an estimated time between isolation (based on blood culture dates) between 0 and 239 days (mean 15 days). Five serial isolates from 4 separate patients/cases showed MIC changes from the earlier sampled isolate (**Table 3**), including 2 increases (CG107A->B +8 ug/mL, CG97B->C +4 ug/mL), 1 decrease (CG84G->H -4 ug/mL,) and 1 large transient increase (CG93A, B, C, D, E = 4 ug/mL; CG93H, I, K >64 ug/mL, CG93K = 4 ug/mL).

Mutations identified between serial isolates were mostly in protein coding sequence (CDS) regions (between 53 and 127 mutations per pairs of serial isolates, collectively adding up to 1,741/1,995 total mutations = 87%), despite protein-coding regions taking up only 7.9 / 12.3 Mb (64%) (**Fig. 6B, 6C**). The remaining serial mutations were within intergenic regions (236 mutations; 12%) and intronic regions (18 mutations; 1%). Intronic regions had the highest count of serial mutations after accounting for the total sequence in introns (**Fig. 6C**), albeit with ≤ 3 found per pair of serial isolates. Hypergeometric tests revealed that the number of mutations in coding

sequence compared with non-coding sequence was higher than expected by chance, suggesting a highly significant enrichment of mutations in protein coding genes (*p* = 3e-120).

To explore the 1,741 microevolutionary changes within coding regions, we categorised them into five groups of newly arising mutations (regardless of prior state): 1) insertions/deletions (indels) (*n* = 362; 21%), 2) synonymous mutations (*n* = 264; 15%), 3) non-synonymous mutations (*n* = 303, 17%), 4) nonsense mutations (*n* = 2), and 5) reversion back to reference base (*n* = 810; 47%). Of the indels, 147/362 (41%) were frameshifts that disrupted 54 genes. Non-synonymous mutations were detected in 139 genes (**Fig. 6D, 6E**).

Enrichment for PFAM/GO-terms of these genes with frameshift and non-synonymous mutations (two-tailed Fisher exact test with false discovery rate (FDR)-corrected *p* values (*q*) of < 0.05) revealed three enriched GO-terms and eight enriched terms (**Table 1**). Both categories (frameshifts and non-synonymous mutations) were enriched for GO:0009986 Cell Surface (*q* = 3.21E-08 and 1.09E-06, respectively), suggesting that *C. glabrata* undergoes rapid mutations in several of its cell surface proteins during prolonged/serial blood stream infections. Enriched PFAM terms included the “RNA polymerase *RPB1* C-terminal repeat” for genes with frameshift indels (*q* = 1.25E-36), GLEYA domains for genes with either frameshift (*q* = 3.15E-12) or non-synonymous mutations (*q* = 3.58E-08). Several repeat associated PFAMs and the “Hyphally regulated cell wall protein N-terminal” domain were enriched for non-synonymous mutations (*q* = 5.87E-04).

Several genes of interest (see Methods: Selection and microevolutionary analysis) had microevolutionary changes (*n* = 29/129) (**Table S8**). Notably, one of the two newly acquired nonsense mutations was identified in *FKS2* (Case 6 J->K), coinciding with a substantial drop in fluconazole MIC (**Table 3**). The other was in the uncharacterised CAGL0K04631g at an earlier time point in the same patient (Case 6 D->E).

Twenty adhesins including *EPA* genes were mutated between serial isolates, including in all seven sets/cases of isolates and at every time point. For example, *EPA3* had 5 indels in Case 1 (A->B), a synonymous mutation in Case 2 (A->B; nucleotide position (pos.) 2304), Case 3 (D->E; pos. 1539), Case 4 (A->F; pos. 1119), Case 5 (F->G; pos. 2259), a non-synonymous mutation (pos. 2224) and large (30nt) insertion in Case 5 (G->H), two large deletions (42nt and 16nt), and two synonymous and one non-synonymous mutations in Case 6 (A->B; pos. 1002, 2319 and 2276 respectively).

The longer 42nt deletion from Case 6 (A->B) reverts back to reference in Case 6 (B->C), suggesting either a) a non-descendent isolate (intra-host variation), b) a false negative reference in 6C or c) a false positive deletion in 6A. The same 42nt deletion, along with a new insertion at the site of the previous synonymous mutation appears in Case 6 (C->D), thereby suggesting the variant is real and option c less likely. That 42nt deletion reverts back to reference in Case 6 (D->E), and appears again in Case 6 (E->H). By Case 6 (H->I), the gene has a new synonymous mutation, and in Case 6 (I->J) it has accumulated a new 15nt deletion. *EPA3* is therefore a hot-spot of variation. Another EPA gene that accumulated a large

number of mutations was *AWP12*, which accumulated five non-synonymous mutations and one synonymous mutation (Case 6H->I).

Other genes that had accumulated mutations between serial isolates included those encoding an aspartic protease *YPS5*, several structural wall proteins belonging to the Srp1/Tip family, regulatory protein *PDR1*, the ergosterol synthesis gene *ERG4* (a non-synonymous mutation in Case 3A->B), and both *FKS1* and *FKS2*. Therefore, *C. glabrata* genes that are antifungal targets and gene families involved in drug-resistance and pathogenicity can therefore undergo rapid mutation within a human host.

Discussion

In this study, we sequenced and analysed the largest panel of *C. glabrata* genomes to date. These isolates were collected from blood-stream infections of patients at several Scottish hospitals in 2012. Our 68 genomes were analysed alongside 83 further publicly available and globally isolated genomes [25,32,55], revealing greater genetic diversity than previously recognised, including a nucleotide diversity of 0.00665, which is much higher than has been calculated for the distantly related *C. albicans* at 0.00298 [54]. Surprisingly, we found that only one of our Scottish isolates had evidence of aneuploidy, despite many having been treated with antifungals, which has previously been correlated [55]. Chromosome C in CG46 had lower depth of coverage compared with the rest of the genome, perhaps due to chromosome loss in a subset of cells. The patient that CG46 was isolated from was initially treated with Fluconazole. Following resistance to Fluconazole being

596 detected, the patient was subsequently treated with Caspofungin, suggesting a
597 potential link between those antifungal treatments and the observed aneuploidy.
598
599 We found that the mitochondrial genome of *C. glabrata* was hyper-diverse
600 compared with its nuclear genome for many isolates, including several long deletions
601 spanning one or more genes, with the potential to impact many important biological
602 functions including drug resistance and persistence [61]. High levels of variation in
603 mitochondrial genomes within the major fungal phyla have previously been noted in
604 terms of gene order, genome size, composition of intergenic regions, presence of
605 repeats, introns, associated ORFs, and evidence for mitochondrial recombination in
606 all fungal phyla [62]. This variation is lacking in Metazoa [62]. Our results suggest
607 some of these types of mitochondrial genetic diversity are likely present within the *C.*
608 *glabrata* population.
609
610 Isolates in this study belonged to twenty-nine separate sequence types (STs)
611 of *C. glabrata,* each of which was separated by large number of variants. However,
612 as many as 193 MLST STs have been documented [47]. Therefore, it is likely that
613 the true genetic diversity of *C. glabrata* is much higher than we have been able to
614 calculate with whole-genome sequences (albeit the largest panel studied to date).
615 Indeed, several further STs may yield further evidence of recombination or lack of,
616 and may ultimately require a new effort to group STs into lineages (also dependent
617 on the frequency of recombination that erode these divisions). The genetic diversity
618 of *C. glabrata* in hospitals around Scotland is extremely high, with representatives
619 from 20 STs. Such high genetic diversity (and many of the same STs) have also
620 been found from genome sequencing and phylogenetic analysis of isolates collected

621 in other countries such as Australia [25], suggesting they must have been
622 transported across or between continents, perhaps by anthropogenic or even natural
623 means (for example its association with birds [16,21] and food [22]). Greater
624 sampling and genotyping of clinical and environmental isolates will be required for
625 understanding ancestry or endemicity.
626
627 *C. glabrata* has long been regarded as a haploid asexual yeast, although
628 evidence has recently emerged of a cryptic sexual cycle [28–30]. Our genome
629 sequencing and population genetics supports this work, revealing compelling
630 evidence that at least 12 sequence types (STs) stem from recombination between
631 other STs. However, further work remains to document and describe individual
632 recombinant isolates. Providing genetic recombination between isolates is naturally
633 occurring, the mechanisms of genetic exchange are also unknown, although likely
634 relate to the conserved mating type locus, which play a central role in the sexual
635 cycle of diverse fungi [63]. Here, we show that the MATα1 gene was absent or
636 partially absent in three isolates belonging to ST6, which could potentially impact or
637 even be a hallmark of a rarer second mating type of *C. glabrata*. Together, genetic
638 recombination among *C. glabrata* isolates appears much more common than
639 previously recognised, and likely contributes to increased genetic diversity.
640
641 The nuclear genome for isolates belonging to every ST (apart from the
642 reference ST15 that was included as a control) included evidence of positive or
643 diversifying selection. Signatures of positive selection were found enriched in genes
644 with diverse functions, including several with repeat domains, as well as EPA genes
645 with GLEYA domains. EPA genes are a large sub-telomeric family of virulence-

related surface glycoprotein-encoding genes encoded by several other pathogens including *Plasmodium*, *Trypanosoma*, and *Pneumocystis* [64]. Such gene differences between STs of *C. glabrata* may result in clinically-relevant phenotypic differences.

In host microevolutionary changes between serial isolates were enriched within coding-sequences, which is a surprise, given the expectation for intergenic regions to be more permissive to mutations due to relaxed selection within intergenic regions and purifying selection within coding sequence. The reason for this abundance of serial mutations in coding sequence is unclear, although it could potentially be technical (e.g. false negative variants within repetitive sequences) or biological (e.g. drug exposure and host immune pressure). Alternatively, enriched mutations in genes could potentially be driven by processes such as DNA polymerase induced mutations, or differences in chromatin states (e.g. heterochromatin could lead to increased exposure to DNA damaging agents resulting in higher mutation rates, or conversely, greater surveillance and correction of mutations in euchromatin regions by cellular DNA repair enzymes [65]).

Selection may explain why we identified similar numbers of non-synonymous mutations to synonymous mutations, given random mutations are expected to be non-synonymous in ~2/3 nucleotides of each codon. Furthermore, accumulations of deleterious mutations could be occurring in the serial isolates due to small population sizes, although population size estimates could not be calculated accurately from the metadata.

Genes with GLEYA domains including EPA genes were significantly enriched for frameshift and non-synonymous mutations in the coding sequence between serial isolates. Combined with our finding of positive selection in EPA genes across STs, suggests that EPA genes are undergoing variation at both longer time frames and microevolutionary time-scales.

Genes encoding several important drug-targets also underwent mutations between serial isolates, including a non-synonymous mutation in the ergosterol biosynthesis pathway gene *ERG4*, and a nonsense mutation in the 1,3-β-glucan synthase component *FKS2* (mutations in these genes can confer resistance to azoles [66] and echinocandins [67] respectively). Indeed, the nonsense mutation in FKS2 coincided with a marked drop in fluconazole MIC for isolate CG93K, suggesting a possible link.

Our study highlights the need for further sampling and genomic analysis of *C. glabrata* in order to better inform the population structure and mechanisms underlying its increasing emergence, pathogenicity and multi-drug resistance. While we have largely focused on differences among the conserved regions of the *C. glabrata* ST15 CBS138 genome using an alignment-based strategy, our discoveries of a hyper-diverse mitochondrial sequence highlight the value for future long-read sequencing and assemblies to characterise the pan-genomes of *C. glabrata* and structural genomic diversity that exists among and perhaps within STs, and to explore the mechanisms driving those changes. Furthermore, given the genetic diversity between STs that we document, it would likely be valuable to sequence and assemble additional high-quality reference sequences for the purposes of increasing

variant-calling accuracy and quantifying gene content between different STs. Given the low and patchy alignment depth across the ST15 CBS138 mitochondrial sequence for that same isolate, a review and update for the published CBS138 mitochondrial genome is likely required as well. Indeed, high (~0.5%–1%) frequencies of structural variation in the nuclear genomes of *C. glabrata* isolates was recently found using *de novo* assemblies from long single-molecule real-time reads [68].

The rapidity that *C. glabrata* can mutate important genes and gene families, both *via* microevolution and putative recombination highlights an obstacle for future drug-development, given that individual gene targets are able to mutate within short time spans, and substantial diversity already present between STs. In addition, the epidemiology of *C. glabrata* is poorly understood. Future sampling and genomic comparison studies are necessary to identify the routes and mechanisms of its spread and evolution.

Data availability

Raw sequences for all haploid isolates of *C. glabrata* from this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA669061.

Acknowledgments

R.A.F, D.W., T.W., and A.W. are supported by the Medical Research Council Centre for Medical Mycology MR/N006364/2. R.A.F. is supported by a Wellcome Trust Seed Award (215239/Z/19/Z). D.W. is supported by a Wellcome Trust Senior

721 Research Fellowship (214317/Z/18/Z). C.A.M is funded by the European Union's
722 Horizon 2020, Innovative Training Network: FunHoMic (grant N° 812969) and
723 consortium ‘Host-Directed Medicine in invasive FUNgal infections’—HDM-FUN
724 (Grant Agreement 847507). The authors thank Zeynab Heidari and Elaina Collie-
725 Duguid of the Centre for Genome-Enabled Biology and Medicine , University of
726 Aberdeen for their support and assistance in this work. We would also like to thank
727 Dr. Lucy van Dorp for valuable suggestions for running ADMIXTURE, and Prof. Ian
728 Stansfield for useful discussions.

729

730 **Competing Interest Statement**

731 The authors declare no competing interests.

732

733 **Figures and Tables**

734

735 Figure 1. *Candida glabrata* isolates were collected across eight health boards across
736 Scotland in 2012, belonging to 20 separate sequence types, including the newly
737 described ST204. Duplicate isolates stemming from the same patient at different
738 time points have been excluded.

739

740 Figure 2. A NeighborNet network using SplitsTree, with sequence types (ST) labels
741 replacing isolate names at the nodes. Green = found in Scotland, purple = not found
742 in Scotland. The scale bar represents nucleotide substitutions per site.

743

744 Figure 3. a) A RAxML phylogenetic tree with 1000 bootstrap support of all
745 Saccharomycetaceae species that had a genome assembly in NCBI or JGI

Mycocosm and nucleotide diversity (π). Note: C. glabrata is calculated from whole genome sequence data presented in this study, while the other species are based on ITS sequences only [54]. b) π (based on ITS sequences only) for all Saccharomycotina and non-Saccharomycotina that are listed in the ISHAM ITS reference DNA barcoding database [54]. c) non-overlapping 5 kb, 10 kb and 20 kb windows of $\bar{x}\pi$ (π for all sites in the genome divided by window length).

Figure 4. Population genetics of *C. glabrata* sequence types (ST). a) Principal-component analysis (PCA) of whole-genome SNPs using SmartPCA revealed little evidence of sub-clustering among STs (isolates are calculated and plotted ndividually, but labelled by their ST alone for clarity). SmartPCA failed to calculate the eigenvalues for some isolates including those belonging to ST4. b) The cross-validation (CV) error from running unsupervised ADMIXTURE for variant-sites across the *C. glabrata* population, testing K values between 1 and 35. K = 20 provided the lowest CV error. c) ADMIXTURE plot for all isolates using K = 20, revealing several isolates with evidence of mixed ancestry. Isolates are ordered according to the neighbor-joining tree constructed with PAUP in Figure S1.

Figure 5. Breadth of coverage and depth of coverage across each of the 37 mitochondrial encoded genes for all 151 isolates compared in this study (Each point represents an isolate). A) Breadth of coverage as a % across each gene. B) The Normalised depth of coverage for each gene (total read depth for each gene / total read depth across both nuclear and mitochondrial genomes). C) Breadth of coverage as a % across each gene, categorised by sequence types (ST)s. D) Normalised depth of coverage for each gene, categorised by ST.

Figure 6. Microevolutionary changes across seven sets of *C. glabrata* isolates. A) a RAxML phylogenetic tree of the serial isolates using the general-time-reversible model and CAT rate approximation with 100 bootstrap support. Branch lengths indicate the mean number of changes per site. B) The number of serial mutations total (All), those within protein-coding sequence (CDS), intergenic and intronic regions. C) Those same serial mutations per kb (calculated as the count of serial mutations divided by the total length of the feature (where All = whole genome) and multiplied by 1000). D) Serial mutations within CDS categorized by their effect on the sequencing: Insertion/Deletion (Indel), synonymous mutation (Syn.), non-synoynmous mutation (Non.Syn.) and nonsense mutation. E) Those same serial mutations within CDS per kb.

Figure S1. Phylogenetic trees of *C. glabrata*. All genomic sites that were either a homozygous reference or SNP in every isolate of *C. glabrata* and *C. bracarensis* AGP [40] for rooting were identified using ECATools and concatenated into a FASTA file. A) A neighbor-joining tree constructed with PAUP. Scale bar indicates the distance based on substitutions per site. B) A maximum likelihood tree constructed using RAxML PThreads v.7.7.8 [39] using the general-time-reversible model and CAT rate approximation with 100 bootstrap support. Branch lengths indicate the mean number of changes per site. The clade according to Carreté L *et al.* 2018 [24], as well as sequence type (ST), country code (AU = Australia, BE = Belgium, DE = Germany, FR = France, GB = Great Britain, IT = Italy, TW = Taiwan, US = United States), MAT and reference are also shown.

Figure S2. Mean F_{ST} values from pairwise comparisons of each sequence type (ST) calculated from a) 10 kb non-overlapping windows and b) 5 kb windows. C) Mean F_{ST} values from 10 kb windows were similar to values calculated from 5 kb windows, with a mean difference of -0.046 per pairwise comparison.

Figure S3. Non-overlapping 10 kb windows showing F_{ST} values for 12 pairwise comparisons that had long genomic regions with lower values.

Figure S4. Five independent runs of ADMIXTURE using K = 20 and time-based seed values, revealing several isolates with evidence of mixed ancestry. Isolates are ordered according to the neighbor-joining tree constructed with PAUP in Figure S1.

Figure S5. Non-overlapping 10 kb windows showing normalized depth of coverage (including GC normalization by percentiles; GC, and excluding ambiguous sites (effective window length)).

Figure S6. Integrated Genome Viewer (IGV) screenshots for the reference isolate CBS138, and all isolates compared in this study, indicating substantial differences between our Illumina sequences and the mitochondrial assembly. Gene features are shown as a track (directionality indicated by arrows), and the read alignment from the BAM files are shown for each isolate, where peaks indicate higher depth, and colors on the peaks indicate discrepancies to the reference base: green = A, blue = C, red = T, brown = G, purple = insertion)

Table 1. GO-term and PFAM enrichment (two-tailed Fisher exact test with false discovery rate (FDR)-corrected P values (q) of <0.05) for genes with *dN*/*dS* (ω) > 1, and genes with either microevolutionary frameshifts or non-synonymous mutations across the seven sets of serial isolates. The relative proportion (Rel. prop) was calculated as (number of terms in set 1 / number of terms in set 2) * (genes with any terms in set 2 / genes with any terms in set 1).

Table 2. Summary of microevolution across seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis. We documented 1,995 mutations between all serial isolates, which were either in protein-coding regions (Coding) or Intergenic and intron regions (Non-coding). Coding mutations were further characterised into Coding Indels, some of which caused frameshifts (Coding Indel (frameshift)), non-synonymous mutations (Coding Non.Syn.), nonsense mutations, (Coding Nonsense), synonymous mutations (Coding Syn.), and bases that reverted back to the ST15 CBS138 reference base (either from a previous microevolutionary change or a pre-existing variant between the initial isolate and the reference ST15 CBS138).

Table 3. Minimum inhibitory concentration (MIC) values of fluconazole for each of the serial isolates.

Table S1. Metadata for clinical Scottish *C. glabrata* sequenced and analysed in this study.

Table S2. Haploid variant call summary based on alignments to the published nuclear and mitochondrial assembly of ST15 CBS138. These variants form the basis for the population genetic and comparative genomics tests.

Table S3. Diploid variant call summary. Diploid variant calls were used to check for any evidence of heterozygosity suggestive of diploidy. All heterozygous sites (single nucleotide heterozygous positions + heterozygous insertions + heterozygous deletions) amounted to <0.0404% (0.4 per Kb) of total positions called per isolate, suggesting these were errors and not evidence of diploidy.

Table S4. Counts of presence/absence (P/A) polymorphisms in each isolate, based on zero reads aligning to the ST15 CBS138 nuclear and mitochondrial genomic regions encoding gene sequences.

Table S5. A summary of *dN*/*dS* (ω) and nonsense mutations found across every gene in isolates representing each of the sequence types (ST).

Table S6. Details of 21 genes (found 67 times across all STs) with *dN*/*dS* (ω) > 1, which belonged to our set of 129 “genes of interest” including adhesions (e.g. EPA genes), aspartic proteases, phospholipases, cell wall biogenesis, structural wall proteins, regulatory, efflux pumps (all genes in Table 1 of [53]), as well as both FKS and 12 ERG pathway genes.

Table S7. Counts of all microevolutionary changes documented between seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis.

Table S8. Details of 29 genes that had microevolutionary changes documented between seven sets of between 2 and 9 *C. glabrata* isolates from recurrent cases of candidiasis, and belonged to our set of 129 “genes of interest” including adhesions (e.g. EPA genes), aspartic proteases, phospholipases, cell wall biogenesis, structural wall proteins, regulatory, efflux pumps (all genes in Table 1 of [53]), as well as both FKS and 12 ERG pathway genes. For each mutation, the information is encoded in a string with details separated by a semi colon. The first detail in variant type (e.g. ref_to_snp, where ref=reference), the second is location in CDS by nucleotide count, the third is the codon position (1, 2 or 3), the fourth is codon found along with amino acid position and type. Finally, a short description is given e.g. INSERTION and DELETION along with the number of nucleotides, or SYN=Synonymous, NSY=Non-synonymous, and NON=Nonsense.

References

1. Daniel H-M, Lachance M-A, Kurtzman CP. 2014 On the reclassification of species assigned to *Candida* and other anamorphic ascomycetous yeast genera based on phylogenetic circumscription. *Antonie van Leeuwenhoek* **106**, 67–84. (doi:10.1007/s10482-014-0170-z)

2. Turner SA, Butler G. 2014 The *Candida* pathogenic species complex. *Cold Spring Harb Perspect Med* **4**. (doi:10.1101/cshperspect.a019778)

3. Perlroth J, Choi B, Spellberg B. 2007 Nosocomial fungal infections: epidemiology, diagnosis, and treatment. *Med Mycol* **45**, 321–346. (doi:10.1080/13693780701218689)

4. Rajendran R, Sherry L, Deshpande A, Johnson EM, Hanson MF, Williams C, Munro CA, Jones BL, Ramage G. 2016 A prospective surveillance study of candidaemia: epidemiology, risk factors, antifungal treatment and outcome in hospitalized patients. *Front Microbiol* **7**, 915. (doi:10.3389/fmicb.2016.00915)

5. Mullick A, Leon Z, Min-Oo G, Berghout J, Lo R, Daniels E, Gros P. 2006 Cardiac failure in C5-deficient A/J mice after *Candida albicans* infection. *Infect Immun* **74**, 4439–4451. (doi:10.1128/IAI.00159-06)

6. Hirano R, Sakamoto Y, Kudo K, Ohnishi M. 2015 Retrospective analysis of mortality and *Candida* isolates of 75 patients with candidemia: a single hospital experience. *Infect Drug Resist* **8**, 199–205. (doi:10.2147/IDR.S80677)

7. Lamoth F, Lockhart SR, Berkow EL, Calandra T. 2018 Changes in the epidemiological landscape of invasive candidiasis. *J Antimicrob Chemother* **73**, i4–i13. (doi:10.1093/jac/dkx444)

8. Yapar N. 2014 Epidemiology and risk factors for invasive candidiasis. *Ther Clin Risk Manag* **10**, 95–105. (doi:10.2147/TCRM.S40160)

9. Odds FC, Hanson MF, Davidson AD, Jacobsen MD, Wright P, Whyte JA, Gow NAR, Jones BL. 2007 One year prospective survey of *Candida* bloodstream infections in Scotland. *J Med Microbiol* **56**, 1066–1075. (doi:10.1099/jmm.0.47239-0)

10. Cowen LE, Anderson JB, Kohn LM. 2002 Evolution of drug resistance in *Candida albicans*. *Annu. Rev. Microbiol.* **56**, 139–165. (doi:10.1146/annurev.micro.56.012302.160907)

11. Pfaller MA, Castanheira M, Lockhart SR, Ahlquist AM, Messer SA, Jones RN. 2012 Frequency of Decreased Susceptibility and Resistance to Echinocandins among Fluconazole-Resistant Bloodstream Isolates of Candida glabrata. *J Clin Microbiol* **50**, 1199–1203. (doi:10.1128/JCM.06112-11)

12. Barber AE *et al.* 2019 Comparative genomics of serial *Candida glabrata* isolates and the rapid acquisition of echinocandin resistance during therapy. *Antimicrob Agents Chemother* **63**. (doi:10.1128/AAC.01628-18)

13. Brunke S, Hube B. 2013 Two unlike cousins: *Candida albicans* and *C. glabrata* infection strategies. *Cell Microbiol* **15**, 701–708. (doi:10.1111/cmi.12091)

14. Gabaldón T *et al.* 2013 Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics* **14**, 623. (doi:10.1186/1471-2164-14-623)

15. Cormack BP, Ghori N, Falkow S. 1999 An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. *Science* **285**, 578–582. (doi:10.1126/science.285.5427.578)

16. Cafarchia C, Romito D, Coccioli C, Camarda A, Otranto D. 2008 Phospholipase activity of yeasts from wild birds and possible implications for human disease. *Medical Mycology* **46**, 429–434. (doi:10.1080/13693780701885636)

17. Bink A, Vandenbosch D, Coenye T, Nelis H, Cammue BPA, Thevissen K. 2011 Superoxide Dismutases Are Involved in Candida albicans Biofilm Persistence against Miconazole ▿ . *Antimicrob Agents Chemother* **55**, 4033–4037. (doi:10.1128/AAC.00280-11)

18. Ibrahim AS, Mirbod F, Filler SG, Banno Y, Cole GT, Kitajima Y, Edwards JE, Nozawa Y, Ghannoum MA. 1995 Evidence implicating phospholipase as a virulence factor of *Candida albicans*. *Infection and Immunity* **63**, 1993–1998. (doi:10.1128/iai.63.5.1993-1998.1995)

19. Martchenko M, Alarco A-M, Harcus D, Whiteway M. 2004 Superoxide Dismutases in *Candida albicans*: transcriptional regulation and functional characterization of the hyphal-induced SOD5 gene. *Mol Biol Cell* **15**, 456–467. (doi:10.1091/mbc.E03-03-0179)

20. Beck-Sagué C, Jarvis WR. 1993 Secular trends in the epidemiology of nosocomial fungal infections in the United States, 1980-1990. National Nosocomial Infections Surveillance System. *J Infect Dis* **167**, 1247–1251. (doi:10.1093/infdis/167.5.1247)

21. Al-Yasiri MH, Normand A-C, L'Ollivier C, Lachaud L, Bourgeois N, Rebaudet S, Piarroux R, Mauffrey J-F, Ranque S. 2016 Opportunistic fungal pathogen *Candida glabrata* circulates between humans and yellow-legged gulls. *Sci Rep* **6**, 36157. (doi:10.1038/srep36157)

22. de Melo Pereira GV, Soccol VT, Pandey A, Medeiros ABP, Andrade Lara JMR, Gollo AL, Soccol CR. 2014 Isolation, selection and evaluation of yeasts for use in fermentation of coffee beans by the wet process. *International Journal of Food Microbiology* **188**, 60–66. (doi:10.1016/j.ijfoodmicro.2014.07.008)

23. Arora P, Singh P, Wang Y, Yadav A, Pawar K, Singh A, Padmavati G, Xu J, Chowdhary A. 2021 Environmental isolation of *Candida auris* from the coastal wetlands of Andaman Islands, India. *mBio* **12**. (doi:10.1128/mBio.03181-20)

24. Carreté L *et al.* 2018 Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr Biol* **28**, 15-27.e7. (doi:10.1016/j.cub.2017.11.027)

25. Biswas C *et al.* 2018 Whole genome sequencing of Australian *Candida glabrata* isolates reveals genetic diversity and novel sequence types. *Front Microbiol* **9**. (doi:10.3389/fmicb.2018.02946)

26. Gabaldón T, Gómez-Molero E, Bader O. 2020 Molecular Typing of *Candida glabrata*. *Mycopathologia* **185**, 755–764. (doi:10.1007/s11046-019-00388-x)

27. Drenth A, McTaggart AR, Wingfield BD. 2019 Fungal clones win the battle, but recombination wins the war. *IMA Fungus* **10**, 18. (doi:10.1186/s43008-019-0020-8)

28. Wong S, Fares MA, Zimmermann W, Butler G, Wolfe KH. 2003 Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata*. *Genome Biol* **4**, R10. (doi:10.1186/gb-2003-4-2-r10)

29. Gabaldón T, Fairhead C. 2019 Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. *Curr Genet* **65**, 93–98. (doi:10.1007/s00294-018-0867-z)

30. Dodgson AR, Pujol C, Pfaller MA, Denning DW, Soll DR. 2005 Evidence for recombination in *Candida glabrata*. *Fungal Genet Biol* **42**, 233–243. (doi:10.1016/j.fgb.2004.11.010)

31. Rajendran R *et al.* 2016 Biofilm formation is a risk factor for mortality in patients with *Candida albicans* bloodstream infection-Scotland, 2012-2013. *Clin Microbiol Infect* **22**, 87–93. (doi:10.1016/j.cmi.2015.09.018)

32. Xu Z, Green B, Benoit N, Schatz M, Wheelan S, Cormack B. 2020 *De novo* genome assembly of *Candida glabrata* reveals cell wall protein complement and structure of dispersed tandem repeat arrays. *Mol Microbiol* **113**, 1209–1224. (doi:10.1111/mmi.14488)

33. Alexander BD, Clinical and Laboratory Standards Institute. 2017 *Reference method for broth dilution antifungal susceptibility testing of yeasts*.

34. McKenna A *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110)

35. Voss K, Auwera GV der, Gentry J. 2017 Full-stack genomics pipelining with GATK4 + WDL + Cromwell. In *18th Annual Bioinformatics Open Source Conference (BOSC 2017)*, (doi:10.7490/f1000research.1114634.1)

36. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*

37. Li H *et al.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

38. Bankevich A *et al.* 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. (doi:10.1089/cmb.2012.0021)

39. Stamatakis A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)

40. Correia A, Sampaio P, James S, Pais C. 2006 *Candida bracarensis sp. nov.*, a novel anamorphic yeast species phenotypically similar to *Candida glabrata*. *Int J Syst Evol Microbiol* **56**, 313–317. (doi:10.1099/ijs.0.64076-0)

41. Swofford DL. In press. *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) Version 4.0 beta (2001).*

42. Huson DH. 1998 SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73.

43. Danecek P *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

44. Purcell S *et al.* 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575. (doi:10.1086/519795)

45. Alexander DH, Novembre J, Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. (doi:10.1101/gr.094052.109)

46. Abraham G, Inouye M. 2014 Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**. (doi:10.1371/journal.pone.0093766)

47. Jolley KA, Bray JE, Maiden MCJ. 2018 Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124. (doi:10.12688/wellcomeopenres.14826.1)

48. Hudson RR, Kaplan NL. 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164. (doi:10.1093/genetics/111.1.147)

49. Agapow P-M, Burt A. 2001 Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes* **1**, 101–102. (doi:10.1046/j.1471-8278.2000.00014.x)

50. Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

51. Yang Z, Nielsen R. 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.

52. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676. (doi:10.1093/bioinformatics/bti610)

53. Weig M, Jänsch L, Gross U, De Koster CG, Klis FM, De Groot PWJ. 2004 Systematic identification *in silico* of covalently bound cell wall proteins and analysis of protein-polysaccharide linkages of the human pathogen *Candida glabrata*. *Microbiology (Reading)* **150**, 3129–3144. (doi:10.1099/mic.0.27256-0)

54. Irinyi L *et al.* 2015 International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med Mycol* **53**, 313–337. (doi:10.1093/mmy/myv008)

55. Carreté L, Ksiezopolska E, Gómez-Molero E, Angoulvant A, Bader O, Fairhead C, Gabaldón T. 2019 Genome comparisons of *Candida glabrata* serial clinical isolates reveal patterns of genetic variation in infecting clonal populations. *Front. Microbiol.* **10**. (doi:10.3389/fmicb.2019.00112)

56. Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH. 2004 Evolution of the MAT locus and its Ho endonuclease in yeast species. *PNAS* **101**, 1632–1637.

57. Srikantha T, Lachke SA, Soll DR. 2003 Three Mating Type-Like Loci in Candida glabrata. *Eukaryot Cell* **2**, 328–340. (doi:10.1128/EC.2.2.328-340.2003)

58. Koszul R *et al.* 2003 The complete mitochondrial genome sequence of the pathogenic yeast *Candida (Torulopsis) glabrata*. *FEBS Lett* **534**, 39–48. (doi:10.1016/s0014-5793(02)03749-3)

59. Kryazhimskiy S, Plotkin JB. 2008 The population genetics of dN/dS. *PLoS Genet* **4**. (doi:10.1371/journal.pgen.1000304)

60. Teunissen AW, Steensma HY. 1995 Review: the dominant flocculation genes of Saccharomyces cerevisiae constitute a new subtelomeric gene family. *Yeast* **11**, 1001–1013. (doi:10.1002/yea.320111102)

61. Garcia-Rubio R, Jimenez-Ortigosa C, DeGregorio L, Quinteros C, Shor E, Perlin DS. In press. Multifactorial Role of Mitochondria in Echinocandin Tolerance Revealed by Transcriptome Analysis of Drug-Tolerant Cells. *mBio* **0**, e01959-21. (doi:10.1128/mBio.01959-21)

62. Aguileta G, de Vienne DM, Ross ON, Hood ME, Giraud T, Petit E, Gabaldón T. 2014 High variability of mitochondrial gene order among fungi. *Genome Biol Evol* **6**, 451–465. (doi:10.1093/gbe/evu028)

63. Fraser JA, Heitman J. 2003 Fungal mating-type loci. *Curr Biol* **13**, R792-795. (doi:10.1016/j.cub.2003.09.046)

64. De Las Peñas A, Pan S-J, Castaño I, Alder J, Cregg R, Cormack BP. 2003 Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev* **17**, 2245–2258. (doi:10.1101/gad.1121003)

65. Makova KD, Hardison RC. 2015 The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213–223. (doi:10.1038/nrg3890)

66. Feng W, Yang J, Xi Z, Qiao Z, Lv Y, Wang Y, Ma Y, Wang Y, Cen W. 2017 Mutations and/or overexpressions of ERG4 and ERG11 genes in clinical azoles-resistant isolates of *Candida albicans*. *Microb Drug Resist* **23**, 563–570. (doi:10.1089/mdr.2016.0095)

67. Katiyar SK, Alastruey-Izquierdo A, Healey KR, Johnson ME, Perlin DS, Edlind TD. 2012 Fks1 and Fks2 are functionally redundant but differentially regulated in Candida glabrata: implications for echinocandin resistance. *Antimicrob Agents Chemother* **56**, 6304–6309. (doi:10.1128/AAC.00813-12)

68. Xu Z, Green B, Benoit N, Sobel JD, Schatz MC, Wheelan S, Cormack BP. 2021 Cell wall protein variation, break-induced replication, and subtelomere dynamics in *Candida glabrata*. *Molecular Microbiology* **116**, 260–276. (doi:10.1111/mmi.14707)

Tables

Table 1

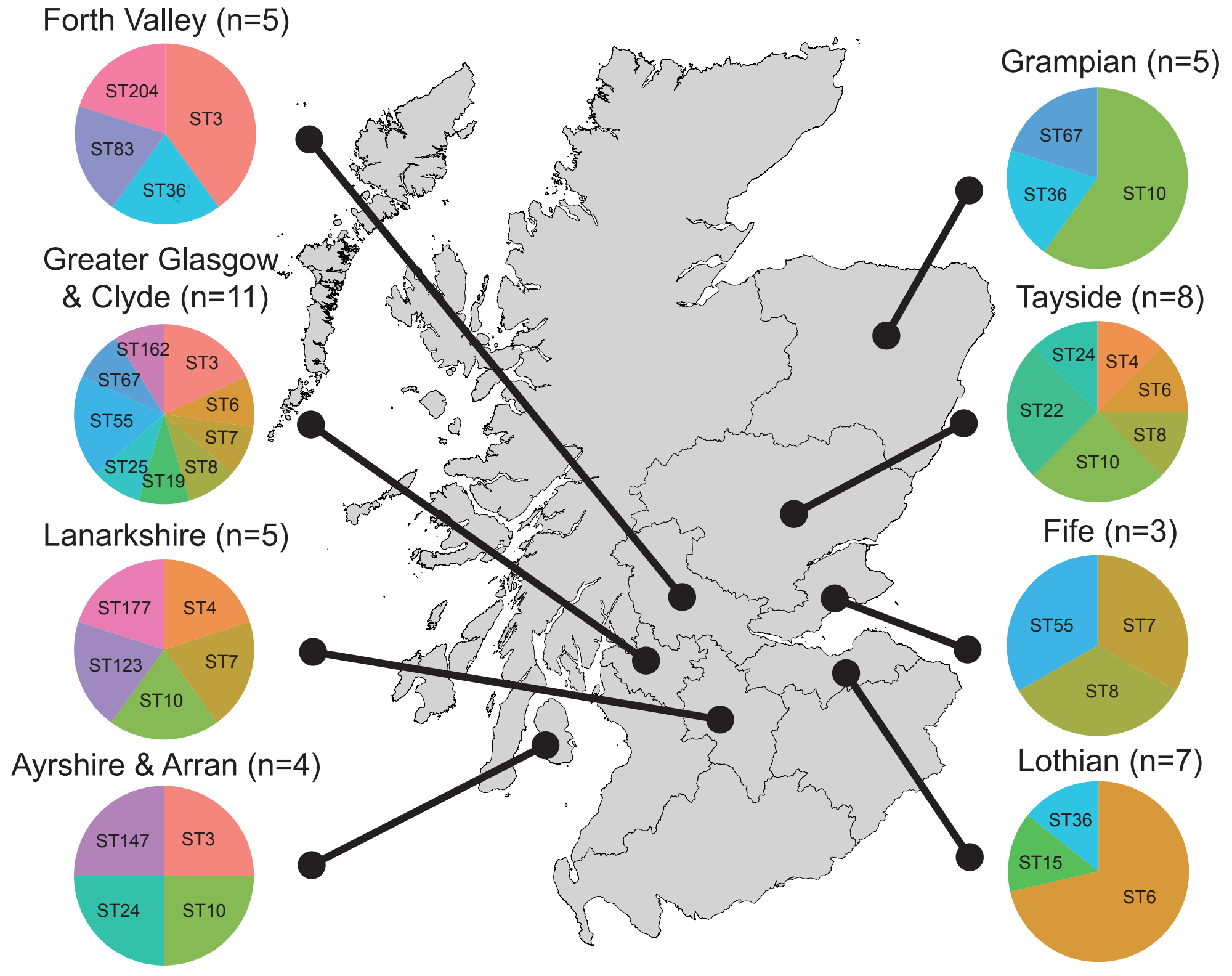
Category	GO/PFAM term	Genes ω< 1	Genes ω> 1	fisher p	q value	Rel. prop	GO/PFAM description
dN/dS>1	GO:0003723	428	25	2.26E-04	3.04E-02	1.95	RNA binding
	GO:0003735	152	3	7.38E-05	1.49E-02	5.78	structural constituent of ribosome
	GO:0003824	1745	155	8.61E-05	1.64E-02	1.28	catalytic activity
	GO:0005515	1100	81	4.48E-06	1.41E-03	1.55	protein binding
	GO:0005740	278	11	5.08E-05	1.06E-02	2.88	mitochondrial envelope
	GO:0005759	160	4	1.90E-04	2.65E-02	4.56	mitochondrial matrix
	GO:0006412	287	11	2.68E-05	6.05E-03	2.98	translation
	GO:0019693	100	1	3.63E-04	4.35E-02	11.4	ribose phosphate metabolic process
	GO:0022626	87	0	1.16E-04	1.99E-02	N/A	cytosolic ribosome
	GO:0022857	277	13	4.08E-04	4.70E-02	2.43	transmembrane transporter activity
	GO:0036094	724	52	2.58E-04	3.31E-02	1.59	small molecule binding
	GO:0043168	689	48	1.89E-04	2.65E-02	1.64	anion binding
	GO:0044281	498	27	1.57E-05	3.86E-03	2.1	small molecule metabolic process
	GO:0044391	143	3	2.32E-04	3.05E-02	5.44	ribosomal subunit
	GO:0071840	1325	113	2.76E-04	3.47E-02	1.34	cellular component organization or biogenesis
	GO:1901362	364	19	1.92E-04	2.65E-02	2.18	organic cyclic compound biosynthetic process
	PF00624.20	4	23	6.53E-20	7.42E-17	0.02	Flocculin repeat
	PF10528.11	12	10	1.87E-05	1.07E-02	0.13	GLEYA domain
	PF00514.25	8	8	5.72E-05	2.17E-02	0.11	Armadillo repeat

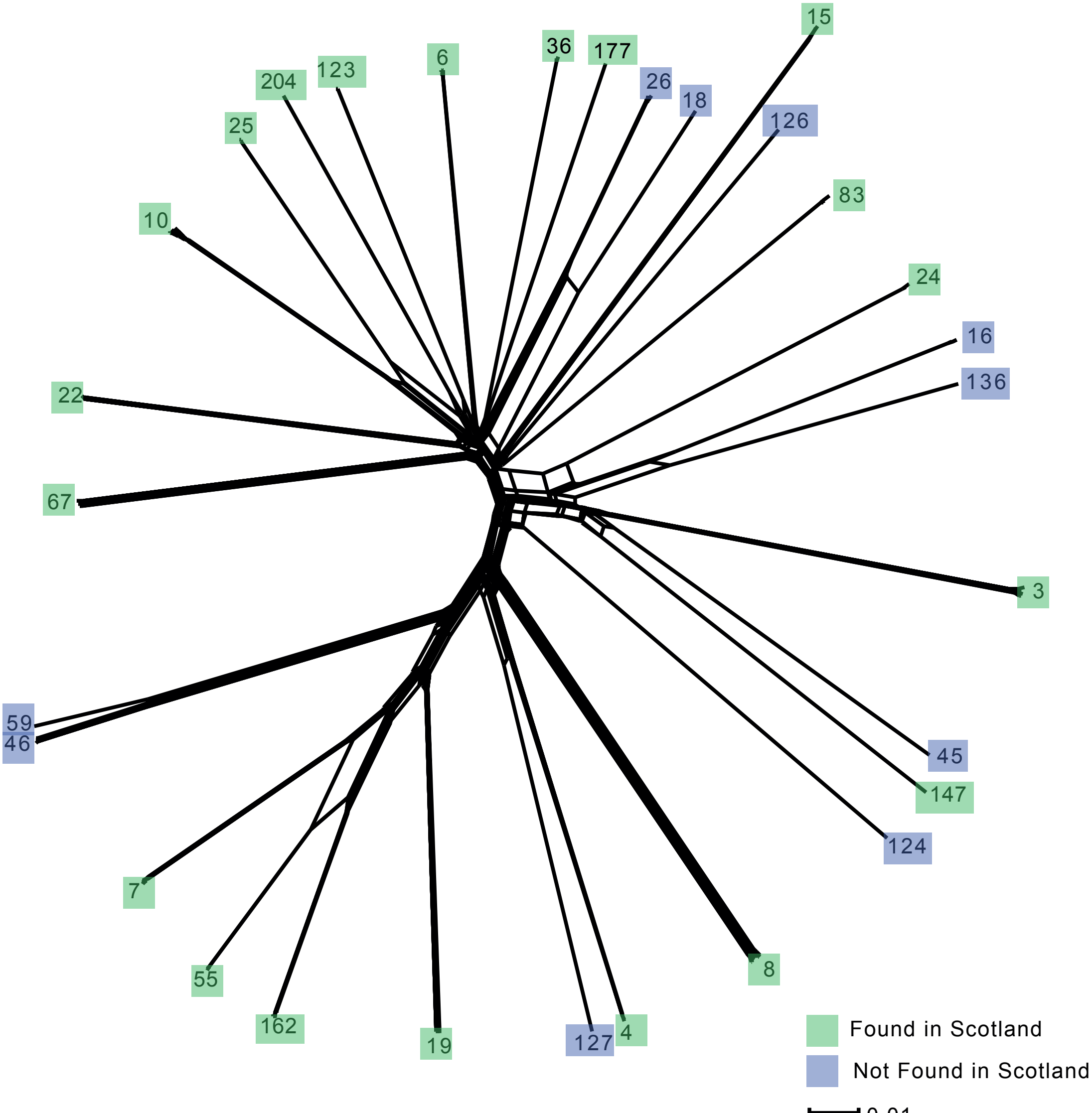
Microevolution (Frameshift)	GO/PFAM term	Genes without frameshift	Genes with frameshift	fisher p	q value	Rel. prop	GO/PFAM description
	GO:0009986	30	9	7.53E-12	4.26E-08	0.03	cell surface
	GO:0009987	3616	17	1.20E-05	3.38E-02	1.7	cellular process
	PF05001.15	0	17	1.10E-39	1.25E-36	0	RNA polymerase Rpb1 C-terminal repeat
	PF10528.11	13	9	5.54E-15	3.15E-12	0.01	GLEYA domain
	PF00399.21	24	9	4.09E-13	1.55E-10	0.02	Yeast PIR protein repeat
	PF08238.14	14	6	2.64E-09	7.49E-07	0.02	Sel1 repeat
	PF11765.10	6	3	2.58E-05	5.86E-03	0.01	Hyphally regulated cell wall protein N-terminal
	PF09770.11	0	2	4.79E-05	9.08E-03	0	Topoisomerase II-associated protein PAT1

Microevolution (Non-synonymous)	GO/PFAM term	Genes without non-synonymous	Genes with non-synonymous	fisher p	q value	Rel. prop	GO/PFAM description
	GO:0009986	28	11	7.88E-10	4.45E-06	0.06	cell surface
	PF10528.11	12	10	3.14E-11	3.58E-08	0.03	GLEYA domain
	PF11765.10	4	5	1.03E-06	5.87E-04	0.02	Hyphally regulated cell wall protein N-terminal

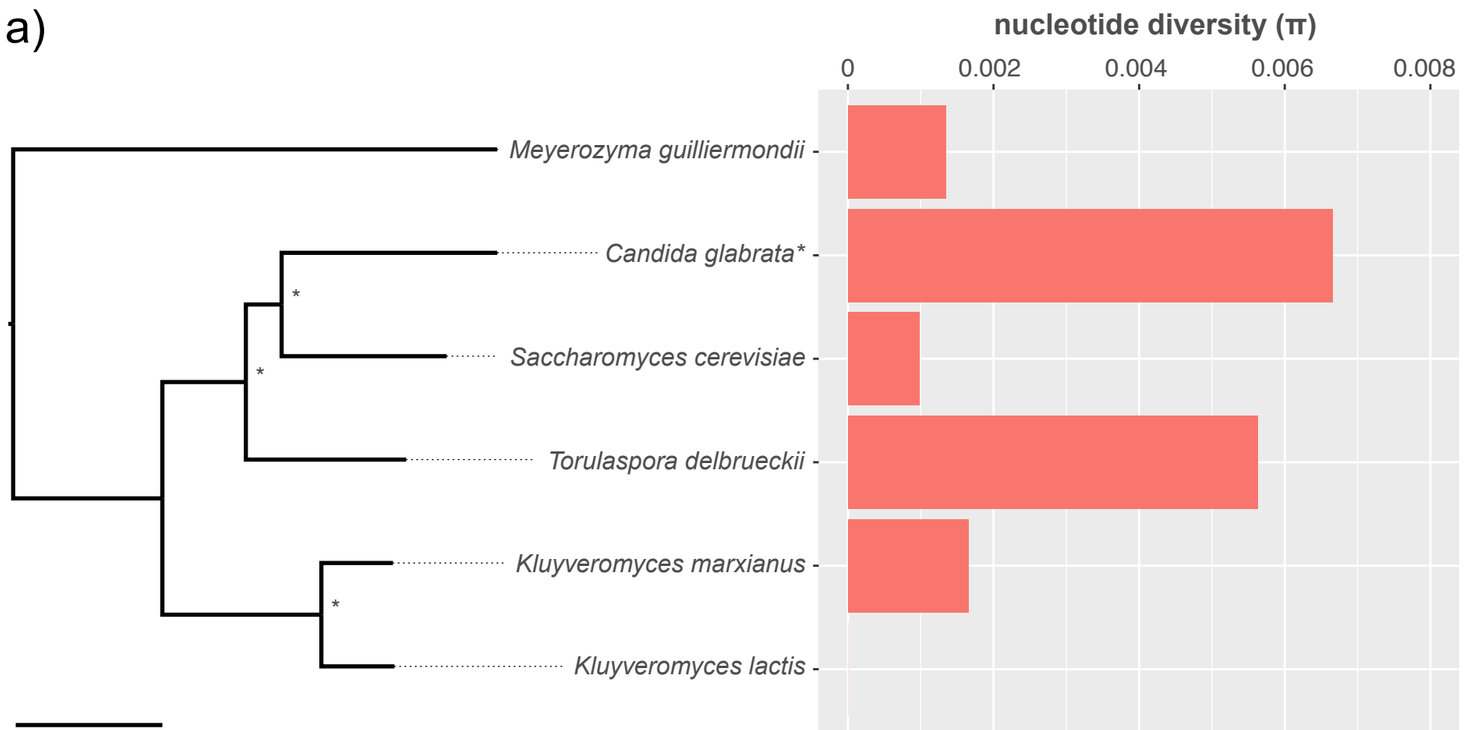
Table 2

Case ID	Initial	Relapse	ST	All Mutations	Coding	Non-coding	Coding Indel	Coding Indel (frameshift)	Coding (Non.Syn.)	Coding Nonsense	Coding (Syn.)	Coding (revert to ref.)
1	CG107A	CG107B	36	97	83	14	18	9	16	0	17	32
2	CG18A	CG18B	10	140	127	13	20	7	24	0	20	63
3	CG191A	CG191B	10	64	53	11	13	4	11	0	8	21
3	CG191B	CG191C	10	78	66	12	14	9	10	0	8	34
3	CG191C	CG191D	10	83	72	11	17	12	14	0	10	31
3	CG191D	CG191E	10	96	85	11	14	4	12	0	9	50
3	CG191E	CG191F	10	87	77	10	23	16	19	0	17	18
4	CG48A	CG48F	7	92	79	13	10	3	11	0	13	45
5	CG84F	CG84G	67	76	64	12	17	5	8	0	5	34
5	CG84G	CG84H	67	71	59	12	21	5	9	0	3	26
6	CG93A	CG93B	162	125	114	11	20	4	15	0	16	63
6	CG93B	CG93C	162	119	105	14	23	10	23	0	18	41
6	CG93C	CG93D	162	124	110	14	19	7	16	0	9	66
6	CG93D	CG93E	162	112	97	15	24	10	20	1	12	40
6	CG93E	CG93H	162	119	105	14	21	5	20	0	18	46
6	CG93H	CG93I	162	116	102	14	20	6	13	0	14	55
6	CG93I	CG93J	162	96	82	14	22	13	13	0	11	36
6	CG93J	CG93K	162	135	120	15	21	11	22	1	28	48
7	CG97A	CG97B	25	79	66	13	14	3	15	0	10	27
7	CG97B	CG97C	25	86	75	11	11	4	12	0	18	34

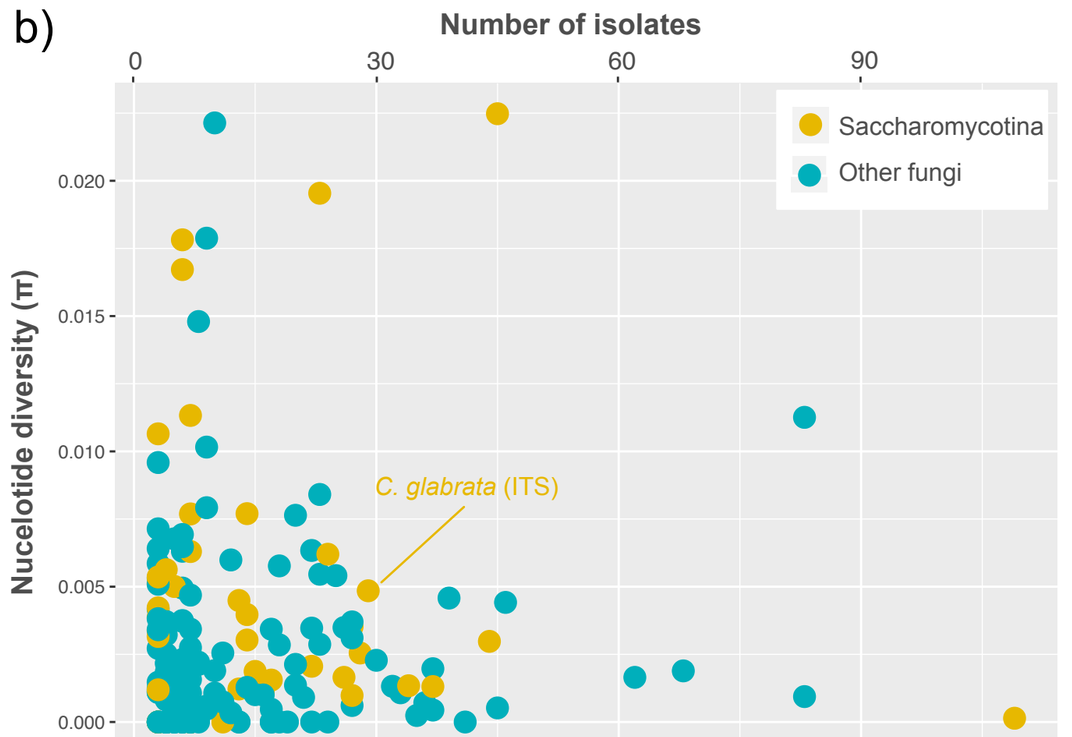

Forth Valley (n=5)
ST204
ST3
ST83
ST36
Greater Glasgow & Clyde (n=11)
ST162
ST3
ST67
ST6
ST55
ST7
ST25
ST8
ST19
Lanarkshire (n=5)
ST177
ST4
ST123
ST7
ST10
Ayrshire & Arran (n=4)
ST147
ST3
ST24
ST10
Grampian (n=5)
ST67
ST36
ST10
Tayside (n=8)
ST24
ST4
ST6
ST22
ST8
ST10
Fife (n=3)
ST55
ST7
ST8
Lothian (n=7)
ST36
ST15
ST6


15
36
177
6
123
204
26
18
25
126
10
83
24
16
22
136
67
3
59
46
45
147
124
7
55
8
162
19
127
4
Found in Scotland
Not Found in Scotland
0.01

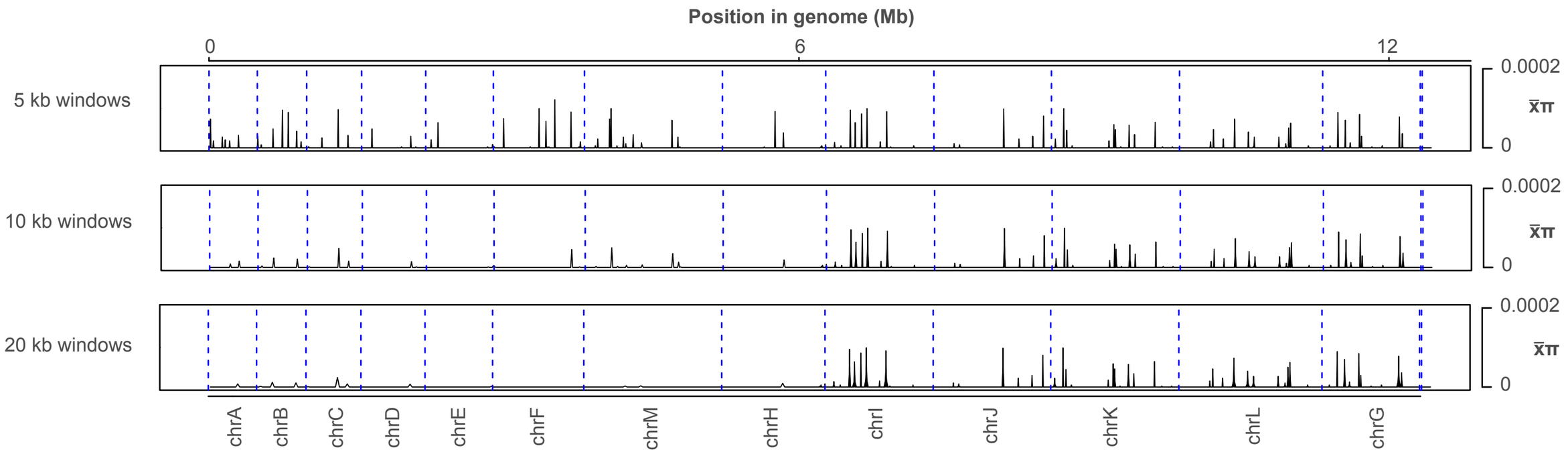
a)

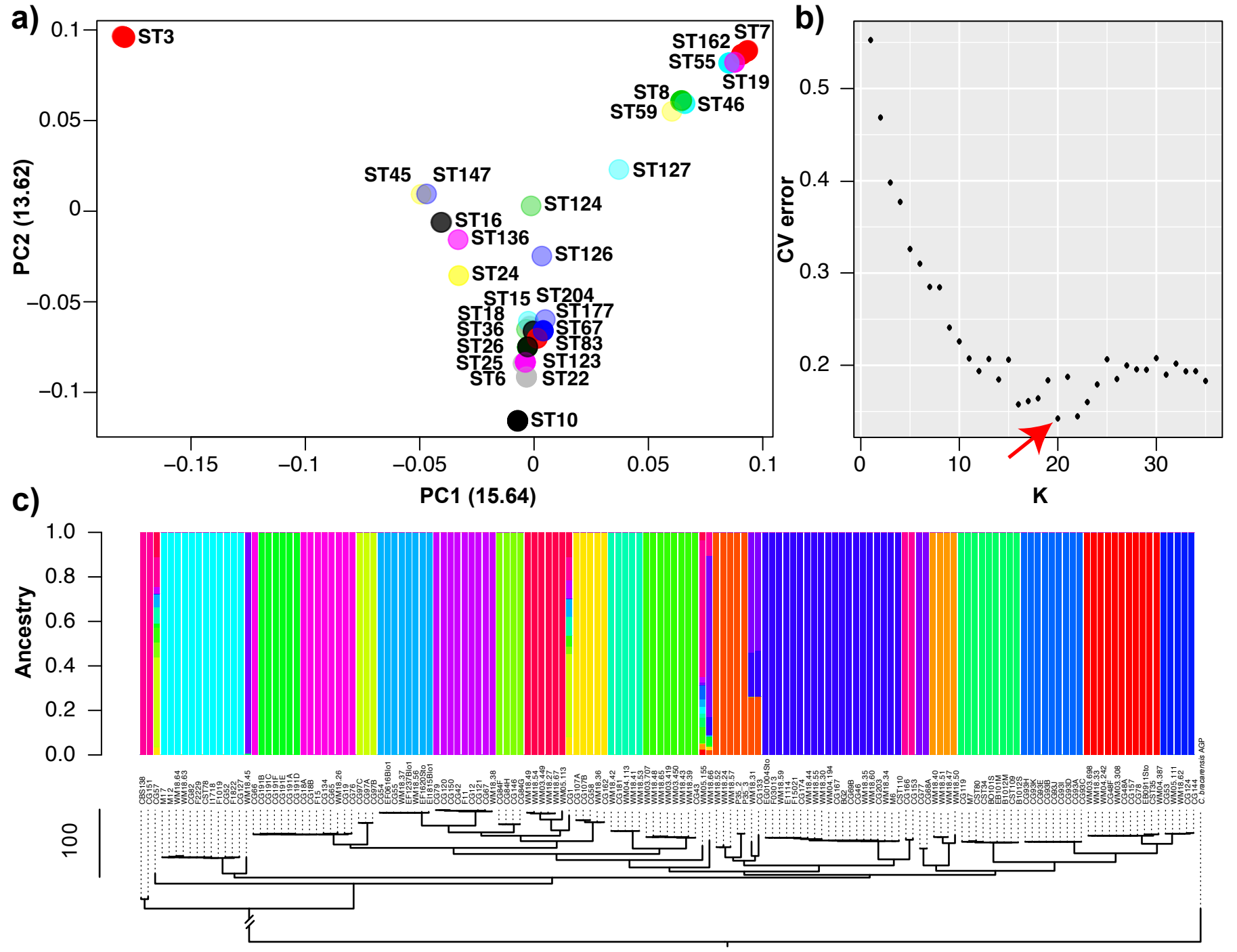

nucleotide diversity (π)
0
0.002
0.004
0.006
0.008
Meyerozyma guilliermondii
Candida glabrata*
Saccharomyces cerevisiae
Torulaspora delbrueckii
Kluyveromyces marxianus
Kluyveromyces lactis
*
*
*
0.2

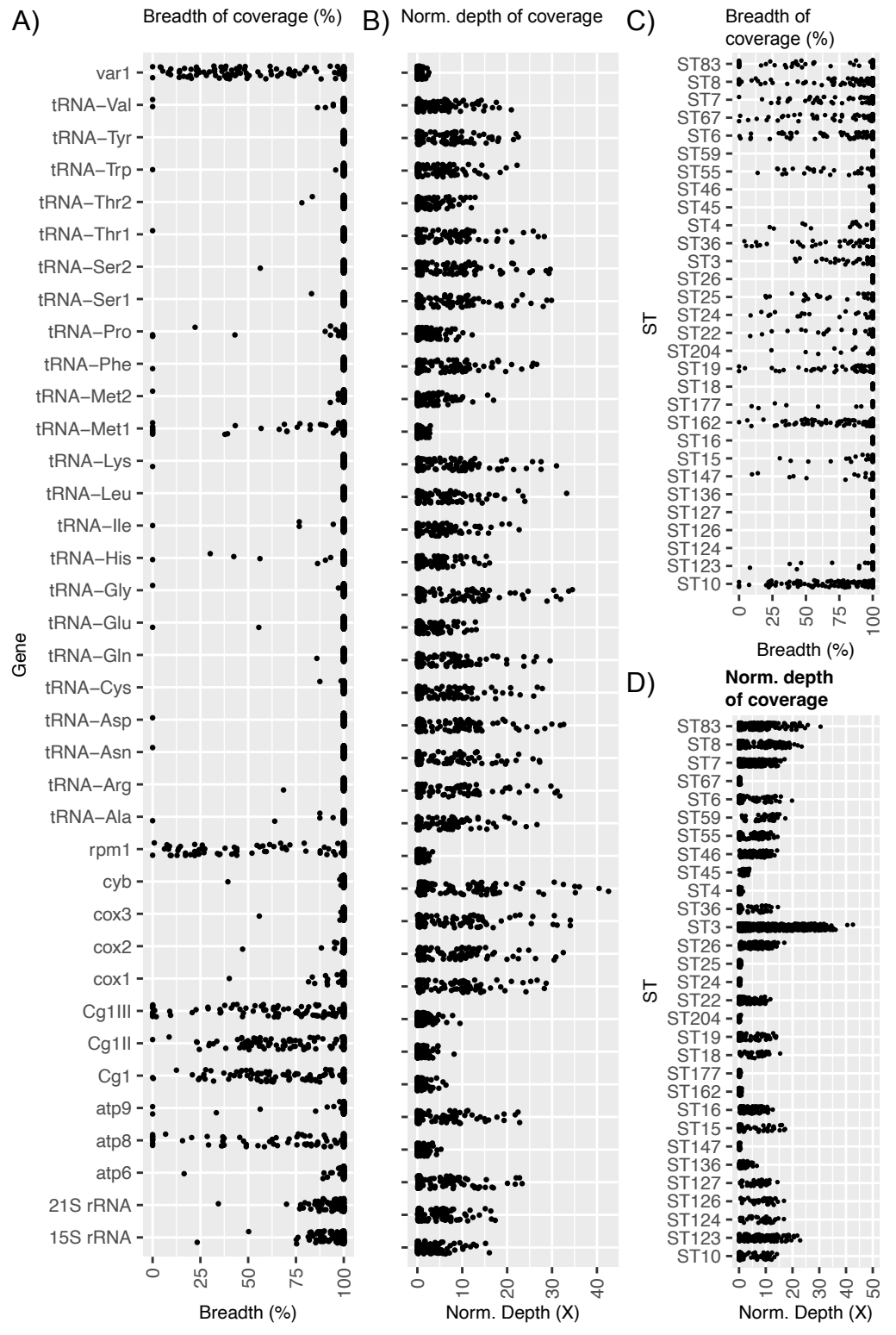
b)

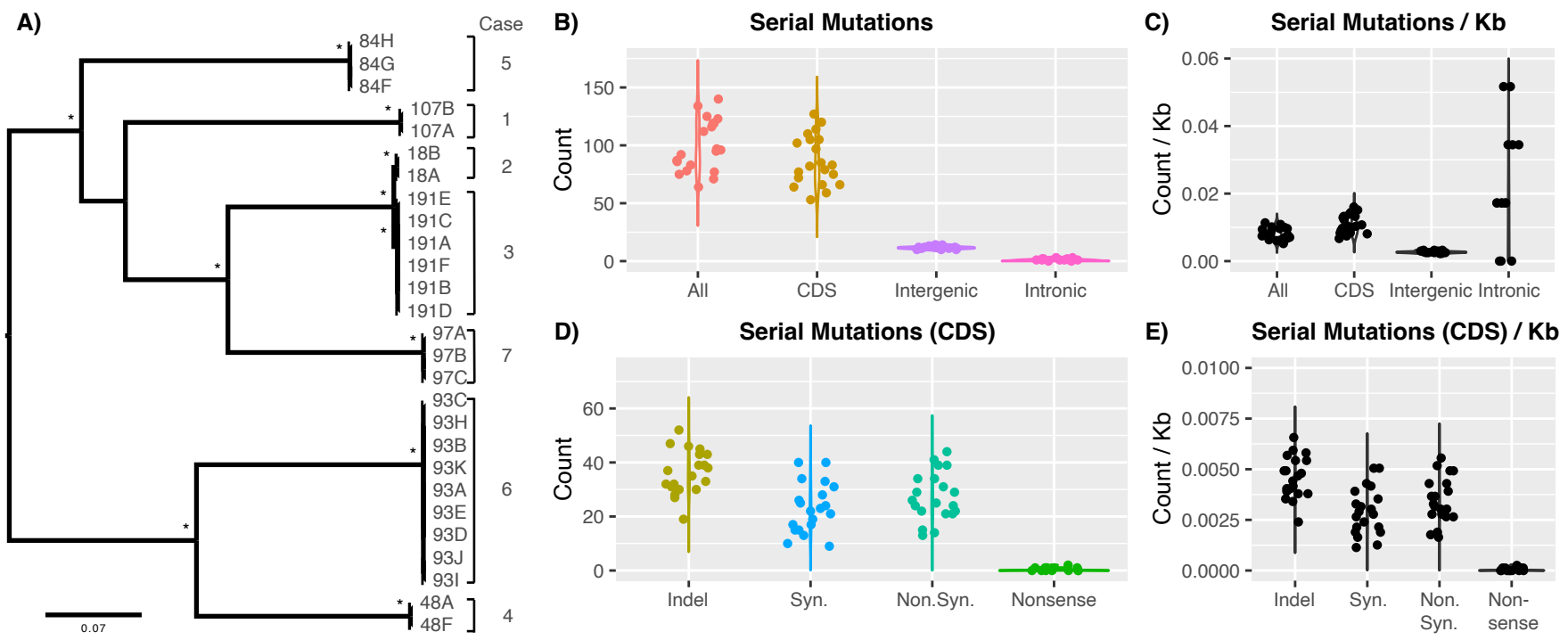

Number of isolates
0
30
60
90
Nucelotide diversity (π)
0.020
0.015
0.010
0.005
0.000
Saccharomycotina
Other fungi
C. glabrata (ITS)

c)


Position in genome (Mb)
0
6
12
5 kb windows
10 kb windows
20 kb windows
0.0002
x̄π
0
chrA
chrB
chrC
chrD
chrE
chrF
chrM
chrH
chrI
chrJ
chrK
chrL
chrG


a)
ST3
ST162
ST7
ST55
ST19
ST8
ST46
ST59
ST127
ST45
ST147
ST124
ST16
ST136
ST126
ST24
ST15
ST204
ST18
ST177
ST36
ST67
ST26
ST83
ST25
ST123
ST6
ST22
ST10
PC2 (13.62)
PC1 (15.64)
b)
CV error
K
c)
Ancestry
100
C. braarensis AGP


A)
Breadth of coverage (%)
B)
Norm. depth of coverage
C)
Breadth of coverage (%)
D)
Norm. depth of coverage
Gene
var1
tRNA–Val
tRNA–Tyr
tRNA–Trp
tRNA–Thr2
tRNA–Thr1
tRNA–Ser2
tRNA–Ser1
tRNA–Pro
tRNA–Phe
tRNA–Met2
tRNA–Met1
tRNA–Lys
tRNA–Leu
tRNA–Ile
tRNA–His
tRNA–Gly
tRNA–Glu
tRNA–Gln
tRNA–Cys
tRNA–Asp
tRNA–Asn
tRNA–Arg
tRNA–Ala
rpm1
cyb
cox3
cox2
cox1
Cg1III
Cg1II
Cg1
atp9
atp8
atp6
21S rRNA
15S rRNA
0
25
50
75
100
Breadth (%)
0
10
20
30
40
Norm. Depth (X)
ST
ST83
ST8
ST7
ST67
ST6
ST59
ST55
ST46
ST45
ST4
ST36
ST3
ST26
ST25
ST24
ST22
ST204
ST19
ST18
ST177
ST162
ST16
ST15
ST147
ST136
ST127
ST126
ST124
ST123
ST10
0
25
50
75
100
Breadth (%)
0
10
20
30
40
50
Norm. Depth (X)


A)
Case
84H
84G
84F
5
107B
107A
1
18B
18A
2
191E
191C
191A
191F
191B
191D
3
97A
97B
97C
7
93C
93H
93B
93K
93A
93E
93D
93J
93I
6
48A
48F
4
0.07
B)
Serial Mutations
Count
150
100
50
0
All
CDS
Intergenic
Intronic
C)
Serial Mutations / Kb
Count / Kb
0.06
0.04
0.02
0.00
All
CDS
Intergenic
Intronic
D)
Serial Mutations (CDS)
Count
60
40
20
0
Indel
Syn.
Non.Syn.
Nonsense
E)
Serial Mutations (CDS) / Kb
Count / Kb
0.0100
0.0075
0.0050
0.0025
0.0000
Indel
Syn.
Non.
Syn.
Non-
sense

Category dN/dS>1	GO/PFAM term	Genes omega < 1	Genes omega > 1	fisher p	q value
	GO:0003723	428	25	2.26E-04	3.04E-02
	GO:0003735	152	3	7.38E-05	1.49E-02
	GO:0003824	1745	155	8.61E-05	1.64E-02
	GO:0005515	1100	81	4.48E-06	1.41E-03
	GO:0005740	278	11	5.08E-05	1.06E-02
	GO:0005759	160	4	1.90E-04	2.65E-02
	GO:0006412	287	11	2.68E-05	6.05E-03
	GO:0019693	100	1	3.63E-04	4.35E-02
	GO:0022626	87	0	1.16E-04	1.99E-02
	GO:0022857	277	13	4.08E-04	4.70E-02
	GO:0036094	724	52	2.58E-04	3.31E-02
	GO:0043168	689	48	1.89E-04	2.65E-02
	GO:0044281	498	27	1.57E-05	3.86E-03
	GO:0044391	143	3	2.32E-04	3.05E-02
	GO:0071840	1325	113	2.76E-04	3.47E-02
	GO:1901362	364	19	1.92E-04	2.65E-02
	PF00624.20	4	23	6.53E-20	7.42E-17
	PF10528.11	12	10	1.87E-05	1.07E-02
	PF00514.25	8	8	5.72E-05	2.17E-02

Microevolution (Frameshift)	GO/PFAM term	Genes without frameshift	Genes with frameshift	fisher p	q value
	GO:0009986	30	9	7.53E-12	4.26E-08
	GO:0009987	3616	17	1.20E-05	3.38E-02
	PF05001.15	0	17	1.10E-39	1.25E-36
	PF10528.11	13	9	5.54E-15	3.15E-12
	PF00399.21	24	9	4.09E-13	1.55E-10
	PF08238.14	14	6	2.64E-09	7.49E-07
	PF11765.10	6	3	2.58E-05	5.86E-03
	PF09770.11	0	2	4.79E-05	9.08E-03

Microevolution (Non-synonymous)	GO/PFAM term	Genes without non-synonymous	Genes with non-synonymous	fisher p	q value
	GO:0009986	28	11	7.88E-10	4.45E-06
	PF10528.11	12	10	3.14E-11	3.58E-08
	PF11765.10	4	5	1.03E-06	5.87E-04

Rel. prop	GO/PFAM description
1.95	RNA binding
5.78	structural constituent of ribosome
1.28	catalytic activity
1.55	protein binding
2.88	mitochondrial envelope
4.56	mitochondrial matrix
2.98	translation
11.4	ribose phosphate metabolic process
N/A	cytosolic ribosome
2.43	transmembrane transporter activity
1.59	small molecule binding
1.64	anion binding
2.1	small molecule metabolic process
5.44	ribosomal subunit
1.34	cellular component organization or biogenesis
2.18	organic cyclic compound biosynthetic process
0.02	Flocculin repeat
0.13	GLEYA domain
0.11	Armadillo repeat

Rel. prop	GO/PFAM description
0.03	cell surface
1.7	cellular process
0	RNA polymerase Rpb1 C-terminal repeat
0.01	GLEYA domain
0.02	Yeast PIR protein repeat
0.02	Sel1 repeat
0.01	Hyphally regulated cell wall protein N-terminal
0	Topoisomerase II-associated protein PAT1

Rel. prop	GO/PFAM description
0.06	cell surface
0.03	GLEYA domain
0.02	Hyphally regulated cell wall protein N-terminal

Case ID	Initial	Relapse	ST	All Mutations	Coding	Non-coding	Coding Indel	Coding Indel (frameshift)	Coding (Non.Syn.)
1	CG107A	CG107B	36	97	83	14	18	9	16
2	CG18A	CG18B	10	140	127	13	20	7	24
3	CG191A	CG191B	10	64	53	11	13	4	11
3	CG191B	CG191C	10	78	66	12	14	9	10
3	CG191C	CG191D	10	83	72	11	17	12	14
3	CG191D	CG191E	10	96	85	11	14	4	12
3	CG191E	CG191F	10	87	77	10	23	16	19
4	CG48A	CG48F	7	92	79	13	10	3	11
5	CG84F	CG84G	67	76	64	12	17	5	8
5	CG84G	CG84H	67	71	59	12	21	5	9
6	CG93A	CG93B	162	125	114	11	20	4	15
6	CG93B	CG93C	162	119	105	14	23	10	23
6	CG93C	CG93D	162	124	110	14	19	7	16
6	CG93D	CG93E	162	112	97	15	24	10	20
6	CG93E	CG93H	162	119	105	14	21	5	20
6	CG93H	CG93I	162	116	102	14	20	6	13
6	CG93I	CG93J	162	96	82	14	22	13	13
6	CG93J	CG93K	162	135	120	15	21	11	22
7	CG97A	CG97B	25	79	66	13	14	3	15
7	CG97B	CG97C	25	86	75	11	11	4	12

Coding Nonsense	Coding (Syn.)	Coding (revert to ref.)
0	17	32
0	20	63
0	8	21
0	8	34
0	10	31
0	9	50
0	17	18
0	13	45
0	5	34
0	3	26
0	16	63
0	18	41
0	9	66
1	12	40
0	18	46
0	14	55
0	11	36
1	28	48
0	10	27
0	18	34

Case ID	Strain	MIC (ug/mL)	Change
1	107a	8	
1	107b	16	8
2	18a	8	
2	18b	8	
3	191a	8	
3	191b	8	
3	191c	8	
3	191d	8	
3	191e	8	
3	191f	8	
4	48a	4	
4	48f	4	
5	84f	8	
5	84g	8	
5	84h	4	-4
6	93a	4	
6	93b	4	
6	93c	4	
6	93d	4	
6	93e	4	
6	93h	>64	>60
6	93i	>64	
6	93j	>64	
6	93k	4	-60
7	97a	4	
7	97b	4	
7	97c	8	4