

RESEARCH ARTICLE

The Genomic Aftermath of Hybridization in the Opportunistic Pathogen *Candida metapsilosis*

Leszek P. Pryszcz^{1,2}, Tibor Németh³, Ester Saus^{1,2}, Ewa Ksiezopolska^{1,2}, Eva Hegedúsová⁴, Jozef Nosek⁴, Kenneth H. Wolfe⁵, Attila Gacser^{3†*}, Toni Gabaldón^{1,2,6†*}

1 Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain, **2** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **3** Department of Microbiology, University of Szeged, Szeged, Hungary, **4** Department of Biochemistry, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia, **5** UCD Conway Institute, School of Medicine & Medical Science, University College Dublin, Dublin, Ireland, **6** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

† AG and TG share senior authorship on this work.
* gacsera@gmail.com (AG); tgabaldon@crg.es (TG)



CrossMark
click for updates

 OPEN ACCESS

Citation: Pryszcz LP, Németh T, Saus E, Ksiezopolska E, Hegedúsová E, Nosek J, et al. (2015) The Genomic Aftermath of Hybridization in the Opportunistic Pathogen *Candida metapsilosis*. *PLoS Genet* 11(10): e1005626. doi:10.1371/journal.pgen.1005626

Editor: Jianzhi Zhang, University of Michigan, UNITED STATES

Received: May 12, 2015

Accepted: October 5, 2015

Published: October 30, 2015

Copyright: © 2015 Pryszcz et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Sequencing data, genome assembly and annotation have been deposited to EBI-ENA (accession: PRJNA238968). Reconstructed trees and alignments were deposited to PhylomeDB (<http://phylomedb.org>) as PhylomeDB ID 243.

Funding: TG's group's research acknowledges support from Spanish Ministry of Economy and Competitiveness grants, 'Centro de Excelencia Severo Ochoa 2013–2017' SEV-2012-0208, and BIO2012-37161 cofounded by European Regional Development Fund (ERDF); from the European

Abstract

Candida metapsilosis is a rarely-isolated, opportunistic pathogen that belongs to a clade of pathogenic yeasts known as the *C. parapsilosis sensu lato* species complex. To gain insight into the recent evolution of *C. metapsilosis* and the genetic basis of its virulence, we sequenced the genome of 11 clinical isolates from various locations, which we compared to each other and to the available genomes of the two remaining members of the complex: *C. orthopsilosis* and *C. parapsilosis*. Unexpectedly, we found compelling genomic evidence that *C. metapsilosis* is a highly heterozygous hybrid species, with all sequenced clinical strains resulting from the same past hybridization event involving two parental lineages that were approximately 4.5% divergent in sequence. This result indicates that the parental species are non-pathogenic, but that hybridization between them formed a new opportunistic pathogen, *C. metapsilosis*, that has achieved a worldwide distribution. We show that these hybrids are diploid and we identified strains carrying loci for both alternative mating types, which supports mating as the initial mechanism for hybrid formation. We trace the aftermath of this hybridization at the genomic level, and reconstruct the evolutionary relationships among the different strains. Recombination and introgression -resulting in loss of heterozygosity- between the two subgenomes have been rampant, and includes the partial overwriting of the MTL_a mating locus in all strains. Collectively, our results shed light on the recent genomic evolution within the *C. parapsilosis sensu lato* complex, and argue for a re-definition of species within this clade, with at least five distinct homozygous lineages, some of which having the ability to form hybrids.

Union and ERC Seventh Framework Programme (FP7/2007-2013) under grant agreements FP7-PEOPLE-2013-ITN-606786 and ERC-2012-StG-310325, and grant from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No H2020-MSCA-ITN-2014-642095. LPP is funded through La Caixa-CRG International Fellowship Program. JN's group was supported by the Slovak Research and Development Agency (APVV 0123-10) and the Scientific Grant Agency (VEGA 1/0333/15). AG is supported by OTKA NN113153, NF84006 and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. 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

Human pathogens belong to different phylogenetic clades and it is clear that the ability to infect humans emerged several times independently. The sequencing and comparison of genomes from pathogenic and non-pathogenic species and strains paves the way to identify what genomic changes underlie the emergence of virulence. In this study we sequenced 11 globally-distributed clinical isolates of *Candida metapsilosis*, an emerging fungal pathogen of growing concern. We found that all isolates were the result of a single hybridization between two unidentified species, which points to hybridization as a mechanism for the origin of a virulent lineage. We found that the hybrids likely originated by sexual reproduction as they were diploids and retained genomic regions of opposite mating types. We reconstructed the aftermath of the genome merging by identifying where recombination led to the removal of one of the parental subgenomes. We finally compare the newly-sequenced genome with those of other pathogens from the *Candida* clade and establish global trends, such as an enriched repertoire in cell-wall proteins in more virulent species. Our results provide insight into how hybridization may play a key role in the emergence of novel pathogenic lineages.

Introduction

Hybridization between species is an important evolutionary mechanism that can drive the origin of novel lineages and adaptation to new environments. Hybridization results in the combination of two diverged genomes, which are subsequently shaped by processes of recombination, deletion, and other genomic rearrangements. Genomics have recently paved the way to investigate the stochastic and adaptive processes that follow genomic hybridization. As compared to metazoans or plants, fungi have lower prezygotic barriers and can reproduce clonally for long periods of time, thus hybridization is thought to have a large impact in the evolution of this clade. Consistently, the presence of hybrids in fungi have been extensively documented and an increasing number of cases are being described in the literature [1–8]. Hybridization has been proposed as a mechanism to drive the origin of new human fungal pathogens [8]. Indeed several hybrid species have been described among several human fungal pathogens [8–10], although in all these cases also non-hybrid strains of the parental species can infect humans. So far, only few studies have focused on population genomics of hybrid species. Here we report a population genomics analysis of an opportunistic pathogen that belongs to the *Candida parapsilosis* species complex.

This species complex comprises opportunistic pathogen species that cause serious infections in immunocompromised patients, and whose incidence has significantly increased in recent years [11]. Three distinct clades within this complex, formerly known as *C. parapsilosis* groups I, II and III, have been re-defined as different species: *C. parapsilosis sensu stricto*, *C. orthopsilosis* and *C. metapsilosis*, respectively [12]. These species differ in their degrees of prevalence and virulence, *C. metapsilosis* being the one with the lowest clinical prevalence, and accounting for only 1.1 to 8.4% of the infections caused by the complex [13]. In addition, the species differ in their degree of prevalence across different types of patients [14–15]. There are few studies investigating the virulence of *C. parapsilosis sensu lato* species, and particularly little is known about the rarer species *C. orthopsilosis* and *C. metapsilosis*. Nonetheless, results of several *in vitro* and *in vivo* studies suggest that, consistent with its lower prevalence, *C. metapsilosis* is the least virulent species of the complex [16–20]. A recent study involving 93 different *C. parapsilosis sensu lato* isolates, revealed that *C. metapsilosis* strains were unable to produce extracellular

lipases and form pseudohyphae, attributes that are both recognized as important virulence factors for *Candida* spp. [18]. Furthermore, it has been shown that *C. metapsilosis* isolates are killed more efficiently by and are less cytotoxic to human primary macrophages [17,18]. Finally, other studies have focused on the antifungal susceptibility of different species within the complex showing that strain-to-strain differences within species are common [13,21,22].

The growing incidence of infections due to *C. parapsilosis sensu lato* spp. underlines the importance of studies investigating the virulence attributes and molecular genetics of these species. Furthermore, the *C. parapsilosis* species complex offers an exquisite system for the study of the evolution of pathogenic yeasts and their adaptation to the human host. The complex belongs to the broader CTG clade of Saccharomycetales, which include species that decode the CUG codon as serine instead of leucine. Although the CTG clade includes other clinically important species such as *C. albicans* and *C. tropicalis* [23], their phylogenetic position (see below) and particular virulence properties indicates that the species within the *Candida parapsilosis* species complex evolved pathogenesis towards humans independently of *C. albicans* and their closest relatives. The sequencing of reference strains for *C. parapsilosis* [24] and *C. orthopsilosis* [25] has been instrumental in assessing the main differences with other pathogens within the CTG clade, particularly with the model yeast pathogen *C. albicans*. Such analyses have revealed that, whereas most *Candida* species display a similar content of families represented by singleton genes, most of the variability is related to copy number differences in multi-gene families, with pathogens having increased number of members in families related to virulence [24]. For instance, initial comparisons found that *C. parapsilosis* has an expanded Hyr/Iff family of virulence-related cell wall genes relative to the less virulent species *C. orthopsilosis* [25]. Subsequent analyses have assessed the genomic diversity among *C. parapsilosis* [26] and *C. orthopsilosis* [8] isolates. These studies have enabled important discoveries, such as the realization of the existence of recombination between clinical and environmental lineages in *C. parapsilosis*, pointing to several recent and recurrent clinical outbreaks from the environment [26], or the discovery of hybrids between differentiated *C. orthopsilosis* subspecies, including the description of one virulent hybrid lineage isolated from two very distant locations [8]. In contrast to the growing awareness on the genomic diversity of other *C. parapsilosis sensu lato* species, for *C. metapsilosis* we lack both a reference genome and a comprehensive insight on the genomic diversity across isolates. This situation precludes understanding the emergence of virulence traits within this complex. To fill in this important gap we undertook the sequencing and analysis of the genomes of eleven *C. metapsilosis* clinical isolates. Unexpectedly, we found that all sequenced isolates, sampled from geographically distant locations, presented highly heterozygous genomes, which we show to result from a single hybridization event between two parental lineages differing by 4.5% at the nucleotide level.

Results and Discussion

Sequencing and assembly reveal a hybrid genome and a newly circularized mitochondrial genome

We used Illumina technology to sequence a panel of eleven *C. metapsilosis* clinical isolates from different geographical locations (Table 1). Initial attempts to assemble the individual strains using standard approaches proved unsuccessful, independently of sequencing coverage or the combined use of libraries of different read lengths and insert sizes (S1 Table). In particular, for the strain PL429 four different libraries were used at varying read lengths and insert sizes, totaling an overall coverage of 1,308x (Table 1), yet this still yielded a highly fragmented assembly with thousands of contigs. This elusive assembly for a relatively small genome was reminiscent of what we had previously observed in a highly heterozygous strain of *C.*

Table 1. Strain and genome sequencing stats. Basic statistics for the genomes sequenced within this work. For each strain the table provides, in this order: strain name; geographical origin; body site of isolation, type of mitochondrial chromosome, sequencing statistics including read length, type of sequencing libraries, insert size and depth of coverage (x fold); number of heterozygous and homozygous SNPs; fraction of the genome in heterozygous blocks (100 bp threshold); and minimal number of LOH events detected in that strain (100 bp threshold). pe: paired-end; mp: mate-pairs; ov: overlapping paired-end reads; pe600: paired-end reads with ~600 bp insert size.

Strain	Place of isolation	Site of isolation	mtDNA	Read length	Library	DoC	Hetero SNPs	Homo SNPs	Hetero	LOH
BP57 (SZMC8022)	Pécs, Hungary	throat	linear	96	pe600	293	360,531	21,133	58.40%	12,693
CP376 (SZMC8098)	Pisa, Italy	faeces	linear	96	pe600	265	348,424	7,609	56.98%	10,796
CP61 (SZMC8093)	Pisa, Italy	nail	linear	96	pe600	273	361,218	26,165	58.63%	12,085
MCO448	Washington, USA	hand	linear	46	pe600	200	308,944	43,421	61.29%	9,231
PL429 (SZMC1548)	Livermore, USA	n.a.	linear	76 46 46 250 822	pe300 pe600 mp5000 pe400ov fosmid	1,308	324,595	48,335	54.53%	11,377
PL448	Washington, USA	hand	circular	46	pe600	212	299,831	57,113	56.08%	9,539
SZMC21154	Cataluna, Spain	blood	circular	100	pe600	569	334,024	44,740	54.80%	10,903
SZMC8029	Debrecen, Hungary	blood	linear	100	pe600	500	345,879	41,406	56.14%	11,181
SZMC8092 (CP43)	Pisa, Italy	lung	linear	100	pe600	568	349,440	42,064	57.37%	11,364
SZMC8094 (CP92)	Pisa, Italy	faeces	linear	100	pe600	467	357,656	258	59.37%	10,460
SZMC8095 (CP231)	Pisa, Italy	nail	linear	100	pe600	575	350,168	38,488	60.00%	10,495

doi:10.1371/journal.pgen.1005626.t001

orthopsilosis [8]. Indeed assemblies from highly heterozygous genomes are highly fragmented and result in a total genome size larger than expected [8,27]. This is because two alternative contigs are recovered for each heterozygous region, while a single, collapsed contig is recovered from each homozygous region. Such assemblies are difficult to scaffold further, as each homozygous contig could be joined, at each side, to either of the two heterozygous contigs. To generate a suitable reference assembled genome for *C. metapsilosis*, we used a previously developed heterozygous genome assembly strategy [8], which we describe here in greater detail (see [Materials and Methods](#)). To obtain an optimal assembly we had to apply this procedure to the combined data from two strains: the one with the highest sequencing coverage and the larger number of libraries, PL429 (SZMC1548), and the one producing the least fragmented *de novo* assembly, SZMC8094 (Table 1, S1 and S2 Tables). Thus the resulting reference assembly is chimeric in nature and comprises sequences from the two isolates. In addition, similar to previous highly heterozygous assemblies [8], only one of the haplotypes for each heterozygous region is represented in the final assembly. We then mapped heterozygous regions in each strain relative to this reference assembly.

The final assembly resulted in a total size of 13.3 Mb in seven putative chromosomes and two unplaced scaffolds, which is similar to the number of bands observed in a Pulsed Field Gel Electrophoresis (PFGE) (S1 Fig). The observed band patterning was also indicative of possible genomic rearrangements in several of the analyzed strains, including the two strains used in the reference assembly. Note that PFGE and short range pair-end reads provide information at different scales and thus integration of both sources of information is difficult. Predicted chromosome ends were enriched in *C. metapsilosis* telomere repeats (GGTTAGGATGTCCAAAG-TATTGA), corresponding to the template domain of the telomerase (TER1) [28] in a region (827,407–829,461) of scaffold 2. The annotation of the genome (see [Materials and Methods](#)) resulted in 5,973 protein-coding genes in *C. metapsilosis*, which is roughly similar to the gene counts in *C. parapsilosis* (5,752) and *C. orthopsilosis* (5,784).

The mitochondrial genome was assembled in a single 21 kb-long contig [29]. Species from the *Candida parapsilosis* complex generally display linear mitochondrial chromosomes, but

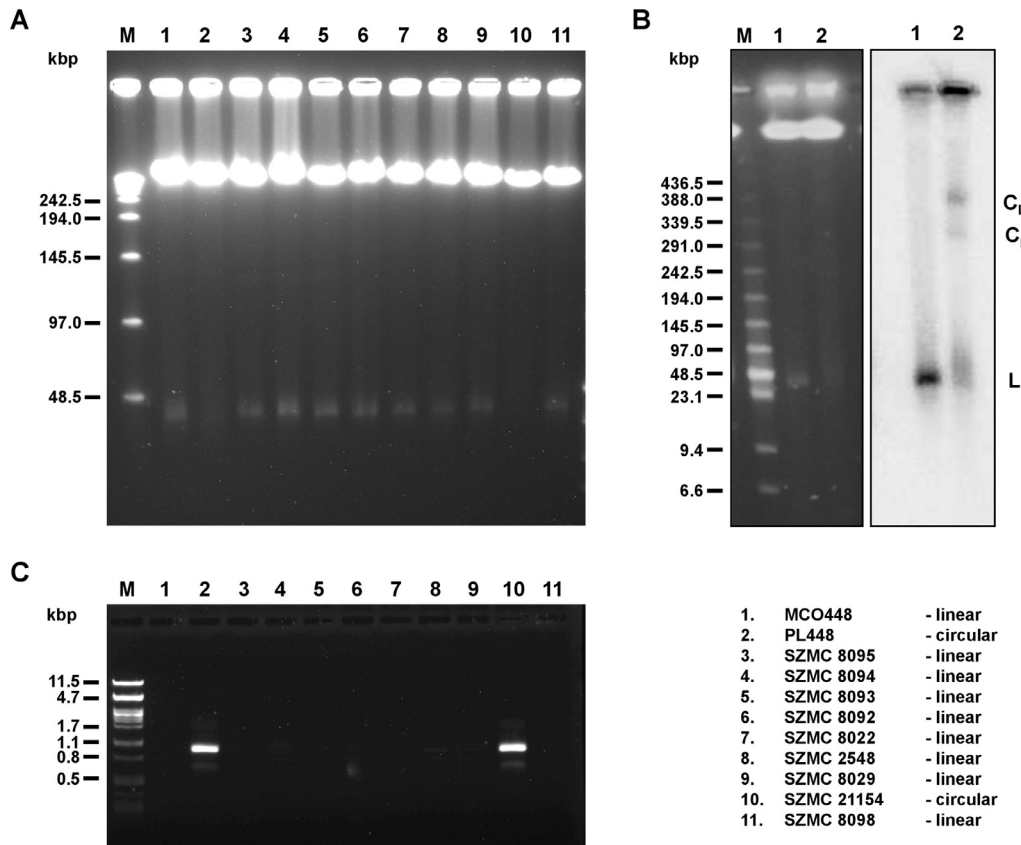


Fig 1. Linear and circular mitochondrial genomes in *C. metapsilosis* isolates. (A) PFGE analysis. DNA was isolated in agarose blocks and separated in PFGE as described in Materials and Methods. Only strains containing linear mitochondrial genome show a sharp band about 25–30 kb long. (B) Southern blot analysis. PFGE separated DNA samples were blotted onto a nylon membrane and hybridized with radioactively labeled mtDNA probe. The probe hybridizes with a sharp band (L) in the strain MCO448 with linear mtDNA (lane 1). In the strain PL448 with circular mtDNA (lane 2) the probe reveals a smeary band (23–100 kb) plus two fractions (labeled as C_I and C_{II}) corresponding to presumed circular replication intermediates of mtDNA. In both strains, the probe also detected mtDNA molecules trapped in wells that contain branched mtDNA structures resulting from recombination processes. (C) PCR analysis. The reactions were performed with primers derived from the subtelomeric genes *nad3* and *atp6* (localized at opposite ends and oriented toward the termini of linear molecules) on total DNA templates and the PCR products were electrophoretically separated. Only strains with circular mitochondrial genome (PL448 and SZMC21154) exhibit a PCR product derived from the end-to-end junction of the mitochondrial genome. MCO448 (lane 1), PL448 (lane 2), SZMC8095 (lane 3), SZMC8094 (lane 4), SZMC8093 (lane 5), SZMC8092 (lane 6), SZMC8022 (lane 7), SZMC2548 (lane 8), SZMC8029 (lane 9), SZMC21154 (lane 10), SZMC8098 (lane 11). M-molecular marker (Lambda Ladder PFG Marker (New England Biolabs) (in A), Low Range PFG Marker (New England Biolabs) (in B), lambda DNA/PstI (in C)).

doi:10.1371/journal.pgen.1005626.g001

rare isolates presenting circularized mitochondrial DNA (mtDNA) have been identified in *C. orthopsilosis* and *C. metapsilosis* [23,29]. Our panel includes five strains whose mitochondrial architecture has been determined earlier (BP57, CP61, CP367, MCO448, PL448 and PL429), with PL448 being the only known case of a *C. metapsilosis* strain bearing a circular mitochondrial chromosome [30]. We here determined the architecture of the mitochondrial chromosomes of six additional strains from the genomic data (See [Materials and Methods](#)). Notably, we found a new case of circular mitochondrial chromosome in the strain SZMC21154, whereas the remaining mtDNAs of the newly-tested strains were predicted to be linear ([S2A Fig](#)). All these *in-silico* predictions were confirmed experimentally through PCR and PFGE tests (see [Materials and Methods, Fig 1](#)). As described in [31] the strain PL448 is a direct derivative of the clinical strain MCO448, and thus the two circularisation events must be necessarily independent evolutionary events. This is consistent with our phylogenetic analysis of the nuclear

genomes of the strains (see below) and with the fact that, although the circularization results from end-to-end fusion events in both strains, it involves different specific sites in each case (S2B Fig). There is extremely low sequence variability of mtDNA among *C. metapsilosis* isolates. Hence, the mtDNA-based phylogeny cannot resolve this issue.

All sequenced *C. metapsilosis* clinical isolates result from a single hybridization event

We next mapped the reads obtained from the sequencing of each of the strains onto the above reference genome, which served to assess sequence variation (Fig 2 and S1 File). All sequenced strains were found to be highly heterozygous (22–26 heterozygous SNPs/kb, Table 1), with divergence between the alleles in heterozygous regions averaging approximately 4.5% (see Materials and Methods). This high divergence made it possible to delimit heterozygous and homozygous regions, and to calculate the fraction of the genome that was occupied by heterozygous blocks.

We will refer to blocks of homozygosity as loss of heterozygosity blocks (LOH, a genome track where heterozygosity has been lost). Such LOH blocks could be the result from several basic mechanisms, including mitotic recombination, break-induced replication or gene conversion [32]. Alternatively an homozygous track of a certain length can occur simply by chance as SNPs in heterozygous regions are not distributed uniformly, but the likelihood of this diminishes quickly with the size of the block. The definition of a threshold is challenging and any arbitrary value will produce false positives or false negatives at different rates. We explored this issue and opted for using a more relaxed (100 bp) and more stringent (200 bp) thresholds for the minimum gap between SNPs in a region designated as an LOH block (see Materials and Methods). Unless indicated otherwise the results at 100 bp threshold are indicated. The fraction of the genome occupied by heterozygous regions varied significantly across strains, ranging from 54.5% to 61.3% of the genome by length when the 100 bp threshold was applied (Table 1), and 63.4–68.5% at the 200 bp threshold. Overall 50% of the homozygous tracks in the genome were in LOH blocks larger than 3,626 bp (LOH-50). The fraction of heterozygous regions is much higher than the 17% found in a *C. orthopsilosis* hybrid (where a 100 bp threshold was used) [8]. This difference may indicate that the *C. metapsilosis* hybridization is a more recent event or, alternatively, that *C. orthopsilosis* lost heterozygosity more rapidly.

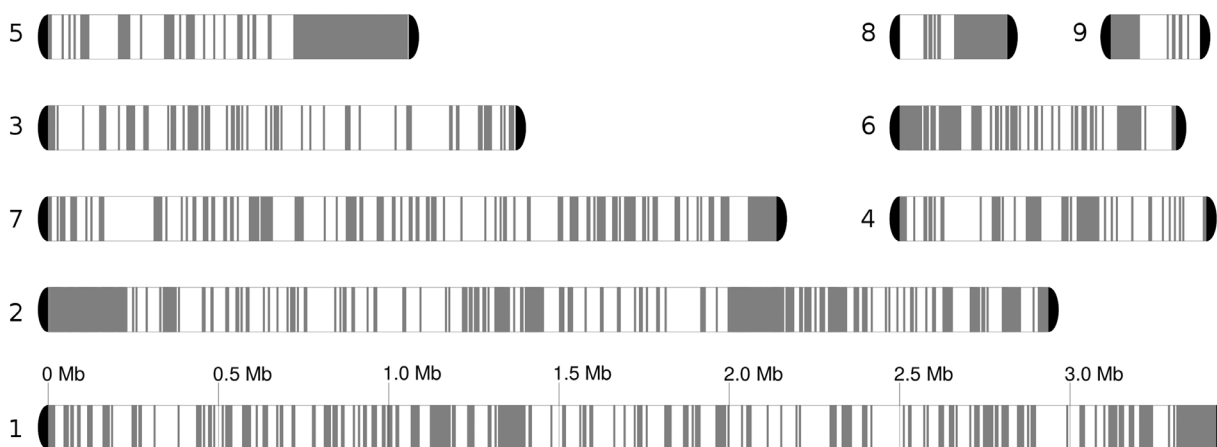


Fig 2. Chromosome blocks for the reference genome, comprising seven chromosomes and two unplaced scaffolds. Heterozygous regions are in white, while homozygous blocks of at least 5 kb are depicted in grey. Loss of heterozygosity (LOH) detection is described in Materials and Methods.

doi:10.1371/journal.pgen.1005626.g002

LOH tracts were generally short, with an average size of 535 bp (1,183 bp at 200 bp threshold) as compared to 2,151 bp in *C. orthopsilosis* (S3 Fig). This difference may be related to the lower level of heterozygosity in the latter, as multiple, partially overlapping or adjacent LOH events will be seen as a single longer homozygous blocks. Indeed in *C. orthopsilosis* 20% of the LOH blocks were longer than 1 kb, whereas in *C. metapsilosis* this fraction was only 7.2%. Interestingly, longer LOH blocks were enriched in sub-telomeric regions (Welch's t-test $P < 3.5e-06$, S4 Fig). Importantly, all *C. metapsilosis* strains shared a significant fraction (~43%) of LOH blocks, with 1,581 LOH blocks having identical boundaries across all sequenced strains (out of an average of 10,920 LOH blocks, Table 1). Even at a stricter threshold of 200 bp, 125 blocks had identical boundaries in all 11 strains. Of these blocks, 29 were larger than 500bp, including 6 longer than 1kb. This indicates that all sequenced strains derive from the same primary hybridization event and shared a number of LOH events before they diverged. The alternative scenario that those regions with exact boundaries were identical in the parental strains differing on average 4.5% is very unlikely. Subsequently other LOH events were shared by only a fraction of the strains, with 19% of LOH blocks being strain-specific in a typical strain.

Contrary to the *C. orthopsilosis* case, where a completely homozygous strain (i.e. one of the putative hybrid parentals) exists [8], none of the sequenced *C. metapsilosis* strains is homozygous. Furthermore, both haplotypes in a given heterozygous region are roughly equally distant to either *C. orthopsilosis* or *C. parapsilosis* out-groups (0.35% and 0.51% difference between the two haplotypes, respectively). This prevents assignment of each haplotype in a heterozygous region to a particular parent, and prevents phasing the homozygous regions (i.e. in Fig 2 we cannot say whether the grey homozygous regions come from one or the other parent). In the *C. orthopsilosis* hybrid both parents were found to be similarly represented among homozygous regions [8]. In the absence of an unequivocal parental mapping for *C. metapsilosis*, we approached the question of parental lineage representation using several indirect strategies. Firstly, our assembly process randomly incorporates one of the two haplotypes in a given heterozygous region. Thus, even if we cannot distinguish among parentals, we can arbitrarily name haplotype A the one included in the reference assembly, and assess for the remaining nine strains how often this or the alternative haplotype (haplotype B) is present in homozygous regions. We applied this procedure to scaffold6 from PL448, which underwent LOH of the entire chromosome (S1 File). This provided an overall estimate of 54.8% and 45.2% representation of the haplotypes A and B, respectively. We stress that these haplotypes do not involve any assignment to a given parental genome. Nevertheless, considering the random incorporation of parental haplotypes in the assembly, this result suggests an approximately balanced presence of both parentals in these homogenized regions. For the rest of the genome, haplotype B is highly underrepresented among LOHs, but this is the result of most LOH blocks being shared among most strains (S1 File).

Finally, to study the pattern of LOH on a local scale, we examined one specific region in four strains by PCR and re-sequencing. This 3.6 kb-long region around the gene g3863.t1 (coding for a protein with a DEAD-box RNA helicase domain), is interesting because it encompasses eight recent LOH blocks that are present specifically in 1 to 3 of the four strains. We can consider these eight LOH blocks to have originated independently and we can assess whether the same or a distinct parental haplotype was introgressed in each of these by selectively amplifying and sequencing each of the DNA molecules. Notably, in all cases LOHs present in a given strain resulted from haplotype A overwriting haplotype B (S5, S6 and S7 Figs). The probability that this occurred by chance assuming equal probability for both haplotypes is 0.0039, which suggests this local region has some preference to lose one of the haplotypes. Altogether these results suggest that, despite a bias towards preferential retention of one parental at the local level in some regions, there is no genome-wide preference for either of the two parental strains.

This result is in line with that obtained for a *C. orthopsilosis* hybrid with similar divergence between parents [8]. This supports the idea that hybrids from less divergent parentals display more balanced inheritance due to a lower risk of genetic incompatibility (Bateson-Dobzhansky-Muller effect). Larger divergence such as the 10% found in *Pichia sorbitophila* has led to more unbalanced inheritance in that species [33].

Genome analyses suggest mating as the hybrid formation mechanism

The distribution of read counts at biallelic single nucleotide polymorphisms (SNPs) suggested a diploid state for the sequenced *C. metapsilosis* strains (S8 Fig). In addition, this analysis revealed partial aneuploidies in two strains, namely triploidy of scaffold5 in PL448 and partial triploidy of scaffold2 in PL448 and SZMC21154 (S8 Fig). These aneuploidies were independently confirmed by depth-of-coverage analyses (S3 Table). FACS (Fluorescence-Activated Cell Sorting) analyses were consistent with the predicted diploidy of 11 *C. metapsilosis* hybrid strains and *C. orthopsilosis* MCO456, although comparison of FACS results across species must be interpreted with caution (Materials and Methods, S9 Fig). Of note PL448 is the strain with more ploidy changes and the one that recently circularized its mitochondrial chromosome (see above). In addition some results suggested some level of heterogeneity in the samples of this strain (e.g. peak of biallelic counts closer to 40% rather than 50% in diploid chromosomes), which may be interpreted as an ongoing genomic instability.

The diploid state of the hybrids suggests mating between two haploid cells with genomes ~4.5% divergent as the probable mechanism of hybridization. It has been observed that *Candida* species respond to mating pheromones of different species within the clade [34]. Based on this, mating was suggested as the mechanism of formation of the previously reported *C. orthopsilosis* diploid hybrid [8]. However, the presence of a single mating type (MTL) idiomorph in the two sequenced *C. orthopsilosis* strains prevented a definitive conclusion. In the current study, however, we found that 10 out of the 11 sequenced *C. metapsilosis* genomes contain both MTL α and MTL α idiomorphs (Fig 3), albeit with MTL α incomplete (see below). This result is consistent with mating as the mechanism for hybrid formation and lends further support to the hypothesis that haploid forms and mating occurs in species of the *C. parapsilosis* clade, a clade which is generally considered asexual [24]. The parasexual cycle, as it occurs in *C. albicans* can also involve strains of opposite mating type, however in that case diploid state is achieved by concerted loss of chromosomes with little recombination between them [25] and thus most chromosomes would be homozygous for one or the other parental lineage. Of note, the proposal of mating as the origin of hybridization does not imply that the hybrid lineage is able to undergo a sexual cycle, and thus we cannot confidently assign the source of LOH to meiotic or mitotic recombination.

Remarkably, the genome sequences indicate that the MTL α idiomorph has partially overwritten the nonhomologous MTL α idiomorph in *C. metapsilosis*. Our reference genome assembly contained an MTL α idiomorph at the MTL locus (Fig 3A) on scaffold 5, with the MTL α genes assembled separately as a small contig (scaffold 28, Fig 3B). In species such as *C. albicans* [35] and *C. orthopsilosis* [36], the MTL α and MTL α idiomorphs are highly divergent in sequence (<50% identity) over a region of ~9 kb. In addition to containing the MTL α 1/ α 2 or MTL α 1/ α 2 genes, which are unrelated in sequence, this idiomorph-specific region also includes three divergent pairs of genes (OBPa/OBP α ; PIKa/PIK α ; PAPA/PAP α) that code for proteins that are not involved in mating and that appear to have diverged into separate α and α isoforms with low amino acid sequence identity because they became trapped in the non-recombining region of the mating-type locus millions of years ago [35]. These three genes are also arranged in a different order in the two idiomorphs. In 10

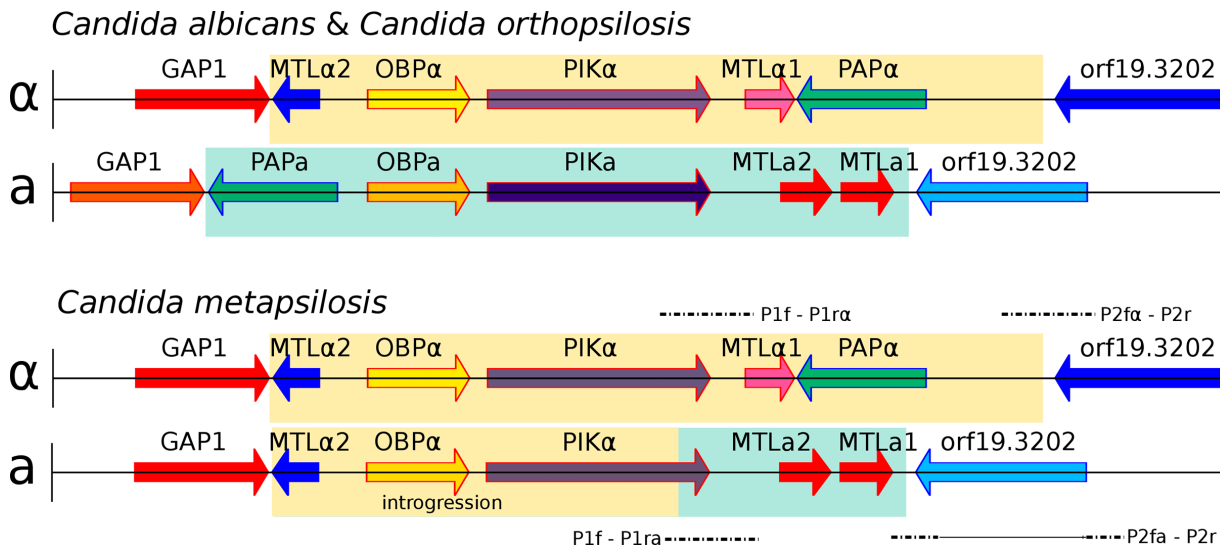


Fig 3. Genomic organisation of mating type locus (MTL) idiormorphs, MTL α (yellow) and MTL α (green), in three *Candida* species. *C. albicans* and *C. orthopsilosis* encode both MTL idiormorphs (A). *C. metapsilosis* encodes complete MTL α and partial MTL α (B). MTL α introgression in *C. metapsilosis* was confirmed with PCR and Sanger sequencing. Alignments of Sanger products are denoted with dashed lines. P1f-P1ra and P2fa-P2r align to MTL α . In contrast, P1f-P1ra and P2fa-P2r align to parts of both MTL α and MTL α . The long, horizontal line in P2fa-P2r represents a partial alignment of this Sanger sequence to MTL α and partial MTL α loci.

doi:10.1371/journal.pgen.1005626.g003

of the 11 *C. metapsilosis* strains, read-depth and mate-pair data indicated that MTL α haplotype appears to have the structure GAP1 –MTL α 2 –OBP α –PIKhybrid–MTL α 2 –MTL α 1–orf19.3202. In other words, a 6-kb region derived from MTL α and containing MTL α 2, OBP α and part of PIK α (in that order) has overwritten a nonhomologous 6 kb region that is normally present in MTL α and contains PAPA, OBPa and PIKa (in that order). Consistent with this introgression happening after the single hybridization event, the divergence between alleles in the introgressed region is very low (0.025%). These results were confirmed by PCR amplification and re-sequencing of five representative strains (SZMC21154, SZMC8094, SZMC8095, DNS94 (CP61), DNS100 (CP376)). The resulting haplotype contains an intact copy of MTL α 2 as well as MTL α 2 and MTL α 1, it has no PAP gene of any kind, and its PIK gene is chimeric (S10A and S10B Fig). Finally, SZMC21154 experienced an additional LOH removing the remaining MTL α cassette (MTL α 1, MTL α 2, PAPA), a situation which may be reminiscent of what may have occurred in the *C. orthopsilosis* hybrid. These results are in contrast with an earlier survey that reported that only the MTL α idiormorph was present in 18 isolates of *C. metapsilosis* [36] based on long-range PCR amplification. However, this discrepancy is likely to result from problems in the PCR in the former study, perhaps attributable to the introgression, because six of these 18 strains were fully sequenced in this study, and in all of them we found intact MTL α 1 and MTL α 2 (Fig 3).

Genomic variation within *C. metapsilosis*

Predicted duplications and deletions (copy number variations, CNVs) in sequenced strains were subjected to restrictive manual curation (see Materials and Methods), resulting in 84 deletions and 87 duplications relative to the reference assembly (S4 Table). Most CNVs affected coding regions, 71 deletions and 85 duplications, but we found no functional enrichment in deleted regions, while nucleic acid binding (GO:0003676) was enriched in duplicated regions. Nearly all CNVs are shared by more than one strain (82 deletions and 86 duplications) and 31

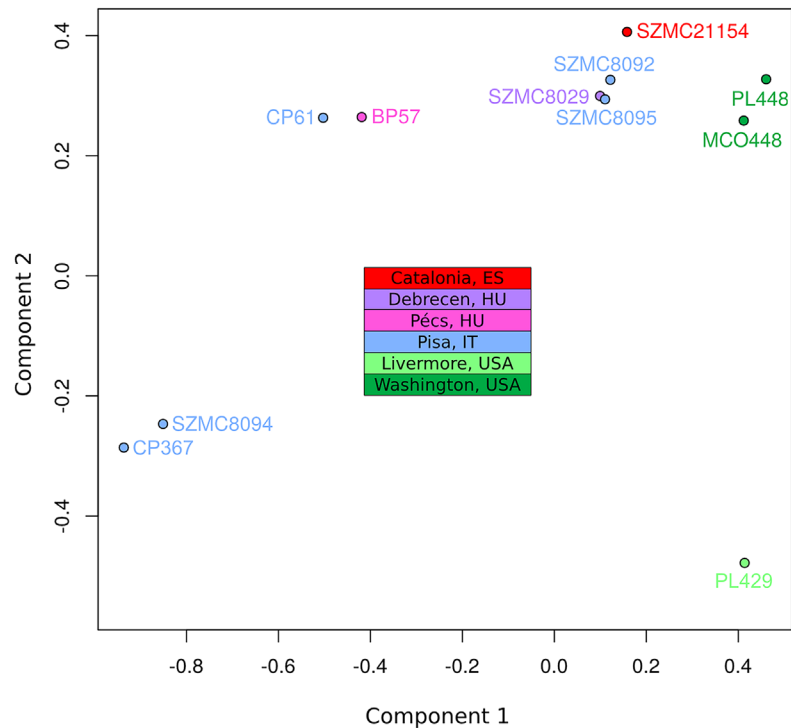


Fig 4. Multiple correspondence analysis (MCA) of genome diversity, based on 618,120 SNPs. The countries of isolation are indicated for all strains. SNPs from four genomic libraries of PL429 (pe300, pe400ov, pe600 and mp5000) were analysed separately, but all four libraries clustered together. For simplicity, the plot shows only PL429 results for one library (pe600).

doi:10.1371/journal.pgen.1005626.g004

deletions and 15 duplications are shared by all strains, which further support a common origin for all strains. The largest duplication (DUP65, 20 kb), spanning 13 genes involved in DNA binding, transport and various enzymatic activities, is shared by eight strains. The largest deletion (DEL8, 13,976 bp) is heterozygous and is shared by all eleven strains. This suggests that the deletion was either present in one parental undergoing hybridization or happened subsequently but prior to the divergence of the strains considered.

To assess the population structure of the species and reconstruct the diversification history of the different *C. metapsilosis* clinical isolates we employed several alternative strategies. First, multiple correspondence analyses (MCA) was conducted using the SNPs (Fig 4). Secondly, SNP based trees were reconstructed using either all SNPs or homozygous SNPs exclusively (S11A, S11B and S11C Fig). Finally, we reconstructed the evolutionary relationships among the strains using LOH blocks and CNVs as characters (S11D and S11E Fig). All of these approaches resulted in a roughly similar clustering of the strains. Particularly, the MCA shows four major groups: SZMC8094 and CP367 from Italy appear close to each other and well separated from the rest; CP61 (Italy) and BP57 (Hungary) appear close together, two strains derived from a single isolate from Washington State (PL448 and MCO448) group together, whereas PL429 from Livermore, California (US) is rather distant from all the rest. The largest cluster is formed by the rest of the strains from various origins (Italy, Hungary, Spain and US), although the two US strains in this cluster appear somewhat more distant from the rest. These broad separations are also apparent in the phylogenetic trees, particularly when branch lengths are considered. The observed clustering is also consistent with earlier reported clusters based on Amplification Fragment Length Polymorphisms (AFLP), for the five strains common in both studies [37].

The significant differences between the hybrid isolates and the certain degree of geographical structure suggest that this lineage is relatively ancient and spread globally a long time ago. This is in stark contrast with what has been found for the *C. orthopsilosis* hybrid lineage, where the two sequenced isolates from distant locations were found to be nearly identical [8]. Given our lack of knowledge on mutation rates, generation time and life cycle of *C. metapsilosis*, we can only speculate on the relative time of when the hybridization occurred. Of note we can only measure divergence among the sequenced strains, while hybridization must have predated this time as indicated by the number of shared events. The low level of differences found between the two most diverged isolates (MCO448 and SZMC8094) suggests that a limited number of point mutations accumulated during this time: 3 SNPs in the mitochondrial genome (0.0124% divergence) and 53 SNPs in the longest common homozygous region (0.01767%).

Comparative genomics of the *Candida parapsilosis* species complex

The availability of the genome sequence of *C. metapsilosis* allows, for the first time, to perform a comprehensive comparative genome analysis of the entire *C. parapsilosis* species complex. Whole genome alignments show that synteny is largely conserved among the three species of the complex (S12 Fig). Overall *C. parapsilosis* and *C. orthopsilosis* are closer to each other (98% of conserved synteny, 154 inversions) than either of them to *C. metapsilosis* (97%/231, and 96%/176 conserved synteny/inversions, respectively). We compared the gene content of the three species and compared them with other 23 sequenced Saccharomycetes by reconstructing the complete collection of evolutionary histories of the genes encoded in their genomes (i.e. the phylome) and establishing orthology and paralogy relationships among them [38]. The phylogenies, alignments and inferred homology relationships are available through phylomeDB [22]. We used 396 conserved, single-copy orthologs to reconstruct the evolutionary relationships among sequenced *Candida* species (Fig 5). This phylogeny was largely congruent with that of a super-tree derived from the whole phylome using a gene tree parsimony approach (S13 Fig). In contrast to earlier analyses based on a smaller sample of genes [25], but in line with the synteny analyses mentioned above and with an earlier phylogenetic analysis of mitochondrial genomes [29,31], our results support a basal position of *C. metapsilosis* to the exclusion of *C. orthopsilosis* and *C. parapsilosis*.

Predicted genes in the three species from the *Candida parapsilosis* species complex were grouped into 5,743 orthologous groups (including orthologs and in-paralogs), of which 5,045 (88%) are present across all three species. Of the widespread groups, 4,574 (91%) were present as one-to-one orthologs, while 226, 107 and 124 groups contained *C. metapsilosis*, *C. orthopsilosis* and *C. parapsilosis* specific paralogs, respectively. Differences in gene content for the three species are presented in (S5 Table). We here limit the discussion to several families considered relevant to explain virulence differences between species (Fig 5). The ability to form pseudohyphae has been associated to virulence in this clade [17]. Unlike *C. parapsilosis* and *C. orthopsilosis*, *C. metapsilosis* does not produce pseudohyphae [18], and this has been related to the lower virulence of the latter. Pseudohyphae production has been associated with two protein families: the cell-wall proteins Hyr/Iff and adhesion cell-surface glycoproteins ALS. Overall, *C. metapsilosis* encodes fewer members of Hyr/Iff family (13), than *C. parapsilosis* (17), but more than *C. orthopsilosis* (3 in 90–125 and 4 in MCO456) (see Phy00767CU tree in PhylomeDB). Interestingly, *C. metapsilosis* encodes one ALS gene, which is very close to ALS6 (CPAG_05054/CPAR2_404790), while the two *C. orthopsilosis* strains encode an ortholog closer to ALS7 (CPAG_05056/CPAR2_404800; see orf19.5736 in phylomeDB, phylome 464). Thus the lack of some ALS genes but not the number of members of Hyr/Iff family correlates with the inability to produce pseudohyphae and consequent lower virulence than in the remaining *C. parapsilosis* complex species.

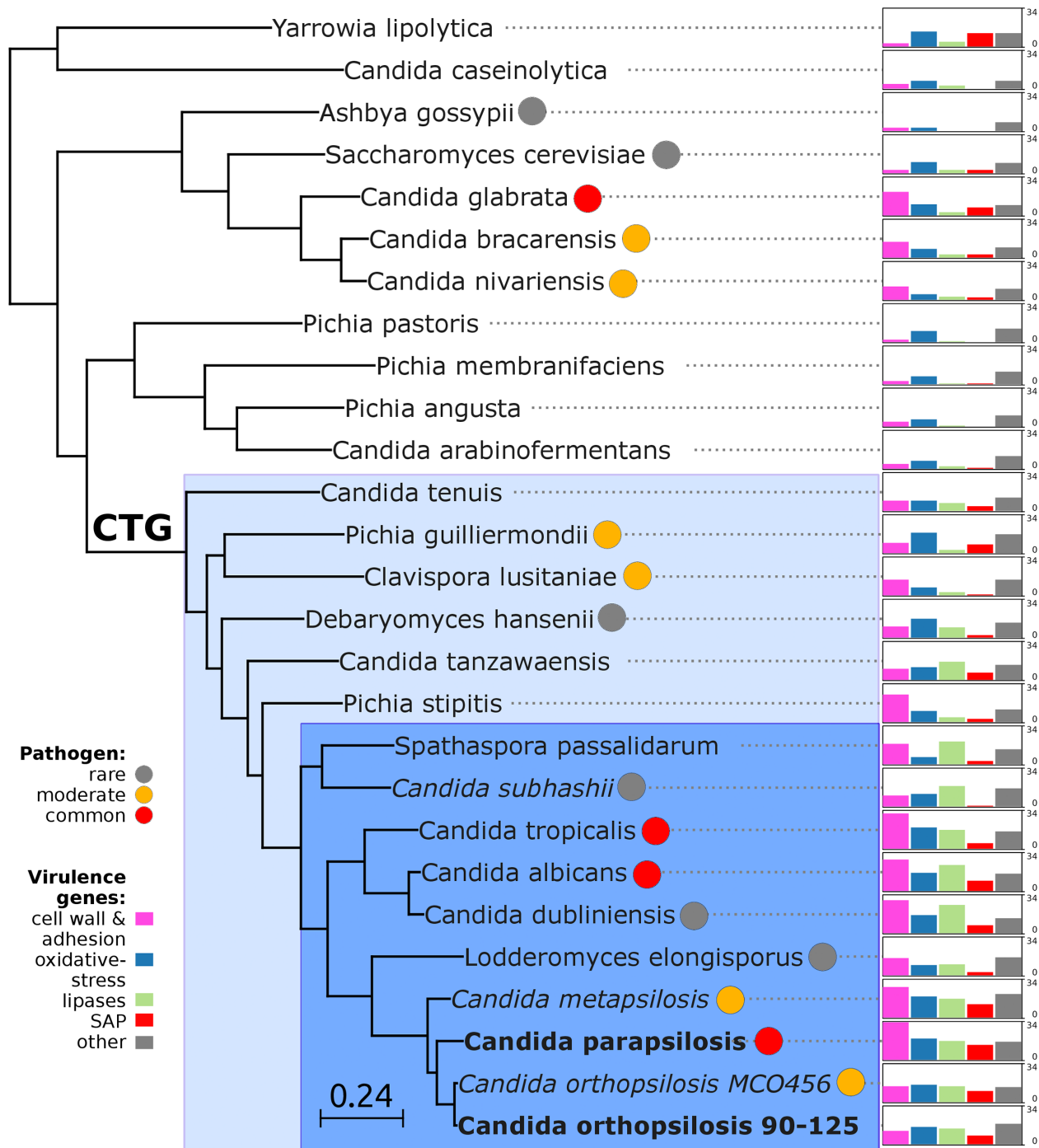


Fig 5. Phylogeny and genome composition in *Candida* and related species. Species translating CUG codon as serine (CTG clade) are denoted with color background: haploid species in light blue and diploid species in dark blue. Common human pathogens (red), moderate pathogens (orange), and rare pathogens (grey) are marked with color circles. Species with heterozygous genome are *in italic*, while extremely homozygous genomes are in bold. Barcharts with number of virulence-related genes for: i) cell wall (HYR/IFF) and adhesion (ALS/EPA) (pink), ii) oxidative-stress response (SOD, CAT, GPX, YBH) (blue), iii) lipases (LIP, PLB, SRR, FOX) (green), iv) secreted aspartic proteinases (SAP) (red) and v) other virulence-related genes (grey) including drug resistance genes (TPS, ERG), regulators (STE20, WH11, PHO100), iron acquisition (FTR1). Barcharts are scaled from 0 to 34 for all species.

doi:10.1371/journal.pgen.1005626.g005

We and others have previously shown that secreted lipases play an important role in the virulence of *C. albicans* and *C. parapsilosis* [18,40–42]. Despite earlier reports suggesting that *C. metapsilosis* strains were unable to produce extracellular lipases [19], we found that its genome actually codes for a similar number of secreted lipases (5) as *C. parapsilosis* and *C. orthopsilosis* (4). This indicates that the observed phenotypic differences may be due to different regulation of the lipase activity rather than to an inherent inability to produce secreted lipases. It has been demonstrated that individual lipase genes are differentially regulated in *C. albicans* during infection [43,44], and although we lack direct evidence on the function of *C. metapsilosis* lipase genes, it is appealing to speculate that a similar phenomenon exists in this species. Secreted aspartic proteases are also considered an important virulence factor in *C. albicans* and *C. parapsilosis* [45,46], and the other two species of the *Candida parapsilosis* complex have also been shown to exhibit this activity [18]. Interestingly, *C. metapsilosis* and *C. parapsilosis* encode more (14) secreted aspartic proteases (SAP) than any other *Candida* spp.: *C. orthopsilosis* (11 in MCO456 and 11 in 90–125) or *C. albicans* (10) (S5 Table and Phy0076724 in PhylomeDB). Although the capacity of *C. orthopsilosis* and *C. metapsilosis* SAPs to affect virulence has yet to be demonstrated, an extended SAP toolkit in species of the *C. parapsilosis* complex, which are not obligate commensals, may represent adaptation to both, environment and host. Similarly, *C. metapsilosis*, and *C. parapsilosis* encode more extracellular CFEM domain proteins (7), that are important for iron acquisition, than *C. albicans* (5). Overall, the broad functional class corresponding to cell wall and adhesion proteins seems to have a tendency to show larger numbers in pathogenic species as compared to less pathogenic ones, and is the only broad functional class that seems also expanded in non-CTG yeast pathogens, such as those in the *Nakaseomyces* clade [47].

Concluding remarks

Our results show compelling evidence that the opportunistic pathogen *C. metapsilosis* is a diploid hybrid species resulting from a single hybridization event, likely through mating, of two parental lineages that were ~4.5% divergent. This hybrid lineage expanded globally and currently isolated strains differ to a significant extent in their genomic background. Divergence has been mostly driven by differential LOH events but also by lineage-specific copy number variations, including large partial aneuploidies. Earlier studies based on AFLP have previously described a high degree of genetic heterogeneity among *C. metapsilosis* strains [37], which is consistent with our observations. In addition, our results provide a mechanistic basis for the source of this large heterogeneity: hybridization followed by differential LOH. It remains to be established, however, whether LOH is achieved mainly through mitotic recombination in clonal reproduction or whether sexual recombination or parasexual cycle also take place. In this respect most (10/11) of the strains harbor at least partial regions of both mating type loci, although in one of the analyzed strains the MTL_a locus has completely overwritten MTL_a. The levels of heterozygosity found for *C. metapsilosis* (22–26 SNPs/kb), are much larger than those described in *C. albicans* (2.5–3 SNPs/kb) [24]. However, considering the trend to lose heterozygosity with time, our findings open the question of whether the heterozygosity in *C. albicans*—and perhaps other highly heterozygous species in the CTG clade—may have originated via a similar process a longer time ago.

Our results argue for a re-definition of the species within the *C. parapsilosis* clade. We propose the existence of at least five different homozygous lineages and at least two hybrid lineages resulting from distinct combinations of the former. Two of the five homozygous lineages would be nowadays represented by *C. parapsilosis* and homozygous *C. orthopsilosis* Type 2 strains such as 90–125, whereas the remaining three are only partially represented by the

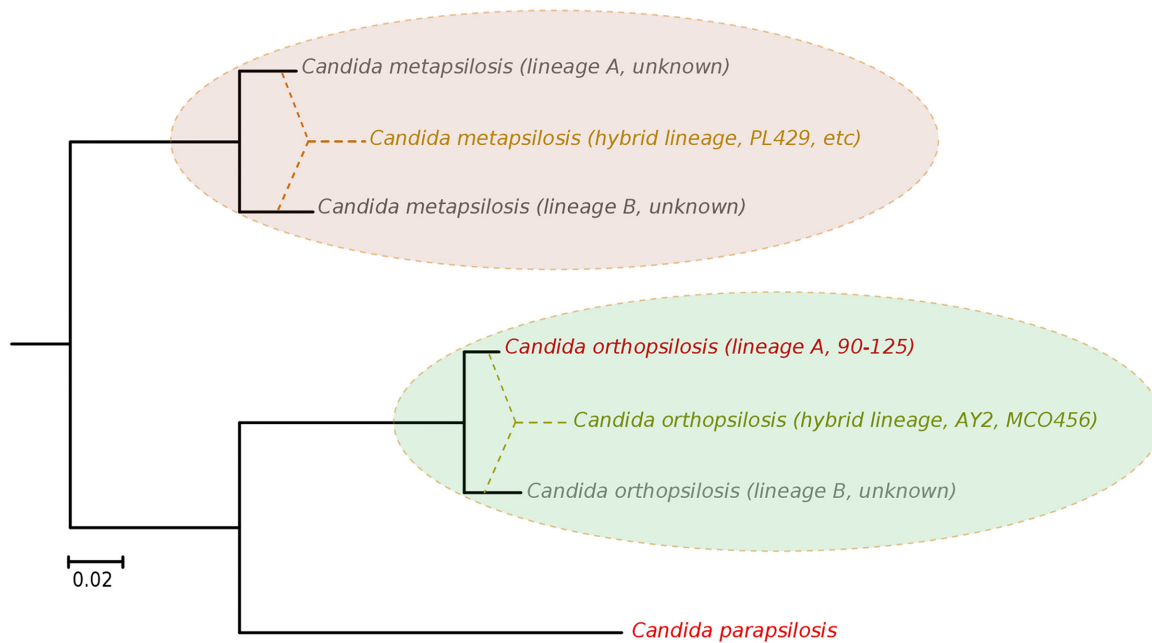


Fig 6. Phylogenetic relationships between *C. parapsilosis* complex lineages. Distance between the distinct lineages in *C. orthopsilosis* and *C. metapsilosis* have been inferred from heterozygous genomic regions in the hybrid strains. The origin of hybrid lineages have been schematically indicated by super-imposing dashed lines connecting different branches in the phylogeny. Strains representing the different lineages are indicated. Lineages whose existence is inferred from the hybrid strains but for which a representative strain is not available are indicated as unknown.

doi:10.1371/journal.pgen.1005626.g006

sequences of *C. orthopsilosis* MCO456 and the *C. metapsilosis* sequences presented in this work (Fig 6). Whether homozygous strains for these three lineages are extinct or still exist remains to be determined. Considering the pervasiveness of hybrid strains across *C. metapsilosis* clinical isolates, it can be suspected that the corresponding homozygous, parental lineages are not able to infect humans. This would imply that hybridization has resulted in the evolutionary emergence of the ability to colonize and infect humans by combining characteristics from two parental species that are not able to do so. Hybridization is known to drive the adaptation to new niches and this work emphasizes the idea that new pathogenic lineages can emerge through hybridization of non-pathogenic parental species. Ability to colonize humans may not necessarily be the key advantageous trait that promoted the survival of a new hybrid lineage, but rather human can be a secondary niche that can be exploited opportunistically. Alternatively, a higher ability to persist in humans may promote the survival of hybrids between species that can only sporadically colonize humans. An interesting idea is whether the stress environment provided by humans to species that are not well adapted may promote activation of mating competence, which in turn may open the way for inter-species crossings. In this respect, metagenomics analyses have identified *C. metapsilosis* in the normal microbiota of one healthy individual, among 20 investigated [48], although in the absence of a genome sequence we do not know whether this commensal strain was homozygous or heterozygous. In addition, the relative divergence between *C. metapsilosis* hybrid strains seems to indicate that this lineage did not emerge recently, and that it did not expand as a clinical outbreak. Rather, recursive infections from an already diverged population seem more plausible. If this is the case, the presence of similar hybrid lineages in the natural environment of *C. metapsilosis* is expected. The earlier reported case of *C. orthopsilosis* hybrid seems to be different, as two distant isolates had nearly identical sequences. Resolving these open questions requires analysis of the diversity present in healthy individuals and environmental samples, a topic which is largely unexplored.

Materials and Methods

DNA extraction

C. metapsilosis cultures were grown overnight in an orbital shaker (200 rpm, 30°C) in 2 ml YPD medium (0.5% (w/v) yeast extract, 1% (w/v) peptone, 1% (w/v) glucose) supplemented with 100 unit/ml penicillin-streptomycin solution (Sigma). Subsequently, cells were centrifuged (850 xg, 5 minutes) and were washed twice with 1x sterile PBS. The pellet was resuspended in 500 µl lysis buffer (1% (w/v) SDS, 50 mM EDTA, 100 mM TRIS pH = 8), 500 µl glass bead was added to the cells and were disrupted by using a vortex for 3 minutes. 275 µl 7M ammonium-acetate was added (65°C, 5 min) and the samples were then cooled on ice for 5 minutes. 500 µl of chloroform-isoamylalcohol (24:1) was added to the mixture, and the samples were centrifuged for 10 minutes at 16000 xg. The upper phase was transferred to a new microcentrifuge tube, and the previous step was repeated. 500 µl isopropanol was mixed with the upper phase in a new microcentrifuge tube, and the mixture was held in a refrigerator at -20°C for 5 minutes. The solution was centrifuged at 16000 xg for 10 minutes. The supernatant was discarded, and the pellet was washed twice with 500 µl 70% ethanol. After the second washing step the pellet was dried, and resuspended in 100 µl sterile bi-distilled water containing 250 µg/ml RN-ase (Sigma).

Genome sequencing

The genome sequences for the 11 strains were obtained at the Ultra-sequencing core facility of the CRG, using Illumina GAIIx, HiSeq2000 and MiSeq sequencing machines. For paired-end libraries, DNA was fragmented by nebulization or in Covaris to a size ~300 bp, ~400 bp, ~600 bp. The ends of the DNA fragments were blunted with T4 DNA polymerase and Klenow fragment (New England Biolabs), after shearing. DNA was purified with a QIAquick PCR purification kit (Qiagen). 3'-adenylation was performed by incubation with dATP and 3'-5'-exo-Klenow fragment (New England Biolabs). DNA was purified using MinElute spin columns (Qiagen) and double-stranded Illumina paired-end adapters were ligated to the DNA using rapid T4 DNA ligase (New England Biolabs). After another purification step, adapter-ligated fragments were enriched, and adapters were extended by selective amplification in an 18-cycle PCR reaction using Phusion DNA polymerase (Finnzymes). Libraries were quantified and loaded into Illumina flow-cells at concentrations of 7–20 pM. Cluster generation was performed in an Illumina cluster station. Sequence runs of 2x50, 2x76, 2x100 or 2x250 cycles were performed on the sequencing instrument. For the preparation of mate-pair libraries 15 micrograms of genomic DNA were sheared using a covaris instrument to the desired size range of 2.5 or 5 kb, respectively. Following size selection on a 0.8% agarose gel, the size fraction of interest was recovered, and library preparation was performed using a modification of the Illumina mate-pair preparation protocol, whereby a biotinylated double-stranded adapter was included in the circularisation reaction [49]. After circularisation, linear background was removed by exonuclease digestion, and the sample further fragmented in the covaris. Fragments that included the biotinylated adapter were enriched using streptavidin beads, and used to prepare an Illumina library. Sequencing was performed on an Illumina HiSeq 2000 sequencer using a 2 x 50 nt paired-end sequencing protocol. Base calling was performed using Illumina pipeline software. In multiplexed libraries, we used 4 bp internal indices (5' indexed sequences). De-convolution was performed using the CASAVA software (Illumina).

Genome assembly

Reads were pre-processed before assembly to trim at the first undetermined base or at the first base having PHRED quality below 10. We filtered out pairs with one (or both) reads shorter

than 31 bases after trimming. SOAPdenovo2 [50] was used to assemble paired-end reads into supercontigs with K-mer ranging from 31 to 91. As the initial assembly was very fragmented (2–4k scaffolds) and nearly twice as large (21–22 Mb) as other *C. parapsilosis* complex species (~13 Mb in 7–8 chromosomes), we assumed the presence of heterozygous regions in our strains. Heterozygous regions (9 Mb in 861 contigs) were removed by Haplomerger version 20120810 [51]. Subsequently, the remaining supercontigs were further scaffolded by SSPACE2 [52] and gaps were filled using GapCloser from the SOAPdenovo2 package. The random incorporation into the assembly of alternative heterozygous contigs was tested by repeating the whole procedure several times, starting from reads that were randomly re-ordered.

Genome annotation

Genes were predicted using Augustus version 2.5.5 [53] and *C. parapsilosis* CDC317 gene models for training [24]. Predicted gene models were curated using RNA-Seq reads to find evidence for exon-intron boundaries and exon skipping. Subsequently, we grouped predicted genes into orthologous groups and transferred functional annotation from one-to-one orthologs in model species i.e. *Candida albicans* or *Saccharomyces cerevisiae*, based on predictions from the MetaPhORs approach [54]. Finally, genes were further annotated using InterProScan 5RC4 [55].

Detection of SNPs, loss of heterozygosity, and divergence estimates

Genomic reads were aligned onto *C. metapsilosis* assembly using Bowtie2 with “very sensitive local alignment” mode [56]. SNPs and INDELS were called using GATK version 2.1–13 [57]. We filtered out clusters of five variants within 20 bases and low quality variants, as described in GATK documentation (QD < 2.0 || MQ < 40 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0). Subsequently, we divided the genome into three categories: unknown, heterozygous, and homozygous. Firstly, regions having lower (<75%) or higher (>125%) coverage than expected were assigned as unknown. Then, we marked heterozygous regions as those having two or more heterozygous sites closer than 100 bases. The remaining regions of the genome were considered homozygous, thus loss of heterozygosity (LOH) regions (100 bp threshold). A stricter threshold (200 bp) was established by considering LOH blocks shorter than 200 bp as heterozygous regions. Note that the two methods differ in expected number of false positive and false negatives. At 4.5% divergence the probability of having by chance a stretch of 100 bp with no heterozygous SNPs is 0.01 (0.955^{100}) and the expected number of LOH blocks of this length or longer in a 13.6 Mb genome is approximately 5,854 blocks, these numbers vary dramatically to 0.0001 and 58 when a 200 bp threshold is applied (Expected number is approximated as $E = Np^k$, where N is genome size in basepairs, p and q are the probabilities of having a SNP or not, respectively and k is the required size of the block). We compared the number of expected blocks of a given size or longer with the observed sizes in the real strains (S14 Fig). This made clear that a large number of short apparent LOH blocks (up to ~59%) may be artefactual when a threshold of 100 bp is applied (~2.9% at 200 bp). However, the stricter threshold would conversely discard many true LOH blocks (up to 97.1% of the discarded blocks would be real if we consider the observed–expected). Given that most differences affect blocks of short length, the effects of the total fraction of the genome assigned to homozygous or heterozygous regions is affected to a lesser extend. We further assessed the accuracy of our method by simulating 20 fully heterozygous genomes (13.6 Mb; 4.5% divergence between haplotypes) to estimate false positive of LOH detection (how often we would detect LOH in perfectly heterozygous genome using current settings). Heterozygous genomes were simulated using fasta2diverged.py v1.0 (https://github.com/Gabaldonlab/ngs_public) with the homozygous chromosome set of *C. metapsilosis*

assuming 4.5% divergence between haplotypes. LOH regions (at least 100bp regions having less than 2 SNPs) appear by chance in all simulations and on average sum up 782,850 bp (5.84% of the genome). Thus we estimate that less than 6% of the regions detected as LOH in our analysis correspond to false positives. Increasing LOH length cut-off from 100bp to 200bp decreases false positive rate to 0.13%, however. Throughout the manuscript we provide estimates based on both of these thresholds. The sequence divergence between the parental haplotypes was calculated as the number of heterozygous sites found in heterozygous regions divided by the cumulative length of heterozygous regions, using a 100 bp threshold.

Discrimination of linear and circular mitochondrial chromosomes

Two main characteristics allow *in silico* differentiation of linear and circular mtDNA chromosomes in *C. metapsilosis*. First, due to the presence of telomeres, consisting of tandem arrays of 620 bp repeat, linear chromosomes are expected to be longer (24,152 bp, NC_006971) than circular ones (22,175 bp, AY391853) [29]. Second, paired-end reads resulting from circular mitochondrial genomes when aligned to the termini of the linear mitochondrial chromosome reference assembly should have their partners aligned in the opposite end of the chromosome with discordant orientation (FF or RR instead for FR). To check this, genomic reads were aligned on the linear mitochondrial chromosome reference of *C. metapsilosis* MCO448 (NC_006971) as indicated above.

Experimental analysis of the mitochondrial genome topology

Mitochondrial DNAs in *C. metapsilosis* isolates were analysed by PFGE and PCR essentially as described in [30,31]. Briefly, whole-cell DNA samples were prepared in agarose blocks and separated in a 1.5% (w/v) agarose gel using a CHEF Mapper XA Chiller System (Biorad) with pulse switching set at 5 to 20 seconds (linear ramping) and 120° angle for 42 hours at 5 V/cm and 100°C (Fig 1A) or in a 1.0% (w/v) agarose gel in a Pulsaphor apparatus (LKB) in contour-clamped homogeneous electric field (CHEF) configuration with pulse switching from 5 to 50 seconds (interpolation) for 24 hours at 150 V and 90°C (Fig 1B). All separations were performed in 0.5x TBE buffer (45 mM Tris-borate, 1 mM EDTA, pH 8.0). Southern blots were hybridized with radioactively labeled probes derived from the mitochondrial genes *cox2* and *nad4*. For PCR analysis, the reactions contained diluted total cell DNA, 0.5 μM upstream and downstream primers (5'-ATTGTTGCTTTTGTGTTGA-3' and 5'-TTAGCTGTTGTTGCTATTACT-3' derived from the subtelomeric genes *nad3* and *atp6*, respectively), 0.2 mM dNTPs each, 1× reaction buffer with 2 mM MgCl₂, and 1 U of *Taq* DNA polymerase (Invitrogen). The amplification was performed using the following cycler profile: 3 min at 95°C; 25× (1 min at 94°C, 45 seconds at 56°C, 1 min at 68°C); 3 min at 72°C. The primers bind to the opposite subterminal regions of the linear mitochondrial genome and allow amplification of about 0.95 kb long fragment derived from the end-to-end junction of circularized genome forms (Fig 1C).

Detection of structural variants

Structural variants were detected using a methodology described and experimentally validated elsewhere [54], which we have implemented in bam2sv.py python script v1.0 (available at https://github.com/Gabaldonlab/ngs_public/). bam2sv.py v1.0 detects duplications, deletion and inversions by means of insert size deviations and incongruent read pairing between paired-end reads. In addition, duplications and deletions are detected from deviations from mean depth of coverage. All detected variants were manually curated. In addition, we generated genome graphs for all chromosomes illustrating copy number variation, as well as, heterozygous and homozygous regions (S1 File).

Experimental validation of a LOH region and MAT locus introgression

To validate the two parental sequences, we performed PCR and Sanger sequencing of the region scaffold2|size2959145:1,570,477–1,574,155 from *C. metapsilosis* genome in four different strains: BP57 (DNS25), CP61 (DNS94), CP376 (DNS100) and PL429. Four PCR primers were designed using Primer3 v4 webtool [58], namely two forward primers and two reverse primers, with four and three different bases among them, respectively, corresponding to allelic differences in each parental sequence (S5 Fig).

FWD_1: 5'-TTGACTGCTGAAGCTGTCTTGG-3';

FWD_2: 5'-TTAACCGCTGAAGTTGTCTTTGA-3';

REV_1: 5'-ATTCCATCTTGGCGCATCTT-3';

REV_2: 5'-GACTCCATCTTGACGAATCTTGG-3'.

Thus, four touchdown PCR reactions (FWD_1+REV_1, FWD_1+REV_2, FWD_2+REV_1 and FWD_2+REV_2) were carried out using Expand Long Range, dNTPack kit (Roche) according to manufacturer's instructions. Briefly, each reaction included primer concentration of 0.3 μM, 10 μl of 5X Buffer with MgCl₂ (final MgCl₂ concentration of 2.5mM), 2.5 μl of PCR Nucleotide Mix, 3% (v/v) of DMSO, 100 ng of DNA and 3.5 U of Enzyme Mix in a final volume of 50 μl. Cycling condition began with a warm-up step of 2 min at 92°C, followed by 15 cycles of 10 seconds at 92°C, 15 seconds at the corresponding annealing temperature per each pair of primers (decreasing 0.5°C each cycle) and 4 min at 68°C. The initial annealing temperatures were: 61.6°C, 64.1°C, 60°C, and 62.5°C for FWD_1+REV_1, FWD_1+REV_2, FWD_2+REV_1 and FWD_2+REV_2, respectively. Then, other 20 cycles of 10 seconds at 92°C, 15 seconds at the corresponding annealing temperature and 4 min at 68°C (increasing the extension time 20 seconds at each cycle) were set up, with a final extension step at 68°C for 7 minutes. PCR products were confirmed by 1.5% agarose gel electrophoresis and were then purified using QIAquick PCR Purification Kit (QIAGEN). Specific PCR products were only obtained when combining the primer sets FWD_1+REV_2, and FWD_2+REV_1 (amplicon size of 3678 bp), while no product was obtained when combining FWD_1+REV_1 or FWD_2+REV_2 primers (S6 Fig). The purified PCR products were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and sequencing products were precipitated and purified using EDTA 125mM, sodium acetate 3M and 100% ethanol. To cover the entire length of the products, apart from the PCR primers described above, further internal primers common to both parental sequences were designed and used for direct sequencing:

INT_FWD: 5'- AGTTTGAAATTGTCATCTTGGAT-3';

INT_REV: 5'- ACTTCTTTAAAATACTAACCACATCCATCTC-3'.

The MTL introgression was checked by means of PCR and Sanger sequencing in following strains of *C. metapsilosis*: SZMC21154, SZMC8092, SZMC8094, SZMC8095, CP61 (DNS94), CP367 (DNS100). We have designed 8 different primers:

MTLα₁_1F: 5'-GCCGCCTGAGAAGTATGAAG-3';

MTLα₁_1R: 5'-TTGGTGACCACAGGAAAACA-3';

MTLα₂_1F: 5'-GTGCTCCTCAAGCACAATCA-3';

MTLα₂_1R: 5'-GTCGCCAGACAACACTCTAGC-3';

MTLa₁_1F: 5'-GCTTGAGTGGGGATTGAGTC-3';

MTLa1_1R: 5'-AATCGTTTTTCGGGGTTTTCT-3';

MTLa2_1F: 5'-TCTCCGATTTCGTTCAATTC-3';

MTLa2_1R: 5'-CTTGACCCCAAAGCTTTCAA-3'.

The introgression was validated by 4 different PCRs (MTL α 1_1F –MTL α 1_1R, MTL α 2_1F –MTL α 2_1R, MTLa1_1F –MTLa1_1R and MTLa2_1F –MTLa2_1R) using Pfu DNA polymerase from PROMEGA. The reaction mixture consisted of 5 μ l of Buffer 10X with MgSO₄ and 1 μ l of 10 mM dNTPs, both provided by the manufacturer, 2 μ l 10 μ M of both forward and reverse primers, 0.4 μ l of Pfu DNA polymerase 3 U/ μ l and water filled up to 50 μ l. The PCR started with initial denaturation at 95°C for 2min, this was followed 30 cycles of 30 seconds at 92°C, 30 seconds at 61.0°C, 30 seconds at 72°C, it was finished with final extension for 5min at 72°C and cooled to 40°C. PCR products were confirmed by 1.5% agarose gel electrophoresis and were then purified using QIAquick PCR Purification Kit (QIAGEN).

Finally, in order to confirm linearity of MTL after introgression, we have amplified and Sanger sequenced genomic regions flanking introgression site. We have designed 6 primers:

P1f: 5'-TATCGGGTTTGAAGCTGCTC -3';

P1 α : 5'-GCTGAATGCTGGTTTTTGGT -3';

P2ra: 5'-TACTCCACGTTGTTTTGTAAAG-3';

P2f α : 5'-TTATGATTGGGATGGGTTGG -3';

P2fa: 5'-CACATTTGAATGGACGTTGG -3';

P2r: 5'-CGATAAATCAGCGCAACAAT-3'.

The linearity of introgression was validated by 4 different PCRs (P1f - P1 α , P1f—P2ra, P2f α —P2r and P2fa—P2r) using Pfu DNA Polymerase from PROMEGA. The reaction mixture consisted of 5 μ l of Buffer 10x with MgSO₄ and 1 μ l of 10 mM dNTPs, both provided by the manufacturer, 2 μ l 10 μ M of both forward and reverse primers, 0.4 μ l of Pfu DNA polymerase 3 U/ μ l and water filled up to 50 μ l. The PCR started with initial denaturation at 95°C for 2min, this was followed 30 cycles of 30 seconds at 92°C, 30 seconds at either 56.2°C (P1f - P1 α), 57.2°C (P1f –P2ra) or 54.2°C (P2f α –P2r and P2fa–P2r); 30 seconds at 72°C, it was finished with final extension for 5 min at 72°C and cooled to 40°C. PCR products were confirmed by 1.5% agarose gel electrophoresis and were then purified using QIAquick PCR Purification Kit (QIAGEN). Sanger sequencing was performed using an ABI Prism 3730xl DNA Analyzer (Applied Biosystems).

Pulsed-field gel electrophoresis

Yeasts cells were cultured for 24 hours, at 30°C, in an oxygen rich environment, in 5ml YPD (0.5% (w/v) yeast extract, 1% (w/v) peptone, 1% (w/v) dextrose) supplemented with 100 unit/ml penicillin-streptomycin (Sigma) in an orbital shaker (180 rpm). After 24 hours 100 μ l suspension was transferred to 5 ml YPD and the cells were incubated for 24 hours again, under the same conditions. Agarose blocks containing the intact chromosomes were prepared using the method of Schwartz & Cantor [59] with the following modifications. 1 ml of the yeast suspension was transferred in a sterile microcentrifuge tube, and was washed two times with 4°C 0.05 M EDTA (pH = 7.5) (Sigma) (2000 xg, 3 min). 1.3x10⁸ cells were resuspended in Isotonic Buffer (0.1 M phosphate–citrate buffer, equipped with 0.7 M sorbitol, 0.3 M mannitol, 0.001 M EDTA pH = 5.8) containing 1 M potassium-thioglycolate (Reanal), and it was incubated for 1 hour, at 30°C in an orbital shaker (180 rpm). The suspension was washed once with Isotonic

Buffer (2000 μ g, 3 min). The pellet was resuspended in 5 ml Isotonic Buffer containing 3% (w/v) Helicase and 0.5% (w/v) NovoZym 234 (Novo BioLabs), and it was incubated at 30°C, in a sterile 15 ml Falcon tube, overnight, in an orbital shaker (180 rpm). The spheroplasts were collected and washed once with Isotonic Buffer (300 μ g, 5 min). The pellet was resuspended in 42°C 0.125 M EDTA, mixed immediately with prewarmed (42°C) 2% (w/v) low-melting-point agarose (Sigma), then placed into a mould chamber. After solidification, inserts were incubated in 2 ml NDS buffer (1% (w/v) N-laurylsarcosine in 0.5 M EDTA, pH = 9.5) supplemented with 1 mg/ml Proteinase K (Sigma) at 50°C. During the two days of incubation the NDS-Proteinase K solution was replaced once. The inserts were washed once with 0.5 M EDTA (pH = 8) overnight, then were stored in 0.5 M EDTA (pH = 8) at 4°C until usage. Finally, yeast chromosomes were separated by CHEF [60] method by using Bio-Rad CHEF-DR II Drive Module, Bio-Rad Pulsefield 760 Modul and Bio-Rad Power Supply. The chromosomal DNA plugs were placed and separated in 0.9% (w/v) agarose gel (Sigma) prepared with filtered 0.5x TBE buffer with the following settings: 60–450 sec switching time, 90 V voltage, 168 h running time. 0.5x TBE was used as a running buffer and the temperature was kept at 10°C during the whole procedure. The buffer was replaced to fresh one 3 times during the running process. The gel was stained for 30 min in a 0.1% ethidium-bromide solution and was destained in distilled water overnight, at 4°C. The results were documented by using UVP Bio-Doc-It System.

Fluorescence-activated cell sorting (FACS) analysis

Ploidy analyses were confirmed using FACS for 12 *Candida* samples. Cells were grown in YPD medium at 30°C (overnight, 200 rpm), harvested, resuspended in deionized distilled water and fixed in ethanol at a 10^7 cells/ml concentration (overnight, 4°C). For the staining of cells with SYBR Green I, cells were first washed and resuspended in 750 μ l of 50 mM sodium citrate buffer, treated with 250 μ l of 1mg/ml RNase A solution for 1 hour at 50°C and finally with 50 μ l of 20 mg/ml proteinase K solution for 1 hour at 50°C. Then, 20 μ l of SYBR Green I (Life Technologies, diluted 1:10 th in Tris-EDTA buffer, pH 8.0) were added to the samples and they were stained overnight at 4°C protected from light. Triton X-100 was added at a final concentration of 0.25% (v/v) and samples were vortex. Finally, samples were sonicated (3 consecutive ultrasound pulsed at 30 W for 2 seconds with intervals of 2 seconds between each pulse) to eliminate most of the cell clumps before FACS analysis with FACScan at the FACS Unit from CRG/UPF.

Synteny analysis

Chromosomes/scaffolds of *C. parapsilosis* CDC317, *C. orthopsilosis* 90–125 and *C. metapsilosis* SZMC8094 were aligned with LAST aligner v189 [61]. Alignments shorter than 1kb were filtered out. Syntenic blocks were defined as contiguous regions in the same chromosome/scaffold of both genomes aligning over 10 kb, percentage of synteny is computed as the total length of syntenic blocks over the length of the genome. Synteny breaks were inferred if fragments longer than 10 kb from a contiguous region in one genome was aligning to at least two different chromosomes in the other species. Inversions in the query sequence were called when alignment direction changed within a given synteny block.

Phylome reconstruction

The evolutionary histories of all *C. metapsilosis* protein-coding genes were reconstructed in the context of 27 Saccharomycotina species (S6 Table), using the PhylomeDB pipeline [62]. In brief this pipeline proceeds as follows: first, homologs were retrieved using Smith-Waterman [63] with E-value cut-off of $1e-05$ and considering only sequences that aligned with at least 50% of their length. Subsequently, homologs were aligned using three programs: MUSCLE v3.8 [64],

MAFFT v6.712b [65] and KALIGN v2.04 [66] in two directions: forward and reverse. The six resulting alignments were combined with M-COFFEE [67] and finally trimmed using trimAl v1.3 [68] applying consistency cutoff of 0.1667 and a gap score cutoff of 0.1. Neighbour Joining trees were reconstructed and the likelihood of obtained topology was computed, allowing branch-length optimisation, using seven different models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff), as implemented in PhyML 3.0 [69]. One evolutionary model best fitting the data was determined for each alignment by comparing the likelihood of the used models according to the AIC criterion. Finally, Maximum Likelihood (ML) trees were inferred for selected models. In all cases a discrete gamma-distribution model with four rate categories plus invariant positions was used, the gamma parameter and the fraction of invariant positions were estimated from the data. Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. All alignments and trees generated have been deposited in PhylomeDB with ID 243 [39]. Orthologs predicted for *C. metapsilosis* genes can be retrieved from MetaPhOrs database [54].

Reconstruction of evolutionary relationships among strains

Five separate analyses were used to reconstruct the evolutionary relationships among the sequenced strains. Alignments were reconstructed from SNP information in three different types of regions: (i) Patterns from homozygous SNPs present in the longest LOH block that is shared by all the strains; (ii) SNP patterns from the whole genome. Heterozygous SNPs were encoded using the base which is alternative to that in the reference, and choosing randomly one of the alternative bases in heterozygous SNPs if both were different from the reference; and (iii) Haplotype patterns based on three-state haplotype (heterozygous, hapA, hapB) assignment in windows of 1 kb, resulting in 1,798 phylogenetically informative patterns; Separately, character-based matrices were reconstructed using (iv) 127 ploidy patterns from 84 deletions and 87 duplications (S4 Table); ploidy state was coded as 0 for null deletion, 1 for heterozygous deletion, 2 for wild-type (no deletion and duplication), 3 for duplication (3 copies of given locus), 4 for duplication (4 copies of given locus), and so on; and (v) 587 LOH presence and absence patterns. Maximum Likelihood phylogenetic trees were reconstructed from these alignments using RAxML 7.2.8 using GTRCAT for sequence alignments (datasets i to iii), and GTRGAMMA for multi-state character matrices (datasets iv and v) [70]. Finally, multiple correspondence analysis (MCA, an extension of Principal Component Analysis to categorical data) of 618,120 SNPs with at least 10x coverage was performed using *ade4* package from R [71].

Reconstruction of species phylogeny

Two different approaches were used to reconstruct a species tree of *Candida* species and relatives. First a parsimony-based super-tree was reconstructed using duptree v1.48 [72] from the 5,780 gene trees from *C. metapsilosis* phylome available at PhylomeDB id 243 [39]. Then a Maximum Likelihood (ML) tree was reconstructed with PhyML v3.0 [69] and the Jones-Taylor-Thornton (JTT) evolutionary model, based on an amino acid super-matrix of 232,760 columns resulting from concatenating the alignments of 396 one-to-one orthologs. The two species trees are nearly identical with Robinson-Foulds symmetric distance of 6, meaning both trees share 50 out of 56 possible splits. All phylogenies were visualised with ETE [73].

Data deposition

Sequencing data, genome assembly and annotation have been deposited to EBI-ENA (accession: PRJNA238968). Reconstructed trees and alignments were deposited to PhylomeDB (<http://phylomedb.org>) as PhylomeDB ID 243.

Supporting Information

S1 Fig. PFGE chromosomal patterns of selected *C. metapsilosis* strains. The chromosomes of several *C. metapsilosis* strains were subjected to Pulse Field Gel Electrophoresis (PFGE). Two *C. parapsilosis* strains were included as reference. Chromosomes with potential rearrangements are denoted with yellow asterisks. Sequenced strains are marked with black asterisks. (PDF)

S2 Fig. Topological analysis of mitochondrial chromosomes. A. Genomic reads from all strains were aligned on the linear mitochondrial chromosome reference of *Candida metapsilosis* MCO448 (NC_006971). Telomeres (marked with blue squares) are expected only in linear chromosomes, thus strains with circular mitochondrial chromosome (PL448 and SZMC21154) miss parts of telomeres. In addition, read pairs aligned on the ends of the reference present discordant pairing (marked in green) in strains with circular mitochondrial chromosome. B. The most of *C. metapsilosis* strains examined in this study contain a linear mitochondrial genome with the length of about 23 kbp (Kosa et al., 2006). The genes *atp6* and *nad3* are located in the right and left subterminal region, respectively, and the telomeres on both ends of linear DNA molecules consist of a subterminal repeat (STR, 358 bp, shown as blue rectangle) and a tandem repeat array (TRA, nx 620 bp, red rectangle). The STR contains a sequence cluster rich in guanine and cytosine residues (GC-box, black rectangle), which is presumably involved in the mitochondrial telomere maintenance via homologous recombination (Nosek et al., 2005; Gerhold et al. 2014). In contrast, the strains PL448 and SZMC21154 contain circular mitochondrial genomes that lack the right telomere and most of the tandem repeat array of the left telomere. The sequence analysis indicates that these genomes represent circularized mutants resulting from end-to-end fusions of originally linear DNA molecules. Since the sites where the fusion events occurred are different, we assume that these mutants emerged independently. (PDF)

S3 Fig. Size distribution of LOH blocks. Histogram of LOH block sizes for *C. metapsilosis* PL429 (blue) and *C. orthopsilosis* MCO 456 (grey). Only LOH shorter than 10 kb are shown. Note, Y axis is log-scaled. (PDF)

S4 Fig. Distribution of LOH block length and chromosome position. Scatter plot representing the size of loss of heterozygosity (LOH) regions and their position on the chromosome given as percentage of chromosome length. Colors represent dots density. Note, two replicas for PL429 (pe300 and pe600) are given in two separate panels. (PDF)

S5 Fig. Strategy to validate LOH with PCR. Details for the independent amplification of an heterozygous region (3678 bp) from two homologous chromosomes. A. Sets of primers (F1, R1 and F2, R2) targeting known heterozygous sequences that flank the region of interest were designated. Forward (F1, F2) and reverse (R1, R2) primers differ by three SNP, so all primers align uniquely to a specific homologous chromosome. Four PCR reactions were performed for each strain. For these only two PCR reactions gave a product (positive controls): P1 obtained from combination of F1 and R2; and P2 from F2 and R1. The other two combinations of primers (F1R1 and F2R2) did not work as the primers anneal on homologous chromosome and not on the same chromosome (negative controls). B. Obtained PCR products are too larger to be sequence by single Sanger reaction. We thus have designed two additional primers (M1F, M1R) that target homozygous region in the middle of PCR products, so they can be used with both PCR products. For each PCR product (P1, P2) we performed four

Sanger sequencing reactions:

- For F1R2 amplification product (P1):
 - pos4_F1 primer amplification (F1),
 - pos4_R2 primer amplification (R2),
 - pos4_M1F (M1F_12),
 - pos4_M1R (M1R_12)
- For F2R1 amplification product:
 - pos4_F2 primer amplification (F2),
 - pos4_R1 primer amplification (R1),
 - pos4_M1F (M1F_21),
 - pos4_M1R (M1R_21)

S6 Fig. PCR of the 3.6kb region (scaffold2|size2959145:1,570,477–1,574,155) from four *C. metapsilosis* strains: BP57 (DNS25), CP61 (DNS94), CP367 (DNS100) and PL429 have been performed. Two forward primers and two reverse primers, with four and three different bases among them, respectively, corresponding to allelic differences in each parental sequence/genotype were designed. Thus, four touchdown PCR reactions (FWD_1+REV_1, FWD_1+REV_2, FWD_2+REV_1 and FWD_2+REV_2) were carried out. Specific PCR products were only obtained when combining the primer sets FWD_1+REV_2, and FWD_2+REV_1 (amplicon size of 3678 bp), while no band was seen when combining FWD_1+REV_1 or FWD_2+REV_2 primers.

S7 Fig. LOH donor. In order to test whether LOH donor is randomly or preferentially selected, we have analysed in detail 3.6 kb genomic region (scaffold2|size2959145:1,570,477–1,574,155) harboring eight LOH in four *C. metapsilosis* strains: BP57 (DNS25), CP61 (DNS94), CP367 (DNS100) and PL429. A) Following features are given for each strain: genomic read density and alignments, followed by aligned Sanger sequences from PCR product 1 (P1) and product 2 (P2). LOH events are marked by red rectangles or red arrow. The recombination donor is marked by green arrow. All recombinations in the analysed region originated from PCR product 2 (P2). B) Schematic representation of LOH events. Six LOH events (#1–#6) are common to three strains (BP57, CP61, PL429), while the remaining two (#7 and #8) are present only in PL429.

S8 Fig. Ploidy analyses based on frequencies of biallelic SNPs. Distributions of frequencies of read counts at biallelic SNPs in *C. metapsilosis* strains. (A) We observed single peak close to 50% for all chromosomes of *C. metapsilosis* PL429, indicating it is diploid. The remaining strains are also diploid, with exceptions in some chromosomes: (B) PL448 scaffold5 is present in three copies (two peaks at 33% and 66%), scaffold6 underwent LOH (very few biallelic SNPs and no peaks) and end of scaffold2 (2–2.9Mb) is present in three copies (two peaks at 33% and 66% visible beside main peak at 50%); and (C) SZMC21154 scaffold2 from 0 to 2Mb is present in three copies while the remaining 0.9 Mb in two copies (two peaks at 33% and 66% beside main peak at 50%).

S9 Fig. Ploidy analyses based on flow cytometry. Twelve *Candida* samples were analyzed using fluorescence-activated cell sorting (FACS): 9 strains of *C. metapsilosis*, 1 *C. parapsilosis*, 1 *C. orthopsilosis* and 1 *C. albicans*. In the histogram, FITC-A values—fluorescence signal of the fluorochrome fluorescein isothiocyanate—corresponding to DNA content versus cell counts are plotted. Peaks around 50 K and 100 K values of FITC-A account for cells in G1 and G2 phases, respectively. In the table, FITC-A G1 and G2 medians are shown per each of the samples, as well as the ratio between them.
(PDF)

S10 Fig. MTL organisation. A. We have confirmed presence of MTL α / α idiomorphs in six *C. metapsilosis* strains by four PCR reactions targeting: *MTL α 1* (150 bp product), *MTL α 2* (188 bp product), *MTL1a* (219 bp product) and *MTL2a* (216 bp product). *MTL α 1* and *MTL α 2* are present in all tested strains. SZMC21154 does not have the *MTL1a* not *MTL2a* fragments (marked with arrows). B. We have confirmed linearity of MTL α / α idiomorphs in two *C. metapsilosis* strains (SZMC8092 and SZMC21154) by four PCR reactions flanking introgression ends: P1f - P1 α (932 bp product), P1f - P1ra (900 bp product), P2f α -P2r (1009 bp product) and P2fa-P2r (1076 bp product). MTL α is present in both strains, thus P1f - P1 α and P2f α -P2r yield PCR product in both strains. P1f - P1ra and P2fa-P2r do not give product in SZMC21154, as MTL α is missing in this strain.
(PDF)

S11 Fig. Phylogenetic relationships of *Candida metapsilosis* strains. Phylogenetic trees were reconstructed from matrices consisting of: A) SNPs in the longest (300kb) homozygous regions (51 patterns), B) concatenated chromosomes with incorporated SNP (3,406 patterns), C) 13,374 three-state (hapA, hapB, heterozygous) haplotypes in 1 kb windows (1,798 patterns), D) 170 multi-state CNVs: 0 for null deletion, 1 for heterozygous deletion, 2 for wild-type (no deletion and duplication), 3 for duplication (3 copies of given locus), 4 for duplication (4 copies of given locus) etc (127 patterns), and E) 8,889 LOH presence / absence profiles (587 patterns). Bootstrap support values are given if lower than 100. Strains are color-coded, accordingly to the place of isolation. Trees were visualised using iTOL (Letunic & Bork, 2011). Note, here we consider patterns as phylogenetically informative loci, this is alleles that are shared by more than one strain, but not present in all of them. Maximum Likelihood phylogenetic trees were reconstructed from these alignments using RAxML 7.2.8 using GTRCAT model for all except multi-state matrices and GTRGAMMA model for multi-state matrices.
(PDF)

S12 Fig. Chromosome alignments. Chromosomes of *Candida parapsilosis* complex species were aligned using nucmer 3.0.7 (Kurtz *et al.*, 2004): A) *C. orthopsilosis* 90–125 against *C. parapsilosis* CDC317, B) *C. metapsilosis* SZMC8094 against *C. parapsilosis* CDC317 and C) *C. metapsilosis* SZMC8094 against *C. orthopsilosis* 90–125. Forward alignments are colored in red, while reverse alignments in blue. Chromosomes of all three species are syntenic with 8–9 translocations (marked with arrows on the right Y axis) and numerous small inversions.
(PDF)

S13 Fig. Species tree of *Candida* spp. Parsimony-based species tree were reconstructed based on the topologies of 5,780 gene trees from *C. metapsilosis* phylome (http://phylomedb.org/phylome_243) using duptree v1.48 (Wehe *et al.*, 2008).
(PDF)

S14 Fig. Expected vs observed LOH block sizes. Distributions of expected (blue dashed line) and observed LOH blocks from 11 strains for given block size cut-off were plotted. Expected

distribution is approximated as $E = N * p * q * k$, where N is genome size in basepairs, p and q are the probabilities of having a SNP or not, respectively and k is the required size of the block. (PDF)

S1 Table. De novo assembly results. For each *C. metapsilosis* strain the table provides: its name, available genomic libraries, read length(s), cumulative depth of coverage, k-mer(s) used for assembly, obtained number of contigs, cumulative assembly size, percentage of GC content, number of contigs longer than 1 kb and the cumulative size of these contigs, N50, N90, the cumulative size of gaps and the length of the longest contigs. (PDF)

S2 Table. Heterozygous assembly statistics. For each step of heterozygous assembly the table provides: name, result file name(s), number of contigs, cumulative assembly size, percentage of GC content, number of contigs longer than 1 kb and the cumulative size of these contigs, N50, N90, the cumulative size of gaps and the length of the longest contigs. (PDF)

S3 Table. Depth of coverage analysis. Estimated ploidy of all chromosomes/scaffolds is provided for all analysed strains. *C. metapsilosis* is diploid, thus wild-type ploidy of 2 is expected. Chromosomes with deviated ploidy are denoted in red (putative duplication) or blue (putative deletion). Note that rDNA cluster is placed in scaffold5, which therefore shows a larger ploidy in all strains. This, however, does not preclude us to detect triploidy of scaffold5 in PL448. Variability in mitochondrial chromosome (scaffold10) copy number is likely reflecting differences in sample preparation and not true biological variability. (PDF)

S4 Table. Copy number differences between *C. metapsilosis* strains. For each event duplication and deletion the table lists: its genomic coordinate and size, number of affected strains and their names; estimations of ploidy for these strains in this region; information whether deletion appears as null in at least one strain; affected genes and their functions. For simplicity, several functional classes are provided with a specific color background. Some selected CNVs among those longer than >5kb are annotated in [S1 File](#). (PDF)

S5 Table. Gene copy number differences between *C. parapsilosis*, *C. orthopsilosis* and *C. metapsilosis*. Clusters of orthologous genes that differ in the number of members between *C. parapsilosis* species complex are given. For each cluster, the table lists: cluster ID, total number of genes in the cluster, number of members and their IDs for *C. metapsilosis*, *C. orthopsilosis* and *C. parapsilosis*; functional annotations for cluster members; difference in number of members between *C. metapsilosis*, *C. parapsilosis* and *C. orthopsilosis*. For simplicity, several functional classes are provided with color background. (PDF)

S6 Table. Table listing the proteomes used to reconstruct *C. metapsilosis* phylome 243. For each species, the table lists: NCBI Taxonomy species identifier, species name, PhylomeDB proteome code, genome source and date when the genome was downloaded. (PDF)

S1 File. *C. metapsilosis* genome graphs. For each chromosome we have plotted: i) coding genes for +/- strand (grey bars) and GC-content in 1kb windows (blue plot) in the bottom track and log2 of observed vs expected value in 1kb windows for depth of coverage (blue) in the top fourteen tracks. In addition, loss of heterozygosity (LOH) regions have been marked in grey, if

the same genotype as reference was kept (hapA), and orange, if alternative genotype was kept (hapB). Four replicas (pe300, pe600, mp500 and pe400ov) were analysed for PL429. *C. metapsilosis* genome is a mixture of heterozygous (light grey), haplotype B (dark grey) and haplotype A (orange) regions. We suspect all analysed strains originate from single hybridisation event, as most of LOH events are shared by all strains. Examples of large LOH, duplications and deletions have been annotated ie. rDNA cluster (scaffold5), scaffold5 triploidy in PL448, partial scaffold2 triploidy in SZMC21154 and PL448, and complete LOH in scaffold6 in PL448. For the sake of simplicity, only some selected duplications and deletions among those longer than 5Kb are represented (annotated as such in Supplementary [S4 Table](#)). rDNA cluster is found on the edge of the largest LOH (over 350kb, scaffold5). Interestingly, we have also found rDNA cluster in long (200kb) LOH track in *C. orthopsilosis* MCO448 (Pryszcz et al., 2014).

(PDF)

S1 Text. Citations for the supplementary material.

(DOC)

Acknowledgments

The authors wish to thank Heinz Himmelbauer and the genomics facility at CRG, as well as Òscar Fornas and the FACS facility at CRG, for their technical assistance. Geraldine Butler is acknowledged for providing sequences from the Fosmid Library.

Author Contributions

Conceived and designed the experiments: TG AG. Performed the experiments: LPP EK ES EH TN TG. Analyzed the data: LPP EK ES EH TN TG KHW JN. Wrote the paper: LPP EK ES EH TN TG KHW JN.

References

- Morales L, Dujon B. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol Mol Biol Rev* MMBR. 2012; 76: 721–739. doi: [10.1128/MMBR.00022-12](https://doi.org/10.1128/MMBR.00022-12) PMID: [23204364](https://pubmed.ncbi.nlm.nih.gov/23204364/)
- González SS, Barrio E, Gafner J, Querol A. Natural hybrids from *Saccharomyces cerevisiae*, *Saccharomyces bayanus* and *Saccharomyces kudriavzevii* in wine fermentations. *FEMS Yeast Res*. 2006; 6: 1221–1234. PMID: [17156019](https://pubmed.ncbi.nlm.nih.gov/17156019/)
- 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 Bethesda Md*. 2012; 2: 299–311.
- Xu J, Luo G, Vilgalys RJ, Brandt ME, Mitchell TG. Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD. *Microbiol Read Engl*. 2002; 148: 203–212.
- Baker E, Wang B, Bellora N, Peris D, Hulfachor AB, Koshalek JA, et al. The Genome Sequence of *Saccharomyces eubayanus* and the Domestication of Lager-Brewing Yeasts. *Mol Biol Evol*. 2015;
- Wendland J. Lager yeast comes of age. *Eukaryot Cell*. 2014; 13: 1256–1265. doi: [10.1128/EC.00134-14](https://doi.org/10.1128/EC.00134-14) PMID: [25084862](https://pubmed.ncbi.nlm.nih.gov/25084862/)
- Marcet-Houben M, Gabaldón T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biol*. 2015; 13: e1002220. doi: [10.1371/journal.pbio.1002220](https://doi.org/10.1371/journal.pbio.1002220) PMID: [26252497](https://pubmed.ncbi.nlm.nih.gov/26252497/)
- Pryszcz LP, Németh T, Gácsér 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. doi: [10.1093/gbe/evu082](https://doi.org/10.1093/gbe/evu082) PMID: [24747362](https://pubmed.ncbi.nlm.nih.gov/24747362/)
- Bovers M, Hagen F, Kuramae EE, Diaz MR, Spanjaard L, Dromer F, et al. Unique hybrids between the fungal pathogens *Cryptococcus neoformans* and *Cryptococcus gattii*. *FEMS Yeast Res*. 2006; 6: 599–607. PMID: [16696655](https://pubmed.ncbi.nlm.nih.gov/16696655/)

10. Hagen F, Khayhan K, Theelen B, Kolecka A, Polacheck I, Sionov E, et al. Recognition of seven species in the *Cryptococcus gattii*/*Cryptococcus neoformans* species complex. *Fungal Genet Biol* FG B. 2015; 78: 16–48. doi: [10.1016/j.fgb.2015.02.009](https://doi.org/10.1016/j.fgb.2015.02.009) PMID: [25721988](https://pubmed.ncbi.nlm.nih.gov/25721988/)
11. Trofa D, Gácser A, Nosanchuk JD. *Candida parapsilosis*, an emerging fungal pathogen. *Clin Microbiol Rev*. 2008; 21: 606–625. doi: [10.1128/CMR.00013-08](https://doi.org/10.1128/CMR.00013-08) PMID: [18854483](https://pubmed.ncbi.nlm.nih.gov/18854483/)
12. Tavanti A, Davidson AD, Gow NAR, Maiden MCJ, Odds FC. *Candida orthopsilosis* and *Candida metapsilosis* spp. nov. to replace *Candida parapsilosis* groups II and III. *J Clin Microbiol*. 2005; 43: 284–292. PMID: [15634984](https://pubmed.ncbi.nlm.nih.gov/15634984/)
13. Cantón E, Pemán J, Quindós G, Eraso E, Miranda-Zapico I, Álvarez M, et al. Prospective multicenter study of the epidemiology, molecular identification, and antifungal susceptibility of *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* isolated from patients with candidemia. *Antimicrob Agents Chemother*. 2011; 55: 5590–5596. doi: [10.1128/AAC.00466-11](https://doi.org/10.1128/AAC.00466-11) PMID: [21930869](https://pubmed.ncbi.nlm.nih.gov/21930869/)
14. Lockhart SR, Messer SA, Pfaller MA, Diekema DJ. Geographic distribution and antifungal susceptibility of the newly described species *Candida orthopsilosis* and *Candida metapsilosis* in comparison to the closely related species *Candida parapsilosis*. *J Clin Microbiol*. 2008; 46: 2659–2664. doi: [10.1128/JCM.00803-08](https://doi.org/10.1128/JCM.00803-08) PMID: [18562582](https://pubmed.ncbi.nlm.nih.gov/18562582/)
15. Oliveira VKP, Paula CR, Colombo AL, Merseguel KB, Nishikaku AS, Moreira D, et al. Candidemia and death by *Candida orthopsilosis* and *Candida metapsilosis* in neonates and children. *Pediatr Neonatol*. 2014; 55: 75–76. doi: [10.1016/j.pedneo.2013.07.006](https://doi.org/10.1016/j.pedneo.2013.07.006) PMID: [24113226](https://pubmed.ncbi.nlm.nih.gov/24113226/)
16. Gácser A, Schäfer W, Nosanchuk JS, Salomon S, Nosanchuk JD. Virulence of *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* in reconstituted human tissue models. *Fungal Genet Biol* FG B. 2007; 44: 1336–1341. PMID: [17391997](https://pubmed.ncbi.nlm.nih.gov/17391997/)
17. Gago S, García-Rodas R, Cuesta I, Mellado E, Alastruey-Izquierdo A. *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* virulence in the non-conventional host *Galleria mellonella*. *Virulence*. 2014; 5: 278–285. doi: [10.4161/viru.26973](https://doi.org/10.4161/viru.26973) PMID: [24193303](https://pubmed.ncbi.nlm.nih.gov/24193303/)
18. Németh T, Tóth A, Szenzenstein J, Horváth P, Nosanchuk JD, Grózer Z, et al. Characterization of virulence properties in the *C. parapsilosis* sensu lato species. *PloS One*. 2013; 8: e68704. doi: [10.1371/journal.pone.0068704](https://doi.org/10.1371/journal.pone.0068704) PMID: [23874732](https://pubmed.ncbi.nlm.nih.gov/23874732/)
19. Orsi CF, Colombari B, Blasi E. *Candida metapsilosis* as the least virulent member of the “*C. parapsilosis*” complex. *Med Mycol*. 2010; 48: 1024–1033. doi: [10.3109/13693786.2010.489233](https://doi.org/10.3109/13693786.2010.489233) PMID: [20507266](https://pubmed.ncbi.nlm.nih.gov/20507266/)
20. Bertini A, De Bernardis F, Hensgens LAM, Sandini S, Senesi S, Tavanti A. Comparison of *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* adhesive properties and pathogenicity. *Int J Med Microbiol IJMM*. 2013; 303: 98–103. doi: [10.1016/j.ijmm.2012.12.006](https://doi.org/10.1016/j.ijmm.2012.12.006) PMID: [23403338](https://pubmed.ncbi.nlm.nih.gov/23403338/)
21. Garcia-Effron G, Canton E, Pemán J, Dilger A, Romá E, Perlin DS. Epidemiology and echinocandin susceptibility of *Candida parapsilosis* sensu lato species isolated from bloodstream infections at a Spanish university hospital. *J Antimicrob Chemother*. 2012; 67: 2739–2748. doi: [10.1093/jac/dks271](https://doi.org/10.1093/jac/dks271) PMID: [22868644](https://pubmed.ncbi.nlm.nih.gov/22868644/)
22. Bonfietti LX, Martins M dos A, Szeszs MW, Pukiskas SBS, Purisco SU, Pimentel FC, et al. Prevalence, distribution and antifungal susceptibility profiles of *Candida parapsilosis*, *Candida orthopsilosis* and *Candida metapsilosis* bloodstream isolates. *J Med Microbiol*. 2012; 61: 1003–1008. doi: [10.1099/jmm.0.037812-0](https://doi.org/10.1099/jmm.0.037812-0) PMID: [22493277](https://pubmed.ncbi.nlm.nih.gov/22493277/)
23. Nosek J, Holesova Z, Kosa P, Gacser A, Tomaska L. Biology and genetics of the pathogenic yeast *Candida parapsilosis*. *Curr Genet*. 2009; 55: 497–509. doi: [10.1007/s00294-009-0268-4](https://doi.org/10.1007/s00294-009-0268-4) PMID: [19662416](https://pubmed.ncbi.nlm.nih.gov/19662416/)
24. 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. doi: [10.1038/nature08064](https://doi.org/10.1038/nature08064) PMID: [19465905](https://pubmed.ncbi.nlm.nih.gov/19465905/)
25. Riccombeni A, Vidanes G, Proux-Wéra E, Wolfe KH, Butler G. Sequence and analysis of the genome of the pathogenic yeast *Candida orthopsilosis*. *PloS One*. 2012; 7: e35750. doi: [10.1371/journal.pone.0035750](https://doi.org/10.1371/journal.pone.0035750) PMID: [22563396](https://pubmed.ncbi.nlm.nih.gov/22563396/)
26. Pryszcz LP, Németh T, Gácser A, Gabaldón T. Unexpected genomic variability in clinical and environmental strains of the pathogenic yeast *Candida parapsilosis*. *Genome Biol Evol*. 2013; 5: 2382–2392. doi: [10.1093/gbe/evt185](https://doi.org/10.1093/gbe/evt185) PMID: [24259314](https://pubmed.ncbi.nlm.nih.gov/24259314/)
27. Safonova Y, Bankevich A, Pevzner PA. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. In: Sharan R, editor. *Research in Computational Molecular Biology*. Springer International Publishing; 2014. pp. 265–279. Available: http://link.springer.com/chapter/10.1007/978-3-319-05269-4_21
28. Gunisova S, Elboher E, Nosek J, Gorkovoy V, Brown Y, Lucier J-F, et al. Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements. *RNA*. 2009; 15: 546–559. doi: [10.1261/rna.1194009](https://doi.org/10.1261/rna.1194009) PMID: [19223441](https://pubmed.ncbi.nlm.nih.gov/19223441/)

29. Valach M, Prysycz LP, Tomaska L, Gacser A, Gabaldón T, Nosek J. Mitochondrial genome variability within the *Candida parapsilosis* species complex. *Mitochondrion*. 2012; 12: 514–519. doi: [10.1016/j.mito.2012.07.109](https://doi.org/10.1016/j.mito.2012.07.109) PMID: [22824459](https://pubmed.ncbi.nlm.nih.gov/22824459/)
30. Rycovska A, Valach M, Tomaska L, Bolotin-Fukuhara M, Nosek J. Linear versus circular mitochondrial genomes: intraspecies variability of mitochondrial genome architecture in *Candida parapsilosis*. *Microbiol Read Engl*. 2004; 150: 1571–1580.
31. Kosa P, Valach M, Tomaska L, Wolfe KH, Nosek J. Complete DNA sequences of the mitochondrial genomes of the pathogenic yeasts *Candida orthopsilosis* and *Candida metapsilosis*: insight into the evolution of linear DNA genomes from mitochondrial telomere mutants. *Nucleic Acids Res*. 2006; 34: 2472–2481. PMID: [16684995](https://pubmed.ncbi.nlm.nih.gov/16684995/)
32. Bennett RJ, Forche A, Berman J. Rapid mechanisms for generating genome diversity: whole ploidy shifts, aneuploidy, and loss of heterozygosity. *Cold Spring Harb Perspect Med*. 2014; 4.
33. 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 Bethesda Md*. 2012; 2: 299–311.
34. Alby K, Bennett RJ. Interspecies pheromone signaling promotes biofilm formation and same-sex mating in *Candida albicans*. *Proc Natl Acad Sci U S A*. 2011; 108: 2510–2515. doi: [10.1073/pnas.1017234108](https://doi.org/10.1073/pnas.1017234108) PMID: [21262815](https://pubmed.ncbi.nlm.nih.gov/21262815/)
35. Hull CM, Johnson AD. Identification of a mating type-like locus in the asexual pathogenic yeast *Candida albicans*. *Science*. 1999; 285: 1271–1275. PMID: [10455055](https://pubmed.ncbi.nlm.nih.gov/10455055/)
36. Sai S, Holland LM, McGee CF, Lynch DB, Butler G. Evolution of mating within the *Candida parapsilosis* species group. *Eukaryot Cell*. 2011; 10: 578–587. doi: [10.1128/EC.00276-10](https://doi.org/10.1128/EC.00276-10) PMID: [21335529](https://pubmed.ncbi.nlm.nih.gov/21335529/)
37. Hensgens LAM, Tavanti A, Mogavero S, Ghelardi E, Senesi S. AFLP genotyping of *Candida metapsilosis* clinical isolates: evidence for recombination. *Fungal Genet Biol FG B*. 2009; 46: 750–758. doi: [10.1016/j.fgb.2009.06.006](https://doi.org/10.1016/j.fgb.2009.06.006) PMID: [19559094](https://pubmed.ncbi.nlm.nih.gov/19559094/)
38. Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*. 2008; 9: 235. doi: [10.1186/gb-2008-9-10-235](https://doi.org/10.1186/gb-2008-9-10-235) PMID: [18983710](https://pubmed.ncbi.nlm.nih.gov/18983710/)
39. Huerta-Cepas J, Capella-Gutiérrez S, Prysycz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res*. 2014; 42: D897–902. doi: [10.1093/nar/gkt1177](https://doi.org/10.1093/nar/gkt1177) PMID: [24275491](https://pubmed.ncbi.nlm.nih.gov/24275491/)
40. Gácser A, Trofa D, Schäfer W, Nosanchuk JD. Targeted gene deletion in *Candida parapsilosis* demonstrates the role of secreted lipase in virulence. *J Clin Invest*. 2007; 117: 3049–3058. PMID: [17853941](https://pubmed.ncbi.nlm.nih.gov/17853941/)
41. Trofa D, Soghier L, Long C, Nosanchuk JD, Gacser A, Goldman DL. A rat model of neonatal candidiasis demonstrates the importance of lipases as virulence factors for *Candida albicans* and *Candida parapsilosis*. *Mycopathologia*. 2011; 172: 169–178. doi: [10.1007/s11046-011-9429-3](https://doi.org/10.1007/s11046-011-9429-3) PMID: [21667319](https://pubmed.ncbi.nlm.nih.gov/21667319/)
42. Gácser A, Stehr F, Kröger C, Kredics L, Schäfer W, Nosanchuk JD. Lipase 8 affects the pathogenesis of *Candida albicans*. *Infect Immun*. 2007; 75: 4710–4718. PMID: [17646357](https://pubmed.ncbi.nlm.nih.gov/17646357/)
43. Schofield DA, Westwater C, Warner T, Balish E. Differential *Candida albicans* lipase gene expression during alimentary tract colonization and infection. *FEMS Microbiol Lett*. 2005; 244: 359–365. PMID: [15766791](https://pubmed.ncbi.nlm.nih.gov/15766791/)
44. Stehr F, Felk A, Gácser A, Kretschmar M, Mähns B, Neuber K, et al. Expression analysis of the *Candida albicans* lipase gene family during experimental infections and in patient samples. *FEMS Yeast Res*. 2004; 4: 401–408. PMID: [14734020](https://pubmed.ncbi.nlm.nih.gov/14734020/)
45. Naglik JR, Challacombe SJ, Hube B. *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. *Microbiol Mol Biol Rev MMBR*. 2003; 67: 400–428, table of contents. PMID: [12966142](https://pubmed.ncbi.nlm.nih.gov/12966142/)
46. Horváth P, Nosanchuk JD, Hamari Z, Vágvölgyi C, Gácser A. The identification of gene duplication and the role of secreted aspartyl proteinase 1 in *Candida parapsilosis* virulence. *J Infect Dis*. 2012; 205: 923–933. doi: [10.1093/infdis/jir873](https://doi.org/10.1093/infdis/jir873) PMID: [22301631](https://pubmed.ncbi.nlm.nih.gov/22301631/)
47. Gabaldón T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, et al. Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics*. 2013; 14: 623. doi: [10.1186/1471-2164-14-623](https://doi.org/10.1186/1471-2164-14-623) PMID: [24034898](https://pubmed.ncbi.nlm.nih.gov/24034898/)
48. 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. doi: [10.1371/journal.ppat.1000713](https://doi.org/10.1371/journal.ppat.1000713) PMID: [20072605](https://pubmed.ncbi.nlm.nih.gov/20072605/)
49. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*. 2014; 505: 546–549. doi: [10.1038/nature12817](https://doi.org/10.1038/nature12817) PMID: [24352233](https://pubmed.ncbi.nlm.nih.gov/24352233/)

50. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 2012; 1: 18. doi: [10.1186/2047-217X-1-18](https://doi.org/10.1186/2047-217X-1-18) PMID: [23587118](https://pubmed.ncbi.nlm.nih.gov/23587118/)
51. Huang S, Chen Z, Huang G, Yu T, Yang P, Li J, et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res*. 2012; 22: 1581–1588. doi: [10.1101/gr.133652.111](https://doi.org/10.1101/gr.133652.111) PMID: [22555592](https://pubmed.ncbi.nlm.nih.gov/22555592/)
52. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma Oxf Engl*. 2011; 27: 578–579.
53. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006; 34: W435–439. PMID: [16845043](https://pubmed.ncbi.nlm.nih.gov/16845043/)
54. Prysycz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res*. 2011; 39: e32. doi: [10.1093/nar/gkq953](https://doi.org/10.1093/nar/gkq953) PMID: [21149260](https://pubmed.ncbi.nlm.nih.gov/21149260/)
55. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005; 33: W116–120. PMID: [15980438](https://pubmed.ncbi.nlm.nih.gov/15980438/)
56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9: 357–359. doi: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) PMID: [22388286](https://pubmed.ncbi.nlm.nih.gov/22388286/)
57. 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. doi: [10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110) PMID: [20644199](https://pubmed.ncbi.nlm.nih.gov/20644199/)
58. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res*. 2012; 40: e115. PMID: [22730293](https://pubmed.ncbi.nlm.nih.gov/22730293/)
59. Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*. 1984; 37: 67–75. PMID: [6373014](https://pubmed.ncbi.nlm.nih.gov/6373014/)
60. Chu G, Vollrath D, Davis RW. Separation of large DNA molecules by contour-clamped homogeneous electric fields. *Science*. 1986; 234: 1582–1585. PMID: [3538420](https://pubmed.ncbi.nlm.nih.gov/3538420/)
61. Frith MC, Hamada M, Horton P. Parameters for accurate genome alignment. *BMC Bioinformatics*. 2010; 11: 80. doi: [10.1186/1471-2105-11-80](https://doi.org/10.1186/1471-2105-11-80) PMID: [20144198](https://pubmed.ncbi.nlm.nih.gov/20144198/)
62. Huerta-Cepas J, Capella-Gutierrez S, Prysycz LP, Denisov I, Kormes D, Marcet-Houben M, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res*. 2011; 39: D556–560. doi: [10.1093/nar/gkq1109](https://doi.org/10.1093/nar/gkq1109) PMID: [21075798](https://pubmed.ncbi.nlm.nih.gov/21075798/)
63. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981; 147: 195–197. PMID: [7265238](https://pubmed.ncbi.nlm.nih.gov/7265238/)
64. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; 32: 1792–1797. PMID: [15034147](https://pubmed.ncbi.nlm.nih.gov/15034147/)
65. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 2008; 9: 286–298. doi: [10.1093/bib/bbn013](https://doi.org/10.1093/bib/bbn013) PMID: [18372315](https://pubmed.ncbi.nlm.nih.gov/18372315/)
66. Lassmann T, Frings O, Sonnhammer ELL. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res*. 2009; 37: 858–865. doi: [10.1093/nar/gkn1006](https://doi.org/10.1093/nar/gkn1006) PMID: [19103665](https://pubmed.ncbi.nlm.nih.gov/19103665/)
67. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 2006; 34: 1692–1699. PMID: [16556910](https://pubmed.ncbi.nlm.nih.gov/16556910/)
68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma Oxf Engl*. 2009; 25: 1972–1973.
69. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010; 59: 307–321. doi: [10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010) PMID: [20525638](https://pubmed.ncbi.nlm.nih.gov/20525638/)
70. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinforma Oxf Engl*. 2005; 21: 456–463.
71. Dray S., Dufour A.B. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw*. 2007; 22: 1–20.
72. Wehe A, Bansal MS, Burleigh JG, Eulenstein O. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinforma Oxf Engl*. 2008; 24: 1540–1541.
73. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*. 2010; 11: 24. doi: [10.1186/1471-2105-11-24](https://doi.org/10.1186/1471-2105-11-24) PMID: [20070885](https://pubmed.ncbi.nlm.nih.gov/20070885/)