


Resource article: genomes explored

Genome analysis of *Candida subhashii* reveals its hybrid nature and dual mitochondrial genome conformations

Verónica Mixão^{1,2†}, Eva Hegedúsová^{3‡}, Ester Saus^{1,2}, Leszek P. Pryszcz⁴, Andrea Cillingová³, Jozef Nosek³, and Toni Gabaldón ^{1,2,5*}

¹Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona, 29, 08034 Barcelona, Spain, ,
²Mechanisms of Disease Department, Institute for Research in Biomedicine (IRB), Barcelona, Spain, ³Faculty of Natural Sciences, Department of Biochemistry, Comenius University in Bratislava, Ilkovičova 6, 842 15 Bratislava, Slovak Republic, ⁴Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain, and ⁵ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*To whom correspondence should be addressed. Tel: +34 93 40 21077. Email: toni.gabaldon.bcn@gmail.com

[†]Present address: Bioinformatics Unit, Infectious Diseases Department, National Institute of Health Dr. Ricardo Jorge, Av. Padre Cruz, 1649-016 Lisbon, Portugal.

[‡]Present address: Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 1160/31, 370 05 České Budějovice, Czech Republic.

Received 15 March 2021; Editorial decision 12 June 2021; Accepted 14 June 2021

Abstract

Candida subhashii belongs to the CUG-Ser clade, a group of phylogenetically closely related yeast species that includes some human opportunistic pathogens, such as *Candida albicans*. Despite being present in the environment, *C. subhashii* was initially described as the causative agent of a case of peritonitis. Considering the relevance of whole-genome sequencing and analysis for our understanding of genome evolution and pathogenicity, we sequenced, assembled and annotated the genome of *C. subhashii* type strain. Our results show that *C. subhashii* presents a highly heterozygous genome and other signatures that point to a hybrid ancestry. The presence of functional pathways for assimilation of hydroxyaromatic compounds goes in line with the affiliation of this yeast with soil microbial communities involved in lignin decomposition. Furthermore, we observed that different clones of this strain may present circular or linear mitochondrial DNA. Re-sequencing and comparison of strains with differential mitochondrial genome topology revealed five candidate genes potentially associated with this conformational change: *MSK1*, *SSZ1*, *ALG5*, *MRPL9* and *OYE32*.

Key words: *Candida subhashii*, genome assembly, mitochondria, hybrid, metabolism of hydroxyaromatic compounds

1. Introduction

Candida species represent a non-monophyletic group of yeasts which comprises important pathogens for human health. *Candida* infections (i.e. candidiasis) are mostly caused by *Candida albicans*,

Candida glabrata, *Candida parapsilosis* and *Candida tropicalis*, which combined account for more than 90% of the cases worldwide.¹ However, over the last years the epidemiology of this disease has been shifting, with non-*albicans* *Candida* species increasing their

relative incidence and with the emergence of new rare *Candida* species that are responsible for several cases of infection or even outbreaks.² The mechanisms involved in the emergence of new pathogens are still unknown, but hybridization has been suggested as a possible evolutionary path for the appearance of novel species with ability to cause infection in humans.^{3–5} Indeed, a majority of hybrid clinical isolates, or a hybrid origin of the species have been described in important pathogenic lineages such as *Candida orthopsilosis*, *Candida metapsilosis*, *Candida inconspicua* and *C. albicans*.^{4,6–10} The comparatively low abundance or absence of the parental species among clinical isolates for many of these hybrids has led to the hypothesis that this pattern results from a higher ability of the hybrids to colonize and infect humans.^{3,4} Of note, hybridization in these lineages only became evident after next-generation sequencing data analysis, showcasing the importance of whole-genome sequencing for our understanding of the evolution of human pathogens.

Candida subhashii was first described in 2009 in Canada as a novel species causing peritonitis and reported as a non-fermentative yeast member of the CUG-Ser clade.¹¹ Later on, this species was isolated from agricultural soil in Switzerland¹² and Japan,¹³ as well as from peatlands in Poland.¹⁴ Although its first description was as a fungal pathogen,¹¹ it was suggested that the soil can possibly be its natural habitat, particularly considering that it is a highly competitive yeast, behaving as antagonist for several filamentous fungi.¹² *Candida subhashii* has been reported to have a linear mitochondrial genome and it was described as the yeast species with the highest GC content on its mitochondrial DNA (mtDNA) (52.7%).¹⁵ Moreover, mitochondrial sequence conservation between Canadian and European *C. subhashii* isolates was reported, suggesting a recent and fast global spreading.¹²

Considering the relevance of genomic analysis for our understanding of the evolution of human pathogens, we decided to sequence and analyze the genome of *C. subhashii* type strain (CBS10753). Our analysis reveals genomic signatures characteristic of yeast hybrids: high levels of heterozygosity organized in non-homogeneously distributed blocks interspersed by regions with low levels of heterozygosity. Thus, *C. subhashii* represents another case in the growing list of hybrid yeast pathogens. We also discovered that this species can harbor both linear and circular mitochondrial genome (see Section 3 for more details) including in two clonal descendants of the same strain, and we used a comparative genomics analysis to identify possible genomic alterations related to the different mitochondrial genome conformation. Moreover, we demonstrate that *C. subhashii* has functional gentisate pathway and two branches of the 3-oxoadipate pathway indicating that this yeast participates in degradation of hydroxyaromatic compounds resulting from lignin decomposition in soil.

2. Materials and methods

2.1. Strains

The *C. subhashii* type strain (FR-392-06, CBS10753) and derived clones (see below) were cultivated in yeast extract–peptone–dextrose (YPD) medium [1% (Wt/Vol) yeast extract, 1% (Wt/Vol) peptone, 2% (Wt/Vol) glucose] at 28°C. The original strain has been subcloned as follows. A culture grown overnight in a YPD medium was diluted in water and plated on a YPD medium containing 2% (Wt/Vol) agar. Several clones were obtained from single colonies growing on the plate. The clones dubbed SUB1, SUB3 and SUB10 were used in further experiments.

The assimilation tests were done in synthetic media [0.17% (Wt/Vol) yeast nitrogen base w/o amino acids and ammonium sulphate (Difco), 0.5% (Wt/Vol) ammonium sulphate, 2% (Wt/Vol) agar] supplemented with an appropriate carbon source [i.e. 2% (Wt/Vol) glucose, 2% (Wt/Vol) glycerol, 2% (Vol/Vol) lactate or 10 mM hydroxyaromatic compound dissolved in dimethyl sulfoxide].

2.2. Mitochondrial genome conformation

Mitochondrial genome conformation has been assayed by pulsed-field gel electrophoresis (PFGE), restriction enzyme mapping of isolated mtDNA and Sanger sequencing essentially as described in Fricova *et al.*¹⁵ PFGE separations were carried out in a Pulsaphor apparatus (LKB) using a contour-clamped homogeneous electric field (CHEF) configuration. DNA samples prepared in agarose plugs were separated in 45 mM Tris-borate, 1 mM EDTA using following conditions/settings: (i) 0.8% (Wt/Vol) agarose gel, pulses 60–600 s (linear interpolation), 100 V, 72 h, 9°C (separation of chromosomal DNA) and (ii) 1.0% (Wt/Vol) agarose gel, pulses 5–50 s (linear interpolation), 150 V, 24 h, 9°C (separation of mtDNA). For Southern blot hybridization, the PFGE separated DNA samples were transferred onto a Hybond N+ membrane (Amersham) using a VacuGene XL Vacuum Blotting System (GE Healthcare) and hybridized overnight at 50°C with 5' [³²P] end-labeled oligonucleotide (5'-TCAGTAGGATCAGTCCCTCTGATAGTCAT-3') derived from the *C. subhashii* mitochondrial gene *nad3* in a buffer containing 5× SSC (750 mM sodium chloride, 75 mM trisodium citrate, pH 7.0), 5× Denhardt's solution [0.1% (Wt/Vol) Ficoll 400, 0.1% (Wt/Vol) polyvinylpyrrolidone, 0.1% (Wt/Vol) bovine serum albumin fraction V] and 0.5% (Wt/Vol) sodium dodecyl sulphate (SDS). After hybridization, the membrane was washed in 2× SSC, 0.1% (Wt/Vol) SDS at room temperature for 5 min, followed by two washes at 45°C for 15 min each. The membrane was exposed to a storage phosphor screen (Kodak) and the signal was detected using a Personal Molecular Imager (Bio-Rad). Restriction enzyme digestion of isolated mtDNAs was performed by *PvuI* endonuclease (New England Biolabs) according to the manufacturer's instructions. Sanger sequencing of circular mitochondrial genome of the SUB1 clone was performed by primer-walking approach on isolated mtDNA template as described by Valach *et al.*¹⁶

2.3. Genomic DNA sequencing

A modified protocol from the MasterPure™ Yeast DNA Purification Kit was used to extract the DNA. In brief, samples were grown overnight in liquid YPD at 30°C. Cells were pelleted and lysed with RNase treatment at 65°C for 15 min. After 5 min of cooling down on ice, samples were purified by the kit reagent by mixing, centrifugation and removal of the debris as described in the kit protocol. Further, samples were left at –20°C with absolute ethanol for at least 2 h after which the DNA was precipitated for 30 min at 4°C. The pellet was washed in 70% (Vol/Vol) ethanol and left to dry. TE buffer (10 mM Tris.Cl, 1 mM EDTA, pH 7.5) was used to resuspend the DNA. Genomic DNA Clean and Concentrator kit was used for the final purification.

Whole-genome sequencing was performed at the Genomics Unit from CRG. Libraries were prepared using the NEBNext® DNA Library Prep Reagent Set for Illumina® kit (New England BioLabs) according to the manufacturer's protocol. Briefly, 1 µg of genomic DNA was fragmented by nebulization in Covaris to ~600 bp and subjected to end repair, addition of 'A' bases to 3' ends and ligation of Truseq adapters. All purification steps were performed using

QIAquick PCR purification columns (Qiagen). Library size selection was done with 2% low-range agarose gels. Fragments with average insert size of 700 bp were cut from the gel, DNA was extracted using QIAquick Gel extraction kit (Qiagen) and eluted in 30 µl EB. Ten microlitres of adapter-ligated size-selected DNA were used for library amplification by PCR using the Truseq Illumina primers. Final libraries were analyzed using Agilent DNA 1000 chip to estimate the quantity and check size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, ref. KK4835) prior to amplification with Illumina's cBot. Libraries were loaded at a concentration of 2 pM onto the flow cell and were sequenced 2 × 125 bp on Illumina's HiSeq 2500.

For *C. subhashii* SUB3 clone, a mate-pair library was also sequenced. For that, DNA was fragmented to sizes between 1 and 20 kb using a transposase that binds biotinylated adapters at the breaking point. Strand displacement was performed to 'repair' the nicks left by the transposase. Fragment sizes of 3–6 kb were then selected on a 0.8% agarose gel and were then circularized. Non-circularized DNA was removed by digestion. The circular DNA was then mechanically sheared to fragments of 100 bp to 1 kb approximately and the fragments containing the biotinylated ends were pulled down using magnetic streptavidin beads and submitted to a standard library preparation. A final size selection on 2% agarose gel was done and fragments of 400–700 bp were selected for the final library. Final libraries were analyzed using Agilent High Sensitivity chip to estimate the quantity and check size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (ref. KK4835, KapaBiosystems) prior to amplification with Illumina's cBot. Libraries were sequenced 2 × 125 bp on Illumina's HiSeq 2500.

2.4. Genome assembly

Next-Generation Sequencing data were inspected with FastQC v0.11.5 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, 31 May 2021, date last accessed). Paired-end reads were filtered for the presence of adapters and for quality below 10 or 4 bp sliding-windows with average base quality of 15, setting a minimum read size of 31 bp with Trimmomatic v0.36.¹⁷ Mate-pair reads were filtered with NxTrim v0.4.1-53c2193.¹⁸ After filtration, only reads clearly identified as paired end or mate pair by the respective programs were used for further work. The K-mer Analysis Toolkit (KAT)¹⁹ was used to count *k*-mer frequency and estimate the expected genome size using default parameters (*k* = 27). SOAPdenovo v2.04²⁰ and SPAdes v3.9 in both SPAdes and dipSPAdes modes^{21,22} were used separately to perform the genome assembly of the paired-end and the mate-pair reads. Afterwards, redundant contigs were removed from each assembly with Redundans v0.13c.²³ The quality of the different assemblies was inspected with Quast v4.5.²⁴ Genome annotation was performed with Augustus v3.5 using *C. albicans* as model organism.²⁵ The assembly completeness was estimated with KAT¹⁹ and BUSCO v3²⁶ (Ascomycota database). The best assembly was chosen based on the assembly completeness, genome size, N50 and number of scaffolds. Mitochondria-related genes were identified with TargetP v1.²⁷ Phylome reconstruction was performed as previously described²⁸ and deposited in the PhylomeDB database (<http://beta.phylomedb.org>, 31 May 2021, date last accessed).²⁹ Phylome results were used to infer orthology and paralogy relations and determine the associated GO terms. Enrichment analysis was performed with FatiGO.³⁰

2.5. Read mapping and variant calling

Read mapping and variant calling were performed with HaploTypo pipeline v1.0.1³¹ using default parameters. Briefly, BWA-MEM³² v0.7.15 was used for read mapping. GATK v4.0.2.1³³ was used to sort the reads by coordinate and to mark duplicates. FreeBayes³⁴ v1.1.0-50-g61527c5 was used for variant calling considering a minimum depth of coverage of 30 reads. Vcfilter (<https://github.com/vcflib/vcflib#vcffilter>, 31 May 2021, date last accessed) set with 'TYPE = snp & QUAL > 1 & QUAL/AO > 10 & SAF > 0 & SAR > 0 & RPR > 1 & RPL > 1' was used to filter variants. Ploidy estimation was performed with nQuire histotest.³⁵ To confirm the mitochondrial conformation of *C. subhashii* clones, SUB1, SUB3 and SUB10 sequencing reads were aligned in the linear mitochondrial genome sequence available in NCBI database (accession number: NC_014337.1).

2.6. Heterozygous and loss of heterozygosity block definition

To determine the presence of heterozygous and loss of heterozygosity (LOH) blocks, heterozygous and homozygous variants were separated. Then, the procedure applied and validated by Prysacz *et al.*⁴ was used. Briefly, bedtools³⁶ merge v2.25.0 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. Moreover, the minimum heterozygous and LOH size was established at 100 bp.

2.7. Data availability

Data generated by this project can be found under the BioProject PRJNA691204 in NCBI database and under the phylome ID 777 in PhylomeDB (<http://beta.phylomedb.org>, 31 May 2021, date last accessed).²⁹

3. Results and discussion

3.1. Genome assembly of *C. subhashii* type strain and phylome reconstruction

To obtain a better understanding of *C. subhashii* genome evolution, we analyzed the type strain of this species (CBS10753), which was isolated in 2006 from a patient with peritonitis in Canada.¹¹ As mentioned above, *C. subhashii* was previously reported to have a linear dsDNA mitochondrial genome with unusually high GC content (52.7%) and proteins covalently bound to the 5' ends of linear mtDNA molecules.¹⁵ Nevertheless, the PFGE analysis of several clones derived from the original culture of CBS10753 revealed that clones harboring circular mitochondrial genome also exist in this particular strain (Supplementary Fig. S1). Restriction enzyme mapping and Sanger sequencing demonstrated that the circular mtDNA represents a mutant presumably formed by end-to-end fusion of linear molecules which was accompanied by deletion of a substantial part of terminal sequences present in the linear mtDNA¹⁵ (see Section 2 for more details, and Supplementary Fig. S1). Of note, this alteration of mitochondrial genome topology in *C. subhashii* did not impair assimilation of respiratory substrates such as glycerol and lactate, but has possibly contributed to a slightly weaker growth of the circular mtDNA clones (Supplementary Fig. S2). Considering this dual mitochondrial genome topology, we initially selected two sub-clones derived from the type strain for whole-genome sequencing, which differ in their mitochondrial genome conformation (SUB1 and

SUB3 with circular and linear mtDNA, respectively; Supplementary Fig. S1). Both samples were sequenced with Illumina paired-end sequencing strategy, and SUB3 was additionally sequenced with Illumina mate-pair sequencing strategy (see Section 2 for more details). During our analyses, we identified read coverage issues in the paired-end sequencing data of SUB3. As this subclone was no longer available in the laboratory, we sequenced a third clone (SUB10) that had been isolated in parallel with SUB3 by plating from the same original culture, and that also presented a linear mitochondrial genome (Supplementary Fig. S3). *K*-mer frequency analysis of the genomic reads of CBS10753 revealed the presence of two peaks of coverage (Fig. 1A), one with half coverage of the other, corresponding to heterozygous and homozygous regions, respectively. This suggests that *C. subhashii* has a highly heterozygous genome. Indeed, nQuire histotest³⁵ estimates that this is a diploid species ($r^2 = 0.92$).

After applying different assembly strategies (see Section 2 for more details), the best assembly was chosen based on different quality parameters such as genome completeness, percentage of mapped reads, N50 and fragmentation. The best genome assembly obtained for *C. subhashii* was the one generated with SPAdes²¹ and comprises 15.4 Mb in 333 contigs and an N50 of 108,455 bp. This assembly had high completeness as judged by high sequencing reads mapping rates (98.6 and 98.8% for SUB1 and SUB10, respectively; Table 1).

As expected for an unphased hybrid genome, the assembly comprised only 58.81% of the *k*-mers present in the sequencing libraries (Table 1). Although at first this number may seem low, the comparison of the *k*-mers present in *C. subhashii* sequencing reads with those present in the genome assembly reveals that all the *k*-mers of the homozygous regions (second peak) are represented in the assembly, and that the missing portion of the genome actually corresponds to ~50% of the first peak, thus 50% of the *k*-mers of the heterozygous regions (Fig. 1A). Therefore, this value indicates the expected consequence of a successful reduction of the heterozygous regions, with only one of the haplotypes being represented in the assembly. Finally, genome annotation predicted 6,178 protein-coding genes, corresponding to an estimated proteome completeness of 99.3% (see Table 1 and Section 2 for more details).

We next reconstructed the phylome, i.e. the complete collection of gene phylogenies,³⁷ of *C. subhashii* including other 26 species (Supplementary File S1), and assessed orthology and paralogy relationships (see Section 2 for more details). Our results revealed that *C. subhashii* has 501 orphan genes, and that genes specifically duplicated in this species are enriched in transmembrane transport activity (Supplementary File S2). This result is similar to what was previously obtained for other *Candida* pathogens, such as *C. inconspicua*, *Diutina (Candida) rugosa* and *Trichomonascus (Candida)*

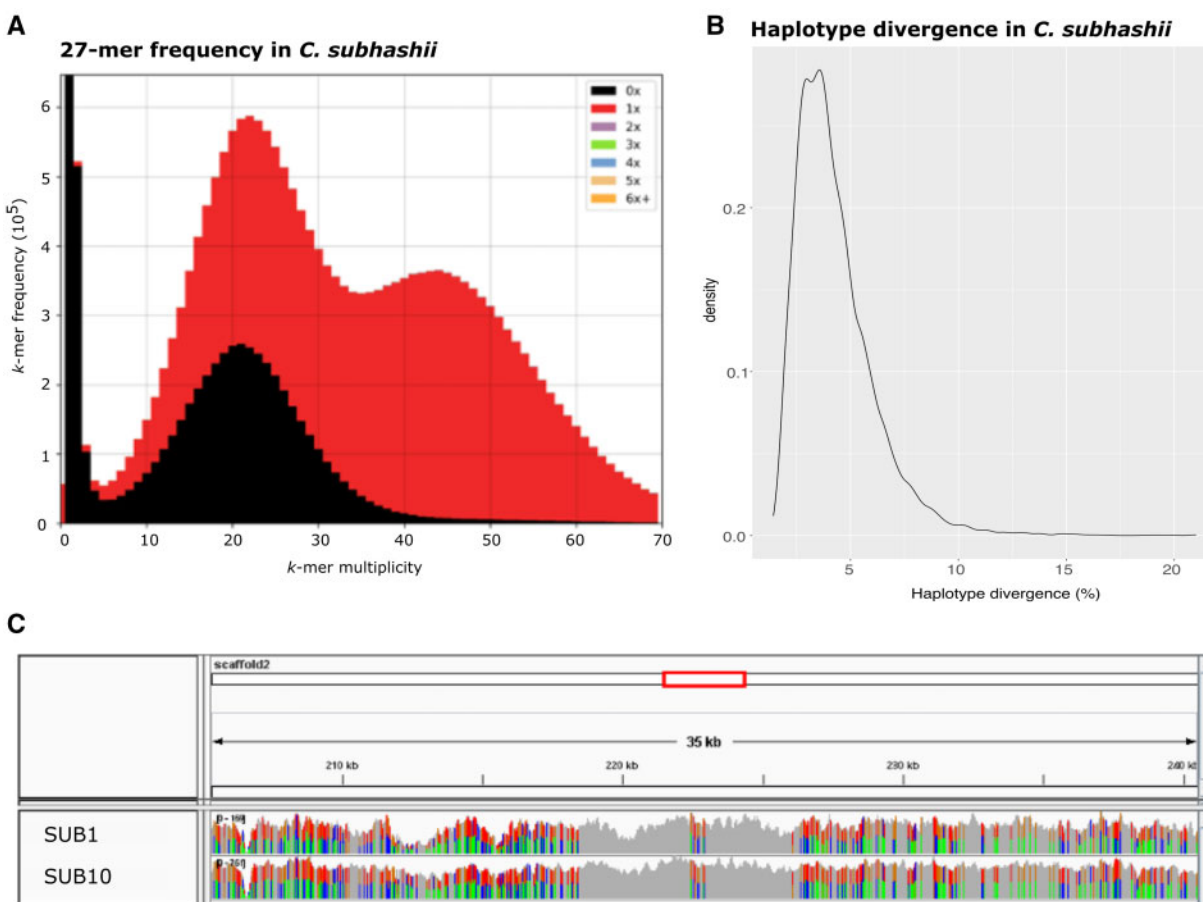


Figure 1. Genomic patterns of *C. subhashii* support its hybrid ancestry. (A) 27-mer frequency in raw sequencing data of CBS10753. The x-axis represents the *k*-mer coverage, and the y-axis the *k*-mer frequency. The density of 27-mers present in the genome assembly is represented in red, and the absence is represented in black.¹⁹ For diploid genomes, two peaks of coverage are expected, the first one (heterozygous) with half coverage of the other (homozygous). When only one of the haplotypes of the heterozygous regions is represented in the assembly, only half of the *k*-mers of the first peak are present. (B) Heterozygous SNP density in heterozygous blocks. (C) IGV screenshot of the genomic patterns of *C. subhashii* in scaffold2.

Table 1. Summary of the genomic features of *C. subhashii* type strain, with indication of estimated and obtained genome size, number of contigs, assembly N50, percentage of GC content, assembly and proteome completeness, number of predicted proteins, percentage of mapped reads, number of SNPs per kilobase (kb), number of LOH blocks, average haplotype divergence in heterozygous blocks, and percentage of loss of heterozygosity (LOH)

<i>Candida subhashii</i> genome assembly with SPAdes	
Estimated genome size (Mb) ^a	15.61
Assembly size (Mb)	15.40
Contigs	333
Contigs > 50 kb	252
N50	108,455
GC	34.56%
Assembly completeness (KAT)	58.81%
Proteome completeness (BUSCO)	99.30%
Number of proteins	6,178
Mapped reads	SUB1—98.63% SUB10—98.82%
SNPs/kb	SUB1—14.26 (14.23 heterozygous) SUB10—14.40 (14.37 heterozygous)
LOH blocks	SUB1—22,705 SUB10—22,679
Divergence (%)	SUB1—4.32% SUB10—4.34%
LOH (%)	SUB1—70.02% SUB10—69.73%

^aEstimated with KAT.¹⁹

ciferrii,^{7,28} and reinforces the previously proposed hypothesis that cell wall composition may play a role in the adoption of an opportunistic pathogenic behaviour by *Candida* species.³⁸ As previous reports suggest that the human body is not the main habitat of *C. subhashii*,¹² we hypothesize that the duplication of cell wall proteins must have been advantageous in non-human environments while perhaps secondarily facilitating opportunistic colonization of humans. Similar exaptation processes (i.e. traits that provide an advantage in niches different from those in which they were originally selected) may be common in other opportunistic pathogens.

3.2. Assimilation of hydroxyaromatic compounds in *C. subhashii*

As soil appears as a natural niche for *C. subhashii*, we investigated whether this yeast is able to contribute to decomposition of lignin by assimilation of compounds such as hydroxyderivatives of benzene and benzoic acid. An assimilation test demonstrated that *C. subhashii* utilizes a range of hydroxybenzenes and hydroxybenzoates as a sole carbon source (Fig. 2A). In other yeasts from the CUG-Ser clade such as *C. parapsilosis* and *C. albicans*, these compounds are metabolized via the 3-oxoadipate and gentisate pathways.^{39,40} The hydroxyhydroquinone (HHQ) branch of the 3-oxoadipate pathway is common to both *C. albicans* and *C. parapsilosis*, which is reflected by their ability to grow with hydroquinone and resorcinol as sole carbon sources. Moreover, as *C. parapsilosis* has longer version of this pathway it also assimilates various hydroxybenzoates (i.e. 4-hydroxybenzoate, 2,4-dihydroxybenzoate and 3,4-dihydroxybenzoate). In contrast, the catechol branch of this pathway is absent in *C. parapsilosis*, and for this reason this species is not able to grow

when phenol or catechol are the exclusive carbon sources.⁴⁰ Nevertheless, *C. parapsilosis* utilizes the gentisate pathway for catabolism of 3-hydroxybenzoate and 2,5-dihydroxybenzoate, which does not occur in *C. albicans*.^{39,40} In the particular case of *C. subhashii*, the assimilation assays demonstrated that, similarly to *C. parapsilosis*, this species is able to utilize 3-hydroxybenzoate and 2,5-dihydroxybenzoate, suggesting the existence of the gentisate pathway, which is confirmed by the presence of the metabolic gene cluster for this pathway (Supplementary File S3) previously described for *C. parapsilosis*.³⁹ Nevertheless, despite being able to grow with phenol, resorcinol, 4-hydroxybenzoate, 2,4-dihydroxybenzoate and 3,4-dihydroxybenzoate as sole carbon sources, which is indicative of the occurrence of both branches of the 3-oxoadipate pathway, *C. subhashii* is not able to grow exclusively in catechol and hydroquinone (Fig. 2A). Therefore, we searched for orthologs of the relevant genes and identified that *C. subhashii* presents orthologs for all the genes involved in both branches of the pathway, except for *PHH2* encoding a phenol monooxygenase (Supplementary File S3). As *C. subhashii* assimilates both phenol and resorcinol, we assume that monooxygenase encoded by the orthologue of *MNX3/PHH1* has a broader substrate specificity and catalyzes hydroxylation of both substrates. In this reaction, phenol and resorcinol/hydroquinone are converted to catechol (catechol branch) and HHQ (HHQ branch), respectively. Moreover, decarboxylation of 4-hydroxybenzoate by monooxygenase *Mnx1* produces hydroquinone (Fig. 2B). Therefore, it is intriguing that *C. subhashii* does not grow in media containing catechol and hydroquinone as carbon sources. Nevertheless, substrate assimilation assays (i.e. the growth on phenol and 4-hydroxybenzoate) and the presence of corresponding gene orthologs (Supplementary File S3) indicate that the growth impairment on these two substrates could rather reflect their inefficient uptake from medium than a defect in the pathway. Taken together, *C. subhashii* has functional catechol and HHQ branches of the 3-oxoadipate pathway as well as the gentisate pathway, which may belong to ancestral inventory of the CUG-Ser clade species. The importance of metabolism of hydroxyaromatic compounds for this yeast is also underlined by the presence of two copies of *MNX1* as well as by two and four paralogues of *HBT2* and *HBT1*, respectively, coding for hydroxybenzoate transporters.⁴¹

3.3. Genomic variability in *C. subhashii*

Based on read mapping and variant calling analysis (see Section 2 for more details), we determined that *C. subhashii* SUB1 and SUB10 have 218,606 and 221,394 SNPs, respectively. Consistently with the high levels of heterozygosity inferred from the *k*-mer frequency analysis and nQuire³⁵ estimations, from these SNPs, 218,242 (14.23 per kb) and 220,958 (14.37 per kb), respectively, are heterozygous (Table 1). Of note, these heterozygous variants are not homogeneously distributed across the genome, but rather form blocks of heterozygosity separated by homozygous regions. These heterozygous blocks have an estimated nucleotide sequence divergence of 4.3% with a single density peak (Fig. 1B and C), suggesting that they all originate at a single time point. These results resemble the genomic patterns previously described for other *Candida* hybrids,^{4,7,8,10} indicating that the heterozygous blocks of *C. subhashii* do not result from the accumulation of mutations with time, but are instead the footprints of a past hybridization.⁸ For the same reason, we conclude that the homozygous regions of this genome actually correspond to blocks of LOH, a common trace of genome shaping after hybridization,³ which in this case represents 70% of the genome. Despite the

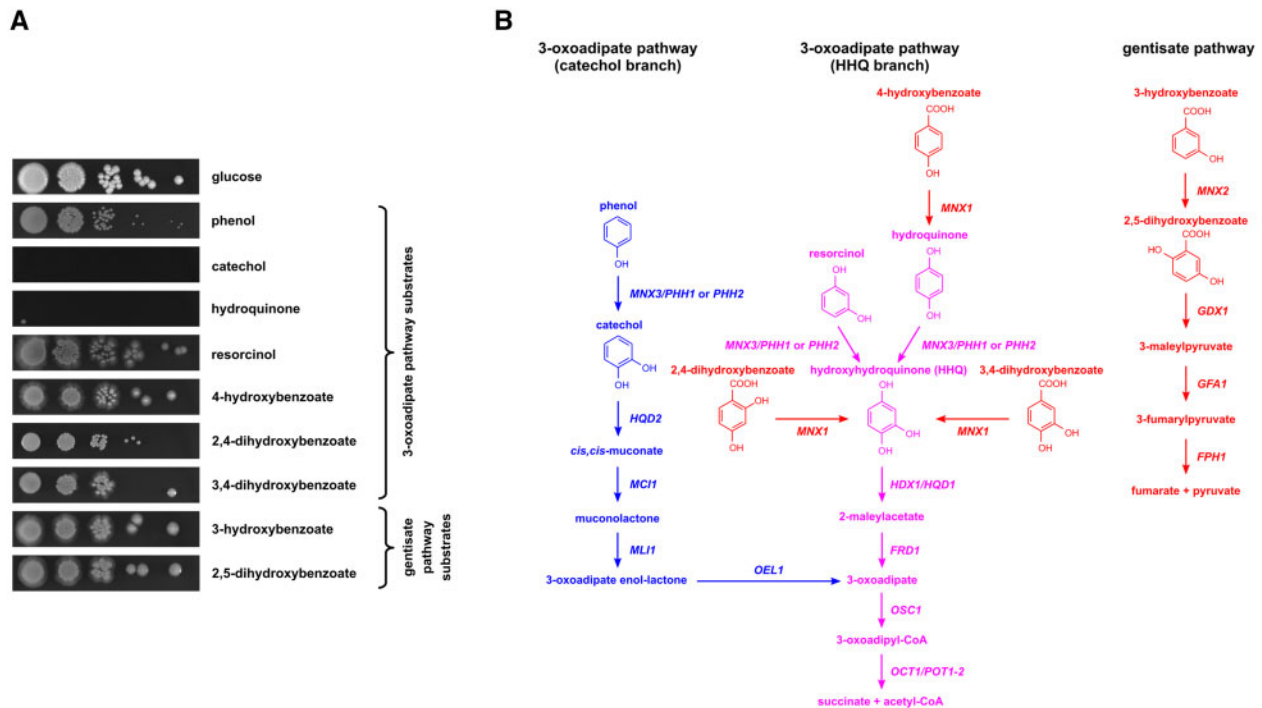


Figure 2. *Candida subhashii* utilizes a range of hydroxyaromatic substrates metabolized via the 3-oxoadipate and gentisate pathway. (A) Assimilation test of the strain CBS10753. These results indicate that *C. subhashii* has active two branches of the 3-oxoadipate pathway as well as the gentisate pathway. The assimilation test of the strain CBS10753 was performed in synthetic media containing indicated substrates as a sole carbon source at 28°C. The cells were pre-grown in a YPD medium, washed with water and serial fivefold dilutions were spotted onto the test plates. (B) A simplified scheme of the 3-oxoadipate (catechol and HHQ branches) and gentisate pathways in *C. subhashii*. Biochemical reactions operating in *C. albicans*, *C. parapsilosis* and both species are shown in blue, red and magenta, respectively. The names of genes for enzymes catalyzing corresponding biochemical reactions are based on the nomenclature from *C. albicans* and *C. parapsilosis* (see [Supplementary File S3](#) for the complete list of *C. subhashii* orthologs). Note that although *C. subhashii* does not grow on plates containing catechol and hydroquinone the corresponding biochemical reactions should be functional as it assimilates phenol and 4-hydroxybenzoate. We assume that the growth on catechol and hydroquinone is impaired due to a defect in the substrate uptake.

fact that the estimated parental divergence of *C. subhashii* (4.3% current haplotype divergence) and the absence of known parental lineages do not allow us to determine if we are in presence of an intra- or interspecific hybrid, it is important to note that the genomic variability and haplotype divergence observed in *C. subhashii* are similar to what was previously described for other hybrids of the CUG-Ser clade, as *C. metapsilosis* and *C. orthopsilosis*.^{4,6,10} *Candida metapsilosis* and *C. orthopsilosis* represent two hybrid pathogens, for which all sequenced strains thus far correspond to clinical isolates.^{4,6,10} The absence of known parental lineages for *C. metapsilosis* among clinical isolates has led to the hypothesis that they are possibly non-pathogenic and that the hybridization event contributed to the emergence of a new pathogenic lineage.^{3,4} However, it remains unclear whether hybrid clones also exist in the environment or whether they are exclusively found in association with humans. Contrary to these hybrid lineages, *C. subhashii* has mostly been isolated from environmental sources, with the only clinical isolate of the species corresponding to the type strain.^{11–14} Therefore, despite the absence of known parental lineages, in this case we can infer that hybridization occurred in the environment. Yet, as in the previous clade, it is still possible that some traits in the hybrids may have also facilitated opportunistic colonization of humans. Therefore, *C. subhashii* adds to an increasing list of hybrid lineages which can adopt a pathogenic behaviour, and raises once again the question of the role of hybridization to the emergence of pathogenic lineages.^{3,4}

3.4. Genomic differences between clones with linear and circular mitochondrial genomes

The topology of the mitochondrial genome is not the same in all the species of the CUG-Ser clade.⁴² For instance, while *C. albicans* or *C. tropicalis* have circular mitochondrial genomes, members of the *C. parapsilosis* species complex have linear mitochondrial genomes terminating with telomeric structures composed of long tandem repeats.⁴³ More interestingly, even within the *C. parapsilosis* clade, specifically in *C. metapsilosis*, two isogenic strains present different mitochondrial genome conformation.⁴ However, the existence of substantial differences between the nuclear genomes of these two isogenic strains has never allowed the identification of possible regions involved in the different mitochondrial genome conformation.⁴ To identify genomic alterations potentially related to the change of mitochondrial genome topology in *C. subhashii* we analysed the genomes of two related strains differing in this trait (SUB1 and SUB10 clones). Our comparative genomics analysis revealed the presence of 317 and 618 exclusive non-synonymous SNPs in SUB1 (circular) and SUB10 (linear), respectively. From these polymorphisms, only one in SUB1 (scaffold81, position 22,972) overlaps a mitochondria-related gene, namely, *MSK1* (Table 2). This gene is involved in tRNA processing and mitochondrial translation, but to the best of our knowledge has not been associated with any mechanism that could lead to alterations in mtDNA conformation. However, in baker's yeast *Saccharomyces cerevisiae*, functional mitochondrial translational machinery is required for the maintenance of intact organellar genome.⁴⁵

Table 2. List of genes identified as having a potential role on the different mitochondrial genome conformation observed in *C. subhashii*, with indication of the mitochondrial genome topology where the relevant genomic alteration was observed, the type of genomic feature associated, and the indication if the gene was identified as a mitochondria-related gene by our analysis

Gene	Mitochondrial genome topology	Genomic feature	Predicted mitochondrial localization (TargetP)
<i>MSK1</i>	linear	Non-Syn SNP	Yes
<i>SSZ1</i>	circular	LOH event	No
<i>ALG5</i> and <i>MRPL9</i>	linear	LOH event	Yes
<i>OYE32</i>	linear	LOH event ^a	No ^b

^aNon-Syn mutations were identified in the clone harbouring circular mitochondrial DNA, suggesting the relevance of the LOH event.

^bHomologue in *S. cerevisiae* is associated with mitochondria.⁴⁴

Considering that *C. subhashii* is a hybrid and LOH blocks represent traces of haplotype shaping to achieve a stage of genome stabilization,³ we decided to inspect the differences between the two *C. subhashii* clones in terms of LOH. Jaccard metric between LOH blocks of SUB1 and SUB10 revealed that they share 99.4% of their LOH regions. Indeed, we only identified three LOH blocks that are not shared by the two clones, and therefore may correspond to relevant targets. One of such blocks is exclusive of SUB1 (circular mtDNA) and overlaps *SSZ1* (Table 2), which is a Hsp70 protein that interacts with Zuo1p (a DnaJ homologue). The second block overlaps both *ALG5* and *MRPL9* genes in SUB10 (linear mtDNA, Table 2). This last gene is a component of the large subunit of the mitochondrial ribosome, which mediates translation in the mitochondrion, and identified in our analysis as a mitochondria-related gene (see Section 2 for more details). Important to note, in this case, the LOH does not affect the promoter region. A third and last LOH event, also exclusive of SUB10 (linear mtDNA), involves *OYE32* (Table 2), which encodes NAD(P)H oxidoreductase family protein whose homologue in *S. cerevisiae* Oye2 is associated with mitochondria.⁴⁴ Interestingly, SUB1 (circular mtDNA) has four non-synonymous SNPs covering the same gene. This suggests that the event of LOH may indeed have led to important differences between the two clones in *OYE32*, and this gene is a possible good target for future studies addressing such a question. Noteworthy, by mapping the mate-pair reads of SUB3 clone to *C. subhashii* genome assembly, we confirmed that SUB3 and SUB10 share the same allelic sequences in these regions, reinforcing the possible role of these genes in the alteration of the mitochondrial conformation of SUB1 (Supplementary File S4).

In summary, *C. subhashii* is a new *Candida* species, whose main niche is likely environmental and as a common member of soil microbial communities it is involved in degradation of lignin-derived hydroxyaromatic compounds. Despite this, *C. subhashii* was the causative agent of a case of fatal peritonitis, indicating that this species can adopt a pathogenic behaviour. This work represents the first genome analysis of this species and reveals that *C. subhashii* results from a hybridization event. The role of hybridization on the emergence of lineages with ability to infect humans is not yet understood, nevertheless next-generation sequencing analyses have allowed the identification of an increasing number of hybrids with possible pathogenic behaviour, and we showed that *C. subhashii* is one of them. Furthermore, similarly to what was observed in *C. metapsilosis* hybrids,⁴ two clones of the same strain of *C. subhashii* present different mitochondrial genome conformation. In this regard, through a comparative genomics analysis we were able to identify five genes

with potential role on mitochondrial genome conformation, namely, *MSK1*, *SSZ1*, *ALG5*, *MRPL9* and *OYE32*. Unfortunately, given the difficulty of genetic manipulation of non-model species, it was not possible to test this association in the laboratory, but future studies may profit from this analysis and other datasets provided by this study as the genome assembly, and the complete collection of gene phylogenies of *C. subhashii*.

Acknowledgements

We would like to thank Heather J. Adam (University of Toronto, Canada) and Subhash Mohan (Mount Sinai Hospital, Toronto, Canada) for providing us a culture of the *C. subhashii* type strain, and Marina Marcet-Houben for the helpful discussions and comments on this work.

Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-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 the EMBL partnership, and grants 'Centro de Excelencia Severo Ochoa 2013-2017' SEV-2012-0208 and BFU2015-67107 co-funded 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 Programme under the grant agreement ERC-2016-724173. T.G. also receives support from an INB Grant (PT17/0009/0023 - ISCIII-SGEFI/ERDF). J.N. group was supported by the Slovak Research and Development Agency (APVV-18-0239) and the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic (VEGA 1/0027/19).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at DNARES online.

References

1. Turner, S.A. and Butler, G. 2014, The *Candida* pathogenic species complex, *Cold Spring Harb. Perspect. Med.*, **4**, a019778.

2. Pfaller, M.A., Diekema, D.J., Gibbs, D.L., et al. 2010, 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.*, **48**, 1366–77.
3. Mixão, V. and Gabaldón, T. 2018, Hybridization and emergence of virulence in opportunistic human yeast pathogens, *Yeast*, **35**, 5–20.
4. Prysacz, L.P., Németh, T., Saus, E., et al. 2015, The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis*, *PLoS Genet.*, **11**, e1005626.
5. Gabaldón, T. 2020, Hybridization and the origin of new yeast lineages, *FEMS Yeast Res.*, **20**, foaa040.
6. Schröder, M.S., Martínez de San Vicente, K., Prandini, T.H.R., et al. 2016, Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species, *PLoS Genet.*, **12**, e1006404.
7. Mixão, V., Hansen, A.P., Saus, E., Boekhout, T., Lass-Florl, C. and Gabaldón, T. 2019, Whole-genome sequencing of the opportunistic yeast pathogen *Candida inconspicua* uncovers its hybrid origin, *Front. Genet.*, **10**, 383.
8. Mixão, V. and Gabaldón, T. 2020, Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*, *BMC Biol.*, **18**, 48.
9. Mixão, V., Saus, E., Boekhout, T. and Gabaldón, T. 2021, Extreme diversification driven by parallel events of massive loss of heterozygosity in the hybrid lineage of *Candida albicans*, *Genetics*, **217**, iyaa004.
10. Prysacz, L.P., Németh, T., Gácsér, A. and Gabaldón, T. 2014, Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies, *Genome Biol. Evol.*, **6**, 1069–78.
11. Adam, H., Groenewald, M., Mohan, S., et al. 2009, Identification of a new species, *Candida subhashii*, as a cause of peritonitis, *Med. Mycol.*, **47**, 305–11.
12. Hilber-Bodmer, M., Schmid, M., Ahrens, C.H. and Freimoser, F.M. 2017, Competition assays and physiological experiments of soil and phyllosphere yeasts identify *Candida subhashii* as a novel antagonist of filamentous fungi, *BMC Microbiol.*, **17**, 4.
13. Tanimura, A., Kikukawa, M., Yamaguchi, S., Kishino, S., Ogawa, J. and Shima, J. 2015, Direct ethanol production from starch using a natural isolate, *Sci. Rep.*, **5**, 9593.
14. Filipowicz, N., Momotko, M., Boczkaj, G., Pawlikowski, T., Wanarska, M. and Ciesliński, H. 2017, Isolation and characterization of phenol-degrading psychrotolerant yeasts, *Water. Air. Soil Pollut.*, **228**, 210.
15. Fricova, D., Valach, M., Farkas, Z., et al. 2010, The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5' termini, *Microbiology (Reading)*, **156**, 2153–63.
16. Valach, M., Tomaska, L. and Nosek, J. 2008, Preparation of yeast mitochondrial DNA for direct sequence analysis, *Curr. Genet.*, **54**, 105–9.
17. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
18. O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M.M., Gormley, N.A. and Cox, A.J. 2015, NxTrim: optimized trimming of Illumina mate pair reads: table 1, *Bioinformatics*, **31**, 2035–7.
19. Mapleson, D., Accinelli, G.G., Kettleborough, G., Wright, J. and Clavijo, B.J. 2016, KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies, *Bioinformatics*, **32**, btw663.
20. Luo, R., Liu, B., Xie, Y., et al. 2012, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, *Gigascience*, **1**, 18.
21. Bankevich, A., Nurk, S., Antipov, D., et al. 2012, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.*, **19**, 455–77.
22. Safonova, Y., Bankevich, A. and Pevzner, P.A. 2015, dipSPAdes: assembler for highly polymorphic diploid genomes, *J. Comput. Biol.*, **22**, 528–45.
23. Prysacz, L.P. and Gabaldón, T. 2016, Redundans: an assembly pipeline for highly heterozygous genomes, *Nucleic Acids Res.*, **44**, e113.
24. Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. 2013, QUILT: quality assessment tool for genome assemblies, *Bioinformatics*, **29**, 1072–5.
25. Stanke, M. and Morgenstern, B. 2005, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, *Nucleic Acids Res.*, **33**, W465–7.
26. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
27. Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. 2000, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, *J. Mol. Biol.*, **300**, 1005–16.
28. Mixão, V., Saus, E., Hansen, A.P., Lass-Florl, C. and Gabaldón, T. 2019, Genome assemblies of two rare opportunistic yeast pathogens: *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*), *G3 Genes|Genomes|Genetics*, **9**, 3921–7.
29. Huerta-Cepas, J., Capella-Gutiérrez, S., Prysacz, L.P., Marcet-Houben, M. and Gabaldón, T. 2014, PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome, *Nucleic Acids Res.*, **42**, D897–902.
30. Al-Shahrour, F., Minguez, P., Tárrega, J., et al. 2007, FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments, *Nucleic Acids Res.*, **35**, W91–6.
31. Pegueroles, C., Mixão, V., Carreté, L., Molina, M. and Gabaldón, T. 2020, HaploTypo: a variant-calling pipeline for phased genomes, *Bioinformatics*, **36**, 2569–71.
32. Li, H. 2013, *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. arXiv:1303.3997.
33. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–303.
34. Garrison, E. and Marth, G. 2012, Haplotype-based variant detection from short-read sequencing, arXiv:1207.3907.
35. Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S. and Burbano, H.A. 2018, nQuire: a statistical framework for ploidy estimation using next generation sequencing, *BMC Bioinformatics*, **19**, 122.
36. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
37. Gabaldón, T. 2008, Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.*, **9**, 235.
38. Gabaldón, T., Naranjo-Ortiz, M.A. and Marcet-Houben, M. 2016, Evolutionary genomics of yeast pathogens in the Saccharomycotina, *FEMS Yeast Res.*, **16**, fow064.
39. Holesova, Z., Jakubkova, M., Zavadiakova, I., Zeman, I., Tomaska, L. and Nosek, J. 2011, Gentsate and 3-oxoadipate pathways in the yeast *Candida parapsilosis*: identification and functional analysis of the genes coding for 3-hydroxybenzoate 6-hydroxylase and 4-hydroxybenzoate 1-hydroxylase, *Microbiology (Reading)*, **157**, 2152–63.
40. Gérecová, G., Neboháčová, M., Zeman, I., et al. 2015, Metabolic gene clusters encoding the enzymes of two branches of the 3-oxoadipate pathway in the pathogenic yeast *Candida albicans*, *FEMS Yeast Res.*, **15**, fov006.
41. Cillingová, A., Zeman, I., Tóth, R., et al. 2017, Eukaryotic transporters for hydroxyderivatives of benzoic acid, *Sci. Rep.*, **7**, 8998.
42. Valach, M., Farkas, Z., Fricova, D., et al. 2011, Evolution of linear chromosomes and multipartite genomes in yeast mitochondria, *Nucleic Acids Res.*, **39**, 4202–19.
43. Rycovska, A., Valach, M., Tomaska, L., Bolotin-Fukuhara, M. and Nosek, J. 2004, Linear versus circular mitochondrial genomes: intraspecies variability of mitochondrial genome architecture in *Candida parapsilosis*, *Microbiology (Reading)*, **150**, 1571–80.
44. Odat, O., Matta, S., Khalil, H., et al. 2007, Old yellow enzymes, highly homologous FMN oxidoreductases with modulating roles in oxidative stress and programmed cell death in yeast, *J. Biol. Chem.*, **282**, 36010–23.
45. Myers, A.M., Pape, L.K. and Tzagoloff, A. 1985, Mitochondrial protein synthesis is required for maintenance of intact mitochondrial genomes in *Saccharomyces cerevisiae*, *Embo J.*, **4**, 2087–92.