Genome plasticity in *Candida albicans* is driven by long repeat sequences

Robert T. Todd[1], Tyler Wikoff[1], Anja Forche[2], Anna Selmecki[1]*

[1]Creighton University Medical School, 2500 California Plaza, Omaha, Nebraska 68178

[2]Bowdoin College, 255 Maine Street, Brunswick, Maine 04011

*Corresponding author

Anna Selmecki

Department of Medical Microbiology and Immunology

Creighton University

2500 California Plaza

Omaha, NE 68178

Phone: (402) 280-4096

FAX: (402) 280-1875

E-mail: annaselmecki@creighton.edu

KEY WORDS

Inverted Repeats, Genome instability, Segmental aneuploidy, Loss of heterozygosity, chromosomal inversion, *in vivo* and *in vitro* evolution, antifungal drug resistance, *Candida albicans*

ABSTRACT

Genome rearrangements resulting in copy number variation (CNV) and loss of heterozygosity (LOH) are frequently observed during the somatic evolution of cancer and promote rapid adaptation of fungi to novel environments. In the human fungal pathogen *Candida albicans*, CNV and LOH confer increased virulence and antifungal drug resistance, yet the mechanisms driving these rearrangements are not completely understood. Here, we unveil an extensive array of long repeat sequences (65-6499bp) that are associated with CNV, LOH, and chromosomal inversions. Many of these long repeat sequences are uncharacterized and encompass one or more coding sequences that are actively transcribed. Repeats associated with genome rearrangements are predominantly inverted and separated by up to ~1.6Mb, an extraordinary distance for homology-based DNA repair/recombination in yeast. These repeat sequences are a significant source of genome plasticity across diverse strain backgrounds including clinical, environmental, and experimentally evolved isolates, and previously uncharacterized variation in the reference genome.

INTRODUCTION

Genome plasticity is surprisingly common in eukaryotes. DNA insertions and deletions (indels), copy number variations (CNV), and loss of heterozygosity (LOH) are frequently described during the evolution of organisms and of disease states such as cancer. In particular, the genome plasticity of fungal pathogens was recognized well before whole genome sequencing was available, including genome copy number variation (polyploidy), inter- and intra- chromosomal rearrangements, and aneuploidy (Chibana et al., 2000; Magee & Magee, 2000; Rustchenko-Bulgac, 1991; Suzuki et al., 1982). Controlled *in vitro* and *in vivo* evolution experiments in combination with whole genome sequencing have further highlighted the speed in which specific genome rearrangements provide a fitness advantage that can be selected for in these fungal pathogens (Araya et al., 2010; Croll et al., 2013; Dunham et al., 2002; Forche et al., 2011; Ford et al., 2015; Gerstein et al., 2015; Hirakawa et al., 2015; Selmecki et al., 2009; Stukenbrock et al., 2010).

Candida albicans is the most prevalent human fungal pathogen, associated with nearly half a million life-threating infections annually, predominantly in immunocompromised individuals (Brown & Netea, 2012). *C. albicans* is a heterozygous diploid yeast capable of mating, yet true meiosis has not been observed. Instead, it undergoes a parasexual process that involves random chromosome loss and rare Spo11-dependent chromosome recombination events (Bennett & Johnson, 2003; Forche et al., 2008; Wang et al., 2018).

The majority of genomic diversity observed in *C. albicans* is attributed to asexual mitotic genome rearrangements (Forche et al., 2011; Lephart & Magee, 2006). Despite this clonal lifestyle, *C. albicans* isolates exhibit extensive genomic diversity in the form of *de novo* base substitutions, indels, ploidy variation (haploid, diploid, and polyploid), karyotypic variation due to segmental and whole chromosome aneuploidies, and allele copy number variation including LOH (Chibana et al., 2000; Forche et al., 2011; Ford et al., 2015; Hickman et al., 2013; Hirakawa et al., 2015; Magee & Magee, 2000; Rustchenko-Bulgac, 1991; Selmecki et al., 2006; Suzuki et al., 1982). Additionally, while *C. albicans* did not undergo an ancient whole genome duplication event like *Saccharomyces cerevisiae* (Butler et al., 2009; Marcet-Houben et al., 2009; Wolfe & Shields, 1997)*,* small-scale duplication events have resulted in gene family expansions, especially in sub-telomeric regions (Anderson et al., 2012; Butler et al., 2009; Dunn et al., 2018). A comprehensive analysis of these duplication events, their evolutionary trajectories and impact on genome stability, remains largely unexplored.

Early comparative studies of the *C. albicans* genome identified diverse repetitive loci that contribute to genotypic and phenotypic plasticity (Braun et al., 2005; Jones et al., 2004). First, repeat analysis in *C. albicans* has characterized at least three major classes of long repetitive sequences: the 23 bp tandem telomeric repeat units and the 14 member telomere-associated (*TLO)* gene family residing in sub-telomeric regions; the Major Repeat Sequences (MRS) found on nearly every *C. albicans* chromosome and formed by a long tandem array of ~2.1 kb RPS units flanking non-repetitive HOK and RBP-2 elements (Chibana et al., 1994; Chindamporn et al., 1998; Lephart & Magee, 2006); and the ribosomal DNA repeats (rDNA) found on ChrR, which are organized into a tandem array of up to ~200 copies of ~12 kb units (Freire-Beneitez et al., 2016; Jones et al., 2004; Rustchenko et al., 1993; Wickes et al., 1991). These long repetitive sequences can undergo both inter- and intra-locus recombination events that rapidly generate chromosome length polymorphisms, chimeric chromosomes, and telomere-telomere chromosomal fusions (Chu et al., 1992; Selmecki et al., 2006, 2010). Secondly, like most eukaryotes, *C. albicans* also encodes many “lone” long terminal repeats (LTRs) and retroelements (Zorro, Tca2, Ty1/Copia) (Goodwin & Poulter, 1998, 2000), however the relative copy number of many of these genes is hypervariable between *C. albicans* isolates and are expanded relative to other Candida species (Butler et al., 2009; Hirakawa et al., 2015). Third, short repeat sequences (short tandem repeats and trinucleotide repeats) are significantly more frequent in protein-coding sequences of *C. albicans* than in *S. cerevisiae* and *S. pombe* (Braun et al., 2005; Jones et al., 2004). Fourth, expansion of multi-gene families (identified by protein alignment) were both more common and larger than the orthologous gene family size found in *S. cerevisiae*. These gene families often encode proteins with roles in commensalism and virulence, including the agglutinin-like sequence (*ALS*) family (eight genes) and other glycosylphosphatidylinositol (GPI)-linked genes that encode large cell-surface glycoproteins (five genes) (Levdansky et al., 2008; Wilkins et al., 2018). Among these gene families, recombination and/or slippage between repeat units yields extensive allelic variation, leading to functional and phenotypic diversity, similar to the *FLO* genes in *S. cerevisiae* (Hoyer et al., 1995; Kunkel, 1993; Pearson et al., 2005; Richard et al., 1999; Verstrepen et al., 2005; Zhang et al., 2003; Zhao et al., 2004). The evolution of different alleles in these repeat-containing ORFs predominantly occurs by the addition, deletion, and rearrangement of repeat units within an ORF and between different ORFs, not by the acquisition of point mutations or indels (Christiaens et al., 2012; Zhang et al., 2010). Importantly, these genomic studies focused on short repeat sequences and repeats found in protein-coding sequences. Less is known about long repeat sequences found throughout the genome, especially those encoding multiple ORFs and intergenic regions.

Over 19 years ago, Wolfe and colleagues showed that the *C. albicans* genome contains thousands of small chromosomal inversion events (~10 genes long) relative to *S. cerevisiae*. These inversions resulted in substantially different gene order between these two species (Seoighe et al., 2000). Similarly, Dujon and colleagues demonstrated that the *C. albicans* genome had the highest rate of genome instability due to micro- and macro-rearrangements of syntenic gene blocks, relative to 11 other hemiascomycete species (Fischer et al., 2006). The loss of synteny primarily resulted from chromosomal rearrangements, not sequence divergence of orthologous regions. A mechanism proposed for this genome instability was a higher incidence of repetitive sequences and/or a less efficient DNA repair process (Fischer et al., 2006).

The genomic diversity of *C. albicans* increases during *in vitro* and *in vivo* exposure to stress. For example, rates of LOH increase during exposure to elevated temperature (37°C), DNA transformation, and antifungal drugs (Bouchonville et al., 2009; Forche et al., 2011; Forche et al., 2018). LOH is also increased during *in vivo* models of infection (Ene et al., 2018; Forche et al., 2008; Forche et al., 2018). LOH events occur due to chromosome nondisjunction leading to whole chromosome LOH or via recombination, in which only part of the chromosome undergoes LOH. Exposure to stress also selects for isolates that have acquired adaptive mutations and genome rearrangements. For example, aneuploidy is found in ~50% of isolates resistant to the most common antifungal drug, fluconazole. The most common and only recurrent aneuploidy in different strain backgrounds is the amplification of the left arm of chromosome 5 (Chr5L), often through acquisition of a novel isochromosome structure (denoted as i(5L)), comprised of two copies of Chr5L separated by the centromere (Selmecki et al., 2006; Selmecki et al., 2008). Acquisition of i(5L) conferred fluconazole resistance via the amplification of two genes, *ERG11* and *TAC1*, encoding the drug target (Erg11) and a transcriptional activator of drug efflux pumps (Tac1) (Selmecki et al., 2008; Selmecki et al., 2009). Importantly, the centromere of Chr5 contains a long inverted repeat sequence, and recombination between these repeats can form homozygous isochromosomes of both the left arm (i(5L)) and right arm of Chr5 (i(5R)) (Selmecki et al., 2006). The role of long repeat sequences in the formation of other segmental aneuploidies and other genome rearrangements has not been comprehensively addressed.

We provide evidence that long repeat sequences are involved in the formation of all observed CNV breakpoints and chromosome inversions, and many LOH breakpoints, across 33 diverse clinical and experimentally evolved isolates. Our comprehensive analysis of long repeat sequences within the *C. albicans* genome identified hundreds of sequences representing novel multicopy repeats, none of which include MRS, rDNA, sub-telomeric repeats, known repeat families (*ALS*, *TLOs*) or known repetitive elements (tRNAs, LTRs, retrotransposons). Long repeats that are associated with genome rearrangements (CNV, LOH, and inversions) have on average higher sequence identity than all long repeats combined. Additionally, long repeats that contain ORFs (including partial ORF sequences, single complete ORF sequences (paralogs), or multiple ORFs and intergenic sequences) are longer and associated with more genome rearrangements than long repeats that contain other genomic features (such as LTRs, retrotransposons, or tRNAs). Additionally, repeat copies involved in genome rearrangements can be located up to ~1.6 Mb apart on the same chromosome, suggesting a non-conventional, long-range mechanism for DNA double-strand break (DSB) repair and somatic genome diversification.

RESULTS

An inverted repeat within *CEN4* is associated with the formation of a novel isochromosome

To identify the mechanisms by which *C. albicans* isolates generate genome plasticity, we performed a comparative genomics analysis of 33 diverse clinical isolates (Supplementary File 1). This set of isolates included 11 that underwent controlled experimental evolution, where a known progenitor isolate was passaged *in vitro* or *in vivo*. Additionally, we performed comparative genomics on newly obtained clinical isolates, and clinical isolates whose genomes were published previously, including the reference genome sequence SC5314.

Given the significant impact of i(5L) on antifungal drug resistance, we focused first on the characterization of a novel segmental aneuploidy detected on Chr4 that arose during *in vitro* evolution in the presence of fluconazole (FLC). Initially, we passaged a FLC-sensitive clinical isolate P78042, which was trisomic for Chr4 (Hirakawa et al., 2015; Lockhart et al., 2002), in the presence of FLC (128 µg/ml) for 100 generations by serial dilution (See Methods). One evolved isolate (AMS3743) was selected, based on increased fitness in FLC (see below), and the whole genome was sequenced. Read depth analysis indicated that this isolate had 4 copies of the right arm of Chr4 (Chr4R), but only two copies of Chr4L, and the copy number breakpoint occurred at the centromere of Chr4 (*CEN4*) (Figure 1A). Wildtype *CEN4*, like *CEN5,* is comprised of a CENP-A-binding core sequence (~3.1 kb) flanked by a long (524 bp) inverted repeat (Burrack et al., 2016; Ketel et al., 2009; Sanyal et al., 2004).

To test the hypothesis that this segmental aneuploidy is an isochromosome structure, we performed CHEF karyotype analysis. Isolate AMS3743 had a novel ~1.2 Mb chromosome band that hybridized to a *CEN4* probe via Southern blot (Figure 1B). This ~1.2 Mb band was twice the size of the right arm of Chr4 (~607 Kb). Consistent with an isochromosome i(4R) structure (a centromere flanked by inverted copies of Chr4R), a single primer amplified a ~4.1 kb product, from Chr4R through *CEN4* and back to Chr4R in the isolate with i(4R) but did not amplify any sequence in the reference (SC5314), or progenitor (P78042) isolates (Figure 1C).

Next, we determined the impact of i(4R) on fitness in the presence and absence of FLC over a 24-hour period. In the presence of FLC, the i(4R) isolate grew significantly better than the progenitor P78042 ($p < 0.0006$, t-test, Figure 1D). Interestingly, in the absence of FLC, the i(4R) isolate grew as well as the progenitor P78042 (Figure 1D). Furthermore, i(4R) was maintained in 12/12 populations for over ~300 generations in the absence of FLC (See Methods). One of the populations, AMS3743_10, appeared to be losing i(4R) by CHEF gel densitometry (See Methods) and was plated for single colonies in the absence of FLC. One colony (out of six) had lost i(4R) (AMS3743_10_S6, Figure 1-figure supplement 1A). To ask if i(4R) was necessary and sufficient for the increased fitness in FLC, fitness was determined in the presence and absence of FLC. The colony that had lost i(4R) had a reduced growth rate in the presence of FLC, similar to the progenitor P78042 (Figure 1-figure supplement 1B).

Overall, these data imply that the long inverted repeat within *CEN4* can generate an independent isochromosome structure comprised of two right arms of Chr4, and that i(4R) is necessary and sufficient for increased fitness in FLC. These results parallel the identification of isochromosomes associated with the long inverted repeat sequence within *CEN5,* which can result in the formation of i(5R) and i(5L), the latter of which confers FLC resistance (Selmecki et al., 2006; Selmecki et al., 2008).

Inverted repeat sequences are associated with inversion of centromere sequences

During our investigation of the i(4R) structure, we unveiled a surprising feature of *CEN4*: the CENP-A-binding core sequence of *CEN4* contained two different alleles. One homologue of Chr4 contained a ~3.1 kb sequence inversion between the inverted repeat associated with *CEN4*. The new, inverted *CEN4* sequence was detected by PCR in the reference strain SC5314, and in the distantly related isolates P78042 and AMS3743 (Figure 1-figure supplement 1C & D). Sanger sequencing indicated that a recombination event occurred between the two arms of the inverted repeat (Figure 1-figure supplement 2). Interestingly, the CENP-A-binding core sequence of *CEN4* is asymmetrically positioned on one side of the inverted repeat sequence (Figure 1-figure supplement 1D, shaded region) (Burrack et al., 2016; Sanyal et al., 2004). Therefore, this inversion caused a separation between the known CENP-A-binding core sequence of *CEN4* that is located to the right and outside of the inverted repeat.

Identification of long repeat sequences throughout the *C. albicans* genome

Given the extensive genome rearrangements observed at the long inverted repeat associated with *CEN4*, we sought to characterize all long repeat sequences within the *C. albicans* reference genome. All long sequence matches within the reference genome SC5314 were identified by aligning the reference genome sequence to itself using the bioinformatics suite MUMmer (Kurtz et al., 2004). First, all exact sequence matches of 20 nucleotides or longer were identified, then all matches were clustered and extended to obtain a maximum-length colinear string of matches, resulting in a final list of long repeat sequences that ranged from 65 bp to 6499 bp (median 318 bp) with sequence identities of ≥80% (See Methods). The genomic position and percent identity of all matched repeats was determined with MUMmer and manually verified using BLASTN and IGV (Robinson et al., 2011; Thorvaldsdottir et al., 2013). After excluding all rDNA, MRS and sub-telomeric repeat sequences, 1974 long repeat matches were identified (Supplementary File 2). The MUMmer analysis identified five ORFs and one gene family with known, complex embedded tandem repeat sequences (*PGA18*, *PGA55, EAP1, orf19.1725, CSA1,* and the *ALS* gene family, herein referred to as 'the complex tandem repeat genes'). The complexity of these repeat sequences prohibited the assignment of exact repeat copy number per genome, and they were removed from analyses when indicated. The remaining long repeat sequences cover 2.87% of the haploid reference genome (See Methods).

Long repeat matches occurred between sequences on the same chromosome (Intra-chromosomal repeats, Figure 2A), on different chromosomes (Inter-chromosomal repeats), or both. The number of all repeat matches per chromosome was correlated with chromosome size ($R^2 = 0.65$, $p < 0.016$, Figure 2B), however regions of high repeat density (e.g. ChrRR near the rDNA) or low repeat density (e.g. Chr7L) were detected on some chromosome arms. This repeat density did not correlate with GC content ($R^2 = 0.063$, $p > 0.32$) or ORF density ($R^2 = 0.02$, $p > 0.59$) on any chromosome arm (Figure 2-source data 1).

We next calculated the orientation and distance between matched intra-chromosomal repeat sequences (Figure 2-figure supplement 1), both important factors for reconstructing the evolutionary history of these duplication events and for analyzing the frequency and outcome of homologous recombination events that occur between repeat sequences (Lobachev et al., 1998; Ramakrishnan et al., 2018). Intra-chromosomal repeats are often generated in tandem by recombination between sister chromatids or replication slippage, and these repeats can move further away from each other by chromosomal rearrangement events (including chromosomal inversions) (Achaz et al.; Reams & Roth). Indeed, intra-chromosomal repeats were predominantly tandem, although inverted and mirrored repeats also occurred (Supplementary File 2). We hypothesized that the distance between matched intra-chromosomal repeats (spacer length) would be predominantly short and that the distribution of spacer lengths on each chromosome would be similar. Strikingly, spacer length ranged from 1 bp to 2,856,212 bp (median ~82.8 kb, excluding the complex tandem repeat genes, See Methods), and was correlated with chromosome size (Figure 2-figure supplement 2A, $R^2 = 0.066$, $p < 0.0001$). Additionally, the distribution of spacer lengths was significantly different between chromosomes (Figure 2-figure supplement 2B, $p < 0.035$, Kruskal-Wallis test with Dunn's multiple comparison) with the larger chromosomes (Chr1 and ChrR) containing many repeat matches that were separated by distances greater than ~1.5 Mb. The increased distance between repeat sequences likely occurred via additional large inversions, insertions or telomere-telomere recombination/fusion events.

We further annotated the long repeat sequences according to the genomic features contained within each repeat (See Methods). The most common long repeats contained lone long terminal repeats (LTRs) (775), followed by ORFs (339, excluding the complex tandem repeat genes), tRNAs (334), and retrotransposons (40). Repeat matches containing ORFs included partial ORF sequences (196/339, 57.8%), single complete ORF sequences (114/339, 33.6%), and multiple ORFs and intergenic sequences (29/339, 8.6%) (Supplementary File 2). Repeat matches containing complete ORFs and multiple ORFs represent paralogs and multi-gene duplication events. Additionally, there were 349 intergenic, unannotated sequences, 231 that shared high sequence identity (> 83%) with an annotated sequence found elsewhere in the genome, including known LTRs, retrotransposons, and ORFs (Supplementary File 2, ‘Unannotated Intergenic Sequence’). For example, an additional 54 LTRs were identified in the reference genome with this analysis. Interestingly, LTR matched repeat pairs were predominantly dispersed on different chromosomes (78%), while ORF matched repeat pairs were predominantly located on a single chromosome (64%, Figure 2C).

Of the matched repeat pairs, the long repeat sequences containing ORFs had the lowest median sequence identity when compared to repeats containing other features (Figure 2-figure supplement 3A, $p < 0.0001$, Kruskal-Wallis followed by Dunn’s multiple comparison test). Conversely, repeats containing ORFs had significantly longer copy length than any other genomic feature ($p < 0.0001$, Kruskal-Wallis followed by Dunn’s multiple comparison test) and was the only feature that had a significant increase in copy length of intra-chromosomal matches relative to inter-chromosomal matches (Figure 2-figure supplement 3B, $p < 0.0001$, Kruskal-Wallis followed by Dunn’s multiple comparison test). The long repeat sequences containing ORFs were predominantly present in only two copies per genome, had pairwise coding sequences with similarly high identity, and therefore represent paralogous gene duplication events (Supplementary File 2). The origin, function, and evolutionary trajectory of these paralogs may provide insight into the evolution of fungal pathogens like *C. albicans* that did not undergo the ancient whole genome duplication event (Butler et al., 2009; Marcet-Houben et al., 2009; Wolfe & Shields, 1997).

The complex tandem repeat genes, for which genome copy number could not be determined, had low sequence identity and were predominantly found on Chr6 (Figure 2-figure supplement 3C). In contrast, the full-length coding sequence of all ORFs that were contained within long repeat sequences, were significantly longer (Median value of 1380 bp *vs*. 1200 bp, Figure 2-figure supplement 3D, p < 0.0008, Kolmogorov-Smirnov test) and had a significantly higher GC content (Median value of 37.22% *vs.* 35.22% Figure 2-figure supplement 3E, p < 0.0001, Kolmogorov-Smirnov test) than the full-length coding sequence of all ORFs not contained within long repeat sequences (genome-wide, excluding the complex tandem repeat genes, See Methods). Interestingly, increased GC content was correlated with increased rates of both mitotic and meiotic recombination events in *S. cerevisiae* (Kiktev et al., 2018).

Identification of CNV breakpoints in isolates with segmental aneuploidies

Next, CNV breakpoints were determined across 13 additional isolates with one or more segmental aneuploidies. Six of these isolates were from *in vitro* evolution experiments in the presence of azole antifungal drugs (FLC or miconazole), 4 were from *in vivo* evolution experiments in a murine model of oropharyngeal candidiasis (OPC) performed in the absence of antifungal drugs, and 3 were human clinical isolates (Supplementary File 1). All segmental aneuploidies arose from a known euploid diploid progenitor (Abbey et al., 2014; Hirakawa et al., 2015), except two clinical isolates with unknown origin and the i(4R) isolate that arose from a trisomic progenitor, described above.

Segmental aneuploidies were initially detected by CHEF karyotype analysis and ddRAD-seq, but the coordinates of the CNV breakpoints were not known (Abbey et al., 2014; Forche et al., 2018; Mount et al., 2018; Ropars et al., 2018). The ploidy of each isolate was measured by flow cytometry and the DNA copy number of all loci was determined using whole genome sequencing (See Methods). Among the 13 diverse isolates, 19 segmental aneuploidies were confirmed, with at least one segmental aneuploidy detected on each of the 8 chromosomes (Figure 3A, Figure 3-figure supplement 1A-J). Segmental amplifications were more frequent (12/19, 63.2%) than segmental deletions (3/19, 15.8%). The remaining segmental aneuploidies (4/19, 21.1%) consisted of more complex rearrangements that resulted in a segmental amplification and a terminal chromosome deletion at the same breakpoint.

All segmental aneuploidies occur at long repeat sequences

The CNV breakpoint of each segmental aneuploidy was determined using both read depth and allele ratio analysis (See Methods). From the 19 segmental aneuploidies, 26 CNV breakpoints were identified because some segmental aneuploidies contained multiple breakpoints. Strikingly, every CNV breakpoint occurred within 2 kb of a long repeat sequence, ranging from 248 bp to ~4.76 kb in length. Observed breakpoints had significantly more overlap with long repeat sequences than expected given the total genome coverage of long repeat sequences ($p < 0.0001$, two-tailed Fishers Exact Test, See Methods). All but one of the repeat sequences were intra-chromosomal and separated by a distance ranging from ~3.1 kb to ~1.62 Mb (Supplementary File 3). Importantly, repeats containing ORFs were significantly more common than all other types of repeats at these breakpoints (18/26 CNV breakpoints, $p < 0.001$, χ^2 Goodness-of-fit test).

Three examples of CNV breakpoints in long repeats containing ORFs were observed in isolates AMS3053, AMS3420 and CEC2871. In both AMS3053 and AMS3420, a long inverted repeat sequence was associated with a complex segmental amplification and a terminal chromosome deletion that resulted in a long-range homozygosis event. In AMS3053, the breakpoint on Chr3L occurred within a ~1.7 kb inverted repeat sequence (>99% identity) separated by ~11.5 kb (Figure 3B). The left side of this inverted repeat contained four uncharacterized ORFs (*orf19.279, orf19.280, orf19.281, orf19.284*) and associated intergenic sequences, while the right side contained three uncharacterized ORFs (*orf19.296, orf19.295, orf19.292*) and one characterized ORF (*orf19.297 DTD2*) plus associated intergenic sequences. Similarly, the OPC-derived isolate AMS3420 underwent a complex segmental amplification and deletion within a ~1.6 kb inverted repeat sequence on Chr1L (91.5% identity) separated by ~26 kb, which contains the high affinity glucose transporters *HGT1* and *HGT2* (Figure 3-figure supplement 1A). Long internal chromosome deletions were also observed. For example, in isolate CEC2871, a ~55 kb deletion resulted from recombination between a ~1.4 kb tandem repeat on ChrR (92.4% identity) containing ORFs of the *PHO* gene family (*PHO112* and *PHO113*, Figure 3C). Proposed models for recombination events that would result in these complex segmental amplifications and deletions are described in the discussion.

Eight CNV breakpoints occurred within other long repeat sequences, including: a ~200 bp microsatellite repeat (1/26), intergenic repeats (1/26), MRS (2/26), LTRs (2/26), and the rDNA repeats (2/26) (Figure 3, Supplementary File 3). Some segmental aneuploidies were comprised of multiple breakpoints, each associated with a different repeat family (e.g. Figure 3-figure supplement 1I & J). Interestingly, both breakpoints that occurred at the rDNA also amplified the ChrR centromere (*CENR*), and everything either to the telomere of the opposite chromosome arm (ChrRL) (Figure 3-figure supplement 1H), or to a microsatellite repeat sequence on ChrRL (AMS3328, Figure 3A).

In summary, all CNV breakpoints in this collection occurred at or within long repeat sequences. Inverted repeat sequences predominantly coincided with segmental amplifications and terminal chromosome deletions, while tandem repeat sequences coincided with internal chromosome deletions. Some aneuploidies were comprised of multiple breakpoints, each associated with a different repeat family. Overall, a repeat homology-associated repair mechanism appears to be driving the formation of segmental aneuploidies. Importantly, the involvement of long repeats in CNV breakpoints is independent of genetic background and environmental selection.

LOH occurs at long inter- and intra-chromosomal repeat sequences

In many of the isolates with segmental aneuploidies, the CNV also was accompanied by LOH (e.g., Figure 3B & C). To ask if long repeat sequences were associated with LOH breakpoints in the absence of detectable CNVs, we selected 20 near-euploid genomes that had at least one long-range homozygous region, but the coordinates of the LOH breakpoint were not known (Ford et al., 2015; Hirakawa et al., 2015; Ropars et al., 2018). These 20 isolates belong to 9 major *C. albicans* clades from different origins (e.g., superficial and invasive human infections, healthy human hosts, and spoiled food) (Figure 4A, Supplementary File 1).

153 LOH breakpoints were identified in the 20 isolates (See Methods, Supplementary File 4). 61/153 LOH breakpoints were found within 2 kb of a long repeat sequence, and, like the CNV breakpoints, these LOH breakpoints could occur on any chromosome (Figure 4A). The copy length of the repeat sequences found at LOH breakpoints ranged from 78 bp to 6499 bp (median 516 bp) with sequence identities ranging from 82.2% to 100% (median of 95.1%). Most of the repeats associated with LOH breakpoints were intra-chromosomal (46/61), in all three orientations (inverted, mirrored, and tandem), and separated by a distance ranging from 903 bp to ~1.6 Mb (median ~35.3 kb). The vast majority of long-range homozygous regions contained only one LOH breakpoint and proceeded from the breakpoint to the proximal telomere, similar to previous analyses (Ene et al., 2018; Forche et al., 2008; Forche et al., 2009; Selmecki et al., 2005). Surprisingly, four isolates had an LOH breakpoint that proceeded from one chromosome arm to the telomere on the opposite chromosome arm, causing centromere homozygosis (three events on ChrR and one event on Chr5).

One isolate, CEC723, had two long-range homozygous regions associated with intra-chromosomal repeat sequences. The first LOH breakpoint on Chr1R was associated with a ~1.1 kb mirrored repeat sequence (>99% identity) separated by ~15 kb (Figure 4B). One copy of the repeat sequence contained a snoRNA (*snR42a*) and the other contained an uncharacterized ORF (*orf19.2800*), which we predict also encodes a second copy of *snR42a*. The second LOH breakpoint on ChrRL was associated with a ~3.2 kb tandem repeat sequence (97.7% identity) separated by ~70 kb (Figure 4C). This breakpoint was flanked by additional long repeat sequences that were associated with CNV in other isolates, indicating that this region is a hotspot for genome rearrangements (Supplementary File 2).

Finally, the reference isolate SC5314 contains a well-known long-range homozygous region on Chr3R. We asked if this LOH breakpoint occurred within a long repeat sequence. Remarkably, the LOH breakpoint occurred in *orf19.5880* near an 8 bp sequence (AACTTCTT) identical to part of the C. *albicans* 23 bp telomere repeat sequence (GGTGTACGGATTGTCT<u>AACTTCTT</u>). Furthermore, a second copy of this same 8 bp sequence was found in an inverted orientation ~3.4 kb away in the adjacent ORF (*orf19.5884*). This long-range LOH event continued to the right telomere of Chr3. While LOH may have resulted from a repair template on the other homolog, an alternative model cannot be ruled out. We previously found that an LOH and CNV breakpoint that caused a segmental Chr5 truncation in the common laboratory strain BWP17 (Selmecki et al., 2005) was initiated at a 9 bp sequence (CTAACTTCT) that is almost identical to the sequence found at this breakpoint (AACTTCTT). We posit that a similar chromosome truncation, followed by reduplication of the monosomic portion of Chr3 (Figure 4-figure supplement 1A & B) may have generated the homozygosis of Chr3. These 8 bp and 9 bp telomere-like sequences occur 2160 and 249 times, respectively, within the non-telomeric portions of the *C. albicans* reference genome (Supplementary

File 5). The presence of such a large number of potential template sequences, especially if including the telomere repeats at each chromosome end, might have driven this two-step model.

Repeat sequences cause sequence inversions and heterozygous islands

As expected, levels of heterozygosity were high within long repeat sequences due to the ability of short-read (Illumina) sequences to map to multiple positions in the genome (e.g. the heterozygous bases within repeat sequences in Figure 4B & C). Unexpectedly, between or adjacent to some long repeat sequences, heterozygous islands were observed in otherwise homozygous regions of the genome. For example, in isolate P75063, an LOH breakpoint on Chr4L was associated with a ~1.7 kb inverted repeat and resulted in a terminal homozygosis of the chromosome (Figure 5A). Adjacent to this homozygous region was an ~32 kb region that had multiple homozygous/heterozygous transitions (5' homozygous-heterozygous-homozygous-heterozygous 3'). We hypothesized that a long sequence inversion, similar to that observed within the repeats flanking *CEN4*, accounted for the multiple heterozygous to homozygous transitions in this region. PCR amplification between unique sequences flanking the inverted repeat revealed a ~32 kb inversion in P75063 and SC5314 and was the only orientation that amplified by PCR; the reference orientation did not amplify, suggesting that the reference genome may be incorrect at this position (Figure 5B).

These two long inversions (at *CEN4* and Chr4L), plus an additional seven potential sequence inversions were identified bioinformatically from a set of 21 clinical isolates (Hirakawa et al., 2015), however none of these inversion breakpoints were characterized or validated by PCR or Sanger sequencing. We found that all potential inversions had breakpoints within long inverted repeats, and these potentially cause chromosomal inversions of ~4.1 kb to ~102.6 kb in length (median ~39.0 kb, Supplementary File 6). All but one sequence inversion (8/9) occurred within repeats containing ORFs and a high median sequence identity (98.3%). In summary, we identified examples of chromosomal inversions that occurred between long repeat sequences and provide the first molecular validation of these inversions in both the reference SC5314 and clinical isolates.

Breakpoints resulting in CNV, LOH, and inversion, occur in the longest repeat sequences with highest homology

Overall, many uncharacterized long repeat sequences exist within the *C. albicans* genome. Repeats associated with breakpoints (CNV, LOH, and inversion) were significantly longer than all other long repeat sequences (median copy length of 785 bp *vs*. 278 bp, $p < 0.0001$, Kolmogorov-Smirnov test), and had a significantly higher percent sequence identity than all other long repeat sequences (median identity of 96.2% *vs*. 94.2%, $p < 0.036$ Kolmogorov-Smirnov test) (Figure 6A). Repeats containing ORFs were longer than repeats containing other genomic features and were the most common repeat identified at breakpoints (33/53, 62.3%, Figure 6B & C). Furthermore, repeats containing ORFs were the only genomic feature with both significantly longer copy length and significantly higher sequence identity at breakpoints than at non-breakpoints ($p < 0.0001$ copy length, $p < 0.0001$ sequence identity Kolmogorov-Smirnov test, Figure 6-figure supplement 1A & B). Additionally, repeat matches that contain multiple ORF sequences represent only 8.6% of all long repeats containing ORFs, yet these extra-long repeats comprise 26.8% of the observed breakpoints (Supplementary File 2). Therefore, at least under selection, genome rearrangements are occurring more often at repeats with high sequence identity, and at repeats with high sequence identity and high copy length, the latter of which includes ORFs.

Nine repeat families were associated with more than one breakpoint type (CNV, LOH, and inversion), and two of these (124 and 151) were associated with all three breakpoint types. Repeat family 124 (Figures 3B & 6A), comprised of 4 ORFs, was one of the longest repeats (~3.2 kb) and had one of the highest percent sequence identities (> 99%). Repeat family 151 flanks *CEN4* and was associated with the formation of the novel isochromosome i(4R), which was necessary and sufficient for increased fitness in the presence of FLC (Figure 1C & Figure 6A). Overwhelmingly, these data support that long repeat sequences found throughout the *C. albicans* genome are utilized to generate segmental aneuploidies, long-range LOH and sequence inversions, and that in at least one environment these rearrangements provide a significant fitness benefit to the organism.

DISCUSSION

Genomic variation caused by CNV, LOH, and sequence inversion can drive rapid adaptation and promote tumorigenesis. Here, we examined the role of genome architecture during the formation of genetic variation in the diploid, heterozygous fungal pathogen, *C. albicans*. Our genome-wide analysis of 33 isolates identified long repeat sequences that had prominent roles in generating genomic diversity. These long repeats included previously uncharacterized repeat sequences, centromeric repeats, repeats found within intergenic sequences, and repeats that span multiple ORFs and intergenic sequences. Importantly, long repeat sequences were found at every CNV and sequence inversion breakpoint observed, and frequently occurred at LOH breakpoints as well. Long repeats that were associated with all breakpoints (CNV, LOH, and inversion) have on average significantly higher sequence identity compared to all repeats identified ($p < 0.036$, Kolmogorov-Smirnov test). Furthermore, repeats containing ORFs had both significantly higher sequence identity and significantly longer copy length at breakpoints than at non-breakpoints (sequence identity $p < 0.0001$, copy length $p < 0.0001$ Kolmogorov-Smirnov test, Figure 6, Figure 6-figure supplement 1A & B). These results were independent of genetic background or source of isolation. Thus, long repeat sequences found across the *C. albicans* genome underlie the formation of significant genome variation that can increase fitness and drive adaptation.

DNA double-strand breaks are repaired using long repeat sequences found across the *C. albicans* genome

The genomic variants described in this study are the result of DNA double-strand breaks (DSBs) and subsequent recombination events resulting in CNVs, LOH, and sequence inversions. While the factors leading to, and the location of the initiating DSBs are unknown, the genomic variants recovered were all selected as viable, and perhaps beneficial, outcomes of the DSB repair process. DSBs are repaired by either non-homologous end-joining (NHEJ) or homologous recombination (HR). HR is thought to be a high-fidelity repair process due to the use of an intact, homologous DNA template. However, recent studies have also implicated HR in an increased rate of mutagenesis and chromosomal rearrangements (Bishop & Schiestl, 2000; Kramara et al., 2018).

We also found that the orientation of repeat copies had a major effect on the outcome of the genome rearrangements observed. Inverted repeat sequences frequently were found within 2 kb of chromosomal amplification events, while tandem repeat sequences frequently were found within 2 kb of long internal chromosomal deletions. We propose two models of HR involved in the production of genome variation observed in this study (Figure 7).

First, we propose that single-strand annealing (SSA) is initiated by the annealing of DNA repeats that become single stranded after a DSB and 5’-3’ DNA resection (Figure 7A-7B) and occurs between both tandem and inverted repeat sequences (Bhargava et al., 2016; Malkova & Haber, 2012; Mehta & Haber, 2014; Ramakrishnan et al., 2018; VanHulle et al., 2007). SSA that occurs between tandem repeats leads to segmental deletion of the sequence located between the repeat sequences (Figure 7C). SSA that occurs between inverted repeats can lead to the formation of complex, often unstable dicentric and ‘fold-back’ chromosomes which then enter the breakage-fusion-bridge cycle leading to further genome instability (Aguilera & Garcia-Muse, 2013; Croll et al., 2013; McClintock, 1939, 1941, 1942; VanHulle et al., 2007) (Figure 7A-7B). Evidence for dicentric chromosomes may exist in several isolates that acquired a segmental amplification of the centromere (Figure 3), however we do not know from these data if the amplification is on the same molecule (generating a dicentric chromosome) or elsewhere in the genome.

The second HR mechanism we propose is break-induced replication (BIR) which is initiated by DSBs that have only one free end available for repair. During BIR, single-strand DNA invades a homologous sequence followed by subsequent DNA synthesis which can copy long, chromosomal-sized DNA segments (Anand et al., 2013; Kramara et al., 2018; Malkova & Ira, 2013; Mehta & Haber, 2014). If templating and synthesis occurs on a homologous chromosome, BIR can lead to long-range homozygosis of a chromosome (Figure 7D). Processes similar to BIR have been proposed for CNV generation in a diverse set of organisms ranging from bacteria to humans (Hastings et al., 2009). These predominantly micro-homology mediated BIR (MMBIR) events use short regions of homology to repair DSBs in a Rad51-independent manner (Hastings et al., 2009). One caveat is that the repeat sequences involved in generating genome rearrangements observed in this study are much longer than those involving MMBIR. While repair by BIR is rare in *S. cerevisiae* model systems, the selective benefit of the resulting genotypes generated by BIR could increase the apparent frequency with which these types of mutations are recovered in certain environments, for instance the acquisition of i(4R) in the presence of FLC (Figure 1).

C. albicans repeat copy length and spacer length

The repeat copy length associated with observed breakpoints in *C. albicans* are similar in copy length to transposable (Ty) elements in *S. cerevisiae* (~6 kb) and long interspersed nuclear elements (LINE) in the human genome (~6-7 kb), which are a major source of genome rearrangements (Chen et al., 2014; Dunham et al., 2002; Gresham et al., 2010; Higashimoto et al., 2013; Selmecki et al., 2015). Both Ty and LINE elements are high copy number repeats; LINE elements are present in thousands of copies in the human genome (Rodić & Burns, 2013). However, beyond the similarly in copy length, we rarely found high copy number repeats, like lone LTRs or retrotransposons, associated with CNV and inversion breakpoints (5.7%, Figure 6). These breakpoints predominantly occurred at repeats containing ORFs that are often present in only two copies per genome (Supplementary File 2). LOH breakpoints, on the other hand, were associated more often with LTRs (22.6%, Figure 6), which may be a result of selection or may suggest a preference for a different repair mechanism when a DSB occurs near these loci.

The repeat copy length and spacer length associated with the observed breakpoints in *C. albicans* are much longer than typically observed in *S. cerevisiae*. Segmental amplification events in *S. cerevisiae* are often mediated by short inverted repeat sequences, for example, 8 bp long and separated by 40 bp (Brewer et al., 2011; Lauer et al., 2018; Payen et al., 2014; Sunshine et al., 2015). The presence of a short, inverted repeat sequence within a replication fork can stimulate ligation between the leading and lagging strands, which results in replication and formation of an extrachromosomal circle. This extra-chromosomal amplification may continue to replicate independently if it contains an origin of replication (defined as origin-dependent inverted-repeat amplification (ODIRA)) (Brewer et al., 2015; Brewer et al., 2011; Payen et al., 2014). It seems unlikely that such a mechanism operates at the long distances observed between repeat sequences in *C. albicans*. However, it is possible that a different origin-dependent mechanism is mediating some of the rearrangements we observed (see centromere discussion below). A future challenge is to determine if/how this occurs.

The spacer length, especially between inverted repeats, has been a major focus of genome instability research. Identification and characterization of inverted repeats in *S. cerevisiae* has primarily focused on those repeats that are separated by very short (~80 bp) spacers (Strawbridge et al., 2010). Inverted repeats that were engineered to have variable repeat spacer lengths identified a correlation between repeat and spacer length and DSB repair. Increasing repeat copy length (from 185 bp to ~1.5

kb) and/or decreasing repeat spacer length (from ~8.5 kb to 0 bp) increases the recombination rate between repeats by up to 17,000-fold (Lobachev et al., 1998). Furthermore, spacer length alone can affect the choice of DSB repair pathway; DSB repair via inter-molecular SSA predominantly occurs with a spacer length of 1 kb, while intra-molecular SSA predominantly occurs with spacer length of 12 bp (Ramakrishnan et al., 2018).

Astoundingly, the *C. albicans* CNV and inversion breakpoints are associated with much longer repeat spacer lengths than those described in *S. cerevisiae*, ranging from ~3.1 kb to ~1.6 Mb (median ~30 kb) and ~3.1 kb to ~94.3 kb (median ~34.6 kb), respectively. Recombination between such long distances requires a naturally occurring, long-distance homology search. It is tempting to speculate that *C. albicans* may have a mechanism for long distance resection, particular chromatin features, or a 3D-nuclear structure that facilitates recombination between inverted repeats separated by long distances.

Inverted repeat sequences directly associated with the CENP-A-binding centromere core sequences facilitate isochromosome formation

Centromeres were common breakpoints for CNV, LOH and inversion. Twelve of the 33 isolates had breakpoint events that occurred within centromeres, including those described at *CEN4* and *CEN5*, as well as two additional centromeres that contain one copy of a long repeat sequence, *CEN2* and *CEN3* (Supplementary File 2). Notably, *C. albicans* centromeres are the earliest firing centers of DNA replication (Koren et al., 2010; Tsai et al., 2014). Therefore, errors in DNA replication may be a common source of DSBs that are repaired via HR between long repeat sequences.

Repair of a DSB within or near a centromere-associated inverted repeat can result in isochromosome formation or centromere inversion (Figure 1, Figure 1-figure supplement 1). Both of the *C. albicans* centromeres that are flanked by long inverted repeat sequences (*CEN4* and *CEN5*) can form isochromosomes (Figure 1 and (Selmecki et al., 2006; Selmecki et al., 2009)). Exposure to the antifungal drug FLC selected for isochromosome formation at both *CEN4* and *CEN5*. If a DSB occurs near the inverted repeat sequence, DNA synthesis via BIR will copy the entire arm of the broken chromosome, resulting in the homozygous isochromosome structures that we observed (Figure 1 and (Selmecki et al., 2010; Selmecki et al., 2009)). Acquisition of either isochromosome i(4R) or i(5L) was both necessary and sufficient for increased fitness in the presence of FLC (Figure 1 and (Selmecki et al., 2006)). Additionally, there was no fitness cost associated with either isochromosome in the absence of FLC: i(4R) was stable for ~300 generations in 12/12 populations in the absence of FLC (Figure 1-figure supplement 1). These data are in contrast to other, often whole chromosome and multiple chromosome aneuploidies that cause significant fitness defects in the absence of selection (Pavelka et al., 2010; Torres et al., 2007), but support observations that aneuploidy in general has less of a fitness cost in diploid and polyploid fungi (Hose et al., 2015; Scott et al., 2017; Selmecki et al., 2015; Tan et al., 2013).

Similarly, repair of a DSB within or near a centromere-associated inverted repeat can result in centromere inversion. Inversions are the result of intra-chromosomal non-allelic homologous recombination (NAHR) between inverted repeats flanking the centromere (Figure 7E). Here we detected an inversion that occurred between inverted repeats flanking *CEN4*. The impact of these inversions on localization of the centromeric histone CENP-A, or of the recombination proteins Rad51 and Rad52, which are thought to recruit CENP-A, are not known. Whether or not inversion of the centromere affects chromosome stability will be important to test in future experiments.

In this study, Illumina short-read datasets were used to identify genomic features that were driving structural and allelic variation across diverse *C. albicans* isolates. The use of both new and previously published short-read datasets highlights the utility of this bioinformatic approach for the analysis of structural variants within this and other species. However, short-read data are unable to provide a key understanding of the molecules containing the long repeat sequences. For example, the definitive structure of chromosomal inversions, including the heterozygous *CEN4* sequence, are difficult to determine with short-read data. PCR enabled rapid validation of these inversions (Figure 1 and 5), however it required knowledge of the repeat location and unique surrounding sequences. Future long-read sequencing is needed to address the definitive structure of existing DNA molecules and potential DNA intermediates involved in recombination and resolution of CNV, LOH, and inversions.

Long repeats containing ORFs were significantly more common at breakpoints resulting in CNV, LOH and inversion than any other genomic feature

One hypothesis is that active transcription may promote DNA DSBs, due to the formation of R-loop structures (Aguilera & Gaillard, 2014; Santos-Pereira & Aguilera, 2015). Additionally, increased transcription in certain environments may increase the probability of a DNA DSB that result in genome rearrangements, as was observed at the *S. cerevisiae CUP1* locus in high copper environments (Adamo et al., 2012; Fogel et al., 1983; Hull et al., 2017; Thomas & Rothstein, 1989). Several indirect results are consistent with this hypothesis in *C. albicans*. First, all ORFs within a long repeat that were associated with a breakpoint were indeed actively transcribed in the reference isolate SC5314 during growth in rich medium (Bruno et al., 2010). Secondly, some breakpoint ORFs have increased expression in the selective environment from which the isolate with the breakpoint was obtained. For example, two different *in vivo* isolates, one bloodstream clinical isolate and one murine OPC-evolved isolate, have the same breakpoint on Chr1 at the inverted repeat that includes *HGT1* and *HGT2* (Supplementary File 2). Both *HGT1* and *HGT2* are induced during OPC, biofilm production and adaptation to serum (Horak, 2013; Nobile et al., 2012; Pitarch et al., 2001). Therefore, increased transcription of these repeat ORFs *in vivo* is a potential source of DNA damage that resulted in DSB repair.

Conclusion

In conclusion, genome rearrangements resulting in segmental aneuploidies, sequence inversions, and LOH are associated with long repeat sequence breakpoints on every chromosome. These genome rearrangements can arise rapidly, both *in vitro* and *in vivo*, and can provide an adaptive phenotype such as improved growth in antifungal drugs. Importantly, long repeat sequences are hotspots for genome variation across diverse selective environments. Indeed, several repeats were involved in all three types of genome rearrangements in different isolates. These data support the idea that the *C. albicans* genome is one of the most rapidly evolving genomes due to disruption of conserved syntenic sequence blocks via genome rearrangements between long repeat sequences (Fischer et al., 2006). Finally, given the frequency of long repeat sequences in the human genome, studies of *C. albicans* genome rearrangements can contribute to understanding the mechanisms that facilitate CNV, LOH, and inversions associated with human disease and cancer.

MATERIALS AND METHODS

Key Resource Table

Reagent type (species) resource	Designation	Source or Reference	Identifiers	Additional Information
strain, strain background (*Candida albicans*)	SC5314	Hirakawa et al., 2015 (doi:10.1101/gr.174623.114)	RRID:SCR_013437	
strain, strain background (*C. albicans*)	P78042	Hirakawa et al., 2015 (doi:10.1101/gr.174623.114)		
strain, strain background (*C. albicans*)	AMS3743	This Study		*In vitro* evolution of P78042 in 128 ug/ml FLC for 100 generations
strain, strain background (*C. albicans*)	AMS3743_10	This Study		*In vitro* evolution of AMS3743 in rich medium for 300 generations
strain, strain background (*C. albicans*)	AMS3743_10_S6	This Study		Single colony from AMS3743_10
antibody	Anti-Digoxigenin-AP Fab Fragments	Roche	11093274910 RRID:AB_2734716	(1:5000)
sequenced-based reagent	PCR Primers	This Study		Supplementary File 7
commercial assay or kit	Illumina Nextera XT Library Prep Kit	Illumina	105032350	
commercial assay or kit	Illumina Nextera XT Index Kit	Illumina	105055294	
commercial assay or kit	Illumina MiSeq v2 Reagent Kit	Illumina	15033625	2x250 cycles
commercial assay or kit	Blue Pippin 1.5% agarose gel dye-free cassette	Sage Science	250 bp - 1.5 kb DNA size range collections, Marker R2	Target of 900 bp
commercial assay or kit	Qubit dsDNA HS kit	Life Technologies	Q32854	
commercial assay or kit	PCR DIG Probe Synthesis Kit	Roche	11636090910	
commercial assay or kit	Agilent 2100 Bioanalyzer High	Agilent Technologies	5067-4626	

	Sensitivity DNA Reagents			
chemical compound, drug	Fluconazole (FLC)	Alfa Aesar	J62015	
software, algorithm	MUMmer Sutie	Kurtz et al., 2004 (doi:10.1186/gb-2004-5-2-r12)	v3.0 RRID:SCR_001200	
software, algorithm	Trimmomatic	Bolger et al., 2014 (doi:10.1093/bioinformatics/btu170)	v0.33 RRID:SCR_011848	
software, algorithm	BWA	Li et al., 2013 (doi:10.1093/bioinformatics/btp324)	v0.7.12 RRID:SCR_010910	
software, algorithm	Samtools	Li et al., 2009 (doi:10.1093/bioinformatics/btp324)	v0.1.19 RRID:SCR_002105	
software, algorithm	Genome Analysis Toolkit	McKenna et al., 2010 (doi:10.1101/gr.107524.110)	v3.4-46 RRID:SCR_001876	
software, algorithm	REPuter	Kurtz et al., 2001 (doi:10.1093/nar/29.22.4633)	V1.0 https://bibiserv.cebitec.uni-bielefeld.de/reputer	
software, algorithm	Yeast Analysis Mapping Pipeline	Abbey et al., 2014 (doi:10.1186/s13073-014-0100-8)	v1.0	
software, algorithm	Graphpad Prism	https://www.graphpad.com	v6.0 RRID:SCR_002798	
software, algorithm	ImageJ	https://imagej.nih.gov/ij/?	v2.0.0-rc-30/1.49s RRID:SCR_003070	
software, algorithm	Integrative Genomics Viewer	Thorvaldsdottir et al., 2013 (doi:10.1093/bib/bbs017)	v2.3.92 RRID:SCR_011793	
software, algorithm	R	https://www.r-project.org	v3.5.2 RRID:SCR_001905	
software, algorithm	Candida Genome Database	http://Candidagenome.org	RRID:SCR_002036	
other	Propidium Iodide	Invitrogen	P3566	25 ug/ml final concentration
other	Ribonuclease A	MP Biomedicals	101076	0.5 mg/ml final concentration

Yeast Isolates and Culture Conditions: All isolates used in this study are shown in Supplementary File 1. Isolates were stored at -80°C in 20% glycerol. Strains were grown at 30°C in YPAD (yeast peptone dextrose medium (Rose, 1990) supplemented with 40 µg ml^{-1} adenine and 80 µg ml^{-1} uridine).

***In vivo* evolution experiments:** OPC isolates were obtained as previously described (Forche et al., 2018; Solis & Filler, 2012). Briefly, mice were orally infected with strain YJB9318 and single colony isolates were obtained from tongue tissue of mice on day 1, 2, 3, and 5 post infection and stored in 50% glycerol at -80°C for further use.

***In vitro* evolution experiments:** Six isolates were obtained from *in vitro* evolution experiments in the presence of antifungal drug (Supplementary File 1). Isolate AMS3053 was obtained on 10 µg/ml Miconazole agar plates as previously described (Mount et al., 2018). Isolates AMS3742, AMS3743, AMS3747, AMS3748, and AMS3744 were obtained from liquid batch culture evolution experiments conducted in 96-well format. Progenitor isolates were plated for single colonies on YPAD and incubated for 48 hours at 30°C. Single colonies were grown to saturation in liquid YPAD at 30°C. A 1:1000 dilution was made in YPAD medium containing either 1 µg/ml or 128 µg/ml of FLC. Plates were covered with BreathEASIER tape (Electron Microscope Science) and cultured in a humidified chamber for 72 hours at 30°C. At each 72-hour time point, cells were resuspended by pipetting and transferred into fresh media via a 1:1000 dilution and cultured for another 72 hours at 30°C, for 10 consecutive passages. After the final transfer, cells were immediately collected for genomic DNA isolation and ploidy analysis by flow cytometry.

To obtain AMS3743 isolates that had lost the i(4R) (Figure 1-figure supplement 1), 12 single colonies of AMS3743 were selected on YPAD plates at 30°C after 48 hours. All 12 single colonies had i(4R) (by PCR) and were used to initiate 12 YPAD-evolved lineages, each cultured for 24 hours in 4 ml liquid YPAD at 30°C with shaking. Every 24 hours, a 1:1000 dilution was inoculated into fresh YPAD medium. Cultures were passaged for 30 days. Cells from all 12 YPAD-evolved lineages were divided into tubes for -80°C storage, genomic DNA isolation, and CHEF analysis. All 12 YPAD-evolved lineages maintained i(4R) by CHEF analysis. CHEF gel densitometry analysis (see below) identified one lineage (AMS3743_10) that had a lighter i(4R) band density relative to the rest of the genome. AMS3743_10 was plated for single colonies on a YPAD plate and incubated at 30°C for 48 hours. Six single colonies were cultured for 24 hours in 4 ml liquid YPAD at 30°C with shaking, and cells were divided into tubes for -80°C storage, genomic DNA isolation, and CHEF analysis. One of the six single colonies lost the i(4R) (AMS3743_10_S6, Figure 1-figure supplement 1).

Contour-clamped homogenous electric field (CHEF) electrophoresis: Samples were prepared as previously described (Selmecki et al., 2005). Cells were suspended in 300 µL 1.5% low-melt agarose (Bio-Rad) and digested with 1.2 mg Zymolyase (US Biological). Chromosomes were separated on a 1% Megabase agarose gel (Bio-Rad) in 0.5X TBE using a CHEF DRIII apparatus. Run conditions as follows: 60 s to 120 s switch, 6 V/cm, 120° angle for 36 hours followed by 120s to 300s switch, 4.5 V/cm, 120° angle for 12 hours.

CHEF gel densitometry: Ethidium bromide stained CHEF gels were imaged using the GelDock XR imaging system (BioRad). Images were exported as .PNG files, converted to 32-bit, and analyzed using ImageJ (v2.0.0-rc-30/1.49s). The total lane density (gray value, area under the curve) was collected for each sample. The density associated with i(4R) was determined by drawing a box around the i(4R) density peak (box distance was from each adjacent minimums). The fraction of i(4R) relative to the entire genome was determined by normalizing the i(4R) density relative to the total lane density.

The population with lowest ratio of i(4R) relative to total genome (AMS3743_10) was used for single colony analysis.

Southern Hybridization: DNA from CHEF gels was transferred to BrightStar Plus nylon membrane (Invitrogen). Probing and detection of the DNA was conducted as previously described (Selmecki et al., 2005; Selmecki et al., 2008; Selmecki et al., 2009). Probes were generated by PCR incorporation of DIG-11-dUTP into target sequences following manufacturer's instructions (Roche). Primer pairs used in probe design are listed in Supplementary File 7.

PCR: All primer sequences were designed to avoid heterozygous or SNP loci in the reference genome SC5314 and clinical isolates. Primers and primer sequences are found in Supplementary File 7. PCR conditions for i(4R) were as follows: 95°C for 3 min, followed by 32 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 5.5 min, and a final extension at 72°C for 10 min. The PCR conditions for the Chr4 inversion (Figure 5) were the same as above, except the annealing temperature was at 53°C and the extension time was for 3.25 min.

Flow Cytometry: Cells were prepared as previously described (Todd et al., 2018). Briefly, cells were grown to a density of 1x10^7 in liquid medium and gently spun down (500 x g) for 3 minutes. The supernatant was removed and cells were fixed with 70% (v/v) ethanol for at least 1 hour at room temperature. Cells were then washed twice with 50 mM sodium citrate and sonicated (Biorupter Fisher Science) for 10-15 s at 30% power to separate the cells. Following sonication, cells were centrifuged and resuspended with 50 mM sodium citrate and incubated for at least 3 hours at 37°C in 0.5 mg ml^{-1} RNase A (MP Biomedicals) + 50 mM sodium citrate (Fisher Scientific). Cells were stained with 25 µg ml^{-1} propidium iodide (Invitrogen) overnight in the dark at 37°C. Cells were sonicated for 5-10 seconds at 15% power, and 30,000 cells were analyzed on a ZE5 cell analyzer (BioRad). Data were analyzed in FlowJo (https://www.flowjo.com/solutions/flowjo/downloads) (v10.4.1).

Growth Curve Analysis: Growth curves were determined using a BioTek Epoch plate reader. Culture medium included YPAD or YPAD+32 µg/ml FLC (Alfa Aesar) Approximately 5x10^3 cells were inoculated into 200 µl culture medium in a clear, flat bottomed 96-well plate (Thermo Scientific). The plate was incubated at 30°C with a double orbital shaking at 256 rpm, and the OD_{600} was measured every 15 minutes. Data were collected with Gen5 Software (BioTek) and exported to Microsoft Excel for downstream analysis. All growth curves were conducted in individual biological triplicate on separate days.

Illumina Whole Genome Sequencing: Genomic DNA was isolated with phenol chloroform as described previously (Selmecki et al., 2006). Libraries were prepared using the NexteraXT DNA Sample Preparation Kit following the manufacturer's instructions (Illumina). DNA fragments between 600 and 1,200 bp were selected for sequencing using a Blue Pippin 1.5% agarose gel dye-free cassette (Sage Science). Library fragments were analyzed with a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies) and Qubit High Sensitivity dsDNA (Life Technologies). Libraries were sequenced using paired-end, 2 x 250 reads on an Illumina MiSeq (Creighton University). Adaptor sequences and low-quality reads were trimmed using Trimmomatic (v0.33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 TOPHRED33) (Bolger et al., 2014). Reads were mapped to the *Candida albicans* reference genome (A21-s02-m09-r08) obtained 7 of October 2015 from the *Candida* Genome Database website:
http://www.candidagenome.org/download/sequence/C_albicans_SC5314/Assembly21/archive/
C_albicans_SC5314_version_A21-s02-m09-r08_chromosomes.fasta.gz). The reads were mapped

using the Burrows-Wheeler Aligner MEM algorithm using default parameters (BWA v0.7.12) (Li, 2013). Duplicate PCR amplicons were removed using Samtools (v0.1.19) (Li et al., 2009), and reads were realigned around possible indels using Genome Analysis Toolkit's RealignerTargetCreator and IndelRealigner (-model USE_READS -targetIntervals) (v3.4-46) (McKenna et al., 2010). All WGS data have been deposited in the National Center for Biotechnology Information Sequence Read Archive database as PRJNA510147. Sequence data obtained from published datasets are noted in Supplementary File 1.

Identification of Aneuploidy and Copy Number Breakpoints: Preliminary identification of chromosomes containing CNVs was conducted using Illumina whole genome sequence data and the Yeast Analysis Mapping Pipeline (YMAP v1.0). Fastq files were uploaded to YMAP and read depth was plotted as a function of chromosome location using the reference genome *Candida albicans* (A21-s02-mo8-r09), with correction for chromosome end bias and GC content (Abbey et al., 2014). The average normalized genome coverage was determined for 45.5 kb non-overlapping windows across each chromosome using the YMAP GBrowse CNV track. The largest absolute difference between the average normalized genome coverage of two consecutive 45.5 kb windows was identified. To further refine CNV breakpoints, fastq files were aligned to the reference genome as above (Illumina Whole Genome Sequencing), read depth was calculated for every base pair in the nuclear genome using Samtools (samtools depth -aa) (v0.1.19), and normalized by read depth of the total nuclear genome using R (v3.5.2). The two consecutive 45.5 kb windows were further sub-divided into 5 kb windows. The average normalized read depth was determined for these 5 kb windows and a rolling mean of every two consecutive 5 kb windows was determined. CNV breakpoint boundaries were identified when 75% of four consecutive means had an average normalized read depth that deviated from the average normalized nuclear genome read depth by more than 25% in tetraploids or 50% in diploids

(Ford et al., 2015). Boundaries were confirmed by visual inspection in Integrative Genomics Viewer (IGV v2.3.92) (Thorvaldsdottir et al., 2013). CNV breakpoints were then determined using visual inspection of total read depth and allele ratio analysis (when the breakpoint was surrounded by heterozygous sequence) within unique, non-repeat sequences. CNV breakpoint positions were compared to Supplementary File 2 and breakpoints were assigned a repeat name if they fell within 2 kb of a long repeat sequence.

Enrichment of CNV Breakpoints at Long Repeat Sequences: Enrichment analysis of CNV breakpoints was conducted using a two-tailed Fisher's Exact Test in Bedtools (Bedtools v2.28.0) with default parameters (Quinlan & Hall, 2010). Briefly, two .bed files were generated with 1) the start and stop positions of all long repeat sequences and, 2) the start and stop positions of all long repeat sequences located within 2 kb of a CNV breakpoint (Supplementary File 2, excluding the complex tandem repeat genes). The overlap of observed breakpoints and long repeat sequences was compared to the expected overlap between CNV breakpoints and long repeat sequences, given the total genome coverage of long repeat sequences. The minimum overlap required was a single base pair between a CNV breakpoint and repeat sequence.

Identification of Long-Range Homozygosity Breakpoints: Illumina whole genome sequence data were analyzed using YMAP (v1.0) and IGV (v2.3.92). First, fastq files were uploaded to YMAP and the density of heterozygous SNPs was determined for non-overlapping 5 kb windows and plotted by chromosomal position in standard SNP/LOH view (default parameters, baseline ploidy was 2N for all isolates except AMS3420, which was 4N). Approximate positions of all long-range homozygous and heterozygous transitions were determined within 20-25 kb. To further refine LOH breakpoints, fastq files were aligned to the reference genome as above (Illumina Whole Genome Sequencing) and visualized in IGV. All heterozygous to homozygous (and vice versa) transitions were recorded when four or more consecutive loci were heterozygous and transitioned to four or more homozygous loci (and vice versa). The minimum distance covered by the four or more consecutive loci was greater than 300 bp and all four of the loci were located within unique, non-repeat sequences. Additionally, all heterozygous loci utilized for breakpoint analysis had an alternate allele frequency greater than or equal to 20%, read depth greater than 10 reads, and both forward and reverse strands that supported the alternate allele (Selmecki et al., 2015). The breakpoints of these long-range homozygous tracks ('LOH breakpoints') were recorded as the last heterozygous locus and the first homozygous locus of the heterozygous>homozygous transition, and vice versa for the homozygous>heterozygous transition. Long-range LOH breakpoints were then compared to Supplementary File 2 and were assigned a repeat number if they fell within 2 kb of a long repeat sequence (Supplementary File 4).

Identification of Inversion Breakpoints: Additional positions of predicted chromosomal inversions were obtained from Hirakawa et al. 2015, Table S13 (Hirakawa et al., 2015). Coordinates corresponding to potential inversions were obtained using BreakDancer or NUCmer (Hirakawa et al., 2015). The distance between the BreakDancer or NUCmer coordinates (start and stop) and the nearest long repeat sequence was determined. If a long repeat sequence occurred within 2 kb of either BreakDancer or NUCmer coordinates, the repeat number and family were recorded. Disagreement between BreakDancer and NUCmer coordinates that coincided with breakpoints in different repeat families (representing more complex chromosome rearrangements or inversions) were removed from the analysis. Additionally, all NUCmer or Breakdancer positions that occurred within *ALS* gene family repeats were removed from the analysis because the BreakDancer and NUCmer coordinates did not support a consistent length of sequence inversion (likely due to mapping errors within and between

ALS repeats). The long repeat sequences identified at these potential inversion breakpoints, including those shared across different isolates, are summarized in Supplementary File 6.

Microsatellite Repeat Identification: Short repetitive sequences found at either copy number breakpoints or allele ratio breakpoints were analyzed using REPuter (Kurtz et al., 2001) with a minimum repeat length of 8 bp. Analysis was conducted using the forward, reverse, complement, and palindromic match direction.

Identification of Long Repeat Sequences: Repeat sequences within the *C. albicans* genome were identified using the MUMmer suite (v3.0) (Kurtz et al., 2004). Whole genome sequence alignment with NUCmer (nucmer --maxmatch --nosimplify) identified all maximum-length matches with 100% sequence identity (minimum match length of 20 bp) within the *Candida albicans* SC5314 reference genome (A21-s02-m09-r08, obtained 7 of October 2015 from the *Candida* Genome Database (CGD): http://www.candidagenome.org/download/sequence/C_albicans_SC5314/Assembly21/archive/ C_albicans_SC5314_version_A21-s02-m09-r08_chromosomes.fasta.gz). All maximum length matches were identified, regardless of their uniqueness (meaning all matches in the genome were identified). Then, all sequence matches were clustered and extended to obtain a maximum-length colinear string of matches if they were separated by no more than 90 nucleotides (NUCmer default parameters). Three repeat matches shared less than 80% sequence identity, therefore an 80% cutoff was used for the final long repeat analysis (Supplementary File 2), similar to previous studies (Achaz et al., 2000; Warren et al., 2014). All sequences that self-aligned to the same genomic position were removed.

Repeat matches were annotated using the reference genome feature file (C_albicans_SC5314_version_A21-s02-m09-r08_Chromosomal_feature file) and repeat tracks obtained from CGD (Skrzypek et al., 2017). To highlight uncharacterized long repeat sequences, repeats associated with the three major classes of repetitive DNA in *C. albicans* were removed, including the rDNA locus, MRS sequences (*RPS*, *HOK*, and *RB2*), telomere-proximal regions, as well as ambiguous sequences (containing poly-N nucleotides). These regions are highly variable and difficult to analyze with short-read sequencing techniques (Chibana et al., 2000; Chibana et al., 1994; Chindamporn et al., 1998; Goodwin & Poulter, 2000; Hoyer & Cota, 2016; Hoyer et al., 1995; Levdansky et al., 2008). Telomere-proximal regions were determined as the region from each chromosome end to the first confirmed, non-repetitive-genome feature, similar to previous studies (Ene et al., 2018; Hirakawa et al., 2015): Chr1: 1-10000, Chr1:3181000-3188548, Chr2: 1-5000, Chr2: 2228650-2232035, Chr3: 1-15000, Chr3: 1787000-1799406, Chr4: 1-2700, Chr4: 1597200-1603443, Chr5: 1-3800, Chr5: 1183000-1190928, Chr6: 1-3000, Chr6: 1031500-1033530, Chr7: 1-75, Chr7: 942300-949616, ChrR: 1-4500, ChrR: 2286355-2286389. Telomere-associated genes, including *TLO* genes, that were not positioned in these telomere-proximal regions were maintained.

All long repeat sequences were verified using BLAST and IGV. Repeat copies that were on the same chromosome were defined as either tandem, mirrored, or inverted using the repeat start and end positions obtained from NUCmer and manually inspected in IGV. Tandem repeat sequences are in the same orientation on the same strand, mirrored repeat sequences are in opposite orientations on the same strand, and inverted repeat sequences are in opposite orientations on the opposite strand. Spacer length was obtained by calculating the shortest distance between repeat matches.

After the post-alignment annotations and filtration, repeats were combined into repeat families if they shared an identical match. For example, if repetitive sequence A was matched with sequence B, sequences A and B were combined into one family. In some instances, a sequence matched with more than one sequence (e.g. A matched with B and C). In these cases, all matched sequences were combined into one family. In total, 230 repeat families were identified with sequence identities of

≥80% (median value of 92.9%) between all copies of the repeat within a family. Of these 230 families, 68 included more than two copies per genome (Supplementary File 2).

The fraction of the genome covered by long repeat sequences was determined by multiplying the average copy length of each repeat family by the number of copies of that repeat family found throughout the genome (excluding the complex tandem repeat genes). The sum of the average copy length of all repeat families (409129 bp) was then divided by the length of the haploid *Candida albicans* SC5314 reference genome (excluding the mt-DNA, 14280189 bp) to determine that 2.87% of the genome is covered by long repeat sequences (Figure 2 – Source Data 1).

Annotation of Repeat Sequences: The long repeat sequences were annotated according to the genomic features contained within each matched repeat sequence using the *C. albicans* genome feature file described above. The genomic features included were: lone long terminal repeats (LTRs) lacking ORFs, retrotransposons, tRNAs, ORFs, and intergenic sequences. Repeat matches containing ORFs included partial ORF sequences, single complete ORF sequences, and multiple ORFs and intergenic sequences. In cases where one repeat copy contained a genome feature, but the other repeat copy contained an intergenic sequence (no genome feature), this later repeat was flagged as "Unannotated Intergenic Sequence" and both repeat copies were assigned the feature found at the annotated repeat copy (Supplementary File 2). All unannotated sequences were verified in both V21 and V22 of the *C. albicans* reference genome (Skrzypek et al., 2017).

Of the known LTRs present within the *C. albicans* genome, only five were not detected in the MUMmer analysis. Analysis of the five undetected LTRs using BLASTN revealed that they lacked an exact match of 20 nucleotides required to establish a matched repeat pair.

All full-length ORF coding sequences within the *C. albicans* reference genome (C_albicans_SC5314_version_A21-s02-m09-r08_chromosomes.fasta.gz) were analyzed for length and GC content using EMBOSS Infoseq (http://imed.med.ucm.es/cgi-bin/emboss.pl?_action=input&_app=infoseq). All full-length ORF coding sequences were divided into coding sequences that were contained within long repeat sequences or coding sequences that were not contained within long repeat sequences (excluding the complex tandem repeat genes, Supplementary File 2, Figure 2-figure supplement 3D & E). If a long repeat sequence contained a partial ORF sequence, the full-length coding sequence was used in the analysis. Similarly, if a long repeat sequence contained multiple ORF sequences, the full-length coding sequence of each ORF were included in the analysis.

Exclusion of Complex Tandem Repeat Genes: Five ORFs and one gene family with known, complex embedded tandem repeats were confirmed by NUCmer (*PGA18*, *PGA55*, *EAP1*, Adhesin-like *orf19.1725*, *CSA1*, and the *ALS* gene family comprised of seven ORFs, Supplementary File 2) (Levdansky et al., 2008; Wilkins et al., 2018). Assignment of a genome copy count was not possible for these tandem repeat genes due to the extreme complexity of matched repeat sequences. For this reason, all repeat copy counts and analysis using copy counts exclude the complex tandem repeat genes listed above and are indicated throughout the text (Supplementary File 2).

Statistical Analyses: For this study, biological replicates are defined as a single, independent culture derived from a frozen -80°C glycerol stock. Data were analyzed using GraphPad Prism v6 and made into graphical representations using RSudio v1.1.463. All p-values below 0.05 were considered significant.

ACKNOWLEDGEMENTS

We thank all members of the Selmecki laboratory, especially Curtis Focht, Alison Guyer, Robert Thomas, and Annette Beach for technical assistance. We thank Dr. Robin Dowell, Dr. Mary Ann Allen, and Dr. Hung-Ji Tsai for feedback on the manuscript and helpful discussions. Support for this research was provided by LB692 NE Tobacco Settlement Biomedical Research Development New Initiative Grant (to AS), NE Established Program to Stimulate Competitive Research (EPSCoR) First Award (to AS), NE Department of Health and Human Services (LB506-2017-55) award (to AS), CURAS Faculty Research Fund Award (to AS), and NIH-NCRR COBRE grant P20RR018788 sub-award (to AS). AF was supported by NIH grant R15 AI090633. The sequencing datasets generated during this study are available in the Sequence Read Archive repository under project PRJNA510147.

FIGURE LEGENDS

Figure 1: Inverted repeat at *CEN4* causes a novel isochromosome leading to increased fluconazole resistance. (A) Whole genome sequence data plotted as a log2 ratio and converted to chromosome copy number (Y-axis) and chromosome location (X-axis) using YMAP, for the progenitor clinical isolate (P78042) and an isolate obtained after 100 generations in FLC (AMS3743). The copy number breakpoint in AMS3743 occurs at *CEN4* (red arrow). **(B)** CHEF karyotype gel stained with ethidium bromide (left panel) identifies a novel band (asterisk) above Chr5. Southern blot analysis (right panel) of the same gel using a DIG-labeled *CEN4* probe identifies the full-length Chr4 homolog in P78042 and AMS3743, and the novel band in AMS3743 that is twice the size of the right arm of Chr4 in an isochromosome structure (asterisk, i(4R)). **(C)** PCR validation of i(4R). Schematic representation of the Chr4 homologue (top) and i(4R), where the location of a single primer sequence (Primer 1, Supplementary File 7) that flanks the *CEN4* inverted repeat is indicated. PCR with Primer 1 amplified the expected product of i(4R) in AMS3743. **(D)** 24-hour growth curves in YPAD (top panel)

and YPAD+32 µg/ml FLC (bottom panel) for P78042 (black line) and AMS3743 (green line). Average slope and standard error of the mean for three biological replicates is indicated. The average maximum slope (n=3) of P78042 and AMS3743 in YPAD was not significantly different (0.046 and 0.046, respectively, $p > 0.75$, t-test). The average maximum slope (n=3) of P78042 and AMS3743 was significantly different in FLC (0.002 and 0.003, respectively, $p < 0.0006$, t-test). OD, optical density (Figure 1 – Source Data 1).

Figure 1-figure supplement 1: Long inverted repeats on Chr4 are associated with a centromere inversion and an isochromosome that confers increased fitness in FLC. (A) CHEF karyotype gel stained with ethidium bromide. Passage of AMS3743_10 for 30 days in YPAD alone followed by single colony selection identified one single colony that had lost the i(4R) band (AMS3743_S6). **(B)** 24-hour growth curves in YPAD (top panel) and YPAD+32 µg/ml FLC (bottom panel) of P78042 (black line), AMS3743 with i(4R) (green line), AMS3734_S1 with i(4R) (blue line), and AMS3743_S6 which lost the i(4R) (red line). There was no significant difference in average max slope between P78042, AMS3743, AMS3743_S1, and AMS3746_S6 in YPAD ($p > 0.96$, one-way ANOVA with Tukey's multiple comparison). The average maximum slope in FLC was significantly higher in isolates containing i(4R) (0.003 for both AMS3743 & AMS3743_S1) than in the isolates not containing i(4R) (0.002 for both P78042 & AMS3742_S6) ($p > 0.05$, one-way ANOVA with Tukey's multiple comparison). OD, optical density (Figure 1 – Source Data 1). **(C)** Location of the *CEN4* inverted repeat (red arrows and lines). Location of the Major Repeat Sequence on Chr4 (black circle). **(D)** *CEN4* is comprised of a ~3.6 kb CENP-A-binding core sequence (hatched box) asymmetrically flanked by a 524 bp inverted repeat sequence (red) separated by ~3.1 kb. PCR analysis with primers anchored outside or inside the inverted repeat (Primers 2, 3, & 4, see Supplementary File 7), identified two different orientations of *CEN4* (denoted Chr4A and Chr4B) that arose due to an inversion between the repeat sequences on one homologue, in the reference strain SC5314 all isolates analyzed.

Figure 1-figure supplement 2: Sanger sequencing of *CEN4* in SC5314. Unique PCR fragments flanking the left side of the *CEN4* inverted repeat were obtained for the reference isolate SC5314. PCR products were amplified for both the reference and inverted orientations of *CEN4*. Primers are indicated as in Figure 1-figure supplement 1D and Supplementary File 7. Sanger sequencing was performed with both forward and reverse primers.

Figure 2: Long repeat sequences are found across the *C. albicans* genome. Detailed results for all long intra- and inter-chromosomal repeat positions, orientations, and gene features are found in Supplementary File 2. Repeats associated with the rDNA, major repeat sequences (MRS), and sub-telomeric repeats were removed prior to the analysis. **(A)** Representative image of the long intra-chromosomal repeat positions (colored lines – not to scale). Each repeat family is assigned a unique color within its respective chromosome. Numbers and symbols below each chromosome indicate chromosomal position (Mb), MRS position (black circles), and rDNA locus (blue circle, ChrR). **(B)** Number of all repeat matches (excluding the complex tandem repeat genes) on each chromosome, ordered by chromosome size ($R^2 = 0.65$, p-value < 0.016, gray indicates 95% confidence interval, Figure 2 – Source Data 1). **(C)** The number of intra-chromosomal (Intra-Chr) and inter-chromosomal (Inter-Chr) repeat matches assigned to each genomic feature: Intergenic, LTR, ORF (excluding the complex tandem repeat genes), retrotransposon (Retro), and tRNA (Figure 2 – Source Data 1).

Figure 2–figure supplement 1: Features of long repeat sequences. Schematic of a previously uncharacterized long repeat sequence (repeat family 124). The repeat sequence (red arrows) is described in terms of copy length (bp) and shared sequence identity (% of exact nucleotide matches) between the two matched sequences. The distance between intra-chromosomal repeat matches is the spacer length and their orientation can be inverted (reverse complement located on the opposite DNA strand), mirrored (reverse complement located on the same DNA strand), or tandem. Long repeat sequences are further characterized by the genomic features contained within the repeat. Long repeats that contain ORFs include partial ORF sequences, single complete ORF sequences (paralogs) or multiple ORFs and intergenic sequences. Repeat family 124 contains four complete ORFs (black arrows) and flanking intergenic sequences in each copy of the long repeat sequence. Other repeat sequences contain lone LTRs, retrotransposons, tRNAs, and intergenic sequences. Details of all repeat sequence matches are found in Supplementary File 2.

Figure 2–figure supplement 2: The intra-chromosomal repeats with the longest spacer length are found on the longer chromosomes. (A) The spacer length for all intra-chromosomal repeat matches (excluding the complex tandem repeat genes) for each chromosome, ordered by chromosome size in bp ($R^2 = 0.06$, $p < 0.0001$, Figure 2 – Source Data 2). **(B)** Distribution of intra-chromosomal spacer length for each of the eight *C. albicans* chromosomes (chromosome ends indicated with a black bar). There is a significant difference in the distributions of repeat spacer lengths between chromosomes ($p < 0.035$, Kruskal-Wallis test with Dunn's multiple comparison), with the longest chromosomes having more repeat matches that are separated by greater spacer lengths than the smallest chromosomes (Figure 2 – Source Data 2).

Figure 2–figure supplement 3: Key features of long repeat sequences in *C. albicans*. The percent shared identity **(A)** and repeat copy length **(B)** of intra-chromosomal (Intra-Chr) or inter-chromosomal (Inter-Chr) repeat matches containing each genomic feature: Intergenic, LTR, ORF (excluding the complex tandem repeat genes), Retrotransposon (Retro), and tRNA (Supplementary File 2). Copy length is significantly different between repeats containing ORFs compared to repeats containing other features ($p < 0.0001$, Kruskal-Wallis with Dunn's multiple comparisons). **(C)** Percent sequence identity of repeat matches for each chromosome both before (pink) and after (blue) removal of the complex tandem repeat genes. The median sequence identity of repeats on Chr6 is significantly increased after removal of the complex tandem repeat genes ($p < 0.0001$, Kruskal-Wallis with Dunn's multiple comparisons). The length **(D)** and percent GC content **(E)** of the full-length ORF coding sequences (CDS) within long repeat sequences (pink) and all other full-length CDSs not contained in long repeat sequences (blue). Dashed lines represent median values. The full-length CDSs contained in long repeats are significantly longer ($p < 0.0008$, Kolmogorov-Smirnov test) and have a significantly higher percent GC content ($p < 0.0001$, Kolmogorov-Smirnov test) than all full-length CDSs not contained in long repeat sequences. *** $p < 0.001$, **** $p < 0.0001$ (See Methods, Figure 2 – Source Data 3).

Figure 3: All copy number breakpoints resulting in segmental aneuploidy occur at repeat sequences.

(A) Whole genome sequence data plotted as a log2 ratio and converted to chromosome copy number (Y-axis) and chromosome location (X-axis) using YMAP. The source of each isolate is indicated in color: *in vivo* evolution experiments in a murine model of oropharyngeal candidiasis (OPC) (green), *in vitro* evolution experiments in the presence of azole antifungal drugs (red), and clinical isolates (blue). Ploidy, determined by flow cytometry, is indicated on the far right. Every copy number breakpoint occurred at a repeat sequence (red arrow), additional details are in Supplementary File 3. Location of the Major Repeat Sequences (black circle) and rDNA array (blue circle) shown below. Example copy number breakpoints for two isolates **(B-C)**. **(B)** Isolate AMS3053 underwent a complex rearrangement on Chr3L at a long inverted repeat (Repeat 124, red lines). Read depth (top panel) and allele frequency (IGV panel) data indicate the copy number breakpoint coincided with LOH (blue region) telomere proximal to the breakpoint. The inverted repeat copies (~3.2 kb, 99.5% sequence identity, separated by ~11.3 kb) each contain four complete ORFs and intergenic sequences. **(C)** Read depth (top panel) and allele frequency (IGV panel) data for isolate CEC2871 shows an internal chromosome deletion on ChrR with copy number breakpoints (red lines) and LOH (blue) that occur between a long tandem repeat (Repeat family 201, red arrows). The tandem repeat copies (~1.4 kb, 93.8% sequence identity, separated by ~55 kb) each contain one ORF.

Figure 3–figure supplement 1: Segmental aneuploidies occur at previously characterized and uncharacterized long repeat sequences. Representative segmental aneuploidy breakpoints from Figure 3. Whole genome sequence data plotted as a log2 ratio and converted to chromosome copy number (Y-axis) and chromosome location (X-axis) using YMAP. Copy number variation breakpoints (red and green arrowheads) are indicated. Each breakpoint is associated with a long repeat sequence (red or green arrow) shown in the gene track, and annotated genomic features are indicated with black arrows, below the gene track **(A-J,** Supplementary File 3**)**. Segmental chromosome aneuploidies from the indicated isolates occur within **(A)** Chr1 repeat family 14, containing ORFs *HGT1* and *HGT2*; **(B)** Chr2 repeat family 93, containing two uncharacterized ORFs; **(C)** Chr3 repeat family 124 containing eight ORFs and associated intergenic sequences; **(D)** *CEN4* repeat family 151; **(E)** *CEN5* repeat family 161, containing two ORFs; **(F)** Chr6 repeat family 137, containing the *ALS* gene family; **(G)** a complex repeat region on Chr7 with both inverted and tandem repeat sequences containing five uncharacterized ORFs; and **(H)** ChrR repeat region containing the rDNA array. Two examples of complex segmental aneuploidies involving more than one repeat family **(I & J)**. **(I)** Chr1 repeat family 65 is associated with a centromere proximal amplification, while repeat family 40 is associated with a chromosome truncation event. **(J)** Example of a segmental aneuploidy flanked by two different repeat families. An internal deletion is flanked by repeat family 14 and family 9 in clinical isolate FH5. Family 9 is the only inter-chromosomal repeat associated with any observed copy number breakpoint.

Figure 4: Many LOH breakpoints occur at long intra- and inter-chromosomal repeat sequences. Whole genome sequence data plotted as a log2 ratio and converted to chromosome copy number (Y-axis) and chromosome location (X-axis) using YMAP. **(A)** All long-range homozygous regions (light blue) that are associated with long repeat sequences (colored arrows) are indicated for 20 diverse *C. albicans* isolates. LOH breakpoints and isolate information are detailed in Supplementary Files 1 & 4. The type of long repeat is indicated with colored arrows: intra-chromosomal (red), inter-chromosomal (yellow), both intra- and inter-chromosomal (green), rDNA repeat (blue), and MRS (black). **(B-C)** Two example LOH breakpoints in isolate CEC723 that occur at long repeats (red arrows) on **(B)** Chr1 (repeat copy length ~1.1 kb), and **(C)** ChrR (repeat copy length ~3.3 kb) and continue to the right telomere of the respective chromosomes. Heterozygous and homozygous allele ratios are indicated in the IGV track. The position, orientation, and spacer length of the long repeat sequence is indicated in the gene track. ORFs (black arrows) contained within the long repeat sequences are indicated above the gene track. The LOH breakpoint on ChrR is within a repeat dense region; additional long repeats in the region are indicated (dashed arrows).

Figure 4–figure supplement 1: Long-track homozygosis occurs on Chr3L at telomere-seed sequences. (A) Homozygosis of the right arm of Chr3 in SC5314 occurred near a telomere repeat sequence. Chromosome plot indicating the location of homozygosis (light blue) on Chr3R in SC5314. An 8 bp unit of the *C. albicans* telomere repeat sequence (5' – AACTTCTT – 3') indicated by the two red arrows. Read depth and allele ratios above the gene track indicates that homozygosis occurs near the 8 bp telomere seed sequence in the 3’ end of *orf19.5880* and continues to the Chr3R telomere. (**B**) Proposed model of telomere addition and subsequent homozygosis of the right arm of Chr3 in SC5314. (**i**) A double-strand DNA break occurs on one homolog of Chr3 (blue) near the 8 bp telomere seed sequence (red arrow). (**ii**) Recombination between the 8 bp telomere seed sequence on Chr3 and a telomere on another chromosome end (**iii**) leads to the formation of a truncated Chr3 capped with a new telomere. (**iv**) A secondary break within the newly added telomere sequence and BIR of the opposite Chr3 homolog results in (**v**) formation of a full-length disomic Chr3 that is homozygous for the right arm.

Figure 5: Long repeat sequences are associated with chromosomal inversions. (A) Whole genome sequence read depth plotted as a log2 ratio and converted to chromosome copy number (Y-axis) and chromosome location (X-axis) using YMAP. Long-range homozygous regions (blue) on Chr4 are indicated for the isolate P75063. IGV allele ratio track indicates multiple homozygous to heterozygous transitions between a long inverted repeat (red arrows, repeat 144, copy length ~1.7 kb). Primers (5, 6, and 7, Supplementary File 7) were designed to unique sequences flanking repeat 144. **(B)** PCR amplification between Primers 6 & 7 identifies a ~32 kb chromosomal inversion in both the reference strain SC5314 and P75063; the reference orientation did not amplify (Primers 5 & 6).

Figure 6: Breakpoints associated with CNV, LOH, and inversion predominantly occur at long repeats that contain ORFs. (A) Scatterplot of percent sequence identity and copy length of all long repeat matches in Supplementary File 2, excluding the complex tandem repeat genes. All long repeats are indicated in gray, and repeats associated with the observed breakpoints are indicated as follows: LOH (blue), CNV (red), and inversion (green). Six repeats (black circle) were associated with more than one type of breakpoint, and two repeats (black star) were associated with all three types of

breakpoints. Solid black lines indicate the median repeat copy length (278 bp, vertical black line) and median percent sequence identity (94.3%, horizontal black line). Repeats associated with LOH, CNV, and inversion breakpoints have a significantly higher median copy length ($p < 0.0001$, Kolmogorov-Smirnov test) and median sequence identity ($p < 0.036$, Kolomogorov-Smirnov) than all other long repeat sequences (excluding the complex tandem repeat genes*,* Figure 6 – Source Data 1). **(B)** Scatterplot as in Figure 6A, where genomic features contained within long repeats are indicated: intergenic sequence (light brown), lone LTR (blue), ORF (pink), retrotransposon (dark brown), and tRNA (green). **(C)** The distribution of genomic features contained within long repeats at LOH, CNV, and inversion breakpoints. Colors indicated as in Figure 6B.

Figure 6–figure supplement 1: Breakpoint-associated repeats containing ORFs have both high sequence identity and long copy length. (A) Percent sequence identity of long repeat matches (excluding the complex tandem repeat genes) associated with an observed breakpoint, or not associated with an observed breakpoint (gray) for each genomic feature contained within the long repeat (intergenic sequence (light brown), lone LTR (blue), ORF (pink), and tRNA (green)). Breakpoint-associated repeats containing intergenic sequences (n=3) have significantly higher identity than all other breakpoint-associated repeats combined ($p < 0.036$, Kruskal-Wallis (K-W)). The sequence identity of breakpoints containing ORFs and intergenic sequence are significantly higher than non-breakpoint associated repeats containing the same genomic features (intergenic sequence $p < 0.023$, ORF $p < 0.0001$, Kolmogorov-Smirnov (K-S)). **(B)** The copy length of repeats associated with an observed breakpoint (color as in A) or not associated with an observed breakpoint (gray) for each genomic feature contained within the long repeat. Breakpoint-associated repeats containing ORFs are significantly longer than all other repeats ($p < 0.0001$, Kruskal-Wallis, Figure 6 – Source Data 1). Breakpoint-associated repeats containing ORFs are significantly longer than non-breakpoint associated repeats containing ORFs ($p < 0.0001$, Kolmogorov-Smirnov). Importantly, breakpoint-associated repeats containing ORFs were the only repeats with both significantly higher identity and significantly longer copy length than non-breakpoint associated repeats (Figure 6-figure supplement 1).

Figure 7: Mechanisms for recombination between long repeats that result in segmental amplification, deletion, LOH, and/or inversion. (A) Intra-molecular single-strand annealing occurs after a double strand break (DSB) on a single DNA molecule undergoes 5'-3' resection exposing two copies of an inverted repeat on the single-stranded 3' overhang. Annealing of the two inverted repeat copies occurs followed by DNA synthesis resulting in a fold-back structure and partial chromosome truncation. **(B)** Inter-molecular single-strand annealing occurs when a DSB occurs on two separate DNA molecules. After 5'-3' resection, annealing between the single-stranded inverted repeat copies of the two different DNA molecules results in the formation of a dicentric chromosome and partial chromosome truncation. **(C)** A single DNA molecule (blue) containing two tandem repeats (red arrows) undergoes a DSB leading to 5'-3' resection that exposes the tandem repeats. The homologous sequences anneal and non-homologous 3' tails are removed. The remaining gap is filled producing an intact chromosome that has undergone an internal deletion. **(D)** Break-Induced-Replication (BIR) induces loss-of-heterozygosity between repeat sequences found on opposite homologs: Two homologs, homolog A (blue) and homolog B (magenta), contain inverted repeat sequences (red arrows). A double strand break occurring on homolog A leads to strand invasion and DNA synthesis. Upon termination of synthesis of both the leading and lagging strands, all sequences to the right of the DSB are homozygous. **(E)** Inversion events occur due to intra-molecular recombination between inverted repeats (red arrows) flanking a unique sequence. The orientation of the reference sequence is indicated above chromosome (1-2-3-4-5). Non-Allelic Homologous Recombination (NAHR) between the inverted repeats leads to an inversion of the sequence between the repeats (1-4-3-2-5).

SUPPLEMENTARY FILES

Supplementary File 1: Strains used in this study

Supplementary File 2: Long repeat sequences in the *Candida albicans* genome

Supplementary File 3: Copy number variation breakpoints in diverse *C. albicans* isolates

Supplementary File 4: Loss of heterozygosity breakpoints in diverse *C. albicans* isolates

Supplementary File 5: Location of telomere-seed sequences throughout the *C. albicans* genome

Supplementary File 6: Predicted inversion breakpoints in diverse *C. albicans* isolates

Supplementary File 7: Primers used in this study

SOURCE DATA FILES

Figure 1 – Source Data 1: Growth curve analysis

Figure 2 – Source Data 1: Distribution, Features, and Coverage of long repeat sequences in *C. albicans*

Figure 2 – Source Data 2: Analysis of long repeat spacer length in *C. albicans*

Figure 2 – Source Data 3: Analysis of key features of long repeat sequences in *C. albicans*

Figure 6 – Source Data 1: Analysis of long repeat sequences associated with CNV, LOH, and sequence inversion

REFERENCES

Abbey, D. A., Funt, J., Lurie-Weinberger, M. N., Thompson, D. A., Regev, A., Myers, C. L., & Berman, J. (2014). YMAP: a pipeline for visualization of copy number variation and loss of heterozygosity in eukaryotic pathogens. *Genome Med, 6*(11), 100. doi:10.1186/s13073-014-0100-8

Achaz, G., Coissac, E., Viari, A., & Netter, P. (2000). Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin. *Mol Biol Evol, 17*(8), 1268-1275. doi:10.1093/oxfordjournals.molbev.a026410

Adamo, G. M., Lotti, M., Tamas, M. J., & Brocca, S. (2012). Amplification of the CUP1 gene is associated with evolution of copper tolerance in Saccharomyces cerevisiae. *Microbiology, 158*(Pt 9), 2325-2335. doi:10.1099/mic.0.058024-0

Aguilera, A., & Gaillard, H. (2014). Transcription and recombination: when RNA meets DNA. *Cold Spring Harb Perspect Biol, 6*(8). doi:10.1101/cshperspect.a016543

Aguilera, A., & Garcia-Muse, T. (2013). Causes of genome instability. *Annu Rev Genet, 47*, 1-32. doi:10.1146/annurev-genet-111212-133232

Anand, R. P., Lovett, S. T., & Haber, J. E. (2013). Break-induced DNA replication. *Cold Spring Harb Perspect Biol, 5*(12), a010397. doi:10.1101/cshperspect.a010397

Anderson, M. Z., Baller, J. A., Dulmage, K., Wigen, L., & Berman, J. (2012). The three clades of the telomere-associated TLO gene family of Candida albicans have different splicing, localization, and expression features. *Eukaryot Cell, 11*(10), 1268-1275. doi:10.1128/ec.00230-12

Araya, C. L., Payen, C., Dunham, M. J., & Fields, S. (2010). Whole-genome sequencing of a laboratory-evolved yeast strain. *BMC Genomics, 11*, 88. doi:10.1186/1471-2164-11-88

Bennett, R. J., & Johnson, A. D. (2003). Completion of a parasexual cycle in Candida albicans by induced chromosome loss in tetraploid strains. *Embo j, 22*(10), 2505-2515. doi:10.1093/emboj/cdg235

Bhargava, R., Onyango, D. O., & Stark, J. M. (2016). Regulation of Single-Strand Annealing and its Role in Genome Maintenance. *Trends Genet, 32*(9), 566-575. doi:10.1016/j.tig.2016.06.007

Bishop, A. J., & Schiestl, R. H. (2000). Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. *Hum Mol Genet, 9*(16), 2427-2334.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bouchonville, K., Forche, A., Tang, K. E., Selmecki, A., & Berman, J. (2009). Aneuploid chromosomes are highly unstable during DNA transformation of Candida albicans. *Eukaryot Cell, 8*(10), 1554-1566. doi:10.1128/ec.00209-09

Braun, B. R., van Het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., Inglis, D. O., Uhl, M. A., Hogues, H., Berriman, M., Lorenz, M., Levitin, A., Oberholzer, U., Bachewich, C., Harcus, D., Marcil, A., Dignard, D., Iouk, T., Zito, R., Frangeul, L., Tekaia, F., Rutherford, K., Wang, E., Munro, C. A., Bates, S., Gow, N. A., Hoyer, L. L., Kohler, G., Morschhauser, J., Newport, G., Znaidi, S., Raymond, M., Turcotte, B., Sherlock, G., Costanzo, M., Ihmels, J., Berman, J., Sanglard, D., Agabian, N., Mitchell, A. P., Johnson, A. D., Whiteway, M., & Nantel, A. (2005). A human-curated annotation of the Candida albicans genome. *PLoS Genet, 1*(1), 36-57. doi:10.1371/journal.pgen.0010001

Brewer, B. J., Payen, C., Di Rienzi, S. C., Higgins, M. M., Ong, G., Dunham, M. J., & Raghuraman, M. K. (2015). Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification. *PLoS Genet, 11*(12), e1005699. doi:10.1371/journal.pgen.1005699

Brewer, B. J., Payen, C., Raghuraman, M. K., & Dunham, M. J. (2011). Origin-dependent inverted-repeat amplification: a replication-based model for generating palindromic amplicons. *PLoS Genet, 7*(3), e1002016. doi:10.1371/journal.pgen.1002016

Brown, G. D., & Netea, M. G. (2012). Exciting developments in the immunology of fungal infections. *Cell Host Microbe, 11*(5), 422-424. doi:10.1016/j.chom.2012.04.010

Bruno, V. M., Wang, Z., Marjani, S. L., Euskirchen, G. M., Martin, J., Sherlock, G., & Snyder, M. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. *Genome Res, 20*(10), 1451-1458. doi:10.1101/gr.109553.110

Burrack, L. S., Hutton, H. F., Matter, K. J., Clancey, S. A., Liachko, I., Plemmons, A. E., Saha, A., Power, E. A., Turman, B., Thevandavakkam, M. A., Ay, F., Dunham, M. J., & Berman, J. (2016). Neocentromeres Provide Chromosome Segregation Accuracy and Centromere Clustering to Multiple Loci along a Candida albicans Chromosome. *PLoS Genet, 12*(9), e1006317. doi:10.1371/journal.pgen.1006317

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., Agrafioti, I., Arnaud, M. B., Bates, S., Brown, A. J., Brunke, S., Costanzo, M. C., Fitzpatrick, D. A., de Groot, P. W., Harris, D., Hoyer, L. L., Hube, B., Klis, F. M., Kodira, C., Lennard, N., Logue, M. E., Martin, R., Neiman, A. M., Nikolaou, E., Quail, M. A., Quinn, J., Santos, M. C., Schmitzberger, F. F., Sherlock, G., Shah, P., Silverstein, K. A., Skrzypek, M. S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M. P., Sudbery, P. E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N. A., Lorenz, M. C., Birren, B. W., Kellis, M., & Cuomo, C. A. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature, 459*(7247), 657-662. doi:10.1038/nature08064

Chen, L., Zhou, W., Zhang, L., & Zhang, F. (2014). Genome architecture and its roles in human copy number variation. *Genomics Inform, 12*(4), 136-144. doi:10.5808/gi.2014.12.4.136

Chibana, H., Beckerman, J. L., & Magee, P. T. (2000). Fine-resolution physical mapping of genomic diversity in Candida albicans. *Genome Res, 10*(12), 1865-1877.

Chibana, H., Iwaguchi, S., Homma, M., Chindamporn, A., Nakagawa, Y., & Tanaka, K. (1994). Diversity of tandemly repetitive sequences due to short periodic repetitions in the chromosomes of Candida albicans. *J Bacteriol, 176*(13), 3851-3858.

Chindamporn, A., Nakagawa, Y., Mizuguchi, I., Chibana, H., Doi, M., & Tanaka, K. (1998). Repetitive sequences (RPSs) in the chromosomes of Candida albicans are sandwiched between two novel stretches, HOK and RB2, common to each chromosome. *Microbiology, 144 (Pt 4)*, 849-857. doi:10.1099/00221287-144-4-849

Christiaens, J. F., Van Mulders, S. E., Duitama, J., Brown, C. A., Ghequire, M. G., De Meester, L., Michiels, J., Wenseleers, T., Voordeckers, K., & Verstrepen, K. J. (2012). Functional divergence of gene duplicates through ectopic recombination. *EMBO Rep, 13*(12), 1145-1151. doi:10.1038/embor.2012.157

Chu, W. S., Rikkerink, E. H., & Magee, P. T. (1992). Genetics of the white-opaque transition in Candida albicans: demonstration of switching recessivity and mapping of switching genes. *J Bacteriol, 174*(9), 2951-2957. doi:10.1128/jb.174.9.2951-2957.1992

Croll, D., Zala, M., & McDonald, B. A. (2013). Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet, 9*(6), e1003567. doi:10.1371/journal.pgen.1003567

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., & Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A, 99*(25), 16144-16149. doi:10.1073/pnas.242624799

Dunn, M. J., Kinney, G. M., Washington, P. M., Berman, J., & Anderson, M. Z. (2018). Functional diversification accompanies gene family expansion of MED2 homologs in Candida albicans. *PLoS Genet, 14*(4), e1007326. doi:10.1371/journal.pgen.1007326

Ene, I. V., Farrer, R. A., Hirakawa, M. P., Agwamba, K., Cuomo, C. A., & Bennett, R. J. (2018). Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proc Natl Acad Sci U S A, 115*(37), E8688-e8697. doi:10.1073/pnas.1806002115

Fischer, G., Rocha, E. P., Brunet, F., Vergassola, M., & Dujon, B. (2006). Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet, 2*(3), e32. doi:10.1371/journal.pgen.0020032

Fogel, S., Welch, J. W., Cathala, G., & Karin, M. (1983). Gene amplification in yeast: CUP1 copy number regulates copper resistance. *Curr Genet, 7*(5), 347-355. doi:10.1007/bf00445874

Forche, A., Abbey, D., Pisithkul, T., Weinzierl, M. A., Ringstrom, T., Bruck, D., Petersen, K., & Berman, J. (2011). Stress alters rates and types of loss of heterozygosity in Candida albicans. *MBio, 2*(4). doi:10.1128/mBio.00129-11

Forche, A., Alby, K., Schaefer, D., Johnson, A. D., Berman, J., & Bennett, R. J. (2008). The parasexual cycle in Candida albicans provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biol, 6*(5), e110. doi:10.1371/journal.pbio.0060110

Forche, A., Cromie, G., Gerstein, A. C., Solis, N. V., Pisithkul, T., Srifa, W., Jeffery, E., Abbey, D., Filler, S. G., Dudley, A. M., & Berman, J. (2018). Rapid Phenotypic and Genotypic Diversification After Exposure to the Oral Host Niche in Candida albicans. *Genetics, 209*(3), 725-741. doi:10.1534/genetics.118.301019

Forche, A., Magee, P. T., Selmecki, A., Berman, J., & May, G. (2009). Evolution in Candida albicans populations during a single passage through a mouse host. *Genetics, 182*(3), 799-811. doi:10.1534/genetics.109.103325

Ford, C. B., Funt, J. M., Abbey, D., Issi, L., Guiducci, C., Martinez, D. A., Delorey, T., Li, B. Y., White, T. C., Cuomo, C., Rao, R. P., Berman, J., Thompson, D. A., & Regev, A. (2015). The evolution of drug resistance in clinical isolates of Candida albicans. *Elife, 4*, e00662. doi:10.7554/eLife.00662

Freire-Beneitez, V., Price, R. J., Tarrant, D., Berman, J., & Buscaino, A. (2016). Candida albicans repetitive elements display epigenetic diversity and plasticity. *Sci Rep, 6*, 22989. doi:10.1038/srep22989

Gerstein, A. C., Fu, M. S., Mukaremera, L., Li, Z., Ormerod, K. L., Fraser, J. A., Berman, J., & Nielsen, K. (2015). Polyploid titan cells produce haploid and aneuploid progeny to promote stress adaptation. *MBio, 6*(5), e01340-01315. doi:10.1128/mBio.01340-15

Goodwin, T. J., & Poulter, R. T. (1998). The CARE-2 and rel-2 repetitive elements of Candida albicans contain LTR fragments of a new retrotransposon. *Gene, 218*(1-2), 85-93.

Goodwin, T. J., & Poulter, R. T. (2000). Multiple LTR-retrotransposon families in the asexual yeast Candida albicans. *Genome Res, 10*(2), 174-191.

Gresham, D., Usaite, R., Germann, S. M., Lisby, M., Botstein, D., & Regenberg, B. (2010). Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. *Proc Natl Acad Sci U S A, 107*(43), 18551-18556. doi:10.1073/pnas.1014023107

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet, 5*(1), e1000327. doi:10.1371/journal.pgen.1000327

Hickman, M. A., Zeng, G., Forche, A., Hirakawa, M. P., Abbey, D., Harrison, B. D., Wang, Y. M., Su, C. H., Bennett, R. J., Wang, Y., & Berman, J. (2013). The 'obligate diploid' Candida albicans forms mating-competent haploids. *Nature, 494*(7435), 55-59. doi:10.1038/nature11865

Higashimoto, K., Maeda, T., Okada, J., Ohtsuka, Y., Sasaki, K., Hirose, A., Nomiyama, M., Takayanagi, T., Fukuzawa, R., Yatsuki, H., Koide, K., Nishioka, K., Joh, K., Watanabe, Y., Yoshiura, K., & Soejima, H. (2013). Homozygous deletion of DIS3L2 exon 9 due to non-allelic homologous recombination between LINE-1s in a Japanese patient with Perlman syndrome. *Eur J Hum Genet, 21*(11), 1316-1319. doi:10.1038/ejhg.2013.45

Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J. M., Greenberg, J. M., Berman, J., Bennett, R. J., & Cuomo, C. A. (2015). Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res, 25*(3), 413-425. doi:10.1101/gr.174623.114

Horak, J. (2013). Regulations of sugar transporters: insights from yeast. *Curr Genet, 59*(1-2), 1-31. doi:10.1007/s00294-013-0388-8

Hose, J., Yong, C. M., Sardi, M., Wang, Z., Newton, M. A., & Gasch, A. P. (2015). Dosage compensation can buffer copy-number variation in wild yeast. *Elife, 4*. doi:10.7554/eLife.05462

Hoyer, L. L., & Cota, E. (2016). Candida albicans Agglutinin-Like Sequence (Als) Family Vignettes: A Review of Als Protein Structure and Function. *Front Microbiol, 7*, 280. doi:10.3389/fmicb.2016.00280

Hoyer, L. L., Scherer, S., Shatzman, A. R., & Livi, G. P. (1995). Candida albicans ALS1: domains related to a Saccharomyces cerevisiae sexual agglutinin separated by a repeating motif. *Mol Microbiol, 15*(1), 39-54.

Hull, R. M., Cruz, C., Jack, C. V., & Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. *PLoS Biol, 15*(6), e2001333. doi:10.1371/journal.pbio.2001333

Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B. B., Newport, G., Thorstenson, Y. R., Agabian, N., Magee, P. T., Davis, R. W., & Scherer, S. (2004). The diploid genome sequence of Candida albicans. *Proc Natl Acad Sci U S A, 101*(19), 7329-7334. doi:10.1073/pnas.0401648101

Ketel, C., Wang, H. S., McClellan, M., Bouchonville, K., Selmecki, A., Lahav, T., Gerami-Nejad, M., & Berman, J. (2009). Neocentromeres form efficiently at multiple possible loci in Candida albicans. *PLoS Genet, 5*(3), e1000400. doi:10.1371/journal.pgen.1000400

Kiktev, D. A., Sheng, Z., Lobachev, K. S., & Petes, T. D. (2018). GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A, 115*(30), E7109-e7118. doi:10.1073/pnas.1807334115

Koren, A., Tsai, H. J., Tirosh, I., Burrack, L. S., Barkai, N., & Berman, J. (2010). Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. *PLoS Genet, 6*(8), e1001068. doi:10.1371/journal.pgen.1001068

Kramara, J., Osia, B., & Malkova, A. (2018). Break-Induced Replication: The Where, The Why, and The How. *Trends Genet, 34*(7), 518-531. doi:10.1016/j.tig.2018.04.002

Kunkel, T. A. (1993). Nucleotide repeats. Slippery DNA and diseases. *Nature, 365*(6443), 207-208. doi:10.1038/365207a0

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res, 29*(22), 4633-4642.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol, 5*(2), R12. doi:10.1186/gb-2004-5-2-r12

Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S. F., & Gresham, D. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. *PLoS Biol, 16*(12), e3000069. doi:10.1371/journal.pbio.3000069

Lephart, P. R., & Magee, P. T. (2006). Effect of the major repeat sequence on mitotic recombination in Candida albicans. *Genetics, 174*(4), 1737-1744. doi:10.1534/genetics.106.063271

Levdansky, E., Sharon, H., & Osherov, N. (2008). Coding fungal tandem repeats as generators of fungal diversity. *Fungal Biology Reviews, 22*(3), 85-96. doi:https://doi.org/10.1016/j.fbr.2008.08.001

Li, H. (2013). Aligning sequence reads, clone sequence and assembly contigs with BWA-MEM. *arxiv*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Lobachev, K. S., Shor, B. M., Tran, H. T., Taylor, W., Keen, J. D., Resnick, M. A., & Gordenin, D. A. (1998). Factors affecting inverted repeat stimulation of recombination and deletion in Saccharomyces cerevisiae. *Genetics, 148*(4), 1507-1524.

Lockhart, S. R., Pujol, C., Daniels, K. J., Miller, M. G., Johnson, A. D., Pfaller, M. A., & Soll, D. R. (2002). In Candida albicans, white-opaque switchers are homozygous for mating type. *Genetics, 162*(2), 737-745.

Magee, B. B., & Magee, P. T. (2000). Induction of mating in Candida albicans by construction of MTLa and MTLalpha strains. *Science, 289*(5477), 310-313.

Malkova, A., & Haber, J. E. (2012). Mutations arising during repair of chromosome breaks. *Annu Rev Genet, 46*, 455-473. doi:10.1146/annurev-genet-110711-155547

Malkova, A., & Ira, G. (2013). Break-induced replication: functions and molecular mechanism. *Current opinion in genetics & development, 23*(3), 271-279. doi:10.1016/j.gde.2013.05.007

Marcet-Houben, M., Marceddu, G., & Gabaldon, T. (2009). Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence. *BMC Evol Biol, 9*, 295. doi:10.1186/1471-2148-9-295

McClintock, B. (1939). The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. *Proc Natl Acad Sci U S A, 25*(8), 405-416.

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics, 26*(2), 234-282.

McClintock, B. (1942). The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. *Proc Natl Acad Sci U S A, 28*(11), 458-463.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Mehta, A., & Haber, J. E. (2014). Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol, 6*(9), a016428. doi:10.1101/cshperspect.a016428

Mount, H. O., Revie, N. M., Todd, R. T., Anstett, K., Collins, C., Costanzo, M., Boone, C., Robbins, N., Selmecki, A., & Cowen, L. E. (2018). Global analysis of genetic circuitry and adaptive mechanisms enabling resistance to the azole antifungal drugs. *PLoS Genet, 14*(4), e1007319. doi:10.1371/journal.pgen.1007319

Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Hernday, A. D., Tuch, B. B., Andes, D. R., & Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans. *Cell, 148*(1-2), 126-138. doi:10.1016/j.cell.2011.10.048

Pavelka, N., Rancati, G., Zhu, J., Bradford, W. D., Saraf, A., Florens, L., Sanderson, B. W., Hattem, G. L., & Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature, 468*(7321), 321-325. doi:10.1038/nature09529

Payen, C., Di Rienzi, S. C., Ong, G. T., Pogachar, J. L., Sanchez, J. C., Sunshine, A. B., Raghuraman, M. K., Brewer, B. J., & Dunham, M. J. (2014). The dynamics of diverse segmental amplifications in populations of Saccharomyces cerevisiae adapting to strong selection. *G3 (Bethesda), 4*(3), 399-409. doi:10.1534/g3.113.009365

Pearson, C. E., Nichol Edamura, K., & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet, 6*(10), 729-742. doi:10.1038/nrg1689

Pitarch, A., Diez-Orejas, R., Molero, G., Pardo, M., Sanchez, M., Gil, C., & Nombela, C. (2001). Analysis of the serologic response to systemic Candida albicans infection in a murine model. *Proteomics, 1*(4), 550-559. doi:10.1002/1615-9861(200104)1:4<550::Aid-prot550>3.0.Co;2-w

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Ramakrishnan, S., Kockler, Z., Evans, R., Downing, B. D., & Malkova, A. (2018). Single-strand annealing between inverted DNA repeats: Pathway choice, participating proteins, and genome destabilizing consequences. *PLoS Genet, 14*(8), e1007543. doi:10.1371/journal.pgen.1007543

Reams, A. B., & Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol, 7*(2), a016592. doi:10.1101/cshperspect.a016592

Richard, G. F., Dujon, B., & Haber, J. E. (1999). Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. *Mol Gen Genet, 261*(4-5), 871-882.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Rodić, N., & Burns, K. H. (2013). Long Interspersed Element–1 (LINE-1): Passenger or Driver in Human Neoplasms? *PLoS Genet, 9*(3), e1003402. doi:10.1371/journal.pgen.1003402

Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C., Ma, L., Schwartz, K., Voelz, K., May, R. C., Poulain, J., Battail, C., Wincker, P., Borman, A. M., Chowdhary, A., Fan, S., Kim, S. H., Le Pape, P., Romeo, O., Shin, J. H., Gabaldon, T., Sherlock, G., Bougnoux, M. E., & d'Enfert, C. (2018). Gene flow contributes to diversification of the major fungal pathogen Candida albicans. *Nat Commun, 9*(1), 2253. doi:10.1038/s41467-018-04787-4

Rose, W., Hieter. (1990). Methods in Yeast Genetics. *COld Spring Harbor Laboratory Press*, 177.

Rustchenko, E. P., Curran, T. M., & Sherman, F. (1993). Variations in the number of ribosomal DNA units in morphological mutants and normal strains of Candida albicans and in normal strains of Saccharomyces cerevisiae. *J Bacteriol, 175*(22), 7189-7199.

Rustchenko-Bulgac, E. P. (1991). Variations of Candida albicans electrophoretic karyotypes. *J Bacteriol, 173*(20), 6586-6596.

Santos-Pereira, J. M., & Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. *Nat Rev Genet, 16*(10), 583-597. doi:10.1038/nrg3961

Sanyal, K., Baum, M., & Carbon, J. (2004). Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. *Proc Natl Acad Sci U S A, 101*(31), 11374-11379. doi:10.1073/pnas.0404318101

Scott, A. L., Richmond, P. A., Dowell, R. D., & Selmecki, A. M. (2017). The Influence of Polyploidy on the Evolution of Yeast Grown in a Sub-Optimal Carbon Source. *Mol Biol Evol, 34*(10), 2690-2703. doi:10.1093/molbev/msx205

Selmecki, A., Bergmann, S., & Berman, J. (2005). Comparative genome hybridization reveals widespread aneuploidy in Candida albicans laboratory strains. *Mol Microbiol, 55*(5), 1553-1565. doi:10.1111/j.1365-2958.2005.04492.x

Selmecki, A., Forche, A., & Berman, J. (2006). Aneuploidy and isochromosome formation in drug-resistant Candida albicans. *Science, 313*(5785), 367-370. doi:10.1126/science.1128242

Selmecki, A., Forche, A., & Berman, J. (2010). Genomic plasticity of the human fungal pathogen Candida albicans. *Eukaryot Cell, 9*(7), 991-1008. doi:10.1128/ec.00060-10

Selmecki, A., Gerami-Nejad, M., Paulson, C., Forche, A., & Berman, J. (2008). An isochromosome confers drug resistance in vivo by amplification of two genes, ERG11 and TAC1. *Mol Microbiol, 68*(3), 624-641. doi:10.1111/j.1365-2958.2008.06176.x

Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B., & Berman, J. (2009). Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet, 5*(10), e1000705. doi:10.1371/journal.pgen.1000705

Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., De, S., Kishony, R., Michor, F., Dowell, R., & Pellman, D. (2015). Polyploidy can drive rapid adaptation in yeast. *Nature, 519*(7543), 349-352. doi:10.1038/nature14187

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., Scherer, S., Tait, E., Shaw, D. J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M. A., Barrell, B. G., & Wolfe, K. H. (2000). Prevalence of small inversions in yeast gene order evolution. *Proc Natl Acad Sci U S A, 97*(26), 14433-14437. doi:10.1073/pnas.240462997

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M., & Sherlock, G. (2017). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res, 45*(D1), D592-d596. doi:10.1093/nar/gkw924

Solis, N. V., & Filler, S. G. (2012). Mouse model of oropharyngeal candidiasis. *Nat Protoc, 7*(4), 637-642. doi:10.1038/nprot.2012.011

Strawbridge, E. M., Benson, G., Gelfand, Y., & Benham, C. J. (2010). The distribution of inverted repeat sequences in the Saccharomyces cerevisiae genome. *Curr Genet, 56*(4), 321-340. doi:10.1007/s00294-010-0302-6

Stukenbrock, E. H., Jørgensen, F. G., Zala, M., Hansen, T. T., McDonald, B. A., & Schierup, M. H. (2010). Whole-Genome and Chromosome Evolution Associated with Host Adaptation and Speciation of the Wheat Pathogen Mycosphaerella graminicola. *PLoS Genet, 6*(12), e1001189. doi:10.1371/journal.pgen.1001189

Sunshine, A. B., Payen, C., Ong, G. T., Liachko, I., Tan, K. M., & Dunham, M. J. (2015). The fitness consequences of aneuploidy are driven by condition-dependent gene effects. *PLoS Biol, 13*(5), e1002155. doi:10.1371/journal.pbio.1002155

Suzuki, T., Nishibayashi, S., Kuroiwa, T., Kanbe, T., & Tanaka, K. (1982). Variance of ploidy in Candida albicans. *J Bacteriol, 152*(2), 893-896.

Tan, Z., Hays, M., Cromie, G. A., Jeffery, E. W., Scott, A. C., Ahyong, V., Sirr, A., Skupin, A., & Dudley, A. M. (2013). Aneuploidy underlies a multicellular phenotypic switch. *Proceedings of the National Academy of Sciences, 110*(30), 12367. doi:10.1073/pnas.1301047110

Thomas, B. J., & Rothstein, R. (1989). Elevated recombination rates in transcriptionally active DNA. *Cell, 56*(4), 619-630.

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Todd, R. T., Braverman, A. L., & Selmecki, A. (2018). Flow Cytometry Analysis of Fungal Ploidy. *Curr Protoc Microbiol, 50*(1), e58. doi:10.1002/cpmc.58

Torres, E. M., Sokolsky, T., Tucker, C. M., Chan, L. Y., Boselli, M., Dunham, M. J., & Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science, 317*(5840), 916-924. doi:10.1126/science.1142210

Tsai, H. J., Baller, J. A., Liachko, I., Koren, A., Burrack, L. S., Hickman, M. A., Thevandavakkam, M. A., Rusche, L. N., & Berman, J. (2014). Origin replication complex binding, nucleosome depletion patterns, and a primary sequence motif can predict origins of replication in a genome with epigenetic centromeres. *MBio, 5*(5), e01703-01714. doi:10.1128/mBio.01703-14

VanHulle, K., Lemoine, F. J., Narayanan, V., Downing, B., Hull, K., McCullough, C., Bellinger, M., Lobachev, K., Petes, T. D., & Malkova, A. (2007). Inverted DNA repeats channel repair of distant double-strand breaks into chromatid fusions and chromosomal rearrangements. *Mol Cell Biol, 27*(7), 2601-2614. doi:10.1128/mcb.01740-06

Verstrepen, K. J., Jansen, A., Lewitter, F., & Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nat Genet, 37*(9), 986-990. doi:10.1038/ng1618

Wang, J. M., Bennett, R. J., & Anderson, M. Z. (2018). The Genome of the Human Pathogen Candida albicans Is Shaped by Mutation and Cryptic Sexual Recombination. *MBio, 9*(5). doi:10.1128/mBio.01205-18

Warren, I. A., Ciborowski, K. L., Casadei, E., Hazlerigg, D. G., Martin, S., Jordan, W. C., & Sumner, S. (2014). Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. *Genome Biol Evol, 6*(7), 1790-1805. doi:10.1093/gbe/evu131

Wickes, B., Staudinger, J., Magee, B. B., Kwon-Chung, K. J., Magee, P. T., & Scherer, S. (1991). Physical and genetic mapping of Candida albicans: several genes previously assigned to chromosome 1 map to chromosome R, the rDNA-containing linkage group. *Infect Immun, 59*(7), 2480-2484.

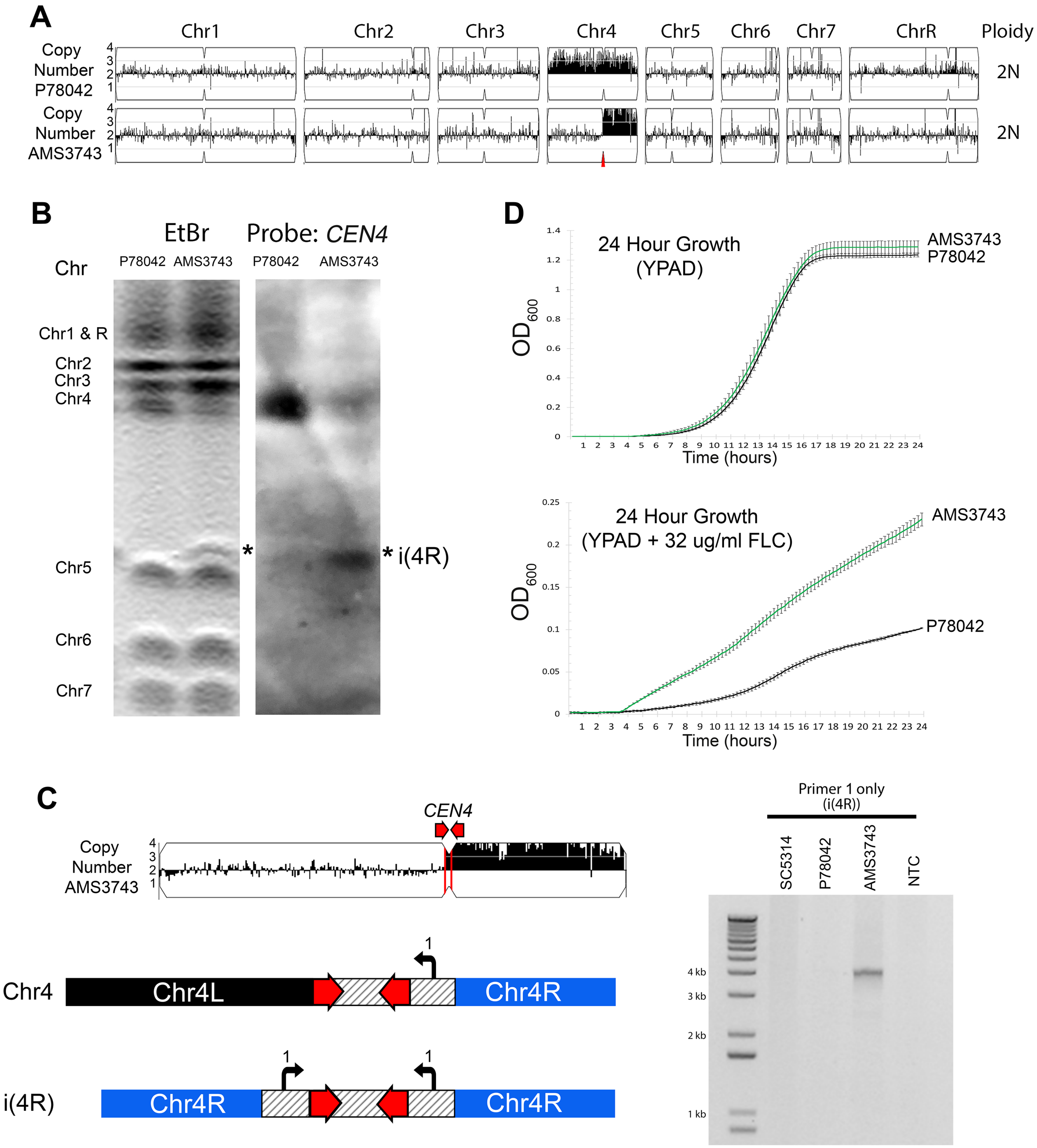
Wilkins, M., Zhang, N., & Schmid, J. (2018). Biological Roles of Protein-Coding Tandem Repeats in the Yeast Candida Albicans. *J Fungi (Basel), 4*(3). doi:10.3390/jof4030078

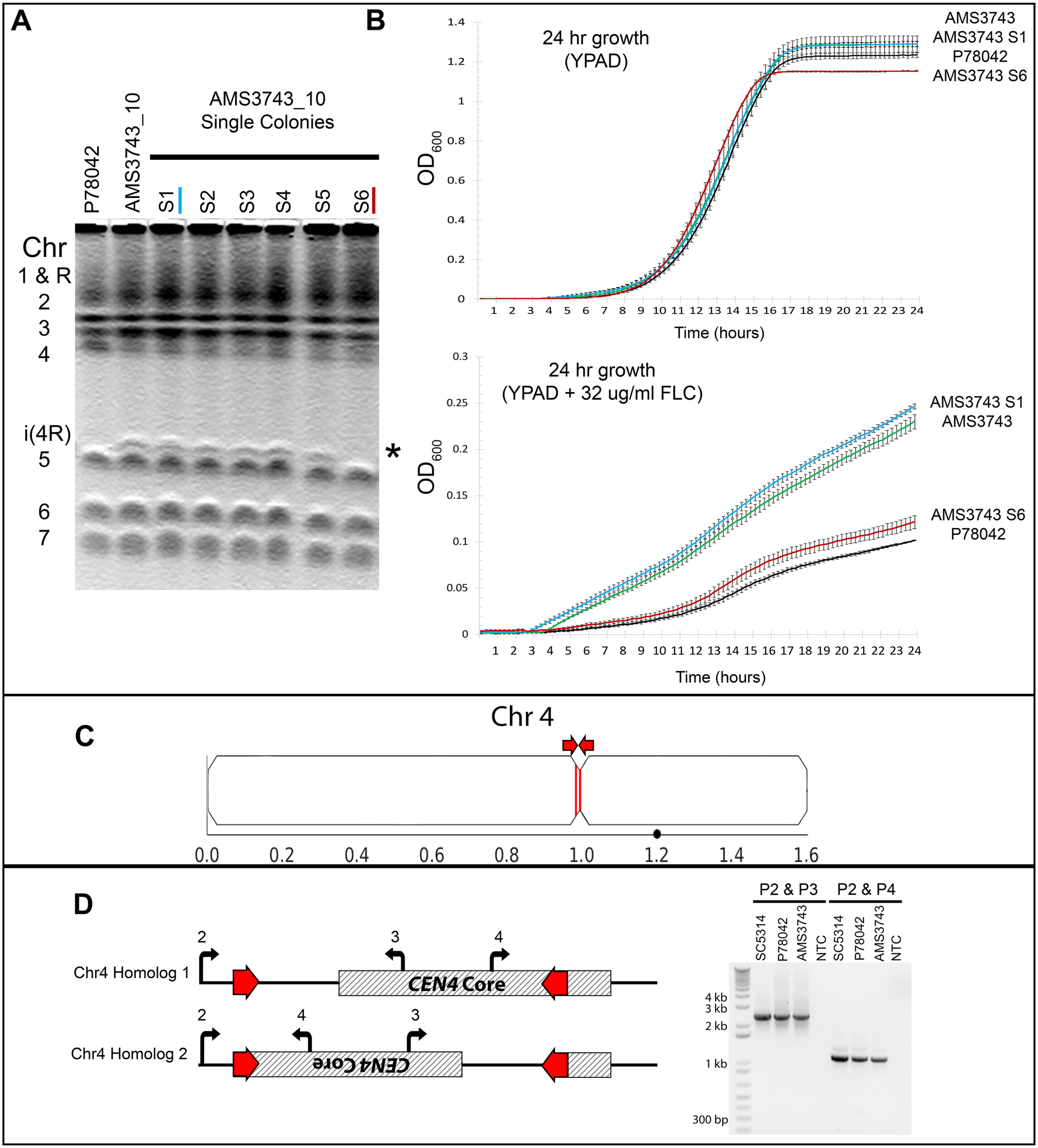
Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature, 387*(6634), 708-713. doi:10.1038/42711

Zhang, N., Cannon, R. D., Holland, B. R., Patchett, M. L., & Schmid, J. (2010). Impact of genetic background on allele selection in a highly mutable Candida albicans gene, PNG2. *PLoS One, 5*(3), e9614. doi:10.1371/journal.pone.0009614

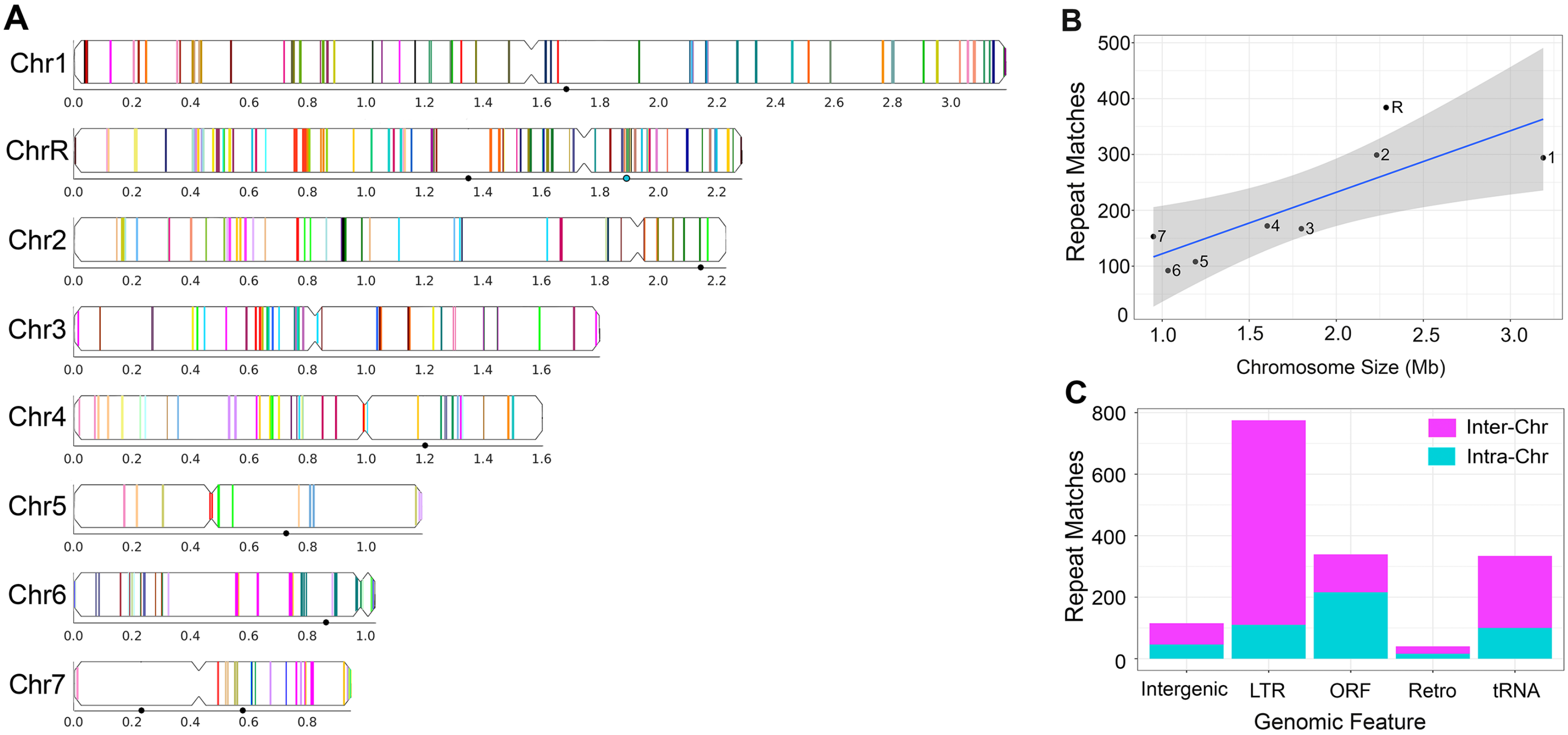
Zhang, N., Harrex, A. L., Holland, B. R., Fenton, L. E., Cannon, R. D., & Schmid, J. (2003). Sixty alleles of the ALS7 open reading frame in Candida albicans: ALS7 is a hypermutable contingency locus. *Genome Res, 13*(9), 2005-2017. doi:10.1101/gr.1024903

Zhao, X., Oh, S. H., Cheng, G., Green, C. B., Nuessen, J. A., Yeater, K., Leng, R. P., Brown, A. J., & Hoyer, L. L. (2004). ALS3 and ALS8 represent a single locus that encodes a Candida albicans adhesin; functional comparisons between Als3p and Als1p. *Microbiology, 150*(Pt 7), 2415-2428. doi:10.1099/mic.0.26943-0


A
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
Ploidy
Copy Number P78042
4
3
2
1
2N
Copy Number AMS3743
2N
B
EtBr
Probe: *CEN4*
Chr
P78042
AMS3743
P78042
AMS3743
Chr1 & R
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
*
* i(4R)
D
24 Hour Growth (YPAD)
OD600
Time (hours)
AMS3743
P78042
24 Hour Growth (YPAD + 32 ug/ml FLC)
AMS3743
P78042
C
CEN4
Copy Number AMS3743
Chr4
Chr4L
Chr4R
1
i(4R)
Chr4R
Chr4R
1
1
Primer 1 only (i(4R))
SC5314
P78042
AMS3743
NTC
4 kb
3 kb
2 kb
1 kb

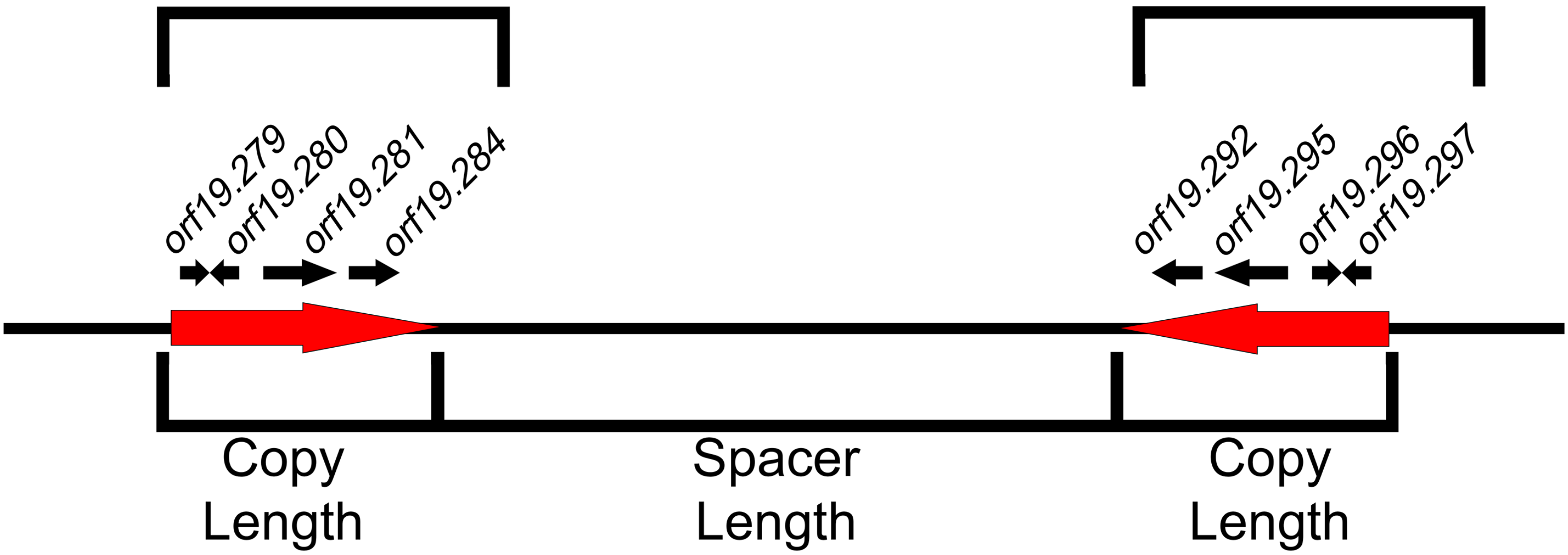

A
P78042
AMS3743_10
AMS3743_10
Single Colonies
S1
S2
S3
S4
S5
S6
Chr
1 & R
2
3
4
i(4R)
5
6
7
*
B
24 hr growth
(YPAD)
AMS3743
AMS3743 S1
P78042
AMS3743 S6
OD_{600}
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Time (hours)
24 hr growth
(YPAD + 32 ug/ml FLC)
AMS3743 S1
AMS3743
AMS3743 S6
P78042
OD_{600}
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Time (hours)
C
Chr 4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
D
Chr4 Homolog 1
2
3
4
CEN4 Core
Chr4 Homolog 2
2
4
3
CEN4 Core
P2 & P3
P2 & P4
SC5314
P78042
AMS3743
NTC
SC5314
P78042
AMS3743
NTC
4 kb
3 kb
2 kb
1 kb
300 bp

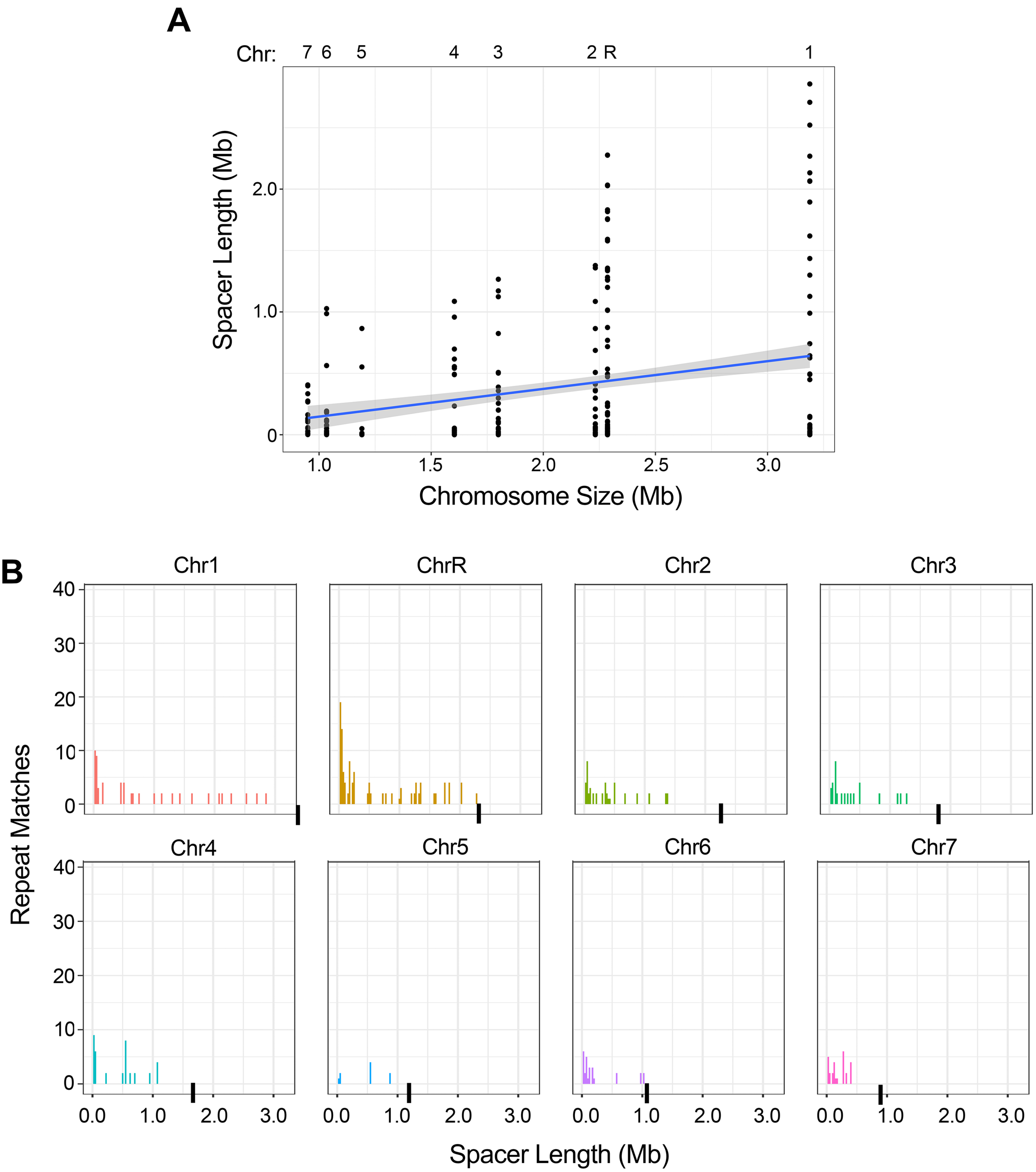
```
>SC5314_CEN4_Reference_PCR_Primer2&3_Sanger_Primer2
TCACAAGTATTCTTCTTCATCATCAATATGGTTTTACTAAAACGGTAATTTACAATAAACCTCAAACGTCTGGAGATATTTCCCAA
ATCGCAAACAAGAATAGCCTCTACCTTCAATTCTGGTCATTTCATCAGTTTAAAATCCAACTCCCACACATCAAAACTGTCAAAGA
GATAGTGACCAATGGAAAATCAATGCAATTAATCCTATAACAAATACCACAGTTCTATGATCAAAAACTCCCAGTTCCAACACAAT
TCCATTCATACCAACCATGCAAAACCTCTGATAGTACAACTAAGAAGAACTCAACGGCACGACTTAAACCCACAACAAAAAGACAA
TTGAATAAATGATCCTCCTGTCAACGACCAACCCTTAATTTGTAACTTCACAAATATCACCAAGAGAAATGTCACCAATATAAATA
GGGACCACCGGAACTCAAGACAAGGCTCACACCGGCCCTATCTCATTTGATTTAGCTCCTATCTCTACCCGCAACCACAGCCAGCT
TGTCTATGCCACACAAAGGGATTCTTTACCTCTCCAAAACGTCTATACCACCGCTGTTAAATTGGTATGTTCATGCGTAATTCTGA
TAACCAAATCAAAACAGTACGCATGCTGGATCTTCTTCCCTTTGGGAAACAACACACGTAGTTTTCTAAATTCCTTCATAACAGCT
ATTTTTTGAAATGTTTCCAAGGTCAATTCTCTTTCTTCTAGCCAAGTGCGGCCACCAAATTACGGGTATCATCATTGACTAGCCTA
CTCTCTAAGCTAGAATTAGCTGAAGTTATACAACCACCTCTATGATCTTAGGACCACAGTAACACAAAAATGCAATCATCATTATC
TAAATGGCAACAGAGTGACAACTCATATAATCTAATACTTCTTTCCCTATATCT

>SC5314_CEN4_Reference_Primer2&3_Sanger_Primer3
TCATAATGTTTTGTAACGCCATGAGAAATCAGGGATATGACTTTCTCATACATTATTTTTAATAAGTCAAACTGTGTGATCGTTGT
GGTTTGATTTTTACAATGGGCATTTCTGATGCAAGATCTTCAAGATCTTCAAGAAAAATATCAAAACTGACTAGATCAATGAGTAT
ACCATTGGCAGAAACAGATACTTTCCTAATAACTTAGGTCATTATCTAGTTGTGCTAGTTAAATCAAAAATTTCTTGAACACATTT
AGTCAGTACTAAGTTTAAGGAAGCGTATGGTGTAGTAACATTATTCAATTGGCGATTAGGATTTGATATCACTGATGGTGTAGGTG
AGCCTAGGCTTAATTTGATCTTTATGGAAATTACAGTATTTGGGACGATATTTTGAACTCCCTAATAATAGTAGATGGTAGAGTTC
CTCCTGCACTAGATATACTGTAGTGGACAGAATTGCCCAAGATTGTATGAAGTACTGAATTTCTGTGATTGCTTGCTTAGAACTGC
CAATTGTTCTGTGGCTCTTGATATACTGGTAGAAAAGTTAATTTTTAATGAGCGGCTTGCAGAATGTATTTAGTTTGGTCCCAAAG
AAGCTCTTTCGGTTGCCTGGCTAATGGGGATTCAGGTGGCAAAGGTTTCAATTTACAGCATATTGATTCGTATGACAATTAATATG
TAATGTTTATAATCGACGTGAAGATACTAGGTCTCAGTGAGTTGTTGTTTGNNNTATAGTATTGGTTGGTATAGTTTAGTTAGAGG
CTTGGATCAGCAGCAATACGTTGATAATTTTTTTCAAATTTGATTTTCTTCTACGAAGGGTAACAGGTTTTCAAATTGTGAAGTAT
CCTGGAATTACAGTAGAGTAGGATTACCTAATCGACGTGAA

>SC5314_CEN4_Inversion_Primer2&4_Sanger_Primer2
CGTCCTCACAAGTATTCTTCTTCNTCATCAATATGGTTTTACTAAAACGGTAATTTACAATAAACCTCAAACGTCTGGAGATATTT
CCCAAATCGCAAACAAGAATAGCCTCTACCTTCAATTCTGGTCATTTCATCAGTTTAAAATCCAACTCCCACACATCAAAACTGTC
AAAGAGATAGTGACCAATGGAAAATCAATGCAATTAATCCTATAACAAATACCACAGTTCTATGATCAAAAACTCCCAGTTCCAAC
ACAATTCCATTCATACCAACCATGCAAAACCTCTGATAGTACAACTAAGAAGAACTCAACGGCACGACTTAAACCCACAACAAAAA
GACAATTGAATAAATGATCCTCCTGTCAACGACCAACCCTTAATTTGTAACTTCACAAATATCACCAAGAGAAATGTCACCAATAT
AAATAGGGACCACCGGAACTCAAGACAAGGCTCACACCGGCCCTATCTCATTTGATTTAGCTCCTATCTCTACCCGCAACCACAGC
CAGCTTGTCTATGCCACACAAAGGGATTCTTTACCTCTCCAAAACGTCTATACCACCGCTGTTAAATTGGTATGTTCATGCGTAAT
TCTGATAACCAAATCAAAACAGTACGCATGCTGGATCTTCTTCCCTTTGGGAAACAACACACGTAGTTTTCTAAATTCCTTCATAA
CAGCTATTTTTTGAAATGTTTCCAAGGTCAATTCTCTTTCTTCTAGCCAAGTGCGGCCACCAAATTACGGGTATCATCATTGACTA
GCCTACTCTCTAAGCTAGAATTAGCTGAAGTTATACAACCACCTCTATGATCTTAGGACCACAGTAACACAAAAATGCTGTCAAAC
CTAGCACGCTAGTATGAGTCAACCACTGATAAGATCATATTTGAAATCTACAGTATATACACAACACCATTCAAAACAGCG

>SC5314_CEN4_Inversion_Primer2&4_Sanger_Primer4
AGAATTGGTTGAGTTGAATTTCGTGATTAATTTAGAATGGTAGTGAATTTTGAAATCTTAACTGTGAGGTAAGATATTTATCGCTG
TTTTGAATGGTGTTGTGTATATACTGTAGATTTCAAATATGATCTTATCAGTGGTTGACTCATACTAGCGTGCTAGGTTTGACAGC
ATTTTTGTGTTACTGTGGTCCTAAGATCATAGAGGTGGTTGTATAACTTCAGCTAATTCTAGCTTAGAGAGTAGGCTAGTCAATGA
TGATACCCGTAATTTGGTGGCCGCACTTGGCTAGAAGAAAGAGAATTGACCTTGGAAACATTTCAAAAAATAGCTGTTATGAAGGA
ATTTAGAAAACTACGTGTGTTGTTTCCCAAAGGGAAGAAGATCCAGCATGCGTACTGTTTTGATTTGGTTATCAGAATTACGCATG
AACATACCAATTTAACAGCGGTGGTATAGACGTTTTGGAGAGGTAAAGAATCCCTTTGTGTGGCATAGACAAGCTGGCTGTGGTTG
CGGGTAGAGATAGGAGCTAAATCAAATGAGATAGGGCCGGTGTGAGCCTTGTCTTGAGTTCCGGTGGTCCCTATTTATATTGGTGA
CATTTCTCTTGGTGATATTTGTGAAGTTACAAATTAAGGGTTGGTCGTTGACAGGAGGATCATTTATTCAATTGTCTTTTTGTTGT
GGGTTTAAGTCGTGCCNTTGAGTTCTTCTTAGTTGTACTATCAGAGGTTTTGCATGGTTGGTATGAATGGAATTGTGTTGGAACTG
GGAGTTTTTGATCATAGAACTGTGGTATTTGTTATAGGATTAATTGCATTGA
```

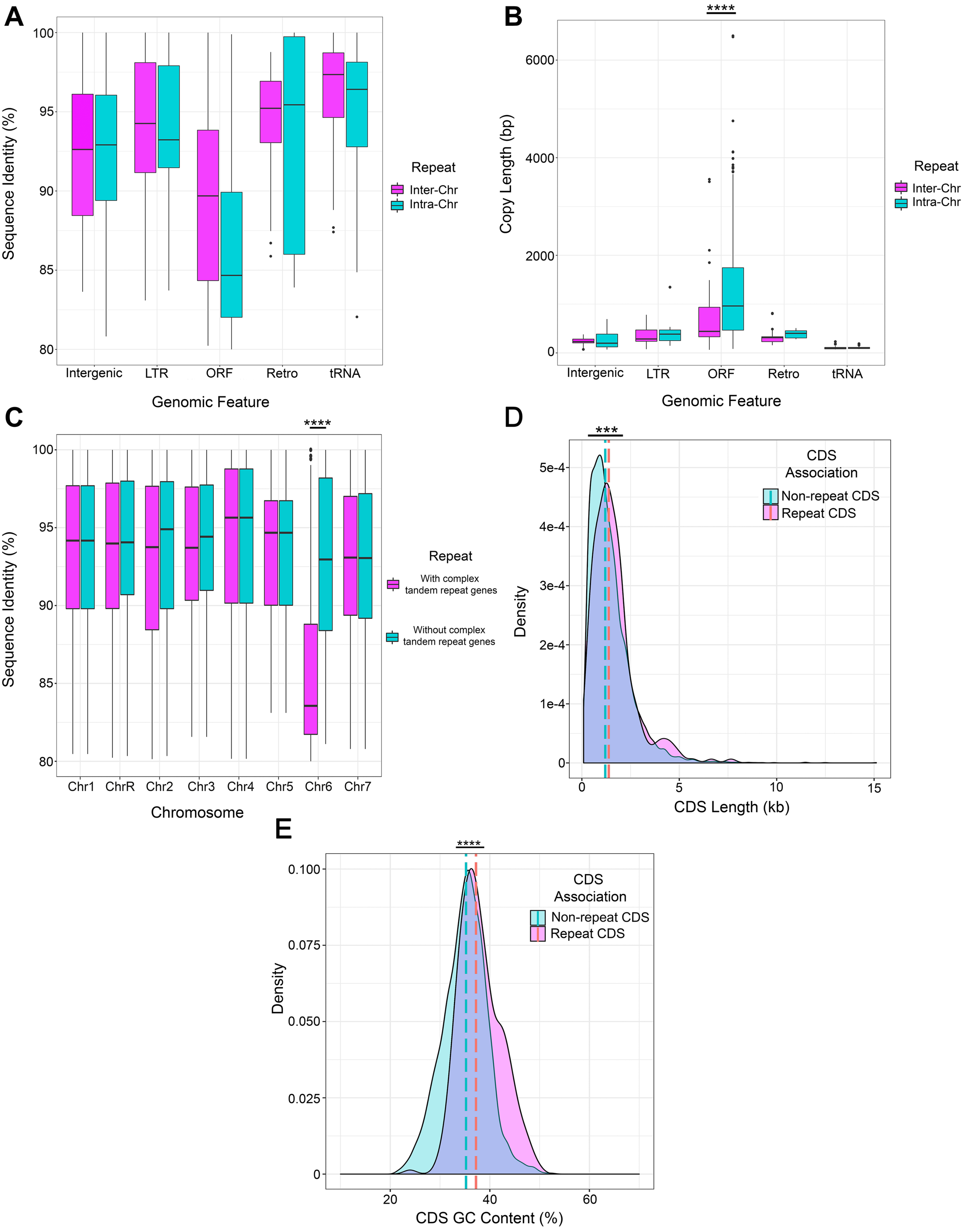

A
Chr1
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
ChrR
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2
Chr2
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2
Chr3
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
Chr4
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
Chr5
0.0 0.2 0.4 0.6 0.8 1.0
Chr6
0.0 0.2 0.4 0.6 0.8 1.0
Chr7
0.0 0.2 0.4 0.6 0.8
B
Repeat Matches
500 400 300 200 100 0
1.0 1.5 2.0 2.5 3.0
Chromosome Size (Mb)
R
2
1
7
4
3
5
6
C
Repeat Matches
800 600 400 200 0
Inter-Chr
Intra-Chr
Intergenic
LTR
ORF
Retro
tRNA
Genomic Feature

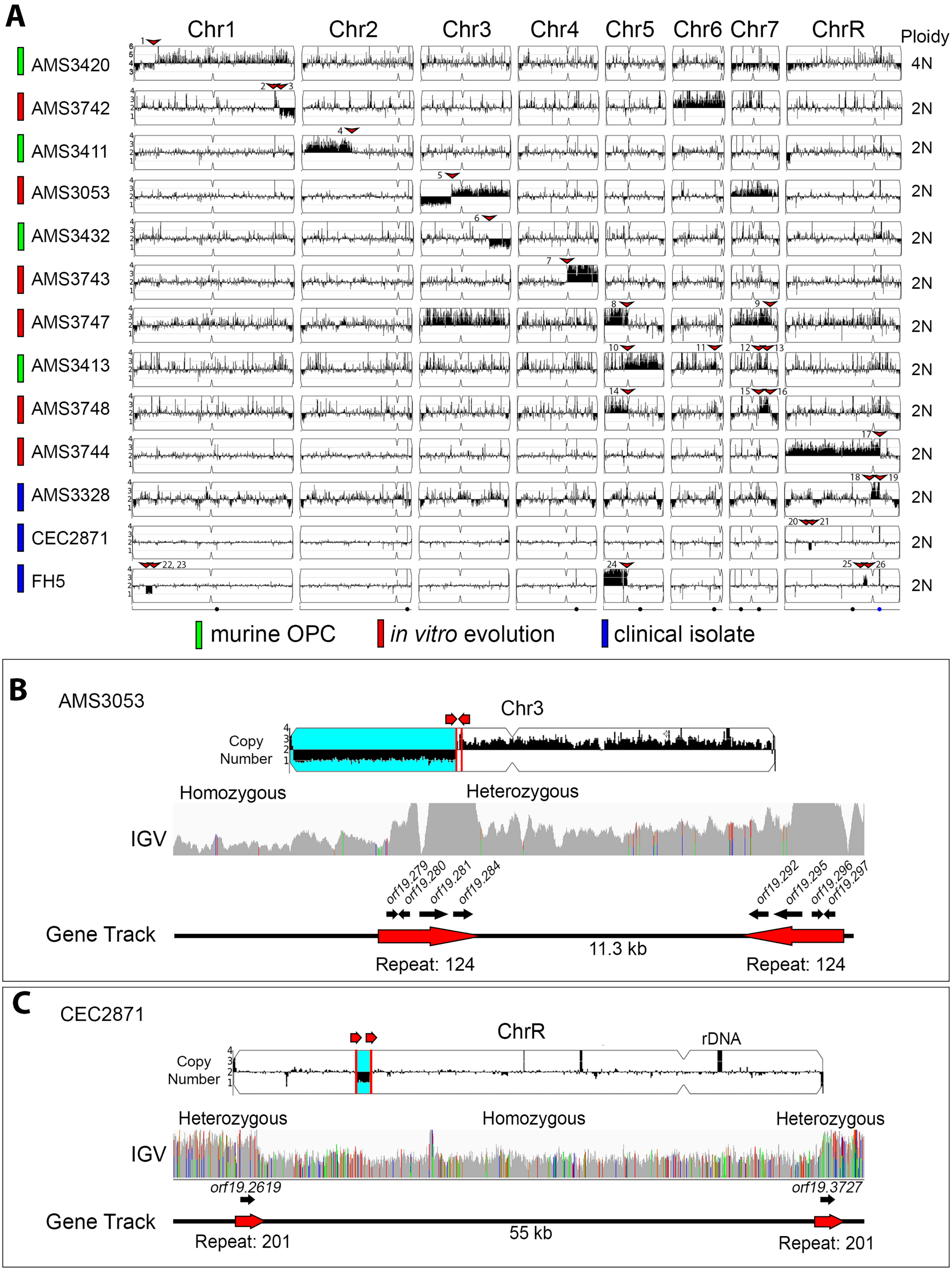
Features of Long Repeat Sequence

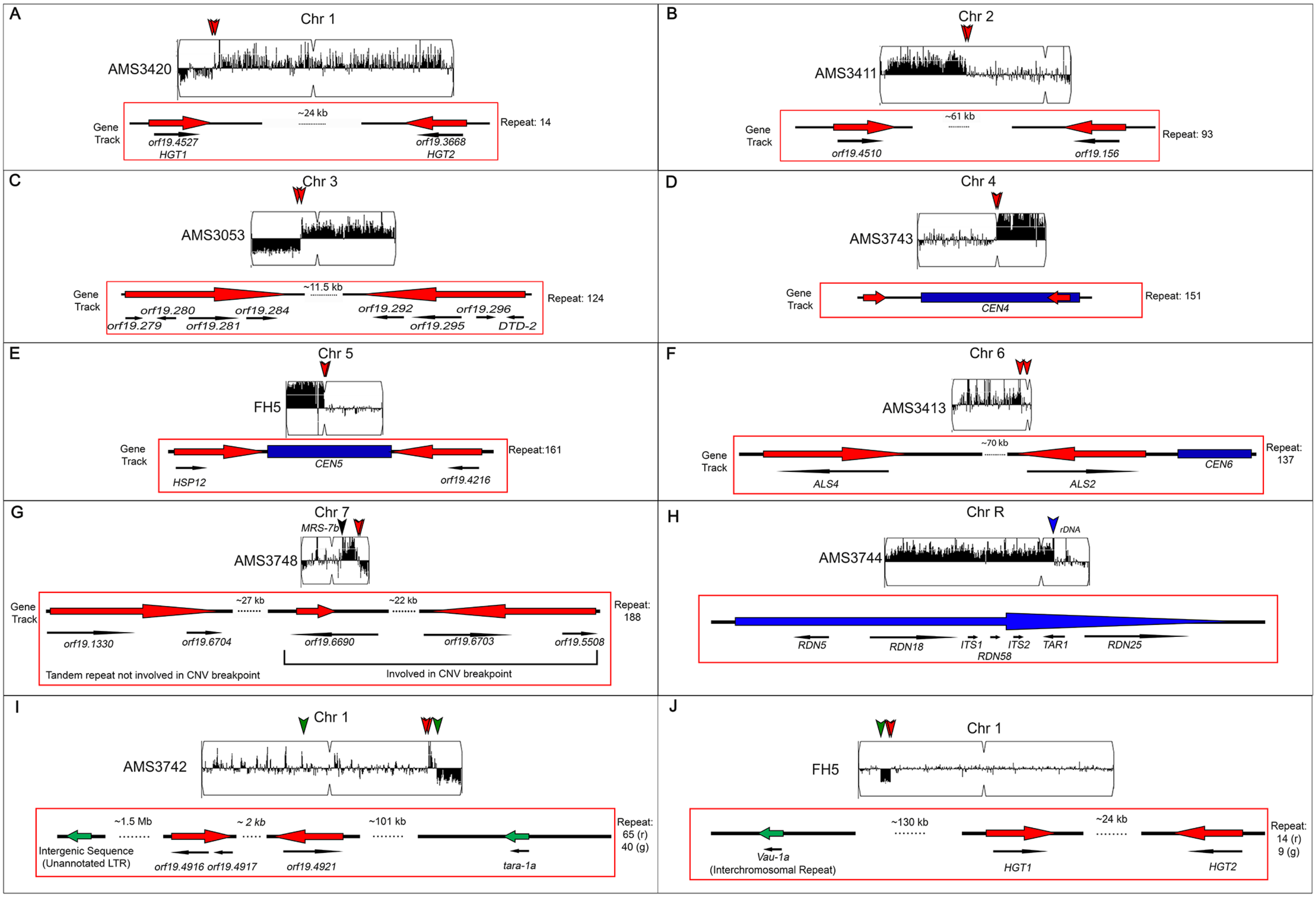
(Intergenic sequence, LTRs, ORFs, Retrotransposons, tRNAs)

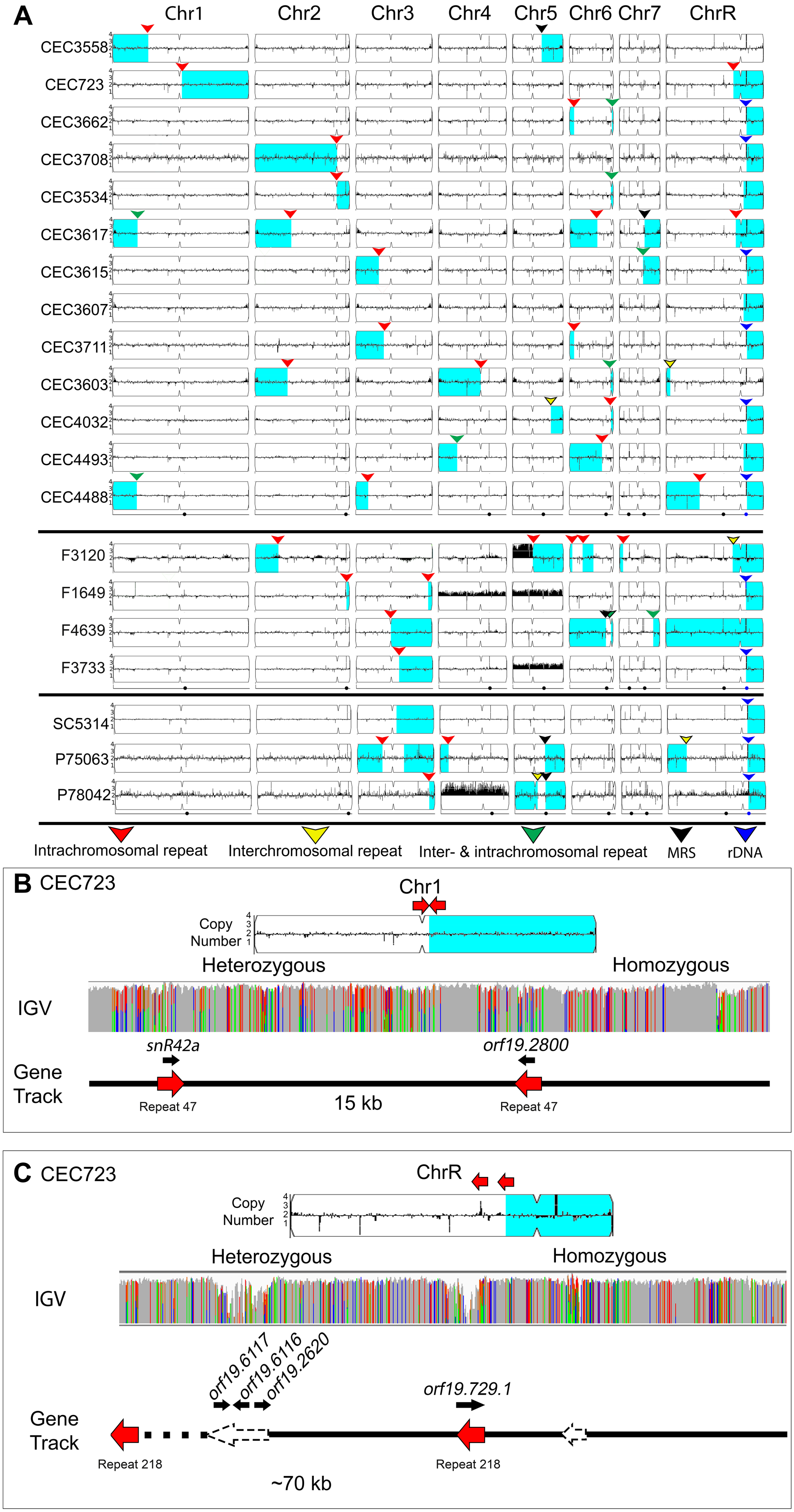


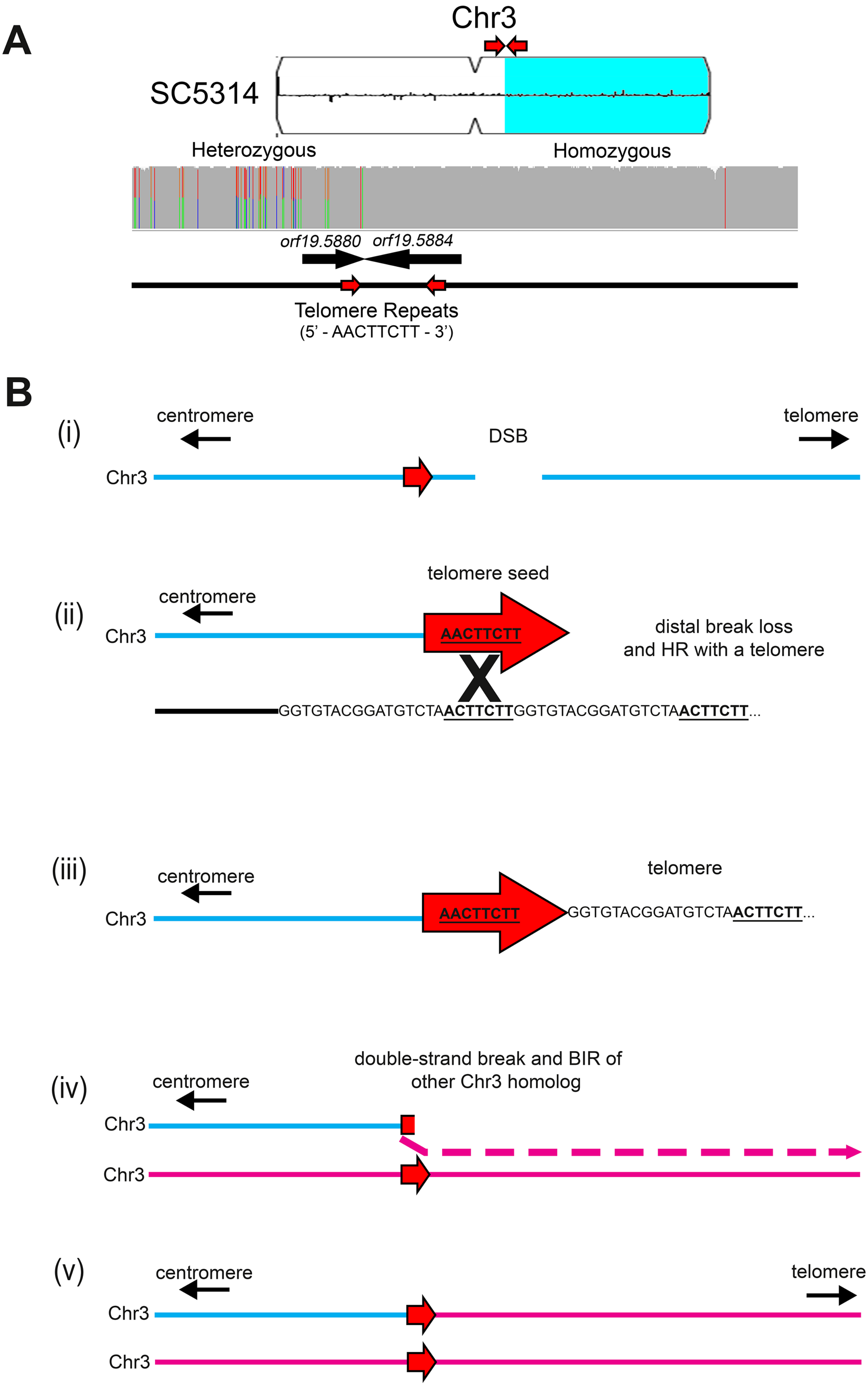

A
Chr:
7
6
5
4
3
2
R
1
Spacer Length (Mb)
2.0
1.0
0
1.0
1.5
2.0
2.5
3.0
Chromosome Size (Mb)
B
Chr1
ChrR
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Repeat Matches
40
30
20
10
0
0.0
1.0
2.0
3.0
Spacer Length (Mb)

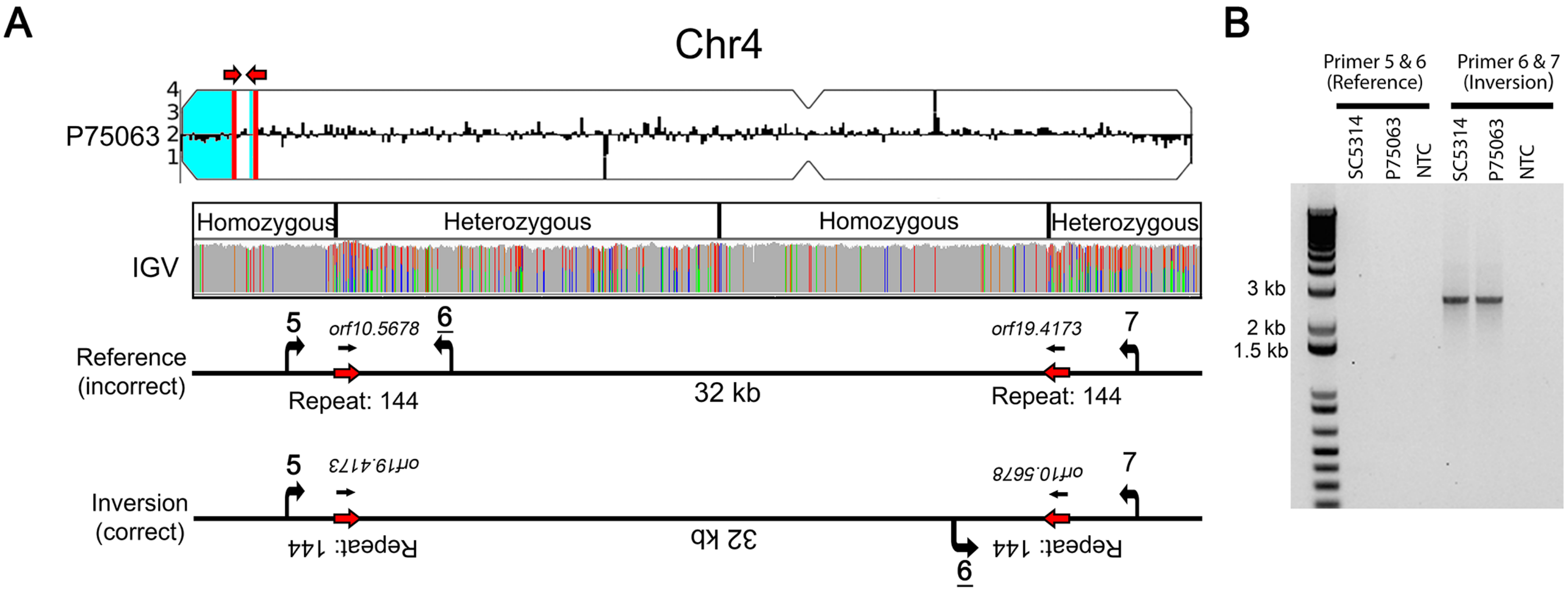

A
Sequence Identity (%)
100
95
90
85
80
Intergenic
LTR
ORF
Retro
tRNA
Genomic Feature
Repeat
Inter-Chr
Intra-Chr
B
****
Copy Length (bp)
6000
4000
2000
0
Intergenic
LTR
ORF
Retro
tRNA
Genomic Feature
Repeat
Inter-Chr
Intra-Chr
C
****
Sequence Identity (%)
100
95
90
85
80
Chr1
ChrR
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chromosome
Repeat
With complex tandem repeat genes
Without complex tandem repeat genes
D
***
Density
5e-4
4e-4
3e-4
2e-4
1e-4
0
0
5
10
15
CDS Length (kb)
CDS Association
Non-repeat CDS
Repeat CDS
E
****
Density
0.100
0.075
0.050
0.025
0
20
40
60
CDS GC Content (%)
CDS Association
Non-repeat CDS
Repeat CDS

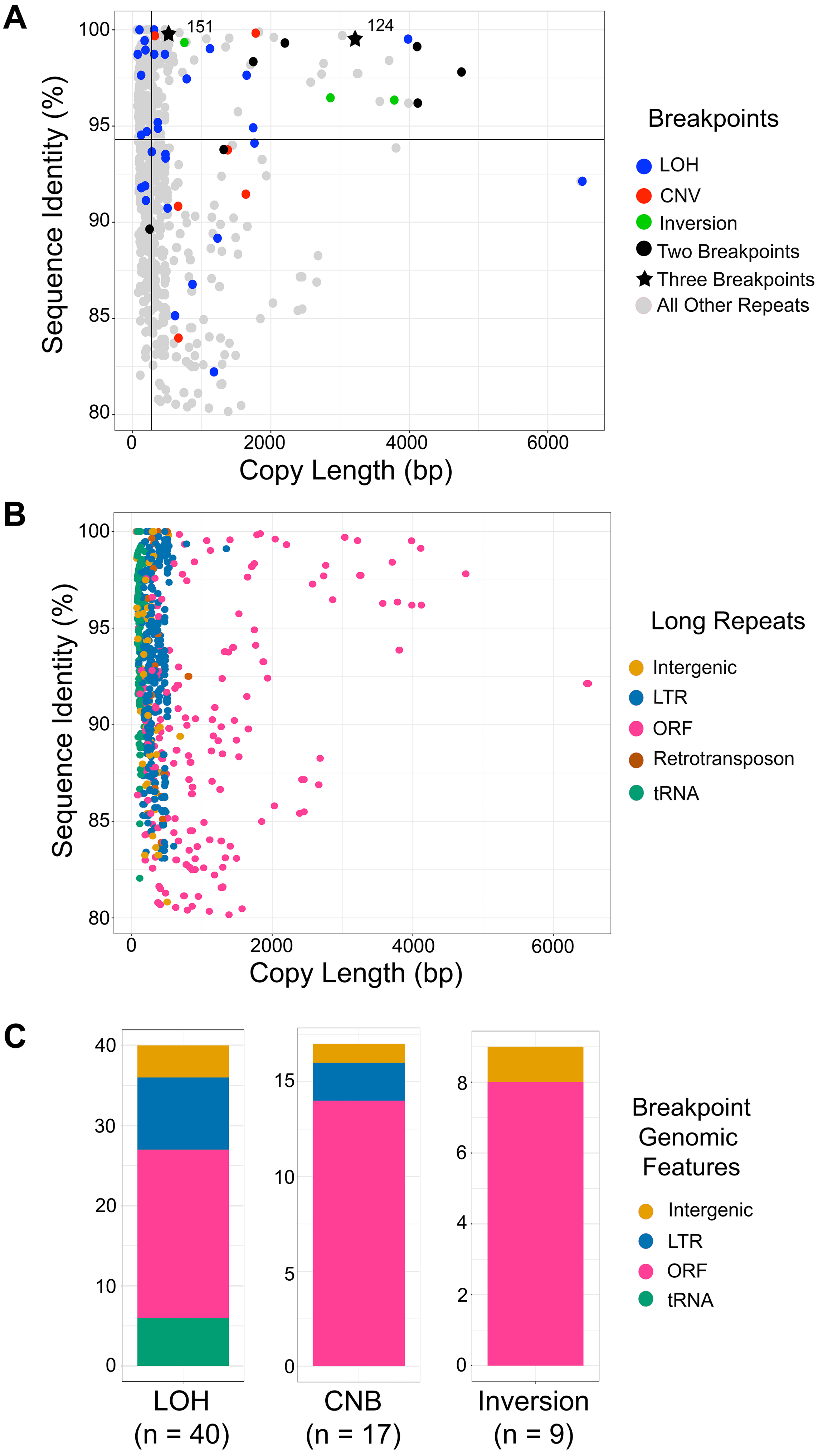

A
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
Ploidy
AMS3420 4N
AMS3742 2N
AMS3411 2N
AMS3053 2N
AMS3432 2N
AMS3743 2N
AMS3747 2N
AMS3413 2N
AMS3748 2N
AMS3744 2N
AMS3328 2N
CEC2871 2N
FH5 2N
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22, 23
24
25
26
murine OPC
in vitro evolution
clinical isolate
B
AMS3053
Chr3
Copy Number
Homozygous
Heterozygous
IGV
orf19.279
orf19.280
orf19.281
orf19.284
orf19.292
orf19.295
orf19.296
orf19.297
Gene Track
11.3 kb
Repeat: 124
Repeat: 124
C
CEC2871
ChrR
rDNA
Copy Number
Heterozygous
Homozygous
Heterozygous
IGV
orf19.2619
orf19.3727
Gene Track
55 kb
Repeat: 201
Repeat: 201

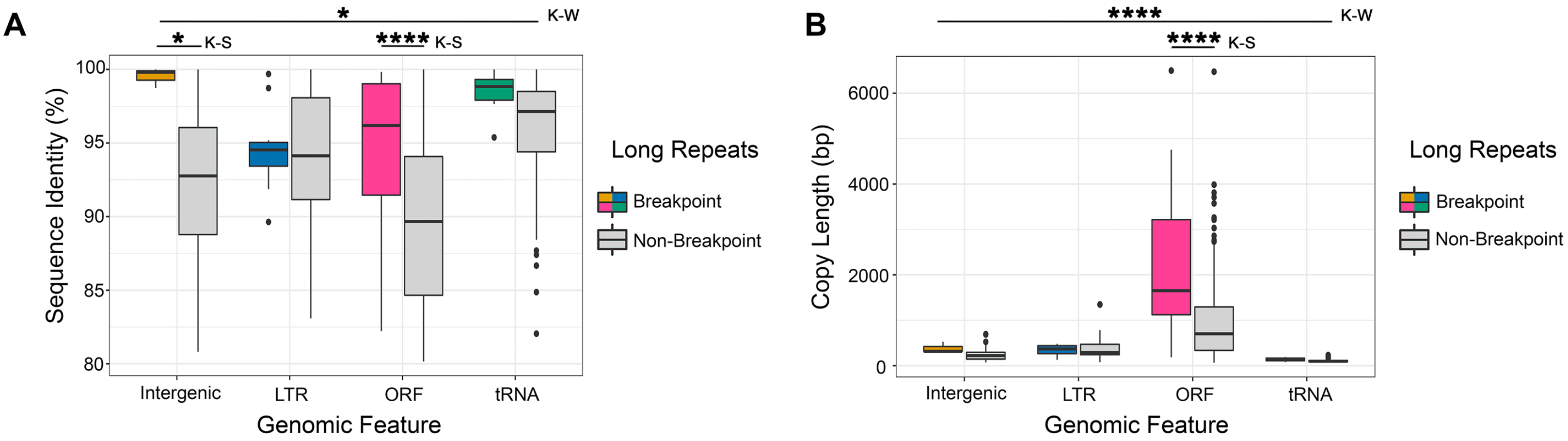

A
Chr 1
AMS3420
Gene Track
~24 kb
orf19.4527
HGT1
orf19.3668
HGT2
Repeat: 14
B
Chr 2
AMS3411
Gene Track
~61 kb
orf19.4510
orf19.156
Repeat: 93
C
Chr 3
AMS3053
Gene Track
~11.5 kb
orf19.280
orf19.284
orf19.279
orf19.281
orf19.292
orf19.296
orf19.295
DTD-2
Repeat: 124
D
Chr 4
AMS3743
Gene Track
CEN4
Repeat: 151
E
Chr 5
FH5
Gene Track
CEN5
HSP12
orf19.4216
Repeat:161
F
Chr 6
AMS3413
Gene Track
~70 kb
CEN6
ALS4
ALS2
Repeat: 137
G
Chr 7
MRS-7b
AMS3748
Gene Track
~27 kb
~22 kb
orf19.1330
orf19.6704
orf19.6690
orf19.6703
orf19.5508
Repeat: 188
Tandem repeat not involved in CNV breakpoint
Involved in CNV breakpoint
H
Chr R
rDNA
AMS3744
RDN5
RDN18
ITS1
ITS2
RDN58
TAR1
RDN25
I
Chr 1
AMS3742
~1.5 Mb
~ 2 kb
~101 kb
Intergenic Sequence (Unannotated LTR)
orf19.4916 orf19.4917
orf19.4921
tara-1a
Repeat: 65 (r) 40 (g)
J
Chr 1
FH5
~130 kb
~24 kb
Vau-1a (Interchromosomal Repeat)
HGT1
HGT2
Repeat: 14 (r) 9 (g)

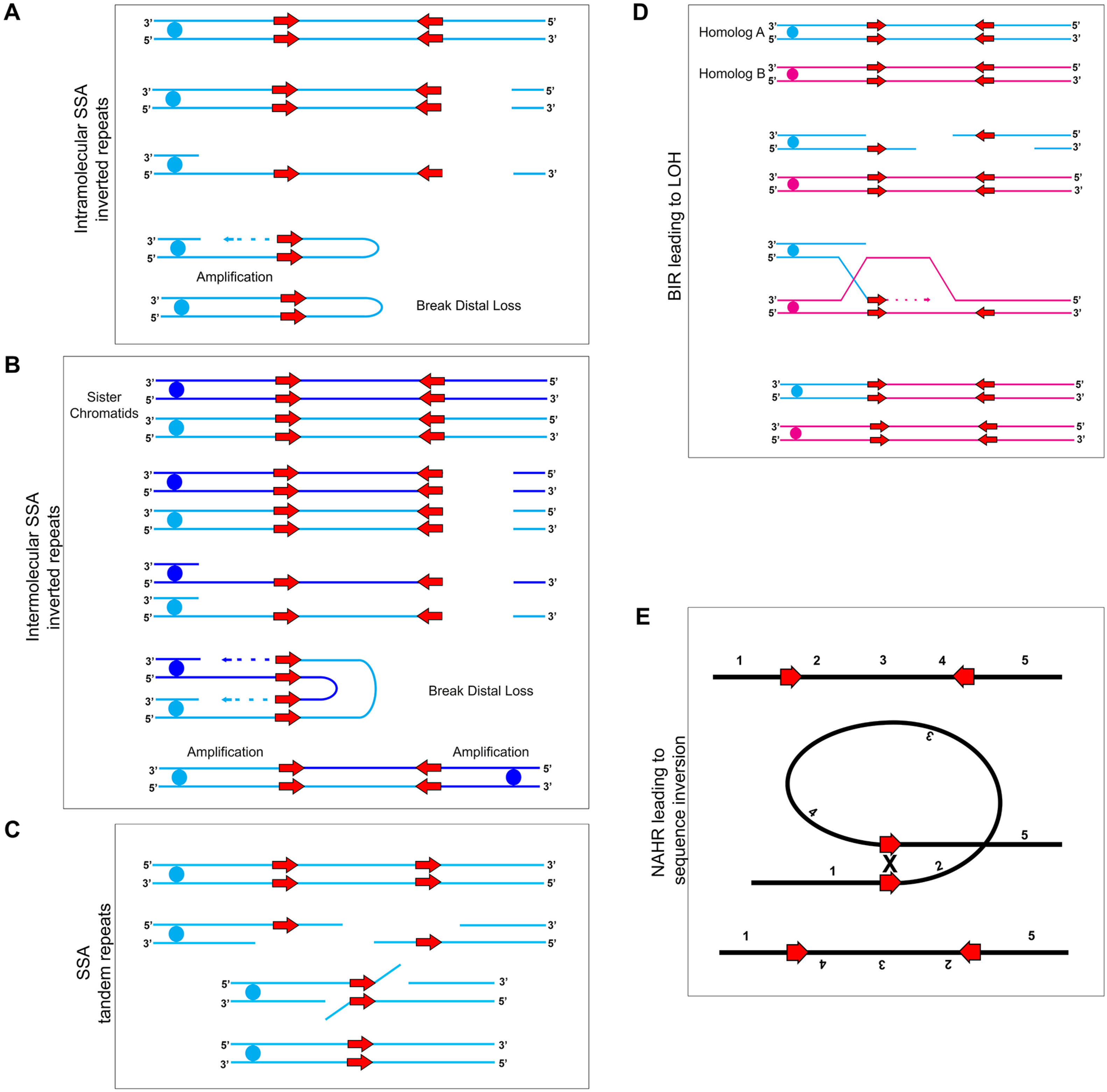

A
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
CEC3558
CEC723
CEC3662
CEC3708
CEC3534
CEC3617
CEC3615
CEC3607
CEC3711
CEC3603
CEC4032
CEC4493
CEC4488
F3120
F1649
F4639
F3733
SC5314
P75063
P78042
Intrachromosomal repeat
Interchromosomal repeat
Inter- & intrachromosomal repeat
MRS
rDNA
B
CEC723
Chr1
Copy Number
4
3
2
1
Heterozygous
Homozygous
IGV
snR42a
orf19.2800
Gene Track
Repeat 47
15 kb
Repeat 47
C
CEC723
ChrR
Copy Number
4
3
2
1
Heterozygous
Homozygous
IGV
orf19.6117
orf19.6116
orf19.2620
orf19.729.1
Gene Track
Repeat 218
~70 kb
Repeat 218


A
Chr3
SC5314
Heterozygous
Homozygous
orf19.5880
orf19.5884
Telomere Repeats
(5' - AACTTCTT - 3')
B
(i)
centromere
telomere
DSB
Chr3
(ii)
telomere seed
centromere
Chr3
AACTTCTT
distal break loss
and HR with a telomere
GGTGTACGGATGTCTAAACTTCTTGGTGTACGGATGTCTAAACTTCTT...
(iii)
centromere
telomere
Chr3
AACTTCTT
GGTGTACGGATGTCTAAACTTCTT...
(iv)
double-strand break and BIR of
other Chr3 homolog
centromere
Chr3
Chr3
(v)
centromere
telomere
Chr3
Chr3


A
Chr4
P75063
4
3
2
1
Homozygous
Heterozygous
Homozygous
Heterozygous
IGV
Reference
(incorrect)
5
orf10.5678
6
orf19.4173
7
Repeat: 144
32 kb
Repeat: 144
Inversion
(correct)
5
orf19.4173
orf10.5678
7
Repeat: 144
32 kb
6
Repeat: 144
B
Primer 5 & 6
(Reference)
Primer 6 & 7
(Inversion)
SC5314
P75063
NTC
SC5314
P75063
NTC
3 kb
2 kb
1.5 kb


A
Sequence Identity (%)
100
95
90
85
80
151
124
Copy Length (bp)
0
2000
4000
6000
Breakpoints
● LOH
● CNV
● Inversion
● Two Breakpoints
★ Three Breakpoints
● All Other Repeats
B
Sequence Identity (%)
Copy Length (bp)
Long Repeats
● Intergenic
● LTR
● ORF
● Retrotransposon
● tRNA
C
40
30
20
10
0
15
10
5
8
6
4
2
LOH
(n = 40)
CNB
(n = 17)
Inversion
(n = 9)
Breakpoint
Genomic
Features
● Intergenic
● LTR
● ORF
● tRNA


A
*
K-W
*
K-S
****
K-S
Sequence Identity (%)
100
95
90
85
80
Intergenic
LTR
ORF
tRNA
Genomic Feature
Long Repeats
Breakpoint
Non-Breakpoint
B
****
K-W
****
K-S
Copy Length (bp)
6000
4000
2000
0
Intergenic
LTR
ORF
tRNA
Genomic Feature
Long Repeats
Breakpoint
Non-Breakpoint


A
Intramolecular SSA
inverted repeats
3'
5'
Amplification
Break Distal Loss
B
Intermolecular SSA
inverted repeats
Sister
Chromatids
Break Distal Loss
Amplification
Amplification
C
SSA
tandem repeats
D
BIR leading to LOH
Homolog A
Homolog B
E
NAHR leading to
sequence inversion
1
2
3
4
5
X
1
5
4
3
2