HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2010 March 08.

Published in final edited form as:

Nature. 2009 June 4; 459(7247): 657–662. doi:10.1038/nature08064.

Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes

A full list of authors and affiliations appears at the end of the article.

Abstract

Candida species are the most common cause of opportunistic fungal infection worldwide. We report the genome sequences of six *Candida* species and compare these and related pathogens and nonpathogens. There are significant expansions of cell wall, secreted, and transporter gene families in pathogenic species, suggesting adaptations associated with virulence. Large genomic tracts are homozygous in three diploid species, possibly resulting from recent recombination events. Surprisingly, key components of the mating and meiosis pathways are missing from several species. These include major differences at the Mating-type loci (*MTL*); *Lodderomyces elongisporus* lacks *MTL*, and components of the a1/alpha2 cell identity determinant were lost in other species, raising questions about how mating and cell types are controlled. Analysis of the CUG leucine to serine genetic code change reveals that 99% of ancestral CUG codons were erased and new ones arose elsewhere. Lastly, we revise the *C. albicans* gene catalog, identifying many new genes.

Four species, *C. albicans*, *C. glabrata*, *C. tropicalis* and *C. parapsilosis*, together account for ~95% of identifiable *Candida* infections1. While *C. albicans* is still the most common causative agent, its incidence is declining and the frequency of other species is increasing. Of these, *C. parapsilosis* is a particular problem in neonates, transplant recipients, and patients receiving parenteral nutrition; *C. tropicalis* is more commonly associated with neutropenia and malignancy. Other *Candida* species, including *C. krusei*, *C. lusitaniae* and *C. guilliermondii*, account for <5% of invasive candidiasis. Almost all *Candida* species, with the exception of *C. glabrata* and *C. krusei*, belong in a single *Candida* clade (Fig. 1) characterized by the unique translation of CUG codons as serine rather than leucine2. Within this, haploid and diploid species occupy two separate subclades (Fig. 1).

To determine the genetic features underlying their diversity of biology and pathogenesis, we sequenced six genomes from the *Candida* clade (Fig. 1). These include a second sequenced isolate of *C. albicans* (WO-1) characterized for white-opaque switching, a phenotypic

†Correspondence and requests to materials should be addressed to C.A.C (cuomo@broad.mit.edu), M.K. (manoli@mit.edu) or G.B. (geraldine.butler@ucd.ie)..

*These authors contributed equally to this work

‡Present address for CDK: 454 Life Sciences, a member of the Roche Group, 20 Commercial St, Branford, CT 06405

Author information Assemblies were submitted to GenBank under the following project accession numbers: *C. albicans* WO-1 (AAFO00000000), *C. tropicalis* (AAFN00000000), *L. elongisporus* (AAPO00000000), *C. guilliermondii* (AAFM00000000), *C. lusitaniae* (AAFT00000000), and *C. parapsilosis* (CABE01000001-CABE01000024).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature

change that correlates with host specificity and mating3, 4. We also sequenced the major pathogens *C. tropicalis* and *C. parapsilosis*, *Lodderomyces elongisporus*, a close relative of *C. parapsilosis* recently identified as a cause of bloodstream infection5, and two haploid emerging pathogens, *C. guilliermondii* and *C. lusitaniae*. We compared these to the previously sequenced *C. albicans* strain (SC5314)6-8, *Debaryomyces hansenii*9, a marine yeast rarely associated with disease, and nine species from the related *Saccharomyces* clade (Fig. 1). These species span a wide evolutionary range and show large phenotypic differences in pathogenicity and mating, allowing us to study the genomic basis for these traits.

Genome sequence and comparative annotation

We found enormous variation in genome size and composition between the *Candida* genomes sequenced (Table 1). Each genome assembly displayed high continuity, ranging from nine to 27 scaffolds (Supplementary table 1). Scaffold number and size largely match pulsed-field gel electrophoresis estimates for all genomes, and telomeric repeat arrays are linked to the ends of nearly all large scaffolds (Supplementary text S2). Genome size ranges from 10.6 to 15.5 Mb, a striking difference of nearly 50%, with haploid species having smaller genomes. GC content ranges from 33% to 45% (Table 1). Transposable elements and other repetitive sequences vary in number and type between assemblies (Supplementary text S6). Regions similar to the Major Repeat Sequence (MRS) of *C. albicans* were found only in *C. tropicalis*, suggesting that MRS-associated recombination could contribute to the observed karyotypic variation between *C. tropicalis* strains10.

Despite the genome size and phenotypic variation among the species, the predicted numbers of protein-coding genes are very similar ranging from 5,733 to 6,318 genes (Table 1). Even the small differences in gene number are not correlated with genome size; the smallest genome, *C. guilliermondii*, has more genes than the largest genome, *L. elongisporus*. Instead, genome size differences are explained by a ∼3-fold variation in intergenic spacing (Table 1). Large syntenic blocks of conserved gene order were detected among the four diploid species and between two of the haploid species, *C. guilliermondii* and *D. hansenii* (Supplementary fig. 5). While synteny blocks have been shuffled by local inversions and rearrangements, these have been primarily intrachromosomal as chromosome boundaries have been largely preserved across the diploid genomes (Supplementary fig. 5).

Given the high conservation of protein-coding genes across the *Candida* clade, we used multiple alignments of the related genomes to revise the annotation of *C. albicans*. We identified 91 new or updated genes, of which 80% are specific to the *Candida* clade (Supplementary text S4). We also corrected existing annotations in *C. albicans*, revealing 222 dubious genes, and also identified 190 likely frame-shifts and 36 nonsense sequencing errors in otherwise well-conserved genes (Supplementary text S4). In each case, manual curation confirmed ∼80% of these predictions.

Polymorphism in diploid genomes

To gain insights into the recent history of *C. albicans*, we compared the two diploid strains, SC5314 and WO-1, which belong to different population subgroups11. Variation in the

karyotype of these strains is primarily due to translocations at MRS sequences (12 and Supplementary fig. 1). The two assemblies are largely co-linear with 12 inversions of 5−94 kb between them, except that in WO-1 some non-homologous chromosomes have recombined at the MRS (Supplementary text S8). We found similar rates of single nucleotide polymorphisms (SNPs) within each strain (1 SNP per 330−390 bases), and twice this level between them, suggesting relatively recent divergence. Polymorphism rates in the other diploids range from 1 SNP per 222 bases in *L. elongisporus* and 1 SNP per 576 bases in *C. tropicalis*, to a remarkably low 1 SNP per 15,553 bases in *C. parapsilosis*, more than 70-fold lower than the closely related *L. elongisporus*.

Striking regions of extended homozygosity are found in three of the four diploid genomes, which may reflect break-induced replication, or recent passage through a parasexual or sexual cycle. *C. albicans*, *C. tropicalis*, and *L. elongisporus* each show large chromosomal regions devoid of SNPs, extending up to ∼1.2 Mb (Fig. 2, Supplementary fig. 6−8). In contrast, the few SNPs in *C. parapsilosis* are randomly distributed across the genome (Supplementary fig. 9A). A total of 4.3 Mb in total (30%) of the WO-1 assembly is homozygous for SNPs, approximately twice that found in SC5314 (Supplementary text S7 and 7, 13, 14). There is at least one homozygous region per chromosome, none of which spans the predicted centromeres, and only one of which starts at a MRS (Fig. 2). While nearly all homogeneous regions are present at diploid levels and are therefore homozygous, WO-1 has lost one copy of a >300 kb region on chromosome 3 comprising nearly 200 genes (Supplementary fig. 10). The pressure to maintain this region as homozygous in both strains is apparently high, as it is diploid but homogenous in SC5314.

Usage and evolution of CUG codons

All *Candida* clade species translate CUG codons as serine instead of leucine15. This genetic code change altered the decoding rules of CUN codons in the *Candida* clade: while *S. cerevisiae* uses two tRNAs which each translate two codons, *Candida* species use a dedicated $tRNA_{CAG}^{Ser}$ for CUG codons and a single $tRNA_{LAG}^{Leu}$ for CUA, CUC, and CUU, as inosine can base pair with A, C, and U (Fig. 3). This remarkable alteration in decoding rules forced the reduced usage of CUG, and also CUA likely due to the weaker wobble, in *Candida* genes (Fig. 3A). CUU and CUC codons do not display the same bias for infrequent usage (Fig. 3B). An additional pressure influencing codon usage may be the GC content, as usage of leucine codons in *Candida* species is correlated with %GC composition (Supplementary table 11).

We also examined the evolutionary fate of ancestral CUG codons and the origin of new CUG codons (Supplementary table 12). CUG codons in *C. albicans* almost never (1%) align opposite CUG codons in *S. cerevisiae*. Instead, CUG serine codons in *C. albicans* align primarily to *Saccharomyces* codons for serine (20%) and other hydrophilic residues (49%). CUG leucine codons in *S. cerevisiae* align primarily to leucine codons in *Candida* (50%) and to other hydrophobic residue codons (30%). This suggests a complete functional replacement of CUG codons in *Candida*.

Gene family evolution

To identify gene families likely to be associated with *Candida* pathogenicity and virulence, we used a phylogenomic approach across seven *Candida* and nine *Saccharomyces* genomes (Supplementary text S10). Among 9,209 gene families, we identified 21 that are significantly enriched in the more common pathogens (Table 2). These include lipases, oligopeptide transporters (Opts), and adhesins, all known to be associated with pathogenicity8, as well as poorly characterized families not previously associated with pathogenesis.

Three cell wall families are enriched in the pathogens: Hyr/Iff16, Als adhesins17, and Pga30-like proteins (Table 2, Supplementary text S11). The Als family (family 17 in Supplementary tables 22 and 23) in *C. albicans* is associated with virulence, and in particular with adhesion to host surfaces18, invasion of host cells19, and iron acquisition20. All these families are absent from the *Saccharomyces* clade species and are particularly enriched in the more pathogenic species (Supplementary table 18). All three families are highly enriched for gene duplications (Supplementary table 19), including tandem clusters of 2−6 genes, and show high mutation rates (fastest 5% of families) (Supplementary text S10d). This variable repertoire of cell wall proteins is likely of profound importance to the niche adaptations and relative virulence of these organisms.

Als17 and Hyr/Iff proteins frequently contain intragenic tandem repeats (ITRs), which modulate adhesion and biofilm formation in *S. cerevisiae*21 (Figure 4, Supplementary Fig. 19, 20). The ITR sequence is conserved at the protein level across species (Supplementary fig. 19−20). Interestingly, two proteins contain both an Als domain and repeats characteristic of the Hyr/Iff family.

Candida clade pathogens show expansions of extracellular enzyme and transmembrane transporter families (Table 2 and Supplementary table 22). These families are either not found in *Saccharomyces* (including amino acid permeases, lipases, and superoxide dismutases), or present in *S. cerevisiae* but significantly expanded in pathogens (including phospholipase B, ferric reductases, sphingomyelin phosphodiesterases, and GPI-anchored yapsin proteases, which have been linked to virulence in *C. glabrata*22). Several groups of cell-surface transporters are also enriched (including Opts, amino acid permeases, and the major facilitator superfamily). Overall these family expansions illustrate the importance of extracellular activities in virulence and pathogenicity. Genes involved in stress response are also variable between species (Supplementary text S12).

C. albicans also showed species-specific expansion of some families, including two associated with filamentous growth, a leucine-rich repeat family and the Fgr6−1 family (Table 2). As *C. albicans* forms hyphae while *C. tropicalis* and *C. parapsilosis* produce only pseudohyphae, these families may contribute to differences in hyphal growth.

We identified 64 families showing positive selection in the highly pathogenic *Candida* species (Supplementary table 32). These are highly enriched for cell wall, hyphal, pseudohyphal, filamentous growth, and biofilm functions (Supplementary text S13). Six of the families have been previously associated with pathogenesis, including *ERG3*, a C-5

sterol desaturase essential for ergosterol biosynthesis, for which mutations can cause drug resistance23.

Structure of the *MTL* locus

Pathogenic fungi may have limited their sexual cycles to maximize their virulence24, and the sequenced *Candida* species show tremendous diversity in their apparent abilities to mate. Among the four diploids, *C. albicans* has a parasexual cycle (mating of diploid cells followed by mitosis and chromosome loss instead of meiosis25), *L. elongisporus* has been described as sexual and homothallic (self-mating)26, while *C. tropicalis* and *C. parapsilosis* have never been observed to mate. Among the three haploids, *C. guilliermondii* and *C. lusitaniae* are heterothallic (cross-mating only) and have a complete sexual cycle, while *D. hansenii* is haploid and homothallic27 (Supplementary text 14c).

To understand the genomic basis for this diversity, we studied the *Candida* mating-type locus (*MTL*), which determines mating type, similar to the *MAT* locus in *S. cerevisiae*. In both *C. albicans* and *S. cerevisiae*, the mating locus has two idiomorphs, **a** and α, encoding the regulators **a**1 and α1/α2 respectively. *C. albicans MTL***a** also encodes **a**2, and both idiomorphs in this species contain alleles of three additional genes without known roles in mating; *PAP*, *OBP*, and *PIK*28. The *MTL*α and *MTL***a** genes, alone or in combination, specify one of the three possible cell-type programs (**a** and α haploid, **a**/α diploid). In *C. albicans*, the alpha-domain protein α1 activates α-specific mating genes, the HMG factor **a**2 activates **a**-specific mating genes, and the **a**1/α2 homeodomain heterodimer represses mating genes in **a**/α cells29, 30.

Despite extended conservation of the genomic context flanking MTL, there is tremendous variability in *MTL* gene content (Fig. 5). *MTL***a**1 has become a pseudogene in *C. parapsilosis* 31, likely a recent loss since target genes retain predicted **a**1/α2 binding sites (Supplementary text S14). *MTL*α2 is missing in both *C. guilliermondii* and *C. lusitaniae* (Reedy, Floyd, and Heitman, submitted). A fused mating type locus containing both **a** and α genes is found in *D. hansenii* and *Pichia stipitis*32, 33.

Most surprisingly, all four mating-type genes are missing in *L. elongisporus*. It contains a site syntenic to *MTL***a** in other species, but this contains only 508 base pairs of apparently noncoding DNA, a length insufficient to encode **a**1 or **a**2 even if they were extensively divergent in sequence. We confirmed this finding in seven other *L. elongisporus* isolates (not shown). The sexual state of *L. elongisporus* has been assumed to be homothallic because asci are generated from identical cells5, 26, but the absence of an **a**-factor pheromone, receptor and transporter (Supplementary table 34) as well as *MTL*, suggests that it may not have a sexual cycle. Alternatively, mating may require only one pheromone and receptor (α), and *L. elongisporus* may be the first identified ascomycete that can mate independently of *MTL/MAT*.

Our discovery that *MTL*α2 and *MTL***a**1 are frequently absent challenges our understanding of how the mating locus operates. The model derived for mating regulation in *C. albicans*, including the role of **a**1/α2 in white-opaque switching3, must differ substantially in the other

sexual species. This is particularly interesting because the major regulators of the white-opaque switch in *C. albicans* (*WOR1*, *WOR2*, *CZF1*, *EFG1*34) are generally conserved in other species, but Wor1 control over white-opaque genes appears to be a recent innovation in *C. albicans* and *C. dubliniensis*35. Our data also suggest that the loss of α2 occurred early in the haploid sexual lineage.

Mating and meiosis

To gain further insights in their diversity in sexual behavior, we examined whether 227 genes required for meiosis in *S. cerevisiae* and other fungi have orthologs in the *Candida* species (Supplementary text S14). A previous report36 that some components of meiosis, such as the major regulator *IME1*, are missing from *C. albicans*, led us to hypothesize that their loss could be correlated with lack of meiosis. Surprisingly however, we find that these genes are missing in all *Candida* species, suggesting that sexual *Candida* species undergo meiosis without them. Conversely, even seemingly non-mating species showed highly conserved pheromone response pathways, suggesting pheromone signaling plays an alternate role such as regulation of biofilm formation37. These findings suggest considerable plasticity and innovation of meiotic pathways in *Candida*.

Moreover, we find that sexual *Candida* species have undergone a recent dramatic change in the pathways involved in meiotic recombination, with loss of the Dmc1-dependent pathway in the heterothallic species *C. lusitaniae* and *C. guilliermondii* (Supplementary text 14c). We also found that mechanisms of chromosome pairing and crossover formation have changed recently in these two species, because they (and to a lesser extent *D. hansenii*) have lost several components of the synaptonemal and synapsis initiation complexes (Supplementary table 35). They have also lost components of the major crossover formation pathway in *S. cerevisiae* (*MSH4*, *MSH5*), while retaining a minor pathway (*MUS81*, *MMS4*)38, 39. Overall, if *Candida* species undergo meiosis it is with reduced machinery, or different machinery, suggesting that unrecognized meiotic cycles may exist in many species, and that the paradigm of meiosis developed in *S. cerevisiae* varies significantly, even among yeasts.

The genome sequences reported here provide a resource that will allow current knowledge of *C. albicans* biology, the product of decades of research, to be applied with maximum effect to the other pathogenic species in the *Candida* clade. They also allow many of the unusual features of *C. albicans* – such as cell wall gene family amplifications, and its apparent ability to undergo mating and a parasexual cycle without meiosis – to be understood in an evolutionary context that shows that the genes involved in virulence and mating have exceptionally dynamic rates of turnover and loss.

Methods Summary

The full methods for this paper are described in Supplementary Information. Here we outline the resources generated by this project.

Assemblies, gene sets, and single nucleotide polymorphisms are available in GenBank and at the Broad *Candida* Database website (http://www.broad.mit.edu/annotation/genome/candida_group/MultiHome.html). The Broad website provides search, visualization,

BLAST, and download of assemblies and gene sets. The *C. parapsilosis* assembly is also available at the Wellcome Trust Sanger Institute website (http://www.sanger.ac.uk/sequencing/Candida/parapsilosis/). The revised annotation of *C. albicans* (SC5314) is available at the *Candida* Genome Database (www.candidagenome.org). Gene families can be accessed on the Broad website via individual gene pages or by searching with family identifiers (CF#####).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Geraldine Butler[1,†], Matthew D. Rasmussen[2], Michael F. Lin[2,3], Manuel A.S. Santos[4], Sharadha Sakthikumar[3], Carol A. Munro[5], Esther Rheinbay[2], Manfred Grabherr[3], Anja Forche[6], Jennifer L. Reedy[7], Ino Agrafioti[8], Martha B. Arnaud[9], Steven Bates[10], Alistair J.P. Brown[5], Sascha Brunke[11], Maria C. Costanzo[9], David A. Fitzpatrick[1], Piet W. J. de Groot[12], David Harris[13], Lois L. Hoyer[14], Bernhard Hube[11], Frans M. Klis[12], Chinnappa Kodira[3,‡], Nicola Lennard[13], Mary E. Logue[1], Ronny Martin[11], Aaron M. Neiman[15], Elissavet Nikolaou[5], Michael A. Quail[13], Janet Quinn[16], Maria C. Santos[4], Florian F. Schmitzberger[9], Gavin Sherlock[9], Prachi Shah[9], Kevin Silverstein[17], Marek S. Skrzypek[9], David Soll[18], Rodney Staggs[17], Ian Stansfield[5], Michael P H Stumpf[8], Peter E. Sudbery[19], Srikantha Thyagarajan[18], Qiandong Zeng[3], Judith Berman[6], Matthew Berriman[13], Joseph Heitman[7], Neil A. R. Gow[5], Michael C. Lorenz[20], Bruce W. Birren[3], Manolis Kellis[2,3,†,*], and Christina A. Cuomo[3,†,*]

Affiliations

[1]UCD School of Biomolecular and Biomedical Science, Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland [2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA [3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA [4]Department of Biology and CESAM, University of Aveiro, 3810−193 Aveiro, Portugal [5]School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen, UK [6]Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN, USA [7]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA [8]Centre for Bioinformatics, Imperial College London, Wolfson Building, South Kensington, London SW7 2AY, UK [9]Department of Genetics, Stanford University Medical School Stanford, CA 94305−5120, USA [10]School of Biosciences, University of Exeter, Exeter EX4 4QD, UK [11]Department of Microbial Pathogenicity Mechanisms, Leibniz Institute for Natural Product Research and Infection Biology- Hans Knoell Institute, Jena, Germany [12]Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands [13]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK CB10 1SA [14]Department of Pathobiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA [15]Department of Biochemistry and Cell

Biology, SUNY Stony Brook, Stony Brook, NY 45215, USA [16]School of Cell and Molecular Biosciences, University of Newcastle upon Tyne, Newcastle upon Tyne NE2 4HH UK [17]Biostatistics and Bioinformatics Group, Masonic Cancer Center, University of Minnesota, MN, USA [18]Department of Biology, The University of Iowa, Iowa City, IA 52242, USA [19]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK [20]Department of Microbiology and Molecular Genetics, The University of Texas Health Science Center at Houston, Houston, TX, 77030

Acknowledgements

We thank the National Human Genome Research Institute (NHGRI) for support under the Fungal Genome Initiative at the Broad Institute. We thank Cletus Kurtzman for providing the sequenced strains of *L. elongisporus*, *C. guilliermondii*, and *C. lusitaniae*, Michael Koehrsen for Broad website support, Danny Park for informatics support, Ken Wolfe and Aviv Regev for comments on the manuscript. We acknowledge the contributions of the Broad Institute Sequencing Platform and Andrew Barron, Louise Clark, Craig Corton, Doug Ormond, David Saunders, Kathy Seeger, Robert Squares from the Wellcome Trust Sanger Institute for the *C. parapsilosis* sequencing and assembly. N.A.G., A.J.P, M.B. and co-workers were supported by the Wellcome Trust; M.G., S.S., Q.Z., B.W.B., and C.A.C. supported by NHGRI and the National Institute of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services; G.B. and co-workers by Science Foundation Ireland; J.B., A.F., J.H., A.M.N. and co-workers by the NIH; M.K., M.R., and M.L. by the NIH, the NSF, and the Sloan Foundation.

References

1. Pfaller MA, Diekema DJ. Epidemiology of invasive candidiasis: a persistent public health problem. Clin Microbiol Rev. 2007; 20:133–63. [PubMed: 17223626]
2. Santos MA, Tuite MF. The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. Nucleic Acids Res. 1995; 23:1481–6. [PubMed: 7784200]
3. Lockhart SR, et al. In *Candida albicans*, white-opaque switchers are homozygous for mating type. Genetics. 2002; 162:737–45. [PubMed: 12399384]
4. Slutsky B, Buffo J, Soll DR. High-frequency switching of colony morphology in *Candida albicans*. Science. 1985; 230:666–9. [PubMed: 3901258]
5. Lockhart SR, Messer SA, Pfaller MA, Diekema DJ. *Lodderomyces elongisporus* masquerading as *Candida parapsilosis* as a cause of bloodstream infections. J Clin Microbiol. 2008; 46:374–6. [PubMed: 17959765]
6. Braun BR, et al. A human-curated annotation of the Candida albicans genome. PLoS Genet. 2005; 1:36–57. [PubMed: 16103911]
7. Jones T, et al. The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci U S A. 2004; 101:7329–34. [PubMed: 15123810]
8. van het Hoog M, et al. Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol. 2007; 8:R52. [PubMed: 17419877]
9. Dujon B, et al. Genome evolution in yeasts. Nature. 2004; 430:35–44. [PubMed: 15229592]
10. Zhang J, Hollis RJ, Pfaller MA. Variations in DNA subtype and antifungal susceptibility among clinical isolates of *Candida tropicalis*. Diagn Microbiol Infect Dis. 1997; 27:63–7. [PubMed: 9147006]
11. Tavanti A, et al. Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. J Clin Microbiol. 2005; 43:5601–13. [PubMed: 16272493]
12. Chu WS, Magee BB, Magee PT. Construction of an SfiI macrorestriction map of the *Candida albicans* genome. J Bacteriol. 1993; 175:6637–51. [PubMed: 8407841]
13. Forche A, Magee PT, Magee BB, May G. Genome-wide single-nucleotide polymorphism map for *Candida albicans*. Eukaryot Cell. 2004; 3:705–14. [PubMed: 15189991]

14. Legrand M, et al. Haplotype mapping of a diploid non-meiotic organism using existing and induced aneuploidies. PLoS Genet. 2008; 4:e1. [PubMed: 18179283]
15. Massey SE, et al. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida* spp. Genome Res. 2003; 13:544–57. [PubMed: 12670996]
16. Bates S, et al. *Candida albicans* Iff11, a secreted protein required for cell wall structure and virulence. Infect Immun. 2007; 75:2922–8. [PubMed: 17371861]
17. Hoyer LL, Green CB, Oh SH, Zhao X. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family--a sticky pursuit. Med Mycol. 2008; 46:1–15. [PubMed: 17852717]
18. Yeater KM, et al. Temporal analysis of *Candida albicans* gene expression during biofilm development. Microbiology. 2007; 153:2373–85. [PubMed: 17660402]
19. Phan QT, et al. Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells. PLoS Biol. 2007; 5:e64. [PubMed: 17311474]
20. Almeida RS, et al. The hyphal-associated adhesin and invasin Als3 of *Candida albicans* mediates iron acquisition from host ferritin. PLoS Pathog. 2008; 4:e1000217. [PubMed: 19023418]
21. Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nat Genet. 2005; 37:986–90. [PubMed: 16086015]
22. Kaur R, Ma B, Cormack BP. A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of *Candida glabrata*. Proc Natl Acad Sci U S A. 2007; 104:7628–33. [PubMed: 17456602]
23. Chau AS, et al. Inactivation of sterol Delta5,6-desaturase attenuates virulence in *Candida albicans*. Antimicrob Agents Chemother. 2005; 49:3646–51. [PubMed: 16127034]
24. Nielsen K, Heitman J. Sex and virulence of human pathogenic fungi. Adv Genet. 2007; 57:143–73. [PubMed: 17352904]
25. Noble SM, Johnson AD. Genetics of *Candida albicans*, a diploid human fungal pathogen. Annu Rev Genet. 2007; 41:193–211. [PubMed: 17614788]
26. van der Walt JP. *Lodderomyces*, a new genus of the Saccharomycetaceae. Antonie Van Leeuwenhoek. 1966; 32:1–5. [PubMed: 5296604]
27. Kurtzman, C.; Fell, J. The Yeasts, a taxonomic study. Elsevier; Amsterdam: 1998.
28. Hull CM, Raisner RM, Johnson AD. Evidence for mating of the “asexual” yeast *Candida albicans* in a mammalian host. Science. 2000; 289:307–10. [PubMed: 10894780]
29. Tsong AE, Miller MG, Raisner RM, Johnson AD. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell. 2003; 115:389–99. [PubMed: 14622594]
30. Tsong AE, Tuch BB, Li H, Johnson AD. Evolution of alternative transcriptional circuits with identical logic. Nature. 2006; 443:415–20. [PubMed: 17006507]
31. Logue ME, Wong S, Wolfe KH, Butler G. A genome sequence survey shows that the pathogenic yeast *Candida parapsilosis* has a defective MTLa1 allele at its mating type locus. Eukaryot Cell. 2005; 4:1009–17. [PubMed: 15947193]
32. Fabre E, et al. Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. Mol Biol Evol. 2005; 22:856–73. [PubMed: 15616141]
33. Jeffries TW, et al. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. Nat Biotechnol. 2007; 25:319–26. [PubMed: 17334359]
34. Zordan RE, Miller MG, Galgoczy DJ, Tuch BB, Johnson AD. Interlocking transcriptional feedback loops control white-opaque switching in Candida albicans. PLoS Biol. 2007; 5:e256. [PubMed: 17880264]
35. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. The evolution of combinatorial gene regulation in fungi. PLoS Biol. 2008; 6:e38. [PubMed: 18303948]
36. Tzung KW, et al. Genomic evidence for a complete sexual cycle in *Candida albicans*. Proc Natl Acad Sci U S A. 2001; 98:3249–53. [PubMed: 11248064]
37. Daniels KJ, Srikantha T, Lockhart SR, Pujol C, Soll DR. Opaque cells signal white cells to form biofilms in *Candida albicans*. Embo J. 2006; 25:2240–52. [PubMed: 16628217]

38. Argueso JL, Wanat J, Gemici Z, Alani E. Competing crossover pathways act during meiosis in *Saccharomyces cerevisiae*. Genetics. 2004; 168:1805–16. [PubMed: 15611158]

39. de los Santos T, et al. The Mus81/Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. Genetics. 2003; 164:81–94. [PubMed: 12750322]

40. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 2006; 440:341–5. [PubMed: 16541074]

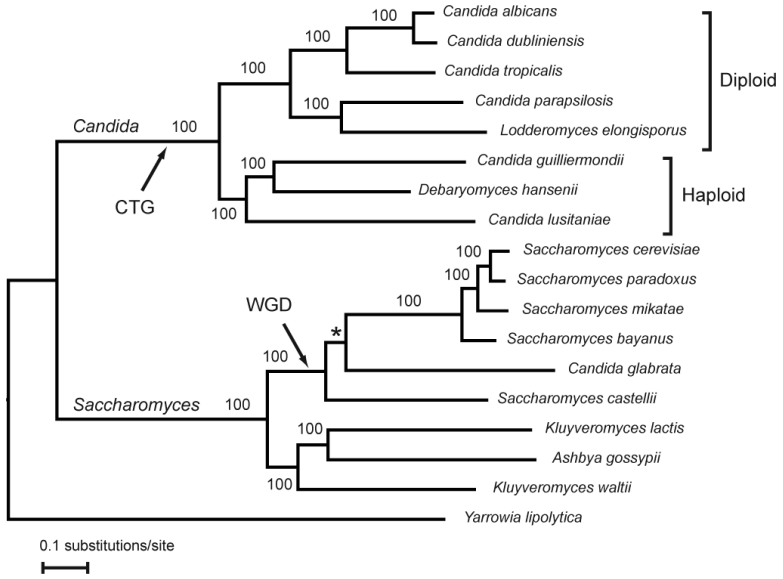

Figure 1.

Phylogeny of sequenced *Candida* and *Saccharomyces* clade species. Tree topology and branch lengths were inferred with MrBayes (see supplementary methods S5). Posterior probabilities are indicated for each branch. The * marks a branch that was constrained based on syntenic conservation40.

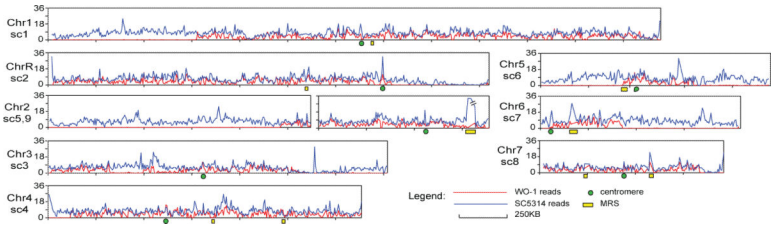

Figure 2.

C. albicans WO-1 is highly homozygous. Red lines show SNPs per kb, normalized by coverage, within WO-1, and blue lines show SNPs per kb between WO-1 and SC5314. While both copies of chromosome 5 have rearranged at the MRS (yellow box) in WO-1, we show this as a single chromosome to allow a haploid reference for polymorphism. Relative to SC5314, chromosomes 1, 4, and 6 are in the opposite orientation (Supplementary figure 6).

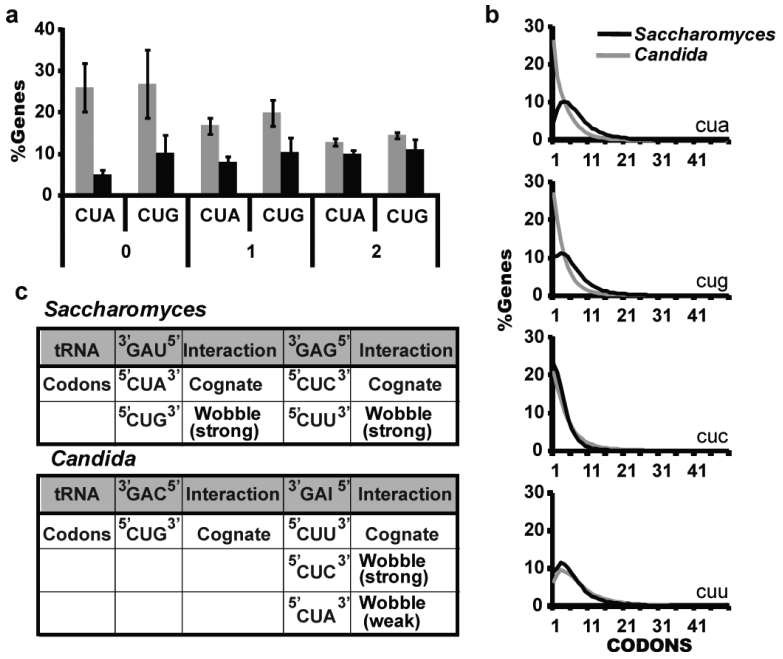

Figure 3.

Evolutionary effects of CUG coding. A. Average percent of genes with 0, 1, or 2 CUG and CUA codons in *Candida* (grey bars) and *Saccharomyces* (black bars). Error bars indicate standard deviations. All differences are significant with $p \leq 0.0004$ (t-test) except for genes with two CUA codons genes (p=0.02). "*Candida*" and "*Saccharomyces*" here refer to the CTG and WGD clades in Figure 1, but including *P. stipitis* and excluding *C. dubliniensis*. B. CUN codon usage for all codon counts. C. Decoding rules for CUN codons in *Saccharomyces* and *Candida*.

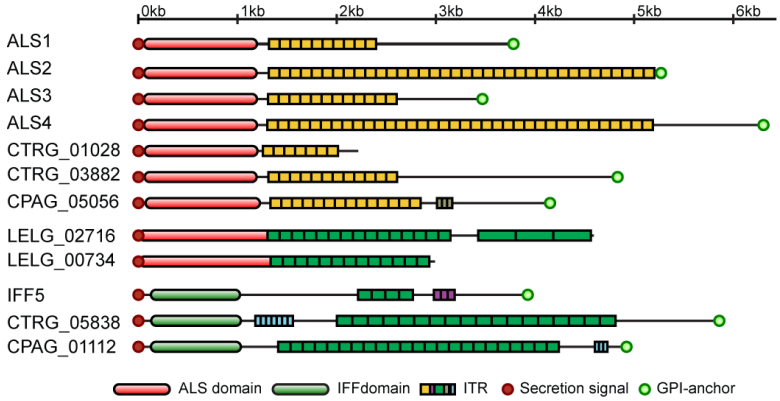

Figure 4.

Conserved domains of Als and Hyr/Iff cell wall families. The N-terminal Als domain (red) and Hyr/Iff domain (green) are shown as ovals. Intragenic tandem repeats (ITRs, see Supplementary text S11) are shown as rectangles, colored to represent similar amino acid sequences.

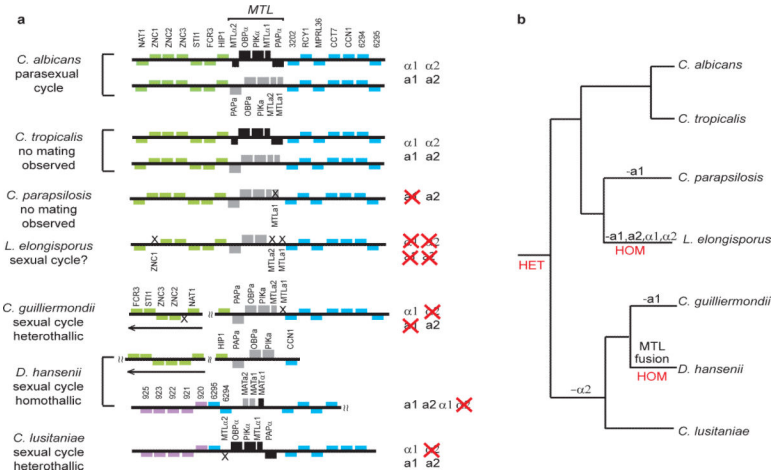

Figure 5.

Organization of *MTL* loci in the *Candida* clade. A. *MTL***a**-specific genes are shown in grey, *MTL*α-specific in black, and other orthologs in color. Two idiomorphs from *C. albicans* and *C. tropicalis* are shown. Arrows indicate inversions relative to *C. albicans*. "X" symbols show gene losses and the *C. parapsilosis MTL***a***1* pseudogene. There is no *MTL* locus in *L. elongisporus*; gene order around the *PAP***a**, *OBP***a** and *PIK***a** genes is shown. For *C. guilliermondii* and *C. lusitaniae* the genome project sequenced one idiomorph, the second was obtained by J. Reedy. In *D. hansenii*, *PAP*, *OBP* and *PIK* are separate from the fused *MTL* locus32. B. Placement of gene losses on the phylogenetic tree.

Table 1

Candida genome features

Species*	Genome size (Mb)	%GC	Genes (#)	Gene size avg. (bp)	Intergenic avg. (bp)	Ploidy	Pathogen†
Candida albicans WO-1	14.4	33.5%	6,159	1,444	921	diploid	++
Candida albicans SC5314	14.3	33.5%	6,107	1,468	858	diploid	++
Candida tropicalis	14.5	33.1%	6,258	1,454	902	diploid	++
Candida parapsilosis	13.1	38.7%	5,733	1,533	752	diploid	++
Lodderomyces elongisporus	15.4	37.0%	5,802	1,530	1,174	diploid	−
Candida guilliermondii	10.6	43.8%	5,920	1,402	426	haploid	+
Candida lusitaniae	12.1	44.5%	5,941	1,382	770	haploid	+
Debaryomyces hansenii	12.2	36.3%	6,318	1,382	550	haploid	−

**C. albicans* SC5314 assembly 21 and gene set dated 28-Jan-2008 downloaded from the Candida Genome Database (www.candidagenome.org); *D. hansenii* assembly from GenBank9. The remaining assemblies are reported as part of this work, and are available in GenBank and at http://www.broad.mit.edu/annotation/genome/candida_group/MultiHome.html

†Relative level of pathogen strength (++, strong pathogen; +, moderate pathogen; −, rare pathogen).

Table 2

Gene families enriched in pathogenic *Candida* sp.

# Annotation	Pathogen genes	Nonpathogen genes	Pval.	Dup.	Loss	Gene rate	C.alb	C.tro	C.par	L.elo	C.gui	C.lus	D.han	C.gla	Yeast (avg.)
1 GPI-family 18 (Hyr/Iff-like)	56	10	1.4E-16	52	11	16.2	11	18	17	9	3	7	1	0	0.0
2 Leucine-rich repeat (IFA/FGR38-like)	34	0	4.2E-16	32	5	18.3	33	1	0	0	0	0	0	0	0.0
3 Ferric reductase family	45	10	1.9E-12	30	25	2.5	12	19	7	7	3	4	2	0	0.1
4 Reductase family	43	11	3.2E-11	31	30	2.3	7	12	9	6	13	2	4	0	0.1
5 GPI-family 17 (ALS-like Adhesins)	31	5	4.4E-10	29	4	20.5	8	16	5	4	2	0	1	0	0.0
6 GPI-family 13 (Pga30-like)	34	7	5.0E-10	25	5	14.8	12	14	6	6	1	1	1	0	0.0
7 Unclassified	20	0	9.0E-10	13	3	15.9	9	9	0	0	2	0	0	0	0.0
8 Cell wall mannoprotein biosynthesis	38	18	7.2E-07	19	34	2.1	8	7	8	8	11	4	9	0	0.1
9 Major facilitator transporters	25	7	9.2E-07	14	17	2.0	3	3	7	3	10	2	4	0	0.0
10 Oligopeptide transporters	31	13	2.2E-06	23	11	6.7	6	9	9	4	4	3	1	0	0.9
11 Unclassified	25	9	6.3E-06	15	6	11.1	7	9	3	5	3	1	4	2	0.2
12 Amino acid permeases	27	11	7.7E-06	11	18	1.7	6	6	6	4	6	3	6	0	0.1
13 Sphingomyelin phosphodiesterases	18	5	3.2E-05	11	9	7.4	4	5	4	2	3	2	1	0	0.2
14 FGR6 family (filamentous growth)	12	1	3.3E-05	7	1	14.5	8	1	1	0	1	1	1	0	0.0
15 Secreted lipases	20	7	4.6E-05	17	8	9.6	10	5	4	4	1	0	3	0	0.0
16 Cytochrome p450 family	34	21	5.5E-05	23	22	6.0	6	8	10	7	5	4	6	1	1.0
17 Amino acid permeases	16	4	5.6E-05	14	10	1.5	2	3	6	3	2	3	1	0	0.0
18 Zinc-finger transcription factors	31	18	6.2E-05	17	14	12.3	5	8	7	7	7	4	11	0	0.0
19 Unclassified	13	2	6.3E-05	8	0	8.1	3	1	6	1	2	1	1	0	0.0
20 Predicted transmembrane family	17	5	7.2E-05	9	2	7.5	4	4	5	3	3	1	2	0	0.0
21 Unclassified secreted family	20	8	1.1E-04	7	6	9.3	4	4	6	4	4	2	4	0	0.0

Pathogen genes = total genes in family for *C. albicans*, *C. tropicalis*, *C. parapsilosis*, *C. guilliermondii*, *C. lusitaniae*, and *C. glabrata*. Nonpathogen genes = total genes in family for *L. elongisporus*, *D. hansenii*, and all *Saccharomyces* clade species (Figure 1) expect *C. glabrata*. Pval.= Pvalue of the hypergeometric test; all families shown above have a false discovery rate less than 0.05 (Supplementary text S10c). Dup., Loss = Duplications and losses (Supplementary Methods 10b). Gene rate = average mutation rate for each family (Supplementary text S10d); the average gene rate across all families is 5.8. Yeast (avg.) = average count for all *Saccharomyces* species.