Genome Report: Whole genome sequence of the heterozygous clinical isolate *Candida krusei* 81-B-5

Christina A. Cuomo¹*, Terrance Shea¹, Bo Yang², Reeta Rao², Anja Forche³*

¹ Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA
² Worcester Polytechnic Institute, Biology & Biotechnology, Worcester, MA 01609 USA
³ Bowdoin College, Department of Biology, Brunswick, ME 04011 USA

*corresponding authors: cuomo@broadinstitute.org, aforche@bowdoin.edu

Data access: All genome sequence data (reads, assembly, and annotation) is available in GenBank under BioProject PRJNA381554.
Running title: Heterozygous genome of *Candida krusei*

Keywords: *Candida krusei*, 81-B-5, heterozygosity, LOH, mating type locus, transporters

Corresponding authors: Christina A. Cuomo, Broad Institute of MIT and Harvard, Cambridge, MA USA, 617-714-7904, cuomo@broadinstitute.org; Anja Forche, Bowdoin College, Department of Biology, Brunswick, ME 04011 USA, 207-725-3365, aforche@bowdoin.edu
Abstract

*Candida krusei* is a diploid, heterozygous yeast that is an opportunistic fungal pathogen in immunocompromised patients. This species also is utilized for fermenting cocoa beans during chocolate production. One major concern in the clinical setting is the innate resistance of this species to the most commonly used antifungal drug fluconazole. Here we report a high-quality genome sequence and assembly for the first clinical isolate of *C. krusei*, strain 81-B-5, into 11 scaffolds generated with PacBio sequencing technology. Gene annotation and comparative analysis revealed a unique profile of transporters that could play a role in drug resistance or adaptation to different environments. In addition, we show that while 82% of the genome is highly heterozygous, a 2.0 Mb region of the largest scaffold has undergone loss of heterozygosity. This genome will serve as a reference for further genetic studies of this pathogen.

Introduction

*Candida krusei* is a diploid, heterozygous yeast with an estimated chromosome number of 6 (Whelan and Kwon-Chung 1988; Samaranayake and Samaranayake 1994; Essayag et al. 1996; Jacobsen et al. 2007). *C. krusei* is an opportunistic fungal pathogen in immunocompromised patients, and unlike other major pathogenic *Candida* species (e.g. *C. albicans*) does not belong to the CUG clade (CTG is translated as Serine rather than Leucine) (Mühlhausen and Kollmar 2014). *Pichia kudriavzevii* (synonym *Issatschenkia orientalis*) is the teleomorphic (sexual) state of *C. krusei* (Kurtzman et al. 1980); it is one of the main fermenters of cocoa beans important for the development of chocolate aroma.

In recent years, human fungal infections caused by *C. krusei* have increased in the clinic mainly due to its innate resistance to theazole class of antifungal drugs specifically to fluconazole (Orozco et al. 1998; Guinea et al. 2006; Desnos-Ollivier et al. 2008; Lamping et al. 2009; Ricardo et al. 2014). Fluconazole is the first line antifungal and is also used as prophylactic treatment in the intensive care unit, and breakthrough Candidemia is increasingly caused by non- *albicans* species including *C. krusei* (Lischewski et al. 1995; Chaudhary et al. 2015; Cuervo et al. 2016). Moreover, there are incidences of resistance to the echinocandin class of antifungals, which are the drug of choice to fight *C. krusei* infections (Forastiero et al. 2015). Therefore, identifying the exact mechanisms that underlie drug resistance, and in particular azole resistance, is of utmost importance.

The mechanisms causing *C. krusei* to be innately resistant to fluconazole are not well understood. Studies have shown that *C. krusei* Erg11p, the drug target, is significantly less susceptible to FLC inhibition than most other fungal Erg11p proteins (Orozco et al. 1998; Fukuoka et al. 2003), and that efflux pumps such as Abc1p are at least partially responsible for the innate fluconazole resistance of *C. krusei* (Lamping et al. 2009). Other studies have shown that overexpression of both *ERG11* and *ABC2* genes might be responsible for resistance to other azole drugs (He et al. 2015).
One approach to examine the basis of drug resistance of *C. krusei* is to mine the genome sequence for genes with potential roles in resistance such as novel drug pumps or transporters. To date, genome sequences have been generated for five environmental strains of *C. krusei* (*P. kudriavzevii*); the only high quality assembly available for strain 129 isolated from fermented masau fruits (Van Rijswijck *et al.* 2017). A genome sequence for clinical isolates is still lacking. Here we report a high-quality genome sequence and assembly for clinical isolate *C. krusei* 81-B-5 (Scherer and Stevens 1987; Beckerman *et al.* 2001) into 11 scaffolds generated with PacBio sequencing technology. Gene annotation and comparative analysis revealed a unique profile of transporters that could play a role in drug resistance or adaptation to different environments. In addition, we show that while 82% of the genome is highly heterozygous, a 2.0 Mb region of the largest scaffold has undergone loss of heterozygosity.

**Methods & Materials**

**Sequencing methods and preparation**

High molecular weight genomic DNA was isolated from *C. krusei* strain 81-B-5 (Scherer and Stevens 1987; Beckerman *et al.* 2001) using a QIAGEN Genomic-tip 500/G kit (catalog # 10262). DNA was adapted using the SMRTbell template prep kit and sequenced using PacBio Technology (P6-C4 chemistry). A total of 3 SMRTcells were run, generating total of 266,621 subreads with mean read length 5758, with a total of 1,535,304,314 bases (~140X coverage). DNA was also adapted for Illumina sequencing, and a total of 16,953,446 paired 101b reads were generated on a HiSeq 2500.
Assembly and Annotation

An initial assembly was generated using HGAP (Chin et al. 2013) version 3 with smrtanalysis-2.3.0; HGAP was run with an estimated genome size of 14 Mb. As the genome was highly heterozygous, we also evaluated Falcon and Falcon-unzip (Chin et al. 2016) assemblies after Quiver polishing (using smrtanalysis-2.3.0). Falcon assembly settings were as follows: length_cutoff=10000; length_cutoff_pr=500;

pa_HPCdaligner_option = -v -dal4 -t16 -e.70 -l1000 -s1000 -M32;

ovlp_HPCdaligner_option = -v -dal4 -t32 -h60 -e.96 -l500 -s1000 -M32;

pa_DBsplit_option = -x500 -s1000; ovlp_DBsplit_option = -x500 -s1000;

falcon_sense_option = --output_multi --min_idt 0.70 --min_cov 2 --max_n_read 15 --n_core 6; overlap_filtering_setting = --max_diff 72 --max_cov 100 --min_cov 2 --bestn 12 --n_core 24. Falcon-unzip was run with default settings other than specifying settings for the SGE compute environment. Quiver (Chin et al. 2013) was then run on both assemblies to improve the consensus accuracy; initial evaluation of assemblies prior to Quiver polishing revealed a high rate of base errors. In both the HGAP and Falcon assemblies, contigs representing the alternative haplotype were identified based on high identity alignments to larger contigs in the assembly and roughly half the sequence depth in these regions; these alternative contigs were removed from both assemblies. Mitochondrial contigs were identified in all assemblies and set aside; the largest mitochondrial contig of 51.3 kb was assembled by HGAP assembly and smaller mitochondrial sequences were also identified in the Falcon or Falcon-unzip assemblies.
All assemblies were annotated to evaluate gene set completeness. An initial gene set was predicted using BRAKER (Hoff et al. 2016) to execute Genemark-ET with the parameter --fungus; tRNAs were predicted using tRNAscan (Lowe and Eddy 1997) and rRNAs predicted using RNAmmer (Lagesen et al. 2007). Genes containing PFAM domains found in repetitive elements or overlapping tRNA/rRNA features were removed. Genes were named and numbered sequentially.

**SNP calling**

Illumina reads were aligned to the HGAP C. krusei genome assembly using the Burrows-Wheeler Aligner (BWA) 0.7.12 mem algorithm (Li 2013) with default parameters. Across the total of 16,306,945 aligned reads, the average depth was 140.0X. BAM files were sorted and indexed using Samtools (Li et al. 2009) version 1.2. Picard version 1.72 was used to identify duplicate reads and assign correct read groups to BAM files. BAM files were locally realigned around INDELs using GATK (Mckenna et al. 2010) version 3.4-46 ‘RealignerTargetCreator’ and ‘IndelRealigner’. SNPs and INDELs were called from all alignments using GATK version 3.4-46 ‘HaplotypeCaller’ in GVCF mode with ploidy = 2, and genotypeGVCFs was used to predict variants in each isolate. Sites were filtered using variantFiltration with QD < 2.0, FS > 60.0, MQ < 40.0, and ReadPosRankSum < -8.0. Individual genotypes were then filtered if the minimum genotype quality < 50, percent alternate allele < 0.8, or depth < 10.

**Repeat analysis**

De novo repetitive elements were identified with RepeatModeler version
open-1.0.7 (www.repeatmasker.org/RepeatModeler.html); this identified only one unclassified element of length 1.3kb and further analysis revealed that this region contains an Arg-tRNA. To identify copies of previously identified elements, RepeatMasker version 4.0.5 (www.repeatmasker.org) was used to identify copies of the RepBase22.04 fungal elements. *Candida albicans* major repeat sequences were retrieved from the Candida Genome Database assembly version 22 (Skrzypek *et al.* 2017). Sequences were compared to the *Candida krusei* assembly using BLAST; no similarity was found at 1e-5, requiring an alignment length of 100 bases or larger.

### Comparative genomic analysis

Gene sets of *C. krusei*, *C. lusitaniae* (Butler *et al*. 2009), *C. albicans* (Jones *et al*. 2004; Van Het Hoog *et al*. 2007), *P. pastoris* (Love *et al*. 2016), *C. glabrata*, and *S. cerevisiae* ((Dujon *et al*. 2004) were compared using BLASTP (e<1e-10) and orthologs identified from the BLASTP hits using Orthomcl (Li *et al*. 2003). For the *CDR/MDR* gene family, protein sequences were aligned using MUSCLE (Edgar 2004) and alignments trimmed using TrimAl (Capella-Gutiérrez *et al*. 2009) with setting –gappyout. The best amino acid replacement model was selected using ProtTest version 3.4.2 (Darriba *et al*. 2011). A phylogeny was inferred using RAxML version 8.2.4 (Stamatakis 2014) with model GAMMALG and 1,000 bootstrap replicates.

### Karyotype analysis

Chromosome plugs were prepared using the CHEF Genomic DNA plug kit (Biorad) with the following modifications: Single colonies were transferred to 5 ml YPD broth (1%
yeast extract, 2% bacto peptone, 2% glucose) and incubated at 30°C for 18 hrs in a roller incubator. The lyticase incubation step was done for 24 hrs, and the CHEF plugs were incubated with Proteinase K for 48 hrs. For the final washing steps, plugs were transferred to 5 ml tubes containing 3 ml of wash buffer. Chromosomes were separated in a 0.8% agarose gel (certified Megabase agarose (Biorad), in 0.5 x TBE buffer) with a DRII pulsed-field gel electrophoresis (PFGE) apparatus (Biorad) using the following run parameters: Block1; 300 s initial and final switch, 3.9 V/cm, at a 120° angle for 24 hrs at 10°C, Block 2; 1000 s initial and final switch at 2.7 V/cm at a 120° angle for 48 hrs at 10°C. The gel was stained with ethidium bromide (0.5 µl/ml) for 15 min, destained in distilled water for 15 min and photographed. *S. cerevisiae* and *Hansenula wingei* (*H. wingei*) chromosome size markers (Biorad) were used for size estimates.

**Phenotypic analyses**

Standard growth and media conditions have been previously described (Chauhan and Kruppa 2009). An Etest was used to determine the MIC for fluconazole (Pfaller *et al.* 2003). Briefly, overnight cultures were grown in YPD, washed and diluted to a final A600 of 0.1. Five hundred microliters were spread onto RPMI1640 agar plates (buffered with MOPS). After a 30 min pre-incubation, an Etest strip was applied and plates were incubated at 30°C for 48 hrs. The susceptibility endpoint reported was read at the first growth inhibition ellipse.

To confirm the non-filamentous phenotype of *C. krusei*, 3 ml of YPD overnight cultures were washed, cells were counted, and $10^3$ cells were transferred to wells of a 12-well petri plate containing 1 ml RPMI1640 with 10% fetal bovine serum. Plates were
incubated at 37°C and microscopic images were taken at 2, 4, and 8 hrs. *C. albicans* (SC5314) and *S. cerevisiae* (S288c) were used for positive (filamenting) and negative (non-filamenting) controls, respectively.

**Results and Discussion**

**Strain sequenced and phenotypic characterization**

The sequenced isolate *C. krusei* 81-B-5 (number 653 in Scherer strain collection) was collected from a clinical source prior 1987 (Scherer and Stevens 1987). To confirm that strain 81-B-5 is resistant to fluconazole, strains were grown in the presence of fluconazole and an Etest was done confirming the drug resistant phenotype with a minimum inhibitory concentration (MIC) of 32 µg/mL (Fig. S1), which is considered highly resistant (Pfaller *et al.* 2003; Espinel-Ingroff *et al.* 2014). To verify the non-filamentous phenotype of *C. krusei*, cells were exposed to serum, a potent inducer of filamentation and microscopically observed over time. Our results confirm that *C. krusei* does not filament as compared to *C. albicans* (Fig. S2).

**Genome sequencing and assembly**

We sequenced the genome of *Candida krusei* using PacBio technology to generate long reads. Early attempts to assemble the genome using Illumina or 454 data had resulted in highly fragmented assemblies ((Chan *et al.* 2012), JQFK00000000, BBOI00000000), and we reasoned that the heterozygosity detected in MLST analyses (Jacobsen *et al.* 2007) could likely complicate short read assembly. In assembling the genome, we compared assemblies generated using three methods, HGAP, Falcon, and Falcon-unzip, and
evaluated metrics for the haploid consensus produced by HGAP and Falcon to the diploid assembly produced by Falcon-unzip. In addition to evaluating assembly metrics, we predicted gene calls on each assembly and evaluated gene set completeness as an additional metric.

While overall assembly statistics were similar, both assembly and gene metrics were superior on the HGAP version (Table S1). The HGAP assembly contained only 11 scaffolds, whereas nearly twice this number were generated by Falcon or in the Falcon-unzip primary contigs. The total assembly size in these assemblies was very similar, with 63kb more sequence in the Falcon-unzip assembly compared to the HGAP assembly. As our prior experience in assembling diploid Candida genomes revealed that consensus errors can result in gene truncations where haplotypes are merged in a haploid assembly (Butler et al. 2009), we compared gene metrics across the three assemblies. Gene sets were compared to Candida albicans to evaluate completeness. By this measure of gene content, the gene set on the HGAP assembly appears to be of higher quality, with 135 more C. albicans orthologs compared to the Falcon assembly and 303 more than the Falcon-unzip. Gene length was also compared and not found to be very different; genes in the Falcon-unzip assembly were 16 bases larger on average than those in the HGAP. We also evaluated gene content on the second haplotype assembled by Falcon-unzip; these scaffolds totaled 2.1 Mb less than the other assemblies, and correspondingly fewer genes were predicted (Table S1). The completeness of the HGAP gene set was also evaluated by comparing to the BUSCO set of 1,438 fungal orthologs (Simão et al. 2015). A total of 1,278 appear complete in the C. krusei gene set. By comparison, this count is
similar to the 1,296 complete orthologs in *C. lusitaniae* but fewer than the 1,412 orthologs identified in the *C. albicans* genome, which has been extensively annotated (Braun *et al.* 2005; Butler *et al.* 2009; Bruno *et al.* 2010; Skrzypek *et al.* 2017). Based on considering both the assembly and gene metrics, we selected the HGAP assembly to represent the genome (Table 1). Compared to a previously reported draft genome (Chan *et al.* 2012), our assembly is more contiguous (11 contigs compared to 626 contigs for the PA12 assembly); the total size and gene number are comparable, with our assembly including 0.5 Mb more of sequence and a slightly higher gene count. A recently reported genome of isolate 129 using a hybrid of PacBio and Illumina in the assembly was also more fragmented (260 contigs) (Van Rijswijk *et al.* 2017); this assembly was larger in terms of total size (0.77 Mb), suggesting that some regions may be represented by both haplotypes in this assembly.

This *Candida krusei* genome shows a high rate of heterozygous SNP variants and one large region of loss of heterozygosity on scaffold 1. Using Illumina sequence, a total of 32,131 heterozygous SNPs were identified, for an average rate of 1 SNP every 340 positions. While SNPs were distributed across the genome assembly, a 2.0 Mb region of scaffold 1 has undergone loss of heterozygosity; the first 0.6 Mb of scaffold 1 has a typical frequency of SNP variants, however very few variants were detected across the remainder of the scaffold (Fig. 1A). This homozygous region is not represented in the alternate haplotype contigs assembled by Falcon-unzip, and this difference explains the smaller assembly size of the Falcon-unzip assembly. All of scaffold 1 is present at diploid levels, and we detect no large regions of aneuploidy in this isolate (Fig. 1B).
The *Candida krusei* genome contains very few repetitive sequences. A search for conserved repetitive elements classified only 0.40% of the assembly as interspersed repeats, with an additional 1.89% of sequence representing simple repeats. There are no regions with significant similarity (BLAST, 1e-5) to the *C. albicans* major repeat sequences (Methods). The average GC content is 38.4%, which is intermediate compared to related species such as *C. albicans* (33.5%) or *C. lusitaniae* (44.5%) (Jones et al. 2004; Van Het Hoog et al. 2007; Butler et al. 2009).

**Chromosome structure**

PFGE was used previously to estimate the number of chromosomes for clinical and environmental isolates of *C. krusei* (Iwaguchi et al. 1990; Doi et al. 1992; Dassanayake et al. 2000; Jespersen et al. 2005). Based on the chromosomal patterns it was estimated that *C. krusei* has a total of 4-6 chromosomes: ~2-4 large chromosomes (~2.8 - 3.5 Mb) and 2 small chromosomes (~1.4 Mb). PFGE for *C. krusei* strain 81-B-5 showed approximately 5 chromosomal bands, which were numbered based on size with 1 being the largest chromosome (Chr1) (Fig. 2). Chromosome sizes were estimated based on the *H. wingei* and *S. cerevisiae* chromosome standards and 3 non-krusei *Candida* species with known chromosome sizes (Doi et al. 1992; Butler et al. 2009): Chr1 (3.1 Mb), Chr2 (2.9 Mb), Chr3 (2.7 Mb), Chr4 (1.4 Mb) Chr5 (1.3 Mb) (Fig. 2). Based on these sizes the estimated genome size is 11.4 Mb, which is in good agreement with the size of the genome assembly. CHEF Southerns will be required to assign each scaffold to its
appropriate chromosome, and additional work would be needed to establish the order and orientation of scaffolds along each chromosome.

By searching for tandem repeats at scaffold ends, we identified a candidate telomeric repeat (ATTGTAACACACCTCGCTCCTAGTTCAT). This repeat is found at 5 scaffold ends, including the start of scaffold 1, end of scaffold 3, both ends of scaffold 4, and start of scaffold 10. This suggests that scaffold 4 is a complete chromosome, and that four other scaffolds extend to the telomeres. rDNA repeats are detected at the end of scaffold 1, across scaffold 11, and end of scaffold 9, suggesting that these scaffolds may be joined in a single chromosome to form a continuous rDNA array.

Comparative genomics

To provide a preliminary view of the genes involved in pathogenesis and drug resistance, we identified orthologs of *C. albicans* genes in the *C. krusei* genome. Overall, gene families involved in pathogenesis in *C. albicans* are present in fewer copies in *C. krusei*. We identified fewer copies of the secreted aspartyl proteases, oligopeptide transporters, and phospholipase B genes (Table S2). In addition we did found no copies of genes similar to the secreted lipase or *ALS* cell surface families of proteins from *C. albicans*. This result is consistent with prior comparison to a wider set of pathogenic *Candida* more closely related to *Candida albicans*, which observed expansion of several of these families in the more commonly pathogenic species (Butler *et al.* 2009). We also identified orthologs of genes noted to be involved in drug resistance in *C. albicans*, via point mutations, increased transcription, or copy number variation. *C. krusei* contains a
single copy of the \textit{ERG11} azole target and of each of the \textit{TAC1} and \textit{UPC2} transcription factors. Several of the sites often subject to drug resistant mutations in \textit{C. albicans} are conserved in \textit{C. krusei} (i.e. Y132, K143, and F126), suggesting no intrinsic azole resistance due to mutation of these sites in \textit{C. krusei}. While we did not identify a copy of the \textit{MDR1} drug transporter, we identified 9 candidate transporters related to \textit{CDR1}, \textit{CDR2}, and related genes (Fig. 3). These include 3 \textit{C. krusei} genes related to \textit{CDR1/CDR2/CDR11/CDR4}, 4 genes related to \textit{SNQ2/PDR18}, and two genes related to \textit{PDR12}. This may suggest a very different capability for drug efflux.

While previous genomic studies have revealed the highly variable content of the mating type locus in pathogenic \textit{Candida} species (Butler et al. 2009), the mating type locus in \textit{C. krusei} appears complete and is more similar to that of Saccharomycetaceae yeasts than the CTG clade \textit{Candida}. The mating type locus in \textit{C. krusei} is found on scaffold 5, and includes the \textit{MTLa1} gene and \textit{MTLa2} located adjacent to \textit{SLA2} (Fig. 4), similar to the configuration in many Saccharomycetaceae yeasts (Gordon et al. 2011). The mating type locus is close to the start of scaffold 7, separated from the end by four genes. Three other genes typically found at the mating locus of CTG clade \textit{Candida} species (Butler et al. 2009) are located on adjacent scaffolds; \textit{PAPI} and \textit{OBPA} are adjacent on scaffold 7 and \textit{PIKA} is on scaffold 2. While the related species \textit{Pichia pastoris} and \textit{Hansenula polymorpha} contain two \textit{MAT} loci (Hanson et al. 2014), only one copy of \textit{MTL1}, \textit{MTLa2}, and \textit{SLA2} were found in the \textit{C. krusei} assembly. This locus is potentially subtelomeric, as the start of the \textit{SLA2} gene is 7.4 kb from the start of scaffold 5. The \textit{MTL} region is heterozygous (Figure 5), as observed in some \textit{MTLa/a} and \textit{MTLα/α} \textit{C. albicans} isolates.
Both of the other assembled genomes of *C. krusei* also contain the MTL\(a\) idiomorph, based on blastp to the available gene set for the 129 assembly or tblastn to the available assembly for M12. This information could guide a search for isolates of the opposite mating type, to begin to study whether *Candida krusei* is capable of sexual reproduction.

**Data availability**

All genome sequence data (reads, assembly, and annotation) is available in GenBank under BioProject PRJNA381554. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NHMM00000000. The version described in this paper is version NHMM01000000.

**Acknowledgements**

We thank the Broad Technology Labs and Broad Genomics Platform for generating the genome sequence for *Candida krusei*. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute and by NIAID grant R15 AI090633 to A. Forche.
References


Chauhan, N., and M. D. Kruppa, 2009 Standard growth media and common techniques for use with *Candida albicans*, pp. 197-201 in *Candida albicans: Methods and


He, X., M. Zhao, J. Chen, R. Wu, J. Zhang et al., 2015 Overexpression of both *ERG11* and *ABC2* genes might be responsible for itraconazole resistance in clinical isolates of *Candida krusei*. PLOS ONE 10: e0136185.


Van Het Hoog, M., T. J. Rast, M. Martchenko, S. Grindle, D. Dignard et al., 2007 Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol 8: R52.

Van Rijswijck, I. M. H., M. F. L. Derks, T. Abbe, D. De Ridder and E. J. Smid, 2017 Genome sequences of Cyberlindnera fabianii 65, Pichia kudriavzevii 129, and
*Saccharomyces cerevisiae* 131 isolated from fermented masau fruits in Zimbabwe. Genome Announcements 5: e00064-00017.

Figure 1. Genome-wide heterozygosity and genome coverage. A. Heterozygous SNP positions are plotted across the assembly scaffolds in windows of 5 kb. B. Normalized read depth is plotted across the assembly scaffolds in windows of 5 kb. Scaffold 11, consisting of ~6 ribosomal DNA repeats, is not depicted.
Figure 2. Karyotype analysis of *C. krusei* strain 81-B-5 reveals 5 chromosomal bands. A. short run to separate chromosomes smaller than 2 Mb, B. long run to separate all chromosomes. The chromosomes for *C. krusei* are labeled 1 through 5. Several other *Candida* species were run as references; *S. cerevisiae* and *H. wingei* standards (Biorad) were used for chromosome size estimation of *C. krusei* chromosomes.
Figure 3. Phylogeny of Cdr and Mdr proteins in *C. krusei* and related species. Cdr and Mdr proteins identified across 6 species were aligned and used to infer a phylogeny using RAxML (Methods). Prefix for each protein corresponds to the species as follows: Ca, *C. albicans*; Cl, *C. lusitaniae*; Ck, *C. krusei*; Pp, *P. pastoris*; Cg, *C. glabrata*; Sc, *S. cerevisiae*.
Figure 4. Mating type locus of *Candida krusei*. Genes adjacent to the mating type locus of *C. krusei* differ from the CTG clade *Candida* species; there is a single copy of MATa1 and MATa2 found in the assembly, adjacent to the SLA2 gene, whereas the OBP, PIK, and PAP genes are found on other scaffolds in the assembly.
Table 1. *Candida krusei* genome statistics

<table>
<thead>
<tr>
<th>Scaffolds</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>11</td>
</tr>
<tr>
<td>Total bases</td>
<td>10,910,993</td>
</tr>
<tr>
<td>Contig N50 length</td>
<td>1.36 Mb</td>
</tr>
<tr>
<td>Contig N90 length</td>
<td>543 kb</td>
</tr>
<tr>
<td>SNP rate</td>
<td>1 SNP/ 340 bases</td>
</tr>
<tr>
<td>GC content</td>
<td>38.42%</td>
</tr>
<tr>
<td>Repeat content</td>
<td>2.15%</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>4,949</td>
</tr>
</tbody>
</table>