Extreme diversification driven by parallel events of massive loss of heterozygosity in the hybrid lineage of Candida albicans

Verónica Mixão,1,2 Ester Saus,1,2 Teun Boekhout,3,4 and Toni Gabaldón1,2,5,*

1Life Sciences Department, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
2Mechanisms of Disease Department, Institute for Research in Biomedicine (IRB), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
3Westerdijk Fungal Biodiversity Institute, 3584 CT Utrecht, The Netherlands
4Institute of Diversity and Ecosystem Dynamics, University of Amsterdam, 1090 GE Amsterdam, The Netherlands
5Institució Catalana de Recerca i Estudis Avancats (ICREA), 08010 Barcelona, Spain

*Corresponding author: Life Sciences Department, Barcelona Supercomputing Centre (BSC-CNS), Jordi Girona, 29, 08034 Barcelona, Spain. toni.gabaldon.bcn@gmail.com

Abstract
Candida albicans is the most commonly reported species causing candidiasis. The taxonomic classification of C. albicans and related lineages is controversial, with Candida africana (syn. C. albicans var. africana) and Candida stellatoidea (syn. C. albicans var. stellatoidea) being considered different species or C. albicans varieties depending on the authors. Moreover, recent genomic analyses have suggested a shared hybrid origin of C. albicans and C. africana, but the potential parental lineages remain unidentified. Although the genomes of C. albicans and C. africana have been extensively studied, the genome of C. stellatoidea has not been sequenced so far. In order to get a better understanding of the evolution of the C. albicans clade, and to assess whether C. stellatoidea could represent one of the unknown C. albicans parental lineages, we sequenced C. stellatoidea type strain (CBS 1905). This genome was compared to that of C. albicans and of the closely related lineage C. africana. Our results show that, similarly to C. africana, C. stellatoidea descends from the same hybrid ancestor as other C. albicans strains and that it has undergone a parallel massive loss of heterozygosity.

Keywords: Candida stellatoidea; Candida albicans; yeast; comparative genomics

Introduction
Candida albicans is one of the most important fungal pathogens, being the Candida species with the highest impact on human health (Turner and Butler 2014). This species comprises a genetically diverse group of isolates, among which we can find Candida africana (syn. C. albicans clade 13 or C. albicans var. africana) and Candida stellatoidea (syn. C. albicans var. stellatoidea) (Meyer 1979; Ropars et al. 2018). These two lineages are genetically similar to C. albicans, but phenotypically different, which has made their taxonomic classification a controversy (Meyer 1979; Kwon-Chung et al. 1988; Jacobsen et al. 2008). Evidence suggests that C. albicans is mostly asexual with clonal reproduction and lacks a normal sexual cycle, possessing instead an alternative parasexual cycle through which different strains can mate (Forche et al. 2008; Alby and Bennett 2010). Therefore, it is difficult to properly establish species boundaries. With the advent of next-generation sequencing technologies, the genomes of multiple isolates of C. albicans and C. africana became available, allowing a better understanding of their evolutionary relationship (Ropars et al. 2018; Bensasson et al. 2019; Mixão and Gabaldón 2020). Indeed, recent comparative genomics analyses have revealed that C. albicans has a hybrid ancestor with unknown parental lineages and that C. africana is a descendent of the same hybridization event, which possibly underwent a massive loss of heterozygosity (LOH) (Mixão and Gabaldón 2020). Nevertheless, due to the absence of genome sequencing data for C. stellatoidea, the role of this lineage in this evolutionary puzzle remains unclear.

Candida stellatoidea was firstly described as Monilia stellatoidea in 1938 (Jones and Martin 1938) and reclassified 1 year later within the genus Candida (Langeron and Guerra 1939). The major phenotypic differences between this lineage and C. albicans are that C. stellatoidea cannot use sucrose as a carbon source and presents a lower growth rate (Kwon-Chung et al. 1988). Moreover, it is less pathogenic in animal models than in C. albicans (Kwon-Chung et al. 1988). In 1979, the analysis of some genetic markers revealed a high similarity between these two lineages and therefore they were proposed to be ranked as the same species (Meyer 1979). Further analyses on C. stellatoidea suggested that it should be divided into two types (type I —comprising the type strain CBS 1905—and type II), which differ in their electrophoretic karyotypes (Kwon-Chung et al. 1988). While type I corresponds to C. stellatoidea sensu stricto, type II corresponds to a sucrose-negative C. albicans which is more virulent than strains of type I (Kwon-Chung et al. 1988). Therefore, depending on the
type of C. stellatoidea used in previous works, the level of similarity with C. albicans could differ. Nevertheless, as chromosomal rearrangements could recover the phenotypes of both C. stellatoidea types, the idea that these lineages correspond to the same species was reinforced (Kwon-Chung et al. 1988; Wickes et al. 1991). Despite all these apparent similarities, more recent studies have shown that C. stellatoidea type I differs substantially from C. albicans at the DNA level, which contributed once again to the uncertainty in the respective taxonomic classification (Pujol et al. 1997; Biswas et al. 2001; Jacobsen et al. 2008).

Although a high number of isolates of the C. albicans clade were sequenced, including C. africana (Ropars et al. 2018; Bensasson et al. 2019), to the best of our knowledge none of them corresponds to C. stellatoidea. To gain a better understanding of the evolutionary relationship between these lineages, and assess the role of C. stellatoidea in the hybridization event at the basis of C. albicans, we sequenced the genome of the C. stellatoidea type strain (CBS 1905) and performed a comparative genomics analysis with C. albicans and C. africana.

Materials and methods

Library preparation and genome sequencing

Genomic DNA extraction was performed using the MasterPure Yeast DNA Purification Kit (Epicentre, USA) following the manufacturer’s instructions. Briefly, CBS 1905 was grown in an orbital shaker overnight (200 rpm, 30°C) in 15 ml of YPD medium. Cells were harvested using 4.5 ml of each culture by centrifugation at 16,000 × g for 10 min to pellet the cellular debris. The supernatant was transferred to a new tube; DNA was precipitated using 100% cold ethanol and centrifuging the samples at 16,000 × g, 30 min, 4°C. The pellet was washed twice with 70% cold ethanol and, once the pellet was dried, the sample was resuspended in 100 μl of TE. The samples were cleaned to remove the remaining RNA using the Genomic DNA Clean & Concentrator Kit (Epicentre) according to the manufacturer’s instructions. Total DNA integrity and quantity of the sample were assessed by means of agarose gel, NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, USA) and Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific).

Whole-genome sequencing was performed at the Genomics Unit from the Centre for Genomic Regulation (CRG) with a HiSeq2500 machine. Libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs, USA) according to the manufacturer’s instructions. All reagents subsequently mentioned are from the NEBNext Ultra DNA Library Prep Kit for Illumina if not specified otherwise. One microgram of gDNA was fragmented by ultrasonic acoustic energy in Covaris to a size of 150–500 bp. After shearing, the ends of the DNA fragments were blunted with the End Prep Enzyme Mix and subsequently amplified by PCR with the NEBNext Q5 Hot Start 2X PCR Master Mix and index primers (12–15 cycles). A further purification step was done using AMPure XP Beads (Agencourt, USA). The final library was analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check size distribution, and it was then quantified by quantitative PCR (qPCR) using the KAPA Library Quantification Kit (KapaBiosystems, USA) prior to amplification with Illumina’s cBot. The library was loaded and sequenced 2 × 125 on Illumina’s HiSeq2500. Base calling was performed using Illumina pipeline software. De-convolution was performed using the CASAVA software (Illumina, USA).

Read mapping and variant calling

Raw sequencing data were inspected with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired-end reads were filtered for quality below 10- or 4-bp sliding windows with average quality per base of 15 and for the presence of adapters with Trimmomatic v0.36 (Bolger et al. 2014). A minimum read size was set to 31 bp. The K-mer Analysis Toolkit v2.4.1 (KAT; Mapleson et al. 2017) was used to get the k-mers frequency of C. stellatoidea sequencing libraries using default parameters (k = 27). This program was also used to inspect the representation of each k-mer in C. albicans SC5314 haplotype A (Muzzey et al. 2013).

The alignment of C. stellatoidea reads to C. albicans SC5314 haplotype A (Muzzey et al. 2013) and posterior variant calling were performed using HaploTypo pipeline v1.0.1 with default parameters (Pegueroles et al. 2020). Briefly, BWA-MEM v0.7.15 (Li 2013) was used to align the reads, and GATK v4.0.2.1 (McKenna et al. 2010) and SAMtools v1.9 (Li et al. 2009) were used for sorting, marking duplicates, creating the index file, and obtaining the mapping statistics. GATK v4.0.2.1 (McKenna et al. 2010) was also used to call variants with HaploTypoCaller, requiring a minimum depth of coverage of 20 reads. The mapping results were visually inspected with IGV version 2.4.14 (Thorvaldsdóttir et al. 2013). In order to determine the number of SNPs/kb, only positions with 20 or more reads were considered for the genome size, and we were determined with bedtools genomcov v2.25.0 (Quinlan and Hall 2010).

Heterozygous and homozygous blocks definition

To define heterozygous and LOH blocks, and considering that the use of a fixed window size cannot properly identify region boundaries and may average their levels of heterozygosity (Pryszcz et al. 2014; Mixão and Gabaldón 2020), we applied a procedure previously used to study the C. albicans clade and other Candida hybrids (Pryszcz et al. 2014; Mixão et al. 2019; Mixão and Gabaldón 2020), and which is an adaptation of the procedure validated by Pryszcz et al. (2015). Briefly, bedtools merge v2.25.0 (Quinlan and Hall 2010) with a distance of 100 bp was used to define heterozygous regions, and by opposite, LOH blocks would be all non-heterozygous regions in the genome. The minimum LOH and heterozygous blocks size was established at 100 bp. The current sequence divergence between the haplotypes in heterozygous regions was calculated by dividing the number of heterozygous positions by the total size of heterozygous blocks.

Comparison of SNPs between C. albicans and C. stellatoidea

To assess whether C. stellatoidea has the same hybrid ancestor as C. albicans, we used the same three groups of C. albicans strains as used in a previous work (group 1 = CECC4942 [clade 1], SC5314 [clade 1], CECC4497 [clade 4], and CECC5026 [clade 9]; group 2 = CECC1426 [clade 11], CECC3544 [clade 1], CECC3716 [clade 4], and CECC3533 [clade 9]; and group 3 = CECC3601 [clade 11], CECC3560 [clade 1], CEC2021 [clade 2], and CECC3557 [clade 3]) (Ropars et al.
Genotype was found in we counted how many ancestral alleles were shared (the same ancestral heterozygous positions (i.e. C. albicans ancestor) previously determined for C. africana in the same position in C. stellatoidea). Gabaldón 2020) and compared their alleles to the allele present responded to one of the haplotypes (haplotype A or B), or were and C. africana sponded to one of the haplotypes (haplotype A or B), or were undefined (none of the previous options was observed). If C. albicans and C. stellatoidea shared a hybrid origin, we would expect a high proportion of positions to be shared, for at least one allele, in the two lineages.

Phylogenetic analysis
For the phylogenetic analysis of C. albicans, C. africana, and C. stellatoidea, we considered the same sample of C. albicans (61) and C. africana (8) strains that was used in the past (Mixão and Gabaldón 2020) and added the type strain of C. stellatoidea. A concatenated alignment of all the positions which are homozygous in all the strains was generated by removing heterozygous positions and indels from the reference genome using bedtools subtract (Quinlan and Hall 2010). A maximum-likelihood tree representative of the final alignment was obtained with RAxML v8.2.8 (Stamatakis 2014), using the GTRCAT model.

In silico reconstruction of a hybrid between C. stellatoidea and C. africana
To discard the possibility that C. stellatoidea and C. africana are the parental lineages of the hybrid ancestor of C. albicans, and the detected mosaic of haplotypes is an artifact generated by chromosomal recombination, we computationally created synthetic hybrid between C. africana CEC4103 and C. stellatoidea CBS 1905. For this, information on homozygous SNPs was used to replace the nucleotide present in C. albicans SC5314 haplotype A for each of the two strains creating two new genomes. Then, genomic reads were simulated for these new genomes using wgsim (https://github.com/lh3/wgsim). Posterior read mapping and variant calling on C. albicans haplotype A (Muzzey et al. 2013) were performed as described in the previous sections. A final comparison between the heterozygous variants of this in silico hybrid and C. albicans strains was used to determine how many C. albicans heterozygous positions could be recovered in the hybrid. As heterozygous positions of the parental lineages could not be used to infer the final genotype of a putative cross, heterozygous positions of CEC4103 and SC5314 were excluded from the analysis. If C. stellatoidea and C. africana were the parental lineages of C. albicans, we would expect the majority of C. albicans ancestral heterozygous positions to be present in the in silico created hybrid genome.

Data availability
Sequencing data generated by this project can be found under the BioProject PRJNA650282 in NCBI database. The remaining sequencing libraries used in this work were generated by Ropars et al. (2018) and can be found under the BioProject PRJNA432584.

Supplementary material is available at figshare DOI: https://doi.org/10.25386/genetics.13207547.

Results and discussion
The genomic patterns of C. stellatoidea reveal its hybrid nature
To get a better understanding of C. albicans clade evolution, we sequenced the genome of the type strain of C. stellatoidea (CBS 1905). Given the expected genetic similarity between this strain and C. albicans haplotypes (Meyer 1979; Jacobsen et al. 2008), the analysis of the respective sequencing data was performed through k-mer frequency analysis, read mapping and posterior variant calling on C. albicans SC5314 haplotype A reference genome (Muzzey et al. 2013). The first inspection of C. stellatoidea genome based on k-mer frequency corroborated the genetic similarity with C. albicans. This analysis revealed the presence of two peaks of coverage, showing that the respective levels of heterozygosity are sufficiently high to detect the first peak (Supplementary Figure S1). These results are compatible with the scenario previously described for C. albicans and C. africana (Mixão and Gabaldón 2020). Indeed, C. stellatoidea has 10.6 variants/kb when compared to C. albicans reference genome haplotype A, from which 2.9 variants/kb correspond to heterozygous positions (Table 1). As was previously observed for C. albicans and C. africana, the heterozygous variants of C. stellatoidea are not spread around the genome, but rather form stretches of blocks of heterozygosity separated by what appears to be blocks of LOH (Supplementary Figure S2). Interestingly, similar to what was observed for C. africana (Mixão and Gabaldón 2020), LOH blocks in C. stellatoidea cover ~95% of the genome, and the haplotype divergence in heterozygous regions is 3.55% with a single density peak (Table 1 and Supplementary Figure S3). As discussed in previous works, the accumulation of mutations in asexual organisms could lead to the occurrence of divergent homolog chromosomes in the same cell (Halkett et al. 2005). Nevertheless, as in C. albicans LOH is expected to simultaneously occur (Ene et al. 2018), such a scenario is expected to result in heterozygous blocks with different haplotype divergence (LOH would erase mutations with time) (Mixão and Gabaldón 2020). This expectation contrasts with our observation of a single density peak. Instead, the observed genomic patterns are similar to what was previously observed in the increasing number of Candida hybrids (Przyucz et al. 2015; Schroeder et al. 2016; Mixão et al. 2019; Mixão and Gabaldón 2020), suggesting that C. stellatoidea also has a hybrid nature but, as C. africana (Ropars et al. 2018; Mixão and Gabaldón 2020), has undergone a massive LOH.

Candida stellatoidea shares a hybrid ancestor with C. albicans and C. africana
To assess whether C. stellatoidea descends from the previously described hybrid ancestor of C. albicans and C. africana (Mixão and Gabaldón 2020), we compared the genomic variability of these three lineages. To this end, we took advantage of the previously applied methodology for comparison of C. albicans and C. africana, in which ancestral heterozygous positions of C. albicans (i.e. those present before the current clades diverged) were identified in the overlap of heterozygous blocks of C. albicans strains (701 heterozygous blocks) and the presence/absence of the respective alleles in the other lineage was quantified (see Materials and methods) (Mixão and Gabaldón 2020). The comparison of C. albicans with C. stellatoidea revealed that, similarly to what was determined in the past for C. africana (Mixão and Gabaldón 2020), 99% of the 2423 identified C. albicans ancestral heterozygous positions have at least one of the alleles present in C. stellatoidea. Furthermore, the allele present in this lineage corresponds 54.5% of the times
to haplotype A and 45.5% of the times to haplotype B. These results indicate that, similarly to C. africana, C. stellatoidea homoyzous regions are a mosaic of C. albicans hybrid ancestor haplotypes, and therefore C. stellatoidea descends from the same hybridization event as C. albicans and C. africana. Of note, depending on the C. albicans strain used for comparison, between 183 and 194 heterozygous blocks had both haplotypes represented in C. stellatoidea, which is compatible with the occurrence of recent chromosome recombination in these lineages (Wang et al. 2018).

Considering the observed levels of similarity between the genomic patterns of C. stellatoidea and C. africana, we hypothesized that they could correspond to the same lineage. To test this hypothesis, we assessed the evolutionary relationship between C. albicans, C. africana, and C. stellatoidea. To this end, we reconstructed a maximum likelihood phylogeny of their homozygous positions including 61 C. albicans strains, eight C. africana, and the type strain of C. stellatoidea (see Materials and methods). As Figure 1 shows, these three lineages are genetically different, forming three different clades, with C. africana and C. stellatoidea being equally distant to C. albicans. These are remarkable findings, as, together with the absence of a clear overlap in the heterozygosity patterns of the two lineages (Supplementary Figure S2), they support the occurrence of two independent massive losses of heterozygosity after a single hybridization event. These results contribute to the understanding of the different phenotypes observed in these three lineages. The phenotypic relevance of massive LOH in C. stellatoidea and C. africana is not explored in this study. However, it is important to note that, from these three lineages, C. albicans is the most virulent one (Kwon-Chung et al. 1988, 1990; Borman et al. 2013), and also the most heterozygous one. This observation is in accordance with previous theories that suggest that high levels of heterozygosity may provide a higher plasticity to survive in hostile environments as the human body (Mixão and Gabaldón 2018). Future studies should address this hypothesis and try to assess whether indeed heterozygosity is relevant for the pathogenicity of some yeasts.

**Table 1** Summary of the genomic patterns of C. stellatoidea type strain and the average numbers previously reported for 61 C. albicans and eight C. africana strains (Mixão and Gabaldón 2020), with indication of lineage name, percentage of mapped reads, all variants per kilobase (kb), heterozygous variants per kb, LOH level (percentage of the genome classified as LOH), and estimated haplotype divergence.

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Mapped reads (%)</th>
<th>Variants/kb</th>
<th>Heterozygous variants/kb</th>
<th>LOH level (&gt;100 bp) (%)</th>
<th>Haplotype divergence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. stellatoidea</td>
<td>98.84</td>
<td>10.64</td>
<td>2.90</td>
<td>94.57</td>
<td>3.55</td>
</tr>
<tr>
<td>C. albicans</td>
<td>98.88</td>
<td>8.60</td>
<td>6.70</td>
<td>85.74</td>
<td>3.52</td>
</tr>
<tr>
<td>C. africana</td>
<td>99.03</td>
<td>10.51</td>
<td>2.64</td>
<td>95.16</td>
<td>3.71</td>
</tr>
</tbody>
</table>

Our analysis revealed that between 60% and 70% of the ancestral heterozygous positions in C. albicans have the same allele represented in C. stellatoidea and C. africana. Moreover, an in silico reconstruction of a hybrid of C. stellatoidea and C. africana could only recover between 21% and 22% of the ancestral heterozygous variants of each C. albicans strain. These results clearly show that C. stellatoidea and C. africana are not the parental lineages of C. albicans, solving any of the previous doubts. Altogether, these results support a scenario in which all the three lineages share a hybrid ancestor, and both C. africana and C. stellatoidea have massively and independently lost their heterozygosity. The eventual identification of the parental lineages involved in this event would undoubtedly provide additional support for this hypothesis.

**Concluding remarks**

_Candida albicans_ is a yeast with medical importance and is used as a model for the study of the reproductive system in the CUG-Ser clade (Alby and Bennett 2010; Turner and Butler 2014). Recent analyses of multiple isolates of this species revealed genomic patterns which are shared between different clades, and an ancient hybridization event was hypothesized as the source of such patterns (Mixão and Gabaldón 2020). Although alternative scenarios, which were discussed by the same authors, could partially explain these observations, the increasing number of unrevealed _Candida_ hybrids and the similarity of their genomic features to those observed in _C. albicans_ additionally support the ancient hybridization event scenario (Prysztz et al. 2015; Schröder et al. 2016; Mixão et al. 2019). Considering this, our results show that C. stellatoidea results from the same hybridization event as _C. albicans_ and _C. africana_ (Mixão and Gabaldón 2020), but currently they represent three distinct lineages (Figure 1). These findings represent an important step toward understanding the evolution of the _C. albicans_ clade. Importantly, it is now clear that _C. albicans_, _C. africana_, and _C. stellatoidea_ represent genetically divergent clades but share a hybrid ancestor.

Despite these new findings, the taxonomic classification of the clade is still unclear due to the difficulties in applying the biological species concept in these yeasts. The fact that _C. albicans_ is mostly asexually reproducing with no evidence for the occurrence of meiosis (Forche et al. 2008; Butler et al. 2009) challenges the identification of species boundaries in the clade. In this context, next-generation sequencing was crucial to reveal a non-vertical evolution event in this clade. The evidence for the occurrence of hybridization in _C. albicans_ (Mixão and Gabaldón 2020) complicates even further the delimitation of species boundaries and raises the question of the reproductive mechanisms underlying such an event. It is important to note that this problem is extended to other _Candida_ species, where the advances of next-generation sequencing technologies are also uncovering several hybrid lineages (Schröder et al. 2016; Mixão et al. 2019). Therefore, there is an urgent need for proper guidelines that should be followed.
to the application of the species concept in these organisms. This clarification may have important repercussions in epidemiological studies, as different lineages have different phenotypes but may still be identified as the same species. For instance, in the particular case of C. albicans clade, it would be relevant to distinguish infections caused by each of the lineages and assess the correlation between the levels of heterozygosity and virulence. This could help us to understand not only the evolution of this clade but also the factors influencing the evolution of pathogens in general.

Acknowledgments

The authors thank all Gabaldón lab members, especially Marina Marce-Houben, for helpful discussions.

Funding

This work was funded by the European Union’s Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie Grant Agreement No. H2020-MSCA-ITN-2014-642095. T.G. group also acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for grant PGC2018-099921-B-I00 co-founded by European Regional Development Fund (ERDF); from the CERCA Programme/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857; and grants from the European Union’s Horizon 2020 Research and Innovation Program under the grant agreements ERC-2016-724173 and MSCA-747607. T.G. also receives support from an INB Grant (PT17/0009/0023—ISCIII-SGEO/ERDF).

Conflicts of interest: None declared.

Literature cited


Wang JM, Bennett RJ, Anderson MZ. 2018. The genome of the human pathogen is shaped by mutation and cryptic sexual recombination. mBio. 9:e01205-18.


Communicating editor: A. Mitchell