RESEARCH ARTICLE

Population genomics of the pathogenic yeast *Candida tropicalis* identifies hybrid isolates in environmental samples

Caoimhe E. O'Brien[1☯], João Oliveira-Pacheco[1☯], Eoin Ó Cinnéide[2], Max A. B. Haase[3], Chris Todd Hittinger[3], Thomas R. Rogers[4], Oscar Zaragoza[5], Ursula Bond[6], Geraldine Butler[1]*

1 School of Biomolecular and Biomedical Science, Conway Institute, University College Dublin, Belfield, Dublin, Ireland, **2** School of Medicine, Conway Institute, University College Dublin, Belfield, Dublin, Ireland, **3** Laboratory of Genetics, Center for Genomic Science Innovation, Wisconsin Energy Institute, DOE Great Lakes Bioenergy Research Center, J.F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **4** Department of Clinical Microbiology, Trinity College Dublin, Dublin, Ireland; Department of Microbiology, St James's Hospital, Dublin, Ireland, **5** Mycology Reference Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III, Carretera Majadahonda-Pozuelo, Km2, Majadahonda, Madrid, Spain, **6** Department of Microbiology, School of Genetics and Microbiology, Trinity College Dublin, Ireland

☯ These authors contributed equally to this work.
* gbutler@ucd.ie

Abstract

Candida tropicalis is a human pathogen that primarily infects the immunocompromised. Whereas the genome of one isolate, *C. tropicalis* MYA-3404, was originally sequenced in 2009, there have been no large-scale, multi-isolate studies of the genetic and phenotypic diversity of this species. Here, we used whole genome sequencing and phenotyping to characterize 77 isolates of *C. tropicalis* from clinical and environmental sources from a variety of locations. We show that most *C. tropicalis* isolates are diploids with approximately 2–6 heterozygous variants per kilobase. The genomes are relatively stable, with few aneuploidies. However, we identified one highly homozygous isolate and six isolates of *C. tropicalis* with much higher heterozygosity levels ranging from 36–49 heterozygous variants per kilobase. Our analyses show that the heterozygous isolates represent two different hybrid lineages, where the hybrids share one parent (A) with most other *C. tropicalis* isolates, but the second parent (B or C) differs by at least 4% at the genome level. Four of the sequenced isolates descend from an AB hybridization, and two from an AC hybridization. The hybrids are *MTL***a**/α heterozygotes. Hybridization, or mating, between different parents is therefore common in the evolutionary history of *C. tropicalis*. The new hybrids were predominantly found in environmental niches, including from soil. Hybridization is therefore unlikely to be associated with virulence. In addition, we used genotype-phenotype correlation and CRISPR-Cas9 editing to identify a genome variant that results in the inability of one isolate to utilize certain branched-chain amino acids as a sole nitrogen source.

OPEN ACCESS

Citation: O'Brien CE, Oliveira-Pacheco J, Ó Cinnéide E, Haase MAB, Hittinger CT, Rogers TR, et al. (2021) Population genomics of the pathogenic yeast *Candida tropicalis* identifies hybrid isolates in environmental samples. PLoS Pathog 17(3): e1009138. https://doi.org/10.1371/journal.ppat.1009138

Editor: Slavena Vylkova, GERMANY

Received: October 28, 2020

Accepted: March 15, 2021

Published: March 31, 2021

Data Availability Statement: All sequencing data is available at NCBI under BioProject accession PRJNA604451 and C. tropicalis genome assembly B and annotation is available under NCBI accession JAFIQD000000000. rDNA sequences are available at accession numbers MW584905- MW584910. Other data sets (i.e. variant calls and images for phenotype analysis) are available at https://doi.org/10.6084/m9.figshare.13128839.v1.

Funding: Corrected FD: This work was supported by grants to GB from European Union's Horizon

2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. H2020-MSCA-ITN-2014-642095, and the Wellcome Trust (wellcome.org; grant number 109167/Z/15/Z to CEO), and Science Foundation Ireland (www.sfi.ie; 19/FFP/6668 to GB). The work from the CTH lab was supported by the National Science Foundation (www.nsf.gov) under Grant No. DEB-1442148 to CTH, in part by the DOE Great Lakes Bioenergy Research Center (www.glbrc.org; DOE BER Office of Science DE-SC0018409 to CTH), the USDA National Institute of Food and Agriculture (nifa.usd.gov; Hatch Project 1020204 to CTH), the Pew Charitable Trusts (Pew Scholar in the Biomedical Sciences to CTH), and the Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (H. I. Romnes Faculty Fellow; CTH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Candida tropicalis is an important fungal pathogen, which is particularly common in the Asia-Pacific and Latin America. There is currently very little known about the diversity of genotype and phenotype of *C. tropicalis* isolates. By carrying out a phylogenomic analysis of 77 isolates, we find that *C. tropicalis* genomes range from very homozygous to highly heterozygous. We show that the heterozygous isolates are hybrids, most likely formed by mating between different parents. Unlike other *Candida* species, the hybrids are more common in environmental than in clinical niches, suggesting that for this species, hybridization is not associated with virulence. We also explore the range of phenotypes, and we identify a genomic variant that is required for growth on valine and isoleucine as sole nitrogen sources.

Introduction

Candida tropicalis is an opportunistic pathogenic yeast, and a cause of both superficial and systemic infections in humans. Although *Candida albicans* remains the most common cause of candidiasis, other *Candida* species such as *C. tropicalis* are increasingly isolated as the cause of invasive *Candida* infections [1–3]. *C. tropicalis* is particularly prevalent in Asia-Pacific and Latin America, where it has been identified as the second- or third-most common cause of candidiasis [1–5]. *C. tropicalis* is particularly associated with infection in patients with hematological malignancies [5,6]. Fluconazole and voriconazole resistance occurs more frequently in clinical isolates of *C. tropicalis* than in clinical isolates of *C. albicans* [1,2]; the frequency of resistant isolates, particularly to fluconazole, ranges from 5–36% [2,7–10]. Notably, more Asia-Pacific isolates are fluconazole-resistant in comparison to isolates from other locales [1–3]. Bloodstream infections by *C. tropicalis* are associated with high mortality rates, ranging from 41–61% [11–13].

C. tropicalis is a member of the CUG-Ser1 clade, a group of species in which the CUG codon is translated as serine instead of the standard leucine [14,15]. The genome of *C. tropicalis* was first sequenced in 2009, revealing a diploid genome of approximately 14.5 Mb [16]. Although once thought to be asexual, it is now known that *C. tropicalis* can mate via a parasexual cycle [17,18]. Cells that are homozygous for either the *MTL***a** or *MTLα* mating idiomorph undergo phenotypic switching to the opaque state, and subsequently mate with cells that are homozygous for the opposite mating type [17,19]. The resulting tetraploid heterozygous *MTL***a**/*α* cells undergo concerted chromosome loss to revert to the diploid state [18]. Same-sex mating (i.e. mating between two cells homozygous for the same mating type) has been observed in this species, but only in the presence of the pheromone from the opposite mating type [19]. The majority of *C. tropicalis* isolates (79–96%) are heterozygous at the *MTL*, implying that the variation conferred by sexual reproduction is largely beneficial [20,21].

To date, there are no population genomics studies of *C. tropicalis* isolates, although multi-locus sequence typing (MLST) suggests that there is a diverse population structure [22,23]. In contrast, analysis of almost 200 genomes from *C. albicans* isolates identified a clonal population structure with high levels of heterozygosity (e.g. single nucleotide polymorphisms, or SNPs) between the haplotypes of isolates in most lineages [24]. There was also some evidence for gene flow between *C. albicans* lineages [24]. Recent analysis suggests that all isolates of *C. albicans* descended from an ancient hybridization event between related parents, followed by extensive loss of heterozygosity [25].

Some other diploid species from the CUG-Ser1 clade with higher levels of heterozygosity than *C. albicans* also arose from hybridization (or mating) between two related but distinct parents [26–28]. Like *C. albicans*, all currently characterized isolates of *Candida metapsilosis* arose from a single hybridization between two unknown parents, followed by rearrangement at the *MTL***a** locus [27]. Similarly, *Millerozyma* (*Pichia*) *sorbitophila* is an interspecific hybrid between one parent that is highly similar to *Millerozyma* (*Pichia*) *farinosa* and a second unidentified parent which has a high degree of synteny with the first parent, but diverges at the sequence level by about 11% [29]. Hybridization appears to be ongoing in *Candida orthopsilosis*, where most isolates descend from one of at least four hybridization events between one known parent with a homozygous genome, and one that differs by about 5% at the genome level [26,28]. In contrast, sequenced isolates of *Candida dubliniensis*, *Candida parapsilosis* and *C. tropicalis* are not hybrids [25].

Hybridization between two genetically divergent parents is hypothesized to drive adaptation of organisms to new or changing environments. For example, hybridization within the *Saccharomyces* species complex is associated with the development of favorable traits, such as cryotolerance in the lager-brewing yeast *Saccharomyces pastorianus*, a hybrid of *Saccharomyces cerevisiae* and *Saccharomyces eubayanus* [30] or increased thermotolerance and cryotolerance in various hybrids of *S. cerevisiae*, *S. eubayanus* and *Saccharomyces kudriavzevii* [31]. Other members of the Saccharomycotina are also hybrids, such as the yeast *Zygosaccharomyces rouxii*, used in the production of soy sauce and balsamic vinegar [32]. Some isolates of this species are haploid, while some are highly heterozygous diploids resulting from the hybridization of two parental *Zygosaccharomyces* species [33–35]. The *Cryptococcus neoformans* species complex, which includes several human pathogens, has also been found to include several hybrids, resulting from multiple recent hybridization events between different serotypes [36,37]. Hybridization has been proposed to drive virulence properties, for species within the CUG-Ser1 clade like *C. metapsilosis* [27], and species outside the clade, like *Candida inconspicua* [38].

Here we carried out a population genomics study of 77 *C. tropicalis* isolates, including some from clinical sources and some isolated from the environment. We found that heterozygosity levels range from 2 to 6 variants per kilobase (kb) in most isolates. However, one isolate is very homozygous, and six isolates have very heterozygous genomes. The heterozygous isolates appear to be the product of hybridization between one parent that is similar to the *C. tropicalis* reference strain MYA-3404, and other parents that differ from the reference strain by 4–4.65%. The hybrid isolates were predominately found in environmental niches, suggesting that hybridization in this species is not associated with virulence. In addition, we characterized the growth phenotypes of the non-hybrid isolates in different conditions, and we associated phenotypic variation with genotypic variation. We found that a deletion of two bases in the gene *BAT22* is associated with the inability of *C. tropicalis* strains to use valine and isoleucine as sole nitrogen sources.

Results

Population study of *C. tropicalis*

The original reference genome sequence of *C. tropicalis* MYA-3404 was sequenced in 2009, resulting in a genome assembly consisting of 23 supercontigs totaling 14.6 Mb with 6,258 annotated genes [16]. We used Illumina data from resequencing of the reference strain to assemble the 23 supercontigs into 16 scaffolds, called Assembly B (see Materials & Methods). The assembly was subsequently further improved as described by Guin et al [39].

77 unique *C. tropicalis* isolates from different geographical locations were collected and sequenced using Illumina technology. For convenience, we named these strains ct01 to ct78,

including only one of two isolates with very similar sequences (S1 Table). Most isolates came from clinical sources from the USA, Spain and Ireland. Twelve environmental isolates were included, eleven collected from soil or compost in the USA and Ireland, and one from coconut water in India. The reference strain *C*. *tropicalis* MYA-3404 (ct11), which was previously sequenced by Sanger sequencing [16], was also resequenced, as were three engineered auxotrophic derivatives in two genetic backgrounds [40,41].

Variants were identified by mapping reads to *C*. *tropicalis* MYA-3404 Assembly B and calling variants with the Genome Analysis Toolkit (GATK) [42]. Analysis of the distribution of allele frequencies in heterozygous biallelic SNPs showed that the majority of isolates are diploid, i.e. the ratio of reference to non-reference allele frequency is 50:50. However one isolate, *C*. *tropicalis* ct66 is triploid (peaks of allele frequency at 0.33 and 0.66), and another isolate, *C*. *tropicalis* ct26, appears to be octaploid (peaks of allele frequency at approximately 0.5, 0.12 and 0.87) (S1 Fig). In addition, we observed single-chromosome aneuploidies in four isolates (S1 Fig). *C*. *tropicalis* ct06 and *C*. *tropicalis* ct18 each have three copies of scaffold 8, and *C*. *tropicalis* ct14 and *C*. *tropicalis* ct15 both have three copies of scaffold 4 (trisomy). *C*. *tropicalis* ct14 (CAY3764) and *C*. *tropicalis* ct15 (CAY3763) were both derived from *C*. *tropicalis* AM2005/0093 and were used as the background to generate gene deletions [41]. Generating gene deletions has been found to induce aneuploidies in *C*. *albicans* [43].

Smaller copy number variants (CNVs) were identified in several scaffolds. The largest of these is a duplication of approximately 253 kb on scaffold 7 of isolates *C*. *tropicali*s ct04 and ct33, from approximately 350 kb to 603 kb (S2 Fig). Several shorter CNVs were also identified. These include a reduction in copy number of a 35 kb region from ~974 kb to 1 Mb on scaffold 4 in seven isolates: ct12, ct14, ct15, ct26, ct33, ct36 and ct69. Three large copy number variants (ranging from 23 to 235 kb in length) described by Guin et al. [39] in chromosomes 4, 5, and R (i.e. scaffolds 3, 8 and 4, respectively) were not observed in the *C*. *tropicalis* isolates described here.

Most isolates have approximately 2–6 heterozygous variants (including SNPs and indels) per kilobase similar to the type strain [16]. This is comparable to the level of heterozygosity seen in *C*. *albicans* (2.5–8.6 SNPs per kilobase) [25,26]. One isolate (*C*. *tropicalis* ct20) is extremely homozygous, with 0.84 heterozygous variants per kilobase. This isolate also has a higher proportion of homozygous variants compared to the reference (83% of total variants are homozygous, compared to an average of 41% in other isolates). However, six isolates have exceptionally high levels of heterozygosity (Fig 1A). These include one clinical isolate from Spain (*C*. *tropicalis* ct25), and five environmental isolates from soil, one from the USA (*C*. *tropicalis* ct42) and four from Ireland (*C*. *tropicalis* ct75, ct76, ct77 and ct78. Ct77 and ct78 were isolated from the same soil sample). These isolates have 36–49 heterozygous variants per kilobase. Phylogenetic analysis shows that most isolates cluster together (Cluster A in Fig 1B). However, the six heterozygous isolates are extremely divergent (Cluster B, Fig 1B). These six isolates separate into two groups, one containing *C*. *tropicalis* ct25, ct42, ct75 and ct76, and a second containing *C*. *tropicalis* ct77 and ct78.

The remaining isolates (Cluster A) are shown in more detail in Fig 1C. There is evidence of some population structure, with at least five well-supported clades identified by principal components analysis (colored ovals in Figs 1C and S3 and S4 Table) and many lineages outside these clades. However, there is little obvious correlation between phylogeny and geography. Two clades contain only isolates from the USA, but this likely reflects the overrepresentation of isolates from the USA in our collection. In addition, although some of the environmental isolates cluster together, others are closely related to clinical isolates (Fig 1C). There is therefore no clear distinction between clinical and environmental isolates.

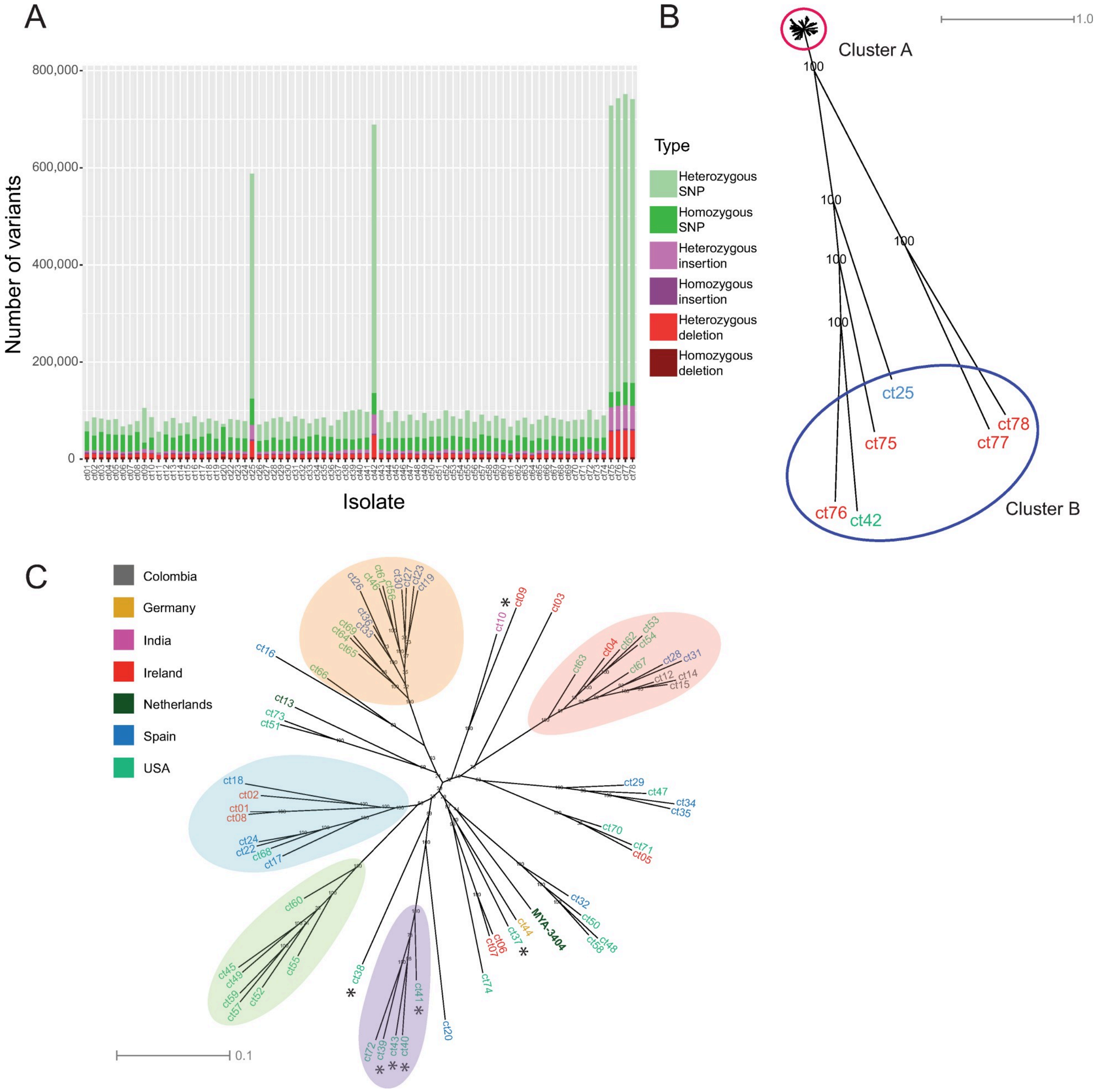

Fig 1. Identification of novel isolates of *C. tropicalis*. (A) Genome variation among *C. tropicalis* isolates. Variants were identified using the Genome Analysis Toolkit HaplotypeCaller and filtered based on genotype quality (GQ) scores and read depth (DP). Variants for all 77 isolates are shown according to variant type. Isolates are labelled on the X-axis by strain ID. One isolate (*C. tropicalis* ct20) has mostly homozygous variants, and six isolates have very high levels of heterozygous variants. **(B) Six isolates of *C. tropicalis* are highly divergent.** Variants were called as in (A). For heterozygous SNPs, a single allele was randomly chosen using RRHS [93] and for homozygous SNPs, the alternate allele to the reference was chosen by default. This process was repeated 100 times and 100 SNP trees were drawn with RAxML using the GTRGAMMA model [94]. The best-scoring maximum likelihood tree was chosen as a reference tree and the remaining 99 trees were used as pseudo-bootstrap trees to generate a supertree. Pseudo-bootstrap values are shown as branch labels. The six divergent isolates (Cluster B) are labelled according to their country of origin (see 1C). **(C) SNP phylogeny of isolates from Cluster A indicates that clade structure is not associated with geography**. The phylogeny of cluster A is shown in detail. Pseudo-bootstrap values are shown as branch labels. Isolates are labelled according to their country of origin, and environmental isolates are indicated with an asterisk. The reference strain, *C. tropicalis* MYA-3404, is labelled. Five putative clades are highlighted with colored bubbles. These clades are supported by principal component analysis (PCA) (S3 Fig). A sixth group was also identified by PCA, encompassing the remainder of isolates in the tree (S3 Fig).

https://doi.org/10.1371/journal.ppat.1009138.g001

Origins of the heterozygous *C. tropicalis* isolates

The levels of heterozygosity in the six divergent *C. tropicalis* isolates are similar to those observed in the hybrid species *C. metapsilosis* and in hybrid isolates of *C. orthopsilosis* [26–28]. This suggests that these *C. tropicalis* isolates may also be hybrids, that is, they may have at least one different parent to most *C. tropicalis* isolates. Hybrid genomes are characterized by regions of heterozygosity, resulting from differences between the homeologous chromosomes, alternating with regions of homozygosity. This results in distinct bimodal patterns of subsequences (*k*-mers) in sequencing reads, which represent the heterozygous and homozygous regions of the genome. Such bimodal *k*-mer patterns are characteristic of heterozygous genomes and have been observed in hybrid isolates of *C. orthopsilosis*, *C. metapsilosis*, *C. inconspicua* and *C. albicans* [25,38].

We find that the *k*-mer frequency distribution of four of the six divergent *C. tropicalis* is also bimodal, with one peak at approximately 100X (the average genome-wide coverage) and one at approximately 50X (half the average genome-wide coverage) (Fig 2A). The full and half coverage peaks represent homozygous regions and heterozygous regions respectively. Approximately half of the heterozygous *k*-mers (i.e. *k*-mers that map to heterozygous regions of the genome) are not represented in the reference genome sequence, because it is a collapsed haploid reference sequence from a non-hybrid isolate (*C. tropicalis* MYA-3404). For the remaining two divergent isolates (*C. tropicalis* ct25 and ct42), a bimodal *k*-mer distribution was not observed, possibly because the sequence coverage was too low. In these two isolates, the peak of *k*-mer multiplicity was less than 20, lower than any other isolates, which may obscure the signal (S4 Fig). This analysis suggests that at least four of the divergent isolates are hybrids, resulting from mating between two related, but distinct, parents. For all four isolates, the heterozygous peak is considerably higher than the homozygous peak, indicating that the hybridization event(s) are recent, and very little loss of heterozygosity (LOH) has occurred.

To further investigate the origins of the six divergent isolates, we attempted to separate the haplotypes of the two parental chromosomes. Approximately 500,000–700,000 heterozygous sites were identified per isolate. The heterozygous sites were placed in phased blocks, using HapCUT2 [44]. On average, 86% of the variants in each isolate were successfully phased, with a total phased span in base pairs of approximately 10–13 Mb (Table 1). The proportion of phased variants was slightly higher in isolates that were sequenced using longer read lengths. For example, 88 and 86% of variants were phased in *C. tropicalis* ct25 and ct42, which were sequenced using 250 bp fragments, compared to 85% in the other isolates, which had read lengths of 150 bp. Longer read lengths likely facilitate more accurate mapping across short homozygous regions.

Variants in the phased genomic regions or "blocks" were assigned to one of two haplotypes. For each phased block greater than 1 kb, the difference of each haplotype in that block to the reference sequence was calculated as the number of variants assigned to the haplotype divided by the length of the block. For the majority of blocks (84–87%), one haplotype (which we refer to as haplotype A) has >99.7% identity to the reference and the second haplotype is more than 1% different to the reference (Fig 2B). The alternative haplotypes were constructed by substituting all variant sites in the reference sequence with alleles that had been assigned to the alternative haplotype. The alternative haplotypes of all six isolates are 4.0–4.6% different from the reference strain. The alternative haplotypes of four of these isolates, *C. tropicalis* ct25, ct42, ct75 and ct76, which we refer to as haplotype B, are approximately 1% different from each other. The alternative haplotypes of the other two, *C. tropicalis* ct77 and ct78, called haplotype C, are approximately 3% different in sequence to the B haplotypes in the other four isolates (and less than 1% different in sequence from each other).

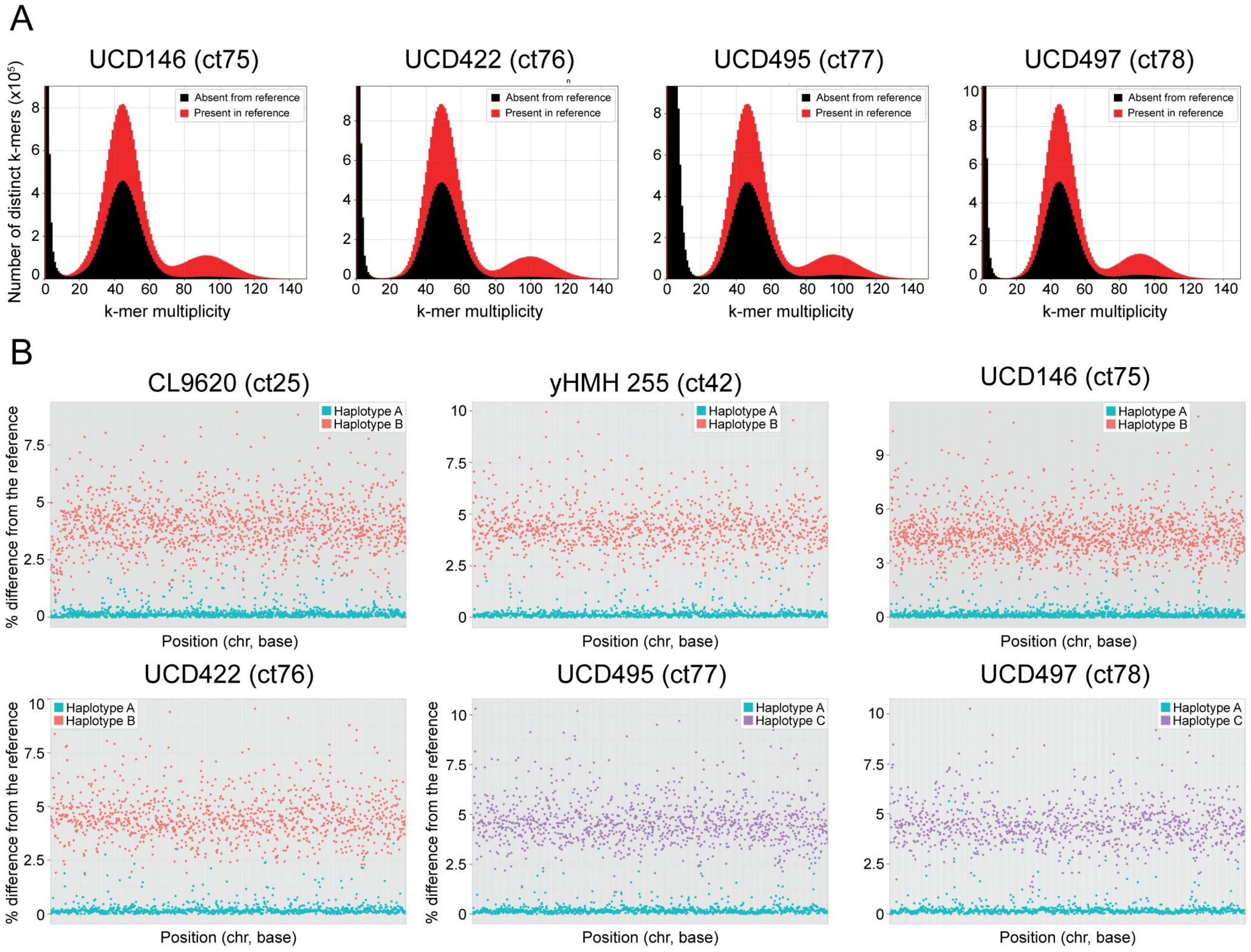

Fig 2. Novel *C. tropicalis* isolates result from hybridization. (A) Analysis of *k*-mer distribution profiles reveals hybrid genomes. *K*-mer analysis of sequencing readsets was performed with the *k*-mer Analysis Toolkit (KAT [82]). For each of four divergent isolates, the number of distinct *k*-mers of length 27 bases (27-mers) is displayed on the Y-axis and *k*-mer multiplicity (depth of coverage) is displayed on the X-axis. *K*-mers that are present in the reference genome are shown in red, and *k*-mers that are absent from the reference genome are shown in black. There are two distinct peaks of *k*-mer coverage at approximately 50X and 100X. This pattern implies that most of the genomes are heterozygous (*k*-mers at 50X coverage) with few homozygous regions (*k*-mers at 100X coverage). Approximately half of the heterozygous *k*-mers in the readsets are not represented in the reference sequence. This pattern has been observed in hybrid isolates from other yeast species [25]. **(B) Analysis of phased variants identifies two distinct haplotypes in divergent isolates of *C. tropicalis*.** Variants were phased using HapCUT2 [44] into blocks covering 10–13 Mb of the genome. For each phased block, percentage difference from the reference strain in each haplotype was calculated as the number of variants divided by the length of the block. For 84–87% of the blocks, one haplotype is <0.3% different to the reference sequence and one haplotype is >1% different to the reference sequence. All phased blocks for each of the six hybrid isolates are shown as haplotype pairs, with the member of the pair more similar to the reference (haplotype A) shown in blue and the member of the pair less similar to the reference shown in orange (haplotype B) or purple (haplotype C). Percentage difference to the reference sequence is displayed on the Y-axis and position in the genome (chromosome, position (bp)) is displayed on the X-axis.

https://doi.org/10.1371/journal.ppat.1009138.g002

Table 1. Results of haplotype phasing.

	CL9620 (ct25)	yHMH255 (ct42)	UCD146 (ct75)	UCD422 (ct76)	UCD495 (ct77)	UCD497 (ct78)
Total number of heterozygous variants	526,189	638,854	691,443	707,685	697,033	685,835
Variants successfully phased	462,386 (88%)	551,867 (86%)	589,165 (85%)	602,663 (85%)	592,497 (85%)	583,248 (85%)
Total phased span (bp)	10,850,562	12,412,152	12,431,473	13,046,231	12,672,096	12,629,063

https://doi.org/10.1371/journal.ppat.1009138.t001

These analyses strongly suggest that the six novel isolates originated from mating or hybridization between related parents, one of which is very similar to the *C*. *tropicalis* reference, and others that are > 4% different. The second parent is not the same for the six divergent isolates. We therefore refer to most *C*. *tropicalis* isolates as AA diploids, to four isolates as AB diploids, and to two isolates as AC diploids. All AB and AC isolates contain only one rDNA locus (D1/D2 region), which is 99% identical to the reference haplotype A. The rDNA sequences were confirmed by PCR amplification and Sanger sequencing (GenBank accession numbers MW584905—MW584910).

Loss of heterozygosity (LOH) analysis in *C*. *tropicalis* hybrid isolates suggests three parental strains

Loss of heterozygosity (LOH) describes tracts of the genome that are essentially homozygous, most likely due to gene conversion or mitotic recombination. We observe a pattern of heterozygous regions alternating with homozygous (LOH) regions in all *C*. *tropicalis* isolates (Fig 3A). We defined heterozygous regions of the genome as regions of at least 100 bp in length containing at least two heterozygous variants; all remaining regions of the genome were classified as homozygous, or LOH, regions, as long as they were at least 100 bp in length.

Only 4% on average of the non-hybrid (AA) genomes are heterozygous. Heterozygous regions in the AA genomes have a mean length of 208 bp and a maximum length of approximately 7.6 kb (S5 Table). In *C*. *tropicalis* ct20 only 0.37% of the genome is heterozygous, with a mean block length of 213 bp. In contrast, on average, 69% of the six hybrid genomes consists of heterozygous regions, with a mean length of approximately 900 bp, and a maximum length of approximately 13.8 kb.

Analysis of heterozygous regions in the six hybrid isolates reveals further support for the hypothesis that they originated from different hybridization events involving different parent strains (B and C). If we assume that the hybrid isolates were derived from a mating event between two parental isolates, we can expect that the heterozygous regions of the genome in the hybrid isolates should be derived equally from the two parent strains. Therefore, if two hybrids originated from hybridization between the same parental strains, the heterozygous regions of their genomes should carry the same variants. However, if two hybrids originated from hybridization between different parental strains, the variants in common heterozygous regions will be different. Shared heterozygous regions were defined as regions of heterozygosity that are common to all isolates. For partially shared heterozygous regions, the portion that was common to all isolates was extracted and analyzed. Shared heterozygous regions in all six hybrid isolates cover 5.8 Mb, with approximately 45% of all heterozygous positions (totaling 485,018 variants) in these regions present in all six. The relatively low proportion of shared variants observed indicates that the six hybrid isolates did not all originate from the same parental strains. However, there is a much higher degree of conservation of variants in shared heterozygous regions among the four AB isolates; 94% of 419,440 heterozygous variants in 6.7 Mb are present in all four. Similarly, the two AC hybrids share 98% of 620,569 variants across 9.6 Mb. This further indicates (in line with our previous analyses) that the four AB isolates share a common origin, and that the two AC isolates share a common origin that is separate from the origin of the AB isolates.

There is extensive LOH in the non-hybrid isolates, covering on average 95% of the genome (S5 Table). In *C*. *tropicalis* ct20, >99% of the genome is in LOH blocks. The average length of LOH blocks across all non-hybrid isolates (excluding *C*. *tropicalis* ct20) is approximately 1.8 kb with a maximum length of 238 kb. In contrast, limited LOH is observed in the six hybrid (AB/AC) isolates, with an average of 13,139 LOH blocks of at least 100 bp, covering between

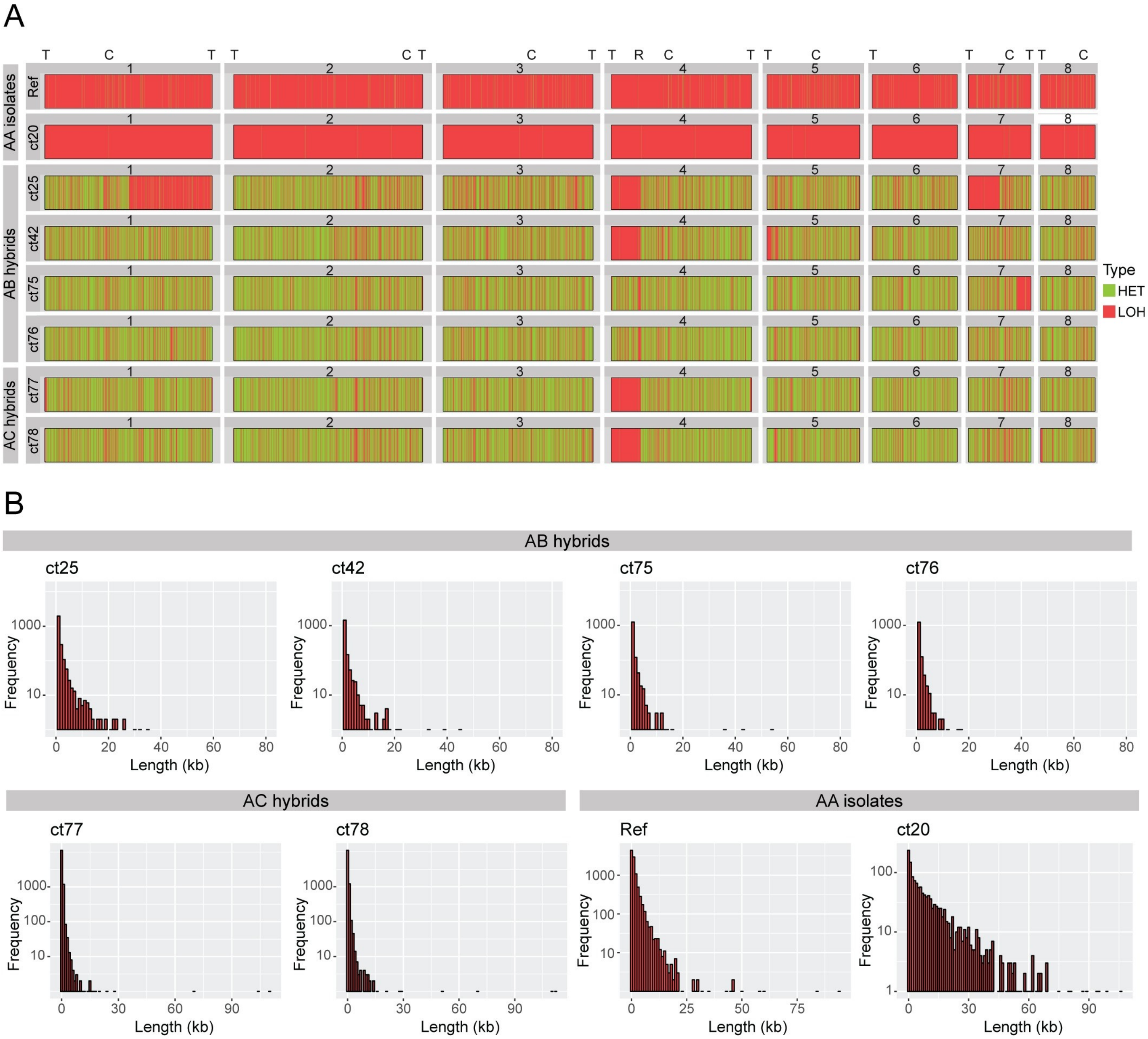

Fig 3. Loss of heterozygosity in *C. tropicalis* isolates. (A) Hybrid and non-hybrid isolates differ in the extent of LOH across the genome. The eight largest scaffolds in the reference genome are displayed horizontally from left to right and labelled from 1 to 8. LOH blocks are shown in pink and heterozygous ("HET") blocks are shown in green. Centromere positions are indicated with "C", telomere positions are indicated with "T" and the rDNA locus is indicated with "R". Isolates are labelled on the left-hand side. The re-sequenced reference strain *C. tropicalis* MYA-3404 (labelled as "Ref") is shown as a representative of the non-hybrid (AA) isolates. The genomes of the AA isolates consist mostly of LOH blocks. The AA isolate *C. tropicalis* ct20 has undergone extensive LOH, covering >99% of the genome. In contrast, in the AB/AC isolates, the majority of the genome consists of heterozygous blocks. **(B) LOH is limited to short tracts of the genome in hybrid isolates.** The histograms show the frequency of LOH blocks of different lengths in the six hybrid isolates and two AA (non-hybrid) isolates the re-sequenced reference strain *C. tropicali*s MYA-3404 (labelled as "Ref") and *C. tropicalis* ct20. Frequency is shown on a log scale on the Y-axis while length in kilobases (kb) is shown on the X-axis, with a bin width of 1000 bp. The average length of LOH blocks in the hybrid isolates ranges from 286–416 bp. A similar pattern is observed in all six hybrid isolates, i.e. a predominance of short LOH blocks, with very few long tracts of LOH. In the non-hybrid isolates (e.g. *C. tropicalis* MYA-3404), LOH blocks are generally longer. *C. tropicalis* ct20 has the longest average LOH block length (~10 kb).

https://doi.org/10.1371/journal.ppat.1009138.g003

25 and 42% of the genome. The average length of LOH blocks in the AB/AC isolates is 330 bp, but can be as long as 112 kb (Fig 3B). Only 1.6% of LOH blocks (equating to 731 LOH blocks or 0.5% of all LOH length in bases) is conserved among all six isolates. There are more shared

LOH regions in the four AB isolates; 17% of LOH blocks (equating to 5,131 LOH blocks) in these isolates are identical. In the AC isolates, 55% of LOH blocks are identical (equating to 8,807 LOH blocks). There is a large LOH block at the start of scaffold 4 (equivalent to Chromosome R [39]) covering approximately 400 kb, that is shared between four of the hybrid isolates (*C. tropicalis* ct25, ct42, ct77 and ct78). The LOH block extends from the telomere to the rDNA locus, although the exact end point differs, and it is interrupted by some small heterozygous regions. A larger LOH block, encompassing this region and extending to the centromere, was identified in a complete, chromosome-scale assembly of *C. tropicalis* and in the related species *Candida sojae* [39]. Two of the AB hybrids (*C. tropicalis* ct75 and ct76) are different, in that only the rDNA locus itself has undergone LOH. The same results were obtained when comparing to the updated *C. tropicalis* reference genome from Guin et al. [39] (S1 Text and S5 Fig and S6 Table).

We considered the possibility that the homozygous isolate *C. tropicalis* ct20 might represent one parent of the hybrid isolates. We therefore compared it with both haplotype A and haplotypes B and C of the six hybrid isolates by computationally reconstructing both subgenomes of each hybrid strain. We constructed a putative A haplotype from *C. tropicalis* ct20 by substituting bases in the reference with homozygous variants identified in this isolate. For the hybrid isolates, the A haplotype was constructed by substituting variants that were originally assigned to haplotype A during haplotype phasing (see Materials & Methods, subsection Haplotype splitting). Similarly, B and C haplotypes were constructed by substituting variants that were assigned to either B or C. The A haplotypes from the hybrids share, on average, approximately 8% of variants with *C. tropicalis* ct20 (i.e. approximately 8% of variants identified in *C. tropicalis* ct20 and a given hybrid isolate are identical). There is even less similarity between the B and C haplotypes and *C. tropicalis* ct20; only 1% of variant sites in *C. tropicalis* ct20 and the hybrid haplotypes B or C are identical. *C. tropicalis* ct20 therefore has an A haplotype, but it is unlikely that it is a parent, or closely related to a parent, of the hybrid isolates.

Mating type-like loci (MTL) in *C. tropicalis* isolates

Most AA isolates (45) are heterozygous at the *MTL*, similar to previous reports [20,21] (S1 Table). Fifteen are homozygous for *MTL***a**/a and seven are homozygous for *MTLα/α*. Three isolates have three copies of the MTL. The triploid isolate *C. tropicalis* ct66 is *MTL* **a/a**/*α*. *C. tropicalis* ct18 is trisomic for scaffold 8, which carries the MTL, and is *MTL* **a**/*α/α*. *C. tropicalis* ct06 is also trisomic for scaffold 8, and has three copies of *MTLα*. The *MTL* idiomorphs of the octaploid isolate, *C. tropicalis* ct26, could not be definitively determined by assembling the Illumina data or by PCR, but it appears to have 7 copies of *MTLα* and one copy of *MTL***a** (S6 Fig).

All six AB and AC isolates contain both *MTL***a** and *MTLα* idiomorphs. In the AB isolates, the *MTL***a** idiomorphs are >99% identical to that of the reference strain (haplotype A) with only three nucleotide changes across the entire locus (8,180 bp). These include synonymous and nonsynonymous substitutions in *PAP***a** and *PIK***a**. In addition, one isolate (*C. tropicalis* ct42) has a nonsynonymous substitution in *MTL***a***1*. Apart from this, the *MTL***a** idiomorphs in the AB isolates are identical. The *MTL***a** idiomorph therefore likely originated from the A parent. The *MTLα* loci are >99% identical in all four AB isolates, and ~7% different to the reference strain, indicating that it was donated by the B parent. All AB isolates therefore most likely resulted from mating between the same parents, an *MTL***a** parent similar to the reference strain (parent A), and an *MTLα* parent which is approximately 4% different (parent B).

In the two AC isolates, the *MTLα* idiomorphs are also identical to each other, and they are >99% identical to the reference strain. The *MTL***a** idiomorphs are identical to each other, and approximately 96% identical to the reference strain. The *MTL***a** idiomorph in the AC isolates therefore originated from the C parent, and the *MTLα* idiomorph originated from the A parent.

Analysis of phenotypic variation in *C*. *tropicalis*

To measure the phenotypic diversity within *C*. *tropicalis*, the growth of 68 AA isolates was tested in 61 different conditions, including alternative carbon sources, stressors (e.g. calcofluor white, congo red), heavy metals (e.g. zinc, cobalt, cadmium) and antifungal drugs (e.g. fluconazole, ketoconazole, caspofungin) (S7 Fig). Because nitrogen and carbon metabolism are important virulence attributes in fungi [45], the ability of *C*. *tropicalis* isolates to use different sole nitrogen sources (e.g. amino acids, gamma-aminobutyric acid (GABA)) was also tested (S7 Fig). The AB and AC isolates and the engineered lab isolates *C*. *tropicalis* ct13, ct14 and ct15 were excluded from the analysis.

The *C*. *tropicalis* isolates show wide variation in their growth characteristics (S7 Fig). We attempted to identify genome variants that are associated with specific growth defects. For this analysis, only conditions that resulted in a growth defect of at least 70% compared to the control condition in at least one strain were included (i.e. 25 conditions using YPD as a base media, and 10 conditions using different nitrogen sources). Reduced growth was scored as 1, and growth similar to the control was scored as 0. Predicted genomic variants were annotated with SnpEff [46] to identify those that were likely to have a major impact on protein function. 390,321 variant sites were identified in total across 68 isolates. The majority of variants (~75%) were SNPs, with the remainder consisting of small insertions and deletions (indels) (S8 Fig). Most variants are found in intergenic regions, or are silent or missense mutations. Only variants that were predicted to have a high impact, including frameshifts, gene fusion events, loss or gain of a stop codon, or variation at splice donor or acceptor sites (9,261 variants, S8 Fig), were included in the genotype-phenotype correlation analysis.

One clinical isolate, *C*. *tropicalis* ct04, identified by cosine similarity analysis [47], has impaired growth when valine or isoleucine (branched chain amino acids) are provided as the sole nitrogen source (Fig 4A). Compared to other isolates, *C*. *tropicalis* ct04 also grows poorly on 2% sodium acetate, 2% starch and in the absence of a carbon source. There are 40 variants unique to this isolate that are predicted to have a high impact on protein function (S7 Table). One of these is a heterozygous deletion of two bases in *CTRG_06204* (*BAT22*), an ortholog of the *S*. *cerevisiae BAT1/2* genes that encode a branched-chain amino acid aminotransferase (BCAT). Amino acid metabolism and acquisition can directly affect pathogenicity, and disruption of branched-chain amino acid metabolism is associated with loss of virulence (reviewed in [48]). BCATs catalyze the final step of biosynthesis and the first step in the degradation of the branched chain amino acids valine, isoleucine and leucine [49]. The deletion results in a frameshift which introduces a premature stop codon at amino acid Gly30 of the Bat22 protein (Fig 4B). We determined if introducing an equivalent change into other genetic backgrounds using CRISPR/Cas9 [50] would result in the same phenotype. A repair template was designed to delete two bases and also to destroy the target of the guide RNA to prevent recutting. The gene was edited in three different *C*. *tropicalis* isolates ct09, ct44 and ct53. All edited strains can no longer use valine or isoleucine as sole nitrogen sources (Fig 4C). However, unlike *C*. *tropicalis* ct04 they have no growth defect on sodium acetate, starch or in the absence of carbon sources, indicating that another variant, or combination of variants, is responsible for these phenotypes.

Analysis of genotype-phenotype correlations becomes more challenging when analyzing variants in intergenic regions or variants with a predicted low or intermediate impact. By

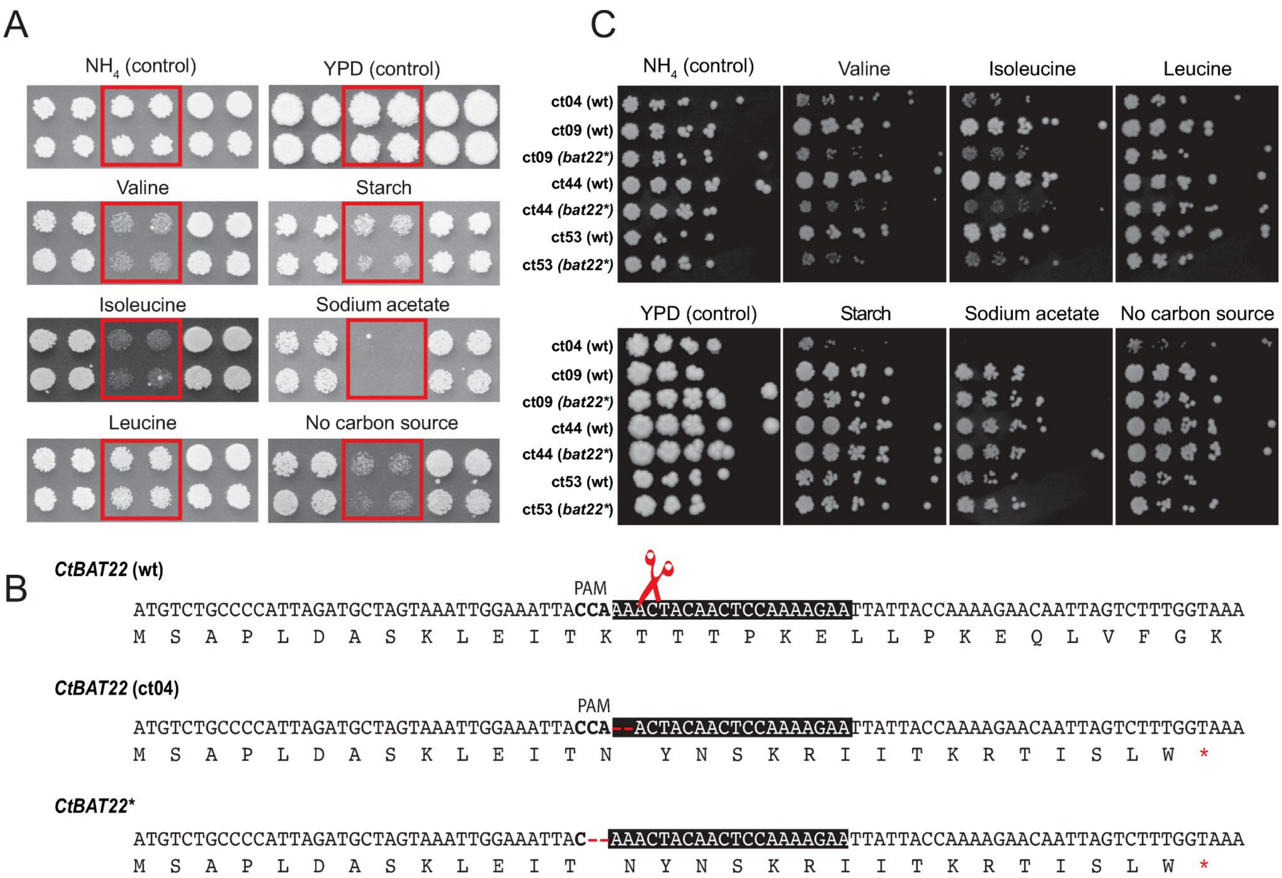
A
NH4 (control)
YPD (control)
Valine
Starch
Isoleucine
Sodium acetate
Leucine
No carbon source
C
NH4 (control)
Valine
Isoleucine
Leucine
ct04 (wt)
ct09 (wt)
ct09 (bat22*)
ct44 (wt)
ct44 (bat22*)
ct53 (wt)
ct53 (bat22*)
YPD (control)
Starch
Sodium acetate
No carbon source
B
CtBAT22 (wt)
PAM
ATGTCTGCCCCATTAGATGCTAGTAAATTGGAAATTACCAAAACTACAACTCCAAAAGAATTATTACCAAAAGAACAATTAGTCTTTGGTAAA
M S A P L D A S K L E I T K T T T P K E L L P K E Q L V F G K
CtBAT22 (ct04)
PAM
ATGTCTGCCCCATTAGATGCTAGTAAATTGGAAATTACCA--ACTACAACTCCAAAAGAATTATTACCAAAAGAACAATTAGTCTTTGGTAAA
M S A P L D A S K L E I T N Y N S K R I I T K R T I S L W *
CtBAT22*
ATGTCTGCCCCATTAGATGCTAGTAAATTGGAAATTAC--AAACTACAACTCCAAAAGAATTATTACCAAAAGAACAATTAGTCTTTGGTAAA
M S A P L D A S K L E I T N Y N S K R I I T K R T I S L W *

Fig 4. Disrupting *BAT22* prevents growth of *C. tropicalis* on branched chain amino acids as a sole nitrogen source. (A) Phenotype analysis of *C. tropicalis* isolates. Growth of *C. tropicalis* ct04 is shown on solid media. Strains were grown in 2x2 arrays; two biological replicates (top and bottom rows), with two technical replicates each (left and right columns), of each strain were tested. *C. tropicalis* ct04 replicates are outlined with red boxes. *C. tropicalis* ct04 cannot utilize valine or isoleucine as a sole nitrogen source and also exhibits a growth defect on solid media with 2% starch or 2% sodium acetate as the sole carbon source, or on solid media without a carbon source provided. **(B) Editing of *BAT22*.** Plasmid pCT-tRNA-BAT22 was generated to edit the wild type sequence of *BAT22* (*CTRG_06204*) using CRISPR-Cas9. The sequences of the reference *C. tropicalis BAT22* (*CtBAT22* (wt)), *BAT22* from *C. tropicalis* ct04 (*CtBAT22* (ct04)) and edited *BAT22* (*CtBAT22**) are shown. The guide sequence is highlighted with a black box, the PAM sequence is shown in bold, and the Cas9 cut site is indicated with a red scissors. *C. tropicalis* isolates ct44, ct09 and ct53 were transformed with pCT-tRNA-BAT22 and a repair template (RT_BAT22_2bpDel_SNP) generated by overlapping PCR using RT_BAT22_2bpDel_SNP-TOP/BOT oligonucleotides. The repair template contains two 60 bp homology arms and deletes two bases in *BAT22* resulting in the same frameshift observed in *C. tropicalis* ct04. **(C) Edited strains have defects in branched-chain amino acid metabolism**. 5-fold serial dilutions of *C. tropicalis* ct04, ct09(wt; bat22*), ct44 (wt; bat22*) and ct53 (wt; bat22*) in the same conditions tested in (A). The edited strains cannot use valine or isoleucine as sole nitrogen sources.

https://doi.org/10.1371/journal.ppat.1009138.g004

restricting our analysis to only variants that were predicted to have a high impact, we were able to narrow our search to a small number of variants, which could be manually validated. We also identified ~69,000 variants in upstream regions (occurring within 300 bp upstream of an annotated gene), which is too many to verify experimentally.

Discussion

Like many opportunistic pathogens of humans, the natural habitat of *C. tropicalis* is unclear. Although *C. tropicalis* is well-adapted to humans, isolates are also commonly isolated from a variety of sources, including soil, sand, animal feces, by-products of industrial food production and the surface of fruits [51–56]. *C. tropicalis* is also a component of the human oral and gastrointestinal mycobiome [57,58] and has been isolated from human skin [59] and the gastrointestinal tracts of mice [60]. Enrichment of *C. tropicalis* in the gastrointestinal tract has been associated with Crohn's disease, potentially due to its invasive abilities [58].

We found little evidence of clade structure associated with geographical origin, suggesting that there may be a high degree of admixture between *C. tropicalis* populations from different regions. This is similar to what has been observed in other diploid CUG-Ser1 clade species, e.g. *C. metapsilosis* [27], *C. orthopsilosis* [28] and *C. albicans*, other than the "*Candida africana*" lineage [24]. Some studies have suggested that population structure in the bakers' yeast *S. cerevisiae* is more related to ecological niche than to geography [61,62], while others found no clear separation between different ecological groups, such as pathogenic and non-pathogenic isolates [63].

Mixao et al [25] suggested that *C. tropicalis* isolates are standard diploids, i.e. that the two parents were closely related. In contrast, *C. metapsilosis* and *C. albicans* isolates descended from ancient hybridizations between two related parents, and hybridization in *C. orthopsilosis* is ongoing [25–28]. We have now shown that six divergent isolates of *C. tropicalis* result from hybridization between one parent that is highly similar in its sequence to the reference genome (parental haplotype A), and other unidentified parents (parental haplotype B or C) that are approximately 4% different in sequence to the reference strain. The low level of LOH in the *C. tropicalis* AB and AC isolates suggests that hybridization has occurred relatively recently. In addition, the isolation of hybrids from different geographical locations, and the identification of multiple hybrids originating from separate hybridization events, indicates that hybridization may be ongoing in this species. This contrasts with *C. albicans* and *C. metapsilosis*, where it is proposed that all known isolates originated from a single hybridization event [25,27], and *C. orthopsilosis*, where several hybridizations have occurred but there has been substantial LOH [28]. In addition, we identified one highly homozygous AA isolate (*C. tropicalis* ct20). This may have resulted from major loss of heterozygosity in a non-hybrid isolate, similar to that proposed for the *C. africana* lineage [25]. It is also possible that homozygous isolates are the parents of hybrid isolates that have not yet been identified.

Ongoing hybridization has been associated with virulence in both plant and animal fungal pathogens [64,65]. In particular, hybridization has been proposed to facilitate the emergence of virulence in species within the CUG-Ser1 clade [66], based on the observation that most isolates of *C. albicans*, *C. orthopsilosis* and *C. metapsilosis* are hybrids [25–28,66]. In addition, clinical isolates of *S. cerevisiae* are more heterozygous than non-clinical isolates, indicating that heterozygous isolates may have an advantage in the human host environment [63]. However, we found that *C. tropicalis* hybrids are rare (6 of 77 isolates), and only one of these was from a clinical setting. In contrast, five of twelve environmental isolates were hybrids, suggesting that hybridization may be advantageous in non-clinical settings. The hybrid isolates we identified are heterozygous at the mating-type like locus, suggesting that they originated by mating [17].

The definition of species is a challenging and controversial topic in biology, particularly so in the case of microorganisms [67]. The level of divergence that we observe between the A and B/C haplotypes in the *C. tropicalis* hybrids is greater than the level of divergence generally observed between strains of the same yeast species. For example, the maximum divergence between strains of *S. cerevisiae* is 1.1% [68], although the divergence between distant isolates of *Saccharomyces paradoxus* or *S. kudriavzevii* can be as high as 4.6% [67]. However, high levels of divergence between parents can be tolerated during hybridization. For example, the parents of the hybrid *M. sorbitophila* are estimated to diverge by approximately 11% [29]. It is clear that species definition in fungi, and in particular in CUG-Ser1 clade yeasts, needs to include hybridization [66]. It has been suggested that the *C. parapsilosis* clade (which currently consists of three species; *C. parapsilosis* sensu stricto, *C. orthopsilosis* and *C. metapsilosis*) should be reorganized to include homozygous lineages (of which there are at least five) and heterozygous lineages (of which there are at least two) [27]. Several of the proposed homozygous lineages are uncharacterized, or only partially characterized. We have shown that *C. tropicalis* isolates can be subdivided into at least three groups; the AA lineage (where either A haplotype may carry the *MTL***a** or *MTLα* idiomorph), the AB lineage (with *MTL***a** from the A haplotype) and the AC lineage (with *MTLα* from the A haplotype). The majority of AA isolates retain some heterozygosity, including at *MTL*. However, one AA isolate (*C. tropicalis* ct20, *MTL***a**/**a**), which has undergone extensive LOH, has approximately one heterozygous variant every 1,190 bases. This is similar to *C. dubliniensis* (approximately one SNP every 1,511 bases [69]), but not quite as homozygous as *C. parapsilosis* (on average, one SNP per 15,553 bases [16]) or homozygous isolates of *C. orthopsilosis* (approximately one heterozygous SNP per 10,692 bases [26]. Further work is required to fully characterize the individual haplotypes of each lineage. For example, long-read sequencing may be useful to produce complete, phased diploid genome sequences of each lineage.

We attempted to correlate genetic variants with phenotypes in the *C. tropicalis* AA isolates. Previous studies using MLST suggested that certain characteristics may be clade-specific in *C. tropicalis*, e.g. increased resistance to antifungals including fluconazole and flucytosine [23,70,71]. There are several difficulties with using genome-wide association studies (GWAS) to identify causative variants in fungi, including small sample sizes (in comparison to human studies), structural variation between isolates, and the influence of population structure [72,73]. In addition, phenotypes are often caused by a complex network of genetic and environmental factors. However, we previously applied cosine similarity to identify phenotype-genotype correlations in the related species *C. orthopsilosis* [47], by converting variants and phenotypes in different growth conditions to binary scores (presence/absence). A similar analysis allowed us to identify a variant in *BAT22* in one *C. tropicalis* isolate that is associated with the inability to use valine or isoleucine as sole nitrogen sources. However, the method has its drawbacks. For example, *C. tropicalis* ct04 has defects in many growth conditions other than valine or isoleucine, and contains at least 40 variants with respect to the reference strain with predicted high impact. The *BAT22* variant was selected based on information available from orthologs in *S. cerevisiae* and *C. albicans*.

S. cerevisiae encodes two BCAT enzymes, Bat1p (found in the mitochondria) and Bat2p (found in the cytosol) [74,75]. *BAT2* is mainly associated with catabolism and *BAT1* with biosynthesis of the branched chain amino acids valine, isoleucine and leucine [49,76,77]. Many *Candida* species, including *C. tropicalis*, also have two BCAT isozymes, which result from a recent gene duplication event [78]. *C. tropicalis* ct04 (*bat22*) has growth defects when either valine or isoleucine are the sole nitrogen source, but not when leucine is the sole nitrogen source. Previous studies have shown that leucine metabolism can occur in *S. cerevisiae* even when BCATs are deleted [49,77]. It has therefore been suggested that there are other unknown transaminases that contribute to leucine metabolism [49,77]. It is possible that in *C. tropicalis* catabolism of leucine requires Bat21 rather than Bat22, or other unknown transaminases.

Our study greatly expands the analyses of genotype and phenotype of *C. tropicalis* isolates. We have described the existence of hybrids for the first time in this species, and we question the hypothesis that hybridization is generally associated with virulence in CUG-Ser1 species. In addition, we have shown that genotype and phenotype correlations can be used to identify causative variants in *C. tropicalis*.

Materials & methods

Strain collection and growth

C. tropicalis isolates were collected from a variety of clinical (anonymized) and environmental sources (S1 Table). Yeast were isolated from soil samples as described in Sylvester et al. [79], or following three passages in YPD (1% yeast extract, 2% peptone, 2% glucose) with chloramphenicol (3% [wt/vol]) and ampicillin (10% [wt/vol]). For phenotype analysis, isolates were inoculated as 2x2 arrays (two independent cultures with one technical replicate of each) into 200 μl of YPD broth in 96-well plates and incubated at 30˚C for 24 h. Stocks were diluted in 96-well plates containing 200 μl of water by dipping a 12x8 pin bolt replicator (V&P Scientific) three times in the culture and then transferring it to the water. Once diluted, the cultures were pinned onto 85 unique media on solid agar plates and incubated at 30˚C for 48 h (S2 Table). For 60 conditions, the base media was YPD, with 2% agar including 2% glucose as a carbon source. Glucose was substituted with different carbon sources where indicated, or compounds were added at the indicated concentrations (S2 Table). To test the ability to use specific nitrogen sources (24 conditions), the base media was 0.19% of YNB (Yeast Nitrogen Base) without ammonium sulfate or amino acids, 2% glucose and 2% agar. Nitrogen sources were added as indicated (S2 Table). Spider media was tested as the 85th condition (S2 Table). Plates were photographed and growth was measured using SGAtools [80]. SGAtools was designed to analyze synthetic genetic interactions and assumes that average growth on a plate does not vary. This was not true for several media, where many strains grew poorly. We therefore compared the growth of each strain on the test media to the growth of the same strain on YPD, or on YNB with ammonium sulfate, as a control, using the raw data extracted from SGAtools. For each strain in each analyzed growth condition, the SGAtools scores (ranging from 0 to 1.8) were converted to a binary score where a growth ratio above 0.3 (no growth defect) was assigned 0, and a ratio below or equal to 0.3 (major growth defect) was assigned 1. These scores were chosen to be very stringent—only conditions which resulted in reducing growth to approximately 30% of that under the control conditions were judged as a defect. We found that SGAtools could not reproducibly identify enhanced growth in these conditions. The raw data for the image analysis is available at https://doi.org/10.6084/m9.figshare.13128839.v1.

Genome sequencing

For most *C. tropicalis* isolates, genomic DNA was isolated by phenol-chloroform extraction followed by purification using the Genomic DNA Cleanup and Concentration kit from Zymo Research (catalogue number D4065). For three isolates (*C. tropicalis* ct76, ct77 and ct78), genomic DNA was extracted and purified using the QIAamp DNA Mini Kit from QIAGEN (catalogue number 51304). For most isolates, library preparation and sequencing was performed at the Earlham Institute, Norwich, UK using the LITE method (Low Input Transposase-Enabled), a custom Nextera-based system. These isolates were sequenced on two lanes of an Illumina HiSeq 2500 generating 2x250 bp paired-end reads. For five isolates (*C. tropicalis* ct51, ct75, ct76, ct77 and ct78), library preparation and sequencing was performed by BGI, Hong Kong, generating 2x150 bp paired-end reads, on an Illumina HiSeq 4000. Our genome sequences of two isolates (*C. tropicalis* ct20 and ct21) were almost identical. These may represent independent isolates of the same strain, or one isolate may have been accidentally sequenced twice. We therefore included only one of these (*C. tropicalis* ct20) in subsequent analysis.

For the 72 unique isolates sequenced using the LITE method, Nextera adapters were removed using TrimGalore v0.4.3 with the parameters "—paired" "—length 35" "—nextera" and "—stringency 3". Custom adapters and low-quality bases were trimmed using Skewer v0.2.2 with the parameters "-m pe" "-l 35" "-q 30" "-Q 30" [81]. For 5 isolates sequenced by BGI, adapters were removed by the sequencing provider and reads were quality trimmed using Skewer. *K*-mer distribution profiles were analyzed using the *k*-mer Analysis Toolkit v2.4.2 using the default *k*-mer length of 27 bases [82]. All genomes were assembled using

SPAdes v3.9.1 with parameters "—careful" "-t 12" "-m 60" [83]. Assembly statistics were assessed using QUAST v4.4 [84]. To confirm the species identity of hybrid isolates, the D1/D2 domain of the large subunit of the ribosomal DNA was amplified using standard universal primers NL-1 and NL-4 (S3 Table).

Mating type-like locus analysis

The *MTL* idiomorph of a subset of isolates was confirmed by PCR using primer pairs *MTL***a***1*F and *MTL***a***1*R to amplify the *MTL***a***1* gene and *MTLα2*F and *MTLα2*R to amplify the *MTLα2* gene, as described in Xie et al. [21]. Colony PCR was performed by boiling single colonies in 5 μl sterile deionized water, then adding 12.5 μl MyTaq Red Mix (2X), 1 μl forward primer (100 μM), 1 μl reverse primer (100 μM) and 5.5 μl deionized water. PCR was run for 1 min at 95˚C; then for 30 cycles of 30 sec at 95˚C, 30 sec at 57˚C, 60 sec at 72˚C; and then a final 2 min at 72˚C.

C. tropicalis reference genome

The *C. tropicalis* reference genome annotation was updated using RNAseq data for three *C. tropicalis* strains downloaded from NCBI under BioProject ID PRJNA290183 [85]. RNAseq data were aligned against the original *C. tropicalis* reference [16] with HISAT2 v2.0.5 with the parameter "—novel-splicesite-outfile" to predict splice sites in the genome [86]. Predicted splice sites were manually validated by examination of transcripts mapping to predicted splice sites. The reference genome sequence was subsequently scaffolded from 23 supercontigs to 16 supercontigs. Areas of overlap between supercontigs in the original reference assembly were identified using Gepard to generate dot matrix plots [87]. Overlapping supercontigs were merged if this arrangement was supported by synteny with other *Candida* species, using the *Candida* Gene Order Browser (CGOB) [78], and by data from Illumina resequencing of the reference strain. The final assembly (also known as Assembly B [39]) contained 16 supercontigs and is available under NCBI accession JAFIQD000000000. The *C. tropicalis* reference was subsequently further improved as described by Guin et al [39].

Variant calling

For isolates sequenced using the LITE method, trimmed reads were aligned to *C. tropicalis* MYA-3404 Assembly B with bwa mem v0.7.11 to generate two BAM files per sample (one for each lane used for sequencing) [88]. BAM files were sorted with SAMtools v1.7 [89], and duplicate reads were marked using GenomeAnalysisToolkit (GATK) v3.7 Mark Duplicates [42]. BAM files from separate lanes were combined for each sample and marked for duplicates again using GATK MarkDuplicates. For isolates sequenced at BGI, Hong Kong, trimmed reads were aligned to the updated *C. tropicalis* MYA-3404 Assembly B with bwa mem v0.7.11 as before, generating only one BAM file per sample (each of these samples was sequenced on only one lane of the sequencer). BAM files were sorted with SAMtools v1.7 [89] and duplicate reads were marked using GenomeAnalysisToolkit (GATK) v3.7 Mark Duplicates [42].

The subsequent steps were applied to all samples. Realignment around indel sites was performed using GATK IndelRealigner and variants were called using GATK HaplotypeCaller in "—genotyping_mode DISCOVERY". Variants were filtered for quality based on genotype quality (GQ) < 20 and read depth (DP) < 10. For SNP trees, gVCFs were generated using GATK HaplotypeCaller with the parameters "—genotyping_mode DISCOVERY" and "—emitRefConfidence GVCF". Joint genotyping was performed using GATK GenotypeGVCFs to produce a single multi-sample gVCF. SNPs were extracted from the multi-sample gVCF using GATK SelectVariants with parameter "-selectType SNP". Variants were filtered based on genotype quality (GQ) < 20 and read depth (DP) < 10. For genotype-phenotype analysis, the presence of a variant at a particular site in each isolate was scored as 1, and absence was scored as 0.

Aneuploidy analysis

To calculate copy number variants based on coverage discrepancies, the *C. tropicalis* MYA-3404 Assembly B genome was split into 1 kb windows using the "makewindows" command from bedtools v2.26.0, with parameters "-i winnum" (label windows sequentially) "-w 1000" (window size 1 kb) [90]. Mean coverage in each 1 kb window was calculated for each sample using the "coverage" command from bedtools [90]. Average whole genome coverage for each strain was calculated using GATK DepthOfCoverage [42]. Coverage ratios for each 1 kb window were calculated as $\log_2$(window coverage/average whole genome coverage). A value of zero was assigned to windows that had zero coverage. The resultant ratios were visualized using the DNACopy package from Bioconductor in R [91]. Ploidy was also visualized using allele frequencies from heterozygous biallelic SNPs extracted from the VCF files using GATK SelectVariants with parameters "-selectType SNP" and "-restrictAllelesTo BIALLELIC". Allele frequency was calculated as allele depth (AD) /read depth (DP). Histograms of allele frequency for each scaffold in each sample were visualized in R using ggplot2 [92].

Phylogeny

SNP trees were drawn from filtered variants, using only those SNPs that passed the filters described in "Variant Calling". To account for heterozygous SNPs, the Repeated Random Haplotype Sampling tool (RRHS) v1.0.0.2 was used to select a random allele at heterozygous SNP sites [93]. This process was performed 100 times to generate 100 SNP profiles for each isolate, thereby encapsulating the full heterozygosity of each isolate. For homozygous variant sites, the alternate allele was chosen by default. 100 maximum likelihood (ML) trees were drawn (one for each SNP profile) using RAxML v8.2.12 [94] with the "GTRGAMMA" model. The best-scoring ML tree was chosen as a reference tree and the remaining 99 ML trees were used as pseudo-bootstrap trees to generate a supertree using RAxML v8.2.12 with options "-f b" (draw bipartition information on a reference tree based on multiple trees (e.g. from a bootstrap)) and the "GTRGAMMA" model. Phylogeny was also examined using principal component analysis (PCA) with the ade4 package in R [95].

Loss of heterozygosity

Loss of heterozygosity (LOH) was calculated in blocks of at least 100 base pairs (bp) across the genome. Heterozygous regions were defined as any region containing at least two heterozygous variants within 100 bp of each other, with a minimum total length of 100 bp. Remaining regions were defined as homozygous, or LOH, regions as long as they were at least 100 bp in length. A similar approach has been used to characterize LOH in other *Candida* hybrids, e.g. *C. metapsilosis* [27], *C. inconspicua* [38] and *C. albicans* [25]. Commonality of variants in heterozygous regions was examined to determine which isolates originated from the same parental strains. Heterozygous regions shared by all isolates were identified using bedops intersect [96]. In the case of heterozygous regions that were partially shared, the portion that was common to all isolates was extracted and analyzed as a shared heterozygous region. The number of common variants in the shared heterozygous regions was counted as the number of variant sites in these regions with the same genotype in all isolates. Shared LOH regions were defined as LOH blocks with identical start and stop coordinates in the relevant isolates. The analysis was repeated using the updated *C. tropicalis* genome assembly from Guin et al. [39].

Haplotype splitting

Hybrid haplotypes were phased using HapCUT2 v0.7 [44]. The filtered variants were used as input for the subcommand "extractHAIRS" (extract haplotype-informative reads) to identify "haplotype-informative reads", i.e. sets of reads that align to the same location in the reference genome but that contain one or more different alleles at variant sites. HapCUT2 was subsequently used to build haplotype blocks from the haplotype-informative reads with parameter "—threshold 30" (Phred-scaled threshold for pruning low-confidence SNPs). Haplotype-informative reads form the basis of these haplotype blocks. Two haplotype-informative reads will be assigned to the same haplotype if they overlap by a certain amount and have matching alleles at variant sites. Thereby, the haplotype block is extended in a continuous manner until no further overlapping reads can be identified. Each block consists of a region of the genome where alleles at variant sites in that region have been assigned to one of two haplotypes. Therefore, each phased block is assigned two sets of variants; one for each haplotype. For each phased block, the percentage difference of each of these two haplotypes was calculated by counting the number of bases in that haplotype that were different from the reference (i.e. the number of variant alleles assigned to the haplotype) and dividing this number by the total number of bases in the block. The difference of each phased block to the reference genome was calculated as the number of SNPs in block/length of block. Haplotypes were subsequently assigned to either the reference haplotype or the alternate haplotype according to their percentage difference; the member of the pair that was more similar to the reference was assigned to haplotype A and the member of the pair that was less similar to the reference was assigned to haplotype B. In the majority of cases, the A haplotype was $< 0.3\%$ different to the reference genome and the B haplotype was $> 4\%$ different to the reference genome.

Analysis of genotype-phenotype correlation

Variants from non-hybrid isolates were further annotated with SnpEff v4.3t to predict the functional effect of variants [46]. High-impact variants (e.g. variation at splice donor or acceptor sites, variants resulting in a gain or loss of stop or start codon, or frameshifts in genes) were extracted and correlated with phenotypes. Variants were converted to binary scores; 1 for the presence of a variant in a given strain, 0 for the absence. Phenotype scores were coded as 1 for a growth defect (score of 0.3 or less), and as 0 for no growth defect (score above 0.3). For each variant-condition pair, two vectors were generated using the binary scores; the first consists of the scores for every strain with respect to the variant, the second consists of the scores for every strain with respect to the condition. For every variant-condition vector pair, the cosine similarity between the two vectors was calculated as $cos\,\theta = \frac{\rightarrow a . \rightarrow b}{\|\rightarrow a\|\|\rightarrow b\|}$. Any variant-condition pair with a cosine similarity of > 0.85 was selected for further analysis.

Editing *BAT22* with CRISPR-Cas9

A 20 bp sequence (guide RNA) targeting *C. tropicalis BAT22* (*CTRG_06204*) was designed using the web tool ChopChop [97]. The guide RNA was generated by annealing of two short oligos (g60BAT22_TOP/BOT, S3 Table), and then cloned into the SapI-digested pCT-tRNA plasmid to generate plasmid pCT-tRNA-BAT22, as previously described in [50]. The repair template carrying the desired modification, including the disruption of the PAM sequence, was generated by primer extension (RT_BAT22_2bpDel_SNP-TOP/BOT) using ExTaq DNA polymerase (Takara Bio, USA). *C. tropicalis* isolates ct09, ct44 and ct53 were transformed with 5 μg pCT-tRNA-BAT22 and 25 μl of unpurified RT- BAT22_2bpDel_SNP using a previously described method [50]. Transformants were selected on YPD agar plates containing 200 μg/ml nourseothricin (NTC), incubated at 30˚C for 48 h. The relevant region was amplified by PCR from two NTC-resistant transformants for each strain using primers bat22_fwd_01/bat22_rev_01 and sequenced using Sanger sequencing. The pCT-tRNA-BAT22 plasmid was cured by growing the cells in the absence of selection on YPD until they failed to grow in the presence of NTC.

Supporting information

S1 Text. LOH analysis using updated *C. tropicalis* assembly.
(DOCX)

S1 Fig. Polyploidy and aneuploidy in *C. tropicalis* isolates. (A) Polyploidy of *C. tropicalis* isolates. The frequency of the non-reference allele for all heterozygous biallelic SNPs across all scaffolds is shown for each of the isolates, with frequency on the Y-axis and alternate (non-reference) allele frequency on the X-axis. For each SNP, allele frequency was calculated as the depth of the alternate allele divided by the total depth at the variant site. Triploidy of *C. tropicalis* ct66 is indicated by peaks of allele frequency at 0.33 and 0.66. Octaploidy of *C. tropicalis* ct26 is indicated by peaks of allele frequency at approximately 0.5, 0.12 and 0.87. Allele frequencies of approximately 0.125 and 0.875 imply that seven chromosomes carry one allele, and one chromosome carries a second allele. In this isolate, we also observe a peak at 0.5, implying that in some cases, four chromosomes carry one allele and four chromosomes carry a second allele. This multimodal distribution (i.e. peaks at 0.125, 0.50 and 0.875) is likely to be the result of loss of heterozygosity (LOH) affecting portions of some scaffolds, leading to a pattern wherein some variant sites have a 4:4 ratio of reference:non-reference allele frequency and some have a 7:1 ratio. **(B) Aneuploidy of *C. tropicalis* isolates.** Single chromosome aneuploidies were identified in four isolates; *C. tropicalis* ct06, a clinical isolate from Dublin, Ireland, *C. tropicalis* ct14 and ct15, both engineered strains from the USA [41], and *C. tropicalis* ct18, a clinical isolate from Madrid, Spain. Aneuploidies were identified by patterns in the distribution of allele frequency in heterozygous biallelic SNPs (shown as red histograms for the relevant scaffold, with frequency on the Y-axis and alternative allele frequency on the X-axis). Allele frequency was calculated as the depth of coverage of the alternate (non-reference) allele divided by the total depth at the variant site. Aneuploidies were confirmed by elevated coverage at the relevant locus (shown as dot plots, with green and black representing alternating scaffolds). Scaffold number is shown on the X-axis and the log2(observed coverage/expected coverage) is shown on the Y-axis (where "expected coverage" is the average genome-wide coverage for that isolate). Scaffolds are listed in decreasing order of size; the eight largest scaffolds are shown. The equivalent chromosomes in the assembly described by Guin et al. [39] are: scaffold 1 and chromosome 3; scaffold 2 and chromosome 1; scaffold 3 and chromosome 4; scaffold 4 and chromosome R; scaffolds 5 and 6 and chromosome 2, scaffold 7 and chromosome 6; and scaffold 8 and chromosome 5.
(PDF)

S2 Fig. CNVs in *C. tropicalis* isolates. (A) CNV in isolates *C. tropicalis* ct04 and *C. tropicalis* ct33. CNVs were visualized as elevated coverage at the relevant locus (shown as dot plots, with green and black representing alternating scaffolds). Scaffold number is shown on the X-axis and the log2(observed coverage/expected coverage) is shown on the Y-axis (where "expected coverage" is the average genome-wide coverage for that isolate). Scaffolds are listed in decreasing order of size; the eight largest scaffolds are shown. A duplication of a region of approximately 253 kb on scaffold 7 is observed in two isolates; *C. tropicalis* ct04, a clinical isolate from Dublin, Ireland, and *C. tropicalis* ct33, a clinical isolate from Madrid, Spain. This CNV

(highlighted with a blue box) spans the region from approximately 350 kb to 603 kb. A score of 1 at this region indicates a doubling in coverage, i.e. a total copy number of four. **(B) CNV in isolates ct12, ct14, ct15, ct26, ct33, ct36 and ct69**. CNVs were visualized as elevated coverage at the relevant locus (shown as dot plots). Position on the chromosome (kb) is shown on the X-axis and the log2(observed coverage/expected coverage) is shown on the Y-axis (where "expected coverage" is the average genome-wide coverage for that isolate). Scaffold 4 only is shown. A small CNV (~35 kb) is visible at the 1 Mb point of scaffold 4 in seven isolates, *C. tropicalis* ct12 (a clinical isolate from Colombia), ct14, ct15 (both engineered isolates from the USA), ct26, ct33, ct36 (three clinical isolates from Madrid, Spain) and ct69 (a clinical isolate from the USA). This CNV (highlighted with a blue box) spans the region from approximately 974 kb to 1.009 Mb on scaffold 4. A score of -1 at this region indicates a relative coverage level of 0.5.
(PDF)

S3 Fig. PCA analysis of *C. tropicalis* genomes. Principal component analysis (PCA) of Cluster A isolates (Fig 1B) was performed using the ade4 package in R [95] (S4 Table). Principal components 1 and 2 are represented on the X- and Y-axes respectively. Six clusters were identified using Ward's method. Clusters one, three, four, five and six are the same as groupings as Fig 1C, except that *C. tropicalis* ct09 is included in Cluster 4 in the PCA analysis only, and *C. tropicalis* ct38 is included in Cluster 1 in the PCA analysis only. Cluster 2 is not clearly separated in the SNP phylogeny.
(PDF)

S4 Fig. *K*-mer distribution profiles of *C. tropicalis* isolates. *K*-mer frequency distribution profiles are shown for all *C. tropicalis* isolates not in Fig 2. *K*-mer analysis was performed with the *k*-mer Analysis Toolkit (KAT [82]). For each isolate, the number of distinct *k*-mers of length 27 bases (27-mers) is displayed on the Y-axis and *k*-mer multiplicity (depth of coverage) is displayed on the X-axis. *K*-mers that are present in the reference genome with a frequency of 1 (i.e. 1X) are shown in red, and *k*-mers that are absent from the reference genome (i.e. 0X) are shown in black. In the non-hybrid (AA) isolates, there is no bimodal pattern observed, unlike in the hybrid isolates.
(PDF)

S5 Fig. Loss of heterozygosity compared to updated genome assembly of *C. tropicalis*. (A) The same patterns of LOH and heterozygosity are observed using an updated reference genome assembly. LOH was re-analyzed using an updated chromosome-level assembly from Guin et al [39]. The seven chromosomes in the reference genome are displayed horizontally from left to right and labelled from 1 to 6, plus chromosome R. Chromosomes in the alternative reference genome map to scaffolds in the original reference genome as follows; chr1:scaffold 2, chr2: scaffolds 5 and 6, chr3: scaffold 1, chr4: scaffold 3, chr5: scaffold 8, chr6: scaffold 7, chrR: scaffold 4. LOH blocks are shown in pink and heterozygous ("HET") blocks are shown in green. Centromere positions are indicated with "C", telomere positions are indicated with "T" and the rDNA locus is indicated with "R". Isolates are labelled on the left-hand side. The re-sequenced reference strain *C. tropicalis* MYA-3404 (labelled as "Ref") is shown as a representative of the non-hybrid (AA) isolates. The same patterns of LOH/heterozygosity are observed in the AA, AB and AC isolates when using the alternative reference as when using the original reference genome. **(B) The length of LOH blocks are unchanged when analyzed using an updated reference genome assembly.** The histograms show the frequency of LOH blocks of different lengths in the six hybrid isolates and two AA (non-hybrid) isolates, the re-sequenced reference strain *C. tropicalis* MYA-3404 (labelled as "Ref") and *C. tropicalis* ct20.

Frequency is shown on a log scale on the Y-axis while length in kilobases (kb) is shown on the X-axis, with a bin width of 1000 bp. The average length of LOH blocks in the hybrid isolates ranges from 289–417 bp, a difference of only a few base pairs from the analysis using the original reference genome. The same patterns are observed in the AA and hybrid isolates when using the updated reference genome.
(PDF)

S6 Fig. Analysis of *MTL* idiomorphs. The gel shows the results of the colony PCR amplification of the *MTL* in eleven *C. tropicalis* isolates (labelled in grey or white boxes). Hyperladder is shown on the left- and right-most column of the gel on both rows, with the sizes of the bottom three markers (200 bp, 400 bp and 600 bp) marked. Two reactions were performed for each isolate—one using primer pairs *MTL***a***1*F and *MTL***a***1*R to amplify the *MTL***a***1* gene (lane marked "a") and *MTLα2*F and *MTLα2*R to amplify the *MTLα2* gene (lane marked "α"), as described in Xie et al. [21]. A band of 253 bp is expected in the "a" lane for isolates with at least one copy of the *MTL***a***1* gene and a band of 525 bp is expected in the "α" lane for isolates with at least one copy of the *MTLα2* gene. Negative control (all components of PCR mix excluding input DNA) is marked as "NC" on the bottom row, with one lane for each primer set (marked "a" and "α"). Most isolates are heterozygous, but *C. tropicalis* ct14 and ct73 are homozygous for *MTL***a**. The octoploid isolate *C. tropicalis* ct26 has a strong positive signal for *MTLα* (lane marked "α") and a weak positive signal for *MTL***a** (lane marked "a"), highlighted with a red box. The genome assembly contains one full copy of *OBP***a**, and partial copies of the remainder of the *MTL***a** genes (*PAP***a**, *PIK***a**, *MTL***a***2* and *MTL***a***1*). The five *MTL***a** genes are scattered across five low-coverage contigs (coverage 1.3X - 2X), most of which are only the length of the gene itself. One gene, *MTL***a***2*, is split across two scaffolds. It is possible that there is one copy of *MTL***a** and up to seven copies of *MTLα*, resulting in low sequencing coverage of the *MTL***a** locus.
(PDF)

S7 Fig. Phenotypic analysis of *C. tropicalis* AA isolates. 68 *C. tropicalis* isolates were grown on YPD (A) or YNB with ammonium (NH_4) (B) solid agar media as a control, and compared to strains growing on solid agar media containing different stressors. Pictures were taken after 48 hours and colony size and growth scores were measured using SGAtools [80]. Heatmaps show the normalized raw colony size in various tested growth conditions. Isolates are represented in rows, and are ordered alphabetically by strain alias. Growth conditions are shown in columns. Increased growth relative to YPD or YNB + NH_4 is shown in green (1–2) and decreased growth is shown in purple (0–1). Major differences are observed between isolates growing in the presence of cell wall stressors (calcofluor white, congo red, sodium dodecyl sulphate, caffeine), and antifungal drugs (ketoconazole, caspofungin, fluconazole). Hybrid isolates and engineered lab isolates were excluded from this analysis.
(PDF)

S8 Fig. Variants in *C. tropicalis* isolates by category. (A) The majority of variants in non-hybrid (AA) *C. tropicalis* isolates are single nucleotide polymorphisms (SNPs). Variants were called in all non-hybrid isolates using the Genome Analysis Toolkit [42] and annotated with SnpEff [46]. Variant type is shown as a barplot, with variant categories on the X-axis and variant count on the Y-axis. Approximately 75% of all annotated variants are SNPs, 12.51% are insertions and 12.57% are deletions. **(B) Identification of high-impact variants.** 9,261 high-impact variants were identified across 68 non-hybrid *C. tropicalis* isolates. Variant classification according to SnpEff is shown as a barplot, with estimated impact level categories on the X-axis and variant count on the Y-axis. Precise counts are shown above each bar. 9,261

variants were annotated as "high impact." These variants are predicted to have a major impact on protein function (e.g. gain or loss of start or stop codon, frameshifts, or splice site variants). These variants were analyzed for potential genotype-phenotype correlations.
(PDF)

S1 Table. List of strains used in this study.
(XLSX)

S2 Table. List of media used for phenotypic testing.
(XLSX)

S3 Table. List of primers used in this study.
(XLSX)

S4 Table. Isolate clusters identified by principal component analysis.
(XLSX)

S5 Table. Summary of LOH and heterozygous blocks in *C. tropicalis* isolates.
(XLSX)

S6 Table. Summary of LOH and heterozygous blocks in *C. tropicalis* isolates using updated reference genome assembly.
(XLSX)

S7 Table. List of phenotype-genotype correlations.
(XLSX)

Acknowledgments

We are grateful to Dr Shawn R. Lockhart from the Mycotic Diseases Branch, Centers for Disease Control and Prevention, United States for providing isolates. Thanks to Elizabeth Boyd, Eric Butler, Jane Kennedy and Aaron McLaughlin for collecting and identifying *C. tropicalis* isolates from soil as part of an undergraduate project at University College Dublin; Quinn K. Langdon and Dana A. Opulente for helping mentor MABH; and the Zasadil Family for collecting samples in the USA.

Author Contributions

Conceptualization: Caoimhe E. O'Brien, Geraldine Butler.

Formal analysis: Caoimhe E. O'Brien, João Oliveira-Pacheco.

Funding acquisition: Chris Todd Hittinger, Geraldine Butler.

Investigation: Caoimhe E. O'Brien, João Oliveira-Pacheco, Eoin Ó Cinnéide, Geraldine Butler.

Resources: Max A. B. Haase, Chris Todd Hittinger, Thomas R. Rogers, Oscar Zaragoza, Ursula Bond.

Supervision: Chris Todd Hittinger, Geraldine Butler.

Writing – original draft: Caoimhe E. O'Brien, João Oliveira-Pacheco, Geraldine Butler.

Writing – review & editing: Caoimhe E. O'Brien, João Oliveira-Pacheco, Chris Todd Hittinger, Geraldine Butler.

References

1. Pfaller MA, Diekema DJ, Gibbs DL, Newell VA, Ellis D, Tullio V, et al. Results from the ARTEMIS DISK Global Antifungal Surveillance Study, 1997 to 2007: a 10.5-year analysis of susceptibilities of *Candida* Species to fluconazole and voriconazole as determined by CLSI standardized disk diffusion. J Clin Microbiol. 2010; 48: 1366–1377. https://doi.org/10.1128/JCM.02117-09 PMID: 20164282

2. Pfaller MA, Diekema DJ, Turnidge JD, Castanheira M, Jones RN. Twenty years of the SENTRY antifungal surveillance program: results for species from 1997–2016. Open Forum Infect Dis. 2019; 6: S79–S94. https://doi.org/10.1093/ofid/ofy358 PMID: 30895218

3. Tan TY, Hsu LY, Alejandria MM, Chaiwarith R, Chinniah T, Chayakulkeeree M, et al. Antifungal susceptibility of invasive *Candida* bloodstream isolates from the Asia-Pacific region. Med Mycol. 2016; 54: 471–477. https://doi.org/10.1093/mmy/myv114 PMID: 26868904

4. Nucci M, Queiroz-Telles F, Alvarado-Matute T, Tiraboschi IN, Cortes J, Zurita J, et al. Epidemiology of candidemia in Latin America: a laboratory-based survey. PLoS One. 2013; 8: e59373. https://doi.org/10.1371/journal.pone.0059373 PMID: 23527176

5. Tan BH, Chakrabarti A, Li RY, Patel AK, Watcharananan SP, Liu Z, et al. Incidence and species distribution of candidaemia in Asia: a laboratory-based surveillance study. Clin Microbiol Infect. 2015. pp. 946–953. https://doi.org/10.1016/j.cmi.2015.06.010 PMID: 26100373

6. Kontoyiannis DP, Vaziri I, Hanna HA, Boktour M, Thornby J, Hachem R, et al. Risk factors for *Candida tropicalis* fungemia in patients with cancer. Clin Infect Dis. 2001; 33: 1676–1681. https://doi.org/10.1086/323812 PMID: 11568858

7. Arendrup MC, Bruun B, Christensen JJ, Fuursted K, Johansen HK, Kjaeldgaard P, et al. National surveillance of fungemia in Denmark (2004 to 2009). J Clin Microbiol. 2011; 49: 325–334. https://doi.org/10.1128/JCM.01811-10 PMID: 20980569

8. Fan X, Xiao M, Liao K, Kudinha T, Wang H, Zhang L, et al. Notable increasing trend in azole non-susceptible *Candida tropicalis* causing invasive candidiasis in China (August 2009 to July 2014): molecular epidemiology and clinical azole consumption. Front Microbiol. 2017;8. https://doi.org/10.3389/fmicb.2017.00008 PMID: 28144237

9. Liu W-L, Huang Y-T, Hsieh M-H, Hii I-M, Lee Y-L, Ho M-W, et al. Clinical characteristics of *Candida tropicalis* fungaemia with reduced triazole susceptibility in Taiwan: a multicentre study. Int J Antimicrob Agents. 2019; 53: 185–189. https://doi.org/10.1016/j.ijantimicag.2018.10.015 PMID: 30722962

10. Hii I-M, Liu C-E, Lee Y-L, Liu W-L, Wu P-F, Hsieh M-H, et al. Resistance rates of non- infections in Taiwan after the revision of 2012 Clinical and Laboratory Standards Institute breakpoints. Infect Drug Resist. 2019; 12: 235–240. https://doi.org/10.2147/IDR.S184884 PMID: 30679913

11. Almirante B, Rodríguez D, Park BJ, Cuenca-Estrella M, Planes AM, Almela M, et al. Epidemiology and predictors of mortality in cases of *Candida* bloodstream infection: results from population-based surveillance, Barcelona, Spain, from 2002 to 2003. J Clin Microbiol. 2005; 43: 1829–1835. https://doi.org/10.1128/JCM.43.4.1829-1835.2005 PMID: 15815004

12. Tortorano AM, Peman J, Bernhardt H, Klingspor L, Kibbler CC, Faure O, et al. Epidemiology of candidaemia in Europe: results of 28-month European Confederation of Medical Mycology (ECMM) hospital-based surveillance study. Eur J Clin Microbiol Infect Dis. 2004; 23: 317–322. https://doi.org/10.1007/s10096-004-1103-y PMID: 15029512

13. Muñoz P, Giannella M, Fanciulli C, Guinea J, Valerio M, Rojas L, et al. *Candida tropicalis* fungaemia: incidence, risk factors and mortality in a general hospital. Clin Microbiol Infect. 2011; 17: 1538–1545. https://doi.org/10.1111/j.1469-0691.2010.03338.x PMID: 20718804

14. Santos MAS, Gomes AC, Santos MC, Carreto LC, Moura GR. The genetic code of the fungal CTG clade. Comptes Rendus Biologies. 2011. pp. 607–611. https://doi.org/10.1016/j.crvi.2011.05.008 PMID: 21819941

15. Krassowski T, Coughlan AY, Shen X-X, Zhou X, Kominek J, Opulente DA, et al. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat Commun. 2018; 9: 1887. https://doi.org/10.1038/s41467-018-04374-7 PMID: 29760453

16. Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, et al. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature. 2009; 459: 657–662. https://doi.org/10.1038/nature08064 PMID: 19465905

17. Porman AM, Alby K, Hirakawa MP, Bennett RJ. Discovery of a phenotypic switch regulating sexual mating in the opportunistic fungal pathogen *Candida tropicalis*. Proc Natl Acad Sci U S A. 2011; 108: 21158–21163. https://doi.org/10.1073/pnas.1112076109 PMID: 22158989

18. Seervai RNH, Jones SK Jr, Hirakawa MP, Porman AM, Bennett RJ. Parasexuality and ploidy change in *Candida tropicalis*. Eukaryot Cell. 2013; 12: 1629–1640. https://doi.org/10.1128/EC.00128-13 PMID: 24123269

19. Du H, Zheng Q, Bing J, Bennett RJ, Huang G. A coupled process of same- and opposite-sex mating generates polyploidy and genetic diversity in *Candida tropicalis*. PLoS Genet. 2018; 14: e1007377. https://doi.org/10.1371/journal.pgen.1007377 PMID: 29734333

20. Porman AM, Hirakawa MP, Jones SK, Wang N, Bennett RJ. MTL-independent phenotypic switching in *Candida tropicalis* and a dual role for Wor1 in regulating switching and filamentation. PLoS Genet. 2013; 9: e1003369. https://doi.org/10.1371/journal.pgen.1003369 PMID: 23555286

21. Xie J, Du H, Guan G, Tong Y, Kourkoumpetis TK, Zhang L, et al. N-acetylglucosamine induces white-to-opaque switching and mating in *Candida tropicalis*, providing new insights into adaptation and fungal sexual evolution. Eukaryot Cell. 2012; 11: 773–782. https://doi.org/10.1128/EC.00047-12 PMID: 22544905

22. Wu Y, Zhou H-J, Che J, Li W-G, Bian F-N, Yu S-B, et al. Multilocus microsatellite markers for molecular typing of *Candida tropicalis* isolates. BMC Microbiol. 2014; 14: 245. https://doi.org/10.1186/s12866-014-0245-z PMID: 25410579

23. Tavanti A, Davidson AD, Johnson EM, Maiden MCJ, Shaw DJ, Gow NAR, et al. Multilocus sequence typing for differentiation of strains of *Candida tropicalis*. J Clin Microbiol. 2005; 43: 5593–5600. https://doi.org/10.1128/JCM.43.11.5593-5600.2005 PMID: 16272492

24. Ropars J, Maufrais C, Diogo D, Marcet-Houben M, Perin A, Sertour N, et al. Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. Nat Commun. 2018; 9: 2253. https://doi.org/10.1038/s41467-018-04787-4 PMID: 29884848

25. Mixão V, Gabaldón T. Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*. BMC Biol. 2020; 18: 48. https://doi.org/10.1186/s12915-020-00776-6 PMID: 32375762

26. Pryszcz LP, Németh T, Gácser A, Gabaldón T. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. Genome Biol Evol. 2014; 6: 1069–1078. https://doi.org/10.1093/gbe/evu082 PMID: 24747362

27. Pryszcz LP, Németh T, Saus E, Ksiezopolska E, Hegedűsová E, Nosek J, et al. The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis*. PLoS Genet. 2015; 11: e1005626. https://doi.org/10.1371/journal.pgen.1005626 PMID: 26517373

28. Schröder MS, Martinez de San Vicente K, Prandini THR, Hammel S, Higgins DG, Bagagli E, et al. Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species. PLoS Genet. 2016; 12: e1006404. https://doi.org/10.1371/journal.pgen.1006404 PMID: 27806045

29. Louis VL, Despons L, Friedrich A, Martin T, Durrens P, Casarégola S, et al. *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. G3. 2012; 2: 299–311. https://doi.org/10.1534/g3.111.000745 PMID: 22384408

30. Libkind D, Hittinger CT, Valério E, Gonçalves C, Dover J, Johnston M, et al. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. Proc Natl Acad Sci U S A. 2011; 108: 14539–14544. https://doi.org/10.1073/pnas.1105430108 PMID: 21873232

31. Belloch C, Orlic S, Barrio E, Querol A. Fermentative stress adaptation of hybrids within the *Saccharomyces* sensu stricto complex. Int J Food Microbiol. 2008; 122: 188–195. https://doi.org/10.1016/j.ijfoodmicro.2007.11.083 PMID: 18222562

32. Solieri L, Landi S, De Vero L, Giudici P. Molecular assessment of indigenous yeast population from traditional balsamic vinegar. J Appl Microbiol. 2006; 101: 63–71. https://doi.org/10.1111/j.1365-2672.2006.02906.x PMID: 16834592

33. Gordon JL, Wolfe KH. Recent allopolyploid origin of *Zygosaccharomyces rouxii* strain ATCC 42981. Yeast. 2008; 25: 449–456. https://doi.org/10.1002/yea.1598 PMID: 18509846

34. Bizzarri M, Cassanelli S, Bartolini L, Pryszcz LP, Dušková M, Sychrová H, et al. Interplay of chimeric Mating-Type Loci impairs fertility rescue and accounts for intra-strain variability in interspecies hybrid ATCC42981. Front Genet. 2019; 10: 137. https://doi.org/10.3389/fgene.2019.00137 PMID: 30881382

35. Bizzarri M, Cassanelli S, Pryszcz LP, Gawor J, Gromadka R, Solieri L. Draft genome sequences of the highly halotolerant strain *Zygosaccharomyces rouxii* ATCC 42981 and the novel allodiploid strain *Zygosaccharomyces sapae* ATB301 obtained Using the MinION platform. Microbiol Resour Announc. 2018;7. https://doi.org/10.1128/MRA.00874-18 PMID: 30533882

36. Xu J, Luo G, Vilgalys RJ, Brandt ME, Mitchell TG. Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD. Microbiology. 2002; 148: 203–212. https://doi.org/10.1099/00221287-148-1-203 PMID: 11782512

37. Xu J, Vilgalys R, Mitchell TG. Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Mol Ecol. 2000; 9: 1471–1481. https://doi.org/10.1046/j.1365-294x.2000.01021.x PMID: 11050543

38. Mixão V, Hansen AP, Saus E, Boekhout T, Lass-Florl C, Gabaldón T. Whole-genome sequencing of the opportunistic yeast pathogen *Candida inconspicua* uncovers its hybrid origin. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00010 PMID: 30815010

39. Guin K, Chen Y, Mishra R, Muzaki SRB, Thimmappa BC, O'Brien CE, et al. Spatial inter-centromeric interactions facilitated the emergence of evolutionary new centromeres. Elife. 2020;9. https://doi.org/10.7554/eLife.58556 PMID: 32469306

40. Mancera E, Porman AM, Cuomo CA, Bennett RJ, Johnson AD. Finding a missing gene: *EFG1* regulates morphogenesis in *Candida tropicalis*. G3. 2015; 5: 849–856. https://doi.org/10.1534/g3.115.017566 PMID: 25758825

41. Mancera E, Frazer C, Porman AM, Ruiz-Castro S, Johnson AD, Bennett RJ. Genetic modification of closely related *Candida* species. Front Microbiol. 2019; 10: 357. https://doi.org/10.3389/fmicb.2019.00357 PMID: 30941104

42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

43. Arbour M, Epp E, Hogues H, Sellam A, Lacroix C, Rauceo J, et al. Widespread occurrence of chromosomal aneuploidy following the routine production of *Candida albicans* mutants. FEMS Yeast Res. 2009; 9: 1070–1077. https://doi.org/10.1111/j.1567-1364.2009.00563.x PMID: 19732157

44. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017; 27: 801–812. https://doi.org/10.1101/gr.213462.116 PMID: 27940952

45. Ries LNA, Beattie S, Cramer RA, Goldman GH. Overview of carbon and nitrogen catabolite metabolism in the virulence of human pathogenic fungi. Mol Microbiol. 2018; 107: 277–297. https://doi.org/10.1111/mmi.13887 PMID: 29197127

46. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 2012; 6: 80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

47. de San Vicente KM, Schröder MS, Lombardi L, Iracane E, Butler G. Correlating genotype and phenotype in the asexual yeast *Candida orthopsilosis* implicates *ZCF29* in sensitivity to caffeine. G3. 2019; 9: 3035–3043. https://doi.org/10.1534/g3.119.400348 PMID: 31352406

48. Garbe E, Vylkova S. Role of amino acid metabolism in the virulence of human pathogenic fungi. Curr Clin Microbiol Rep. 2019; 6: 108–119.

49. Takpho N, Watanabe D, Takagi H. Valine biosynthesis in *Saccharomyces cerevisiae* is regulated by the mitochondrial branched-chain amino acid aminotransferase Bat1. Microb Cell Fact. 2018; 5: 293–299. https://doi.org/10.15698/mic2018.06.637 PMID: 29850466

50. Lombardi L, Oliveira-Pacheco J, Butler G. Plasmid-based CRISPR-Cas9 gene editing in multiple *Candida* species. mSphere. 2019;4. https://doi.org/10.1128/mSphere.00125-19 PMID: 30867327

51. Vogel C, Rogerson A, Schatz S, Laubach H, Tallman A, Fell J. Prevalence of yeasts in beach sand at three bathing beaches in South Florida. Water Res. 2007; 41: 1915–1920. https://doi.org/10.1016/j.watres.2007.02.010 PMID: 17382990

52. Lord ATK, Mohandas K, Somanath S, Ambu S. Multidrug resistant yeasts in synanthropic wild birds. Ann Clin Microbiol Antimicrob. 2010; 9: 11. https://doi.org/10.1186/1476-0711-9-11 PMID: 20307325

53. Yang Y-L, Lin C-C, Chang T-P, Lauderdale T-L, Chen H-T, Lee C-F, et al. Comparison of human and soil *Candida tropicalis* isolates with reduced susceptibility to fluconazole. PLoS One. 2012; 7: e34609. https://doi.org/10.1371/journal.pone.0034609 PMID: 22496832

54. de Oliveira TB, Lopes VCP, Barbosa FN, Ferro M, Meirelles LA, Sette LD, et al. Fungal communities in pressmud composting harbour beneficial and detrimental fungi for human welfare. Microbiology. 2016; 1147–1156. https://doi.org/10.1099/mic.0.000306 PMID: 27170376

55. Lo H-J, Tsai S-H, Chu W-L, Chen Y-Z, Zhou Z-L, Chen H-F, et al. Fruits as the vehicle of drug resistant pathogenic yeasts. J Infect. 2017; 75: 254–262. https://doi.org/10.1016/j.jinf.2017.06.005 PMID: 28648496

56. Opulente DA, Langdon QK, Buh KV, Haase MAB, Sylvester K, Moriarty RV, et al. Pathogenic budding yeasts isolated outside of clinical settings. FEMS Yeast Res. 2019;19. https://doi.org/10.1093/femsyr/foz032 PMID: 31076749

57. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, et al. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog. 2010; 6: e1000713. https://doi.org/10.1371/journal.ppat.1000713 PMID: 20072605

58. Hoarau G, Mukherjee PK, Gower-Rousseau C, Hager C, Chandra J, Retuerto MA, et al. Bacteriome and mycobiome interactions underscore microbial dysbiosis in familial Crohn's Disease. MBio. 2016; 7: e21050–16 https://doi.org/10.1128/mBio.01250-16 PMID: 27651359

59. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, et al. Topographic diversity of fungal and bacterial communities in human skin. Nature. 2013; 498: 367–370. https://doi.org/10.1038/nature12171 PMID: 23698366

60. Iliev ID, Funari VA, Taylor KD, Nguyen Q, Reyes CN, Strom SP, et al. Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science. 2012; 336: 1314–1317. https://doi.org/10.1126/science.1221789 PMID: 22674328

61. Malgoire JY, Bertout S, Renaud F, Bastide JM, Mallié M. Typing of *Saccharomyces cerevisiae* clinical strains by using microsatellite sequence polymorphism. J Clin Microbiol. 2005; 43: 1133–1137. https://doi.org/10.1128/JCM.43.3.1133-1137.2005 PMID: 15750073

62. Muller LAH, Lucas JE, Georgianna DR, McCusker JH. Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. Mol Ecol. 2011; 20: 4085–4097. https://doi.org/10.1111/j.1365-294X.2011.05225.x PMID: 21880084

63. Muller LAH, McCusker JH. Microsatellite analysis of genetic diversity among clinical and nonclinical *Saccharomyces cerevisiae* isolates suggests heterozygote advantage in clinical environments. Mol Ecol. 2009; 18: 2779–2786. https://doi.org/10.1111/j.1365-294X.2009.04234.x PMID: 19457175

64. Stukenbrock EH. The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology. 2016; 106: 104–112. https://doi.org/10.1094/PHYTO-08-15-0184-RVW PMID: 26824768

65. Samarasinghe H, Xu J. Hybrids and hybridization in the *Cryptococcus neoformans* and *Cryptococcus gattii* species complexes. Infect Genet Evol. 2018; 66: 245–255. https://doi.org/10.1016/j.meegid.2018.10.011 PMID: 30342094

66. Mixão V, Gabaldón T. Hybridization and emergence of virulence in opportunistic human yeast pathogens. Yeast. 2018. pp. 5–20. https://doi.org/10.1002/yea.3242 PMID: 28681409

67. Liti G, Barton DBH, Louis EJ. Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. Genetics. 2006; 174: 839–850. https://doi.org/10.1534/genetics.106.062166 PMID: 16951060

68. Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature. 2018; 556: 339–344. https://doi.org/10.1038/s41586-018-0030-5 PMID: 29643504

69. Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, et al. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Res. 2009; 19: 2231–2244. https://doi.org/10.1101/gr.097501.109 PMID: 19745113

70. Chou H, Lo H-J, Chen K-W, Liao M-H, Li S-Y. Multilocus sequence typing of *Candida tropicalis* shows clonal cluster enriched in isolates with resistance or trailing growth of fluconazole. Diagn Microbiol Infect Dis. 2007; 58: 427–433. https://doi.org/10.1016/j.diagmicrobio.2007.03.014 PMID: 17509791

71. Desnos-Ollivier M, Bretagne S, Bernède C, Robert V, Raoux D, Chachaty E, et al. Clonal population of flucytosine-resistant *Candida tropicalis* from blood cultures, Paris, France. Emerg Infect Dis. 2008; 14: 557–565. https://doi.org/10.3201/eid1404.071083 PMID: 18394272

72. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. Nature. 2009; 458: 337–341. https://doi.org/10.1038/nature07743 PMID: 19212322

73. Connelly CF, Akey JM. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. Genetics. 2012; 191: 1345–1353. https://doi.org/10.1534/genetics.112.141168 PMID: 22673807

74. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997; 387: 708–713. https://doi.org/10.1038/42711 PMID: 9192896

75. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 2004; 428: 617–624. https://doi.org/10.1038/nature02424 PMID: 15004568

76. Colón M, Hernández F, López K, Quezada H, González J, López G, et al. *Saccharomyces cerevisiae* Bat1 and Bat2 aminotransferases have functionally diverged from the ancestral-like *Kluyveromyces lactis* orthologous enzyme. PLoS One. 2011; 6: e16099. https://doi.org/10.1371/journal.pone.0016099 PMID: 21267457

77. Schoondermark-Stolk SA, Tabernero M, Chapman J, Ter Schure EG, Verrips CT, Verkleij AJ, et al. Bat2p is essential in *Saccharomyces cerevisiae* for fusel alcohol production on the non-fermentable carbon source ethanol. FEMS Yeast Res. 2005; 5: 757–766. https://doi.org/10.1016/j.femsyr.2005.02.005 PMID: 15851104

78. Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. BMC Genomics. 2010; 11: 290. https://doi.org/10.1186/1471-2164-11-290 PMID: 20459735

79. Sylvester K, Wang Q-M, James B, Mendez R, Hulfachor AB, Hittinger CT. Temperature and host preferences drive the diversification of *Saccharomyces* and other yeasts: a survey and the discovery of eight new yeast species. FEMS Yeast Res. 2015; 15: fov002. https://doi.org/10.1093/femsyr/fov002 PMID: 25743785

80. Wagih O, Usaj M, Baryshnikova A, VanderSluis B, Kuzmin E, Costanzo M, et al. SGAtools: one-stop analysis and visualization of array-based genetic interaction screens. Nucleic Acids Res. 2013; 41: W591–6. https://doi.org/10.1093/nar/gkt400 PMID: 23677617

81. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014; 15: 182. https://doi.org/10.1186/1471-2105-15-182 PMID: 24925680

82. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017; 33: 574–576. https://doi.org/10.1093/bioinformatics/btw663 PMID: 27797770

83. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

84. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 PMID: 23422339

85. Anderson MZ, Porman AM, Wang N, Mancera E, Huang D, Cuomo CA, et al. A multistate toggle switch defines fungal cell fates and Is regulated by synergistic genetic cues. PLoS Genet. 2016; 12: e1006353. https://doi.org/10.1371/journal.pgen.1006353 PMID: 27711197

86. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12: 357–360. https://doi.org/10.1038/nmeth.3317 PMID: 25751142

87. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007; 23: 1026–1028. https://doi.org/10.1093/bioinformatics/btm039 PMID: 17309896

88. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio. GN]. 2013. Available: http://arxiv.org/abs/1303.3997

89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

90. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

91. Seshan VE, Olshen A, Seshan MVE, biocViews Microarray C. Package "DNAcopy." 2013. Available: https://bioconductor.statistik.tu-dortmund.de/packages/3.0/bioc/manuals/DNAcopy/man/DNAcopy.pdf

92. Wickham H, Navarro D, Lin Pederson T. ggplot2: Elegant graphics for data analysis. Springer; 2016. Available at https://ggplot2-book.org/.

93. Lischer HEL, Excoffier L, Heckel G. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of *Microtus* voles. Mol Biol Evol. 2014; 31: 817–831. https://doi.org/10.1093/molbev/mst271 PMID: 24371090

94. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

95. Dray S, Dufour A-B, Others. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. 2007; 22: 1–20.

96. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012; 28: 1919–1920. https://doi.org/10.1093/bioinformatics/bts277 PMID: 22576172

97. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gkz365 PMID: 31106371