Spatial proximity of homologous centromere DNA sequences facilitated karyotype diversity and seeding of evolutionary new centromeres

Krishnendu Guin[1], Yao Chen[2], Radha Mishra[1], Siti Rawaidah B. M. Muzaki[2], Bhagya C. Thimmappa[1], Caoimhe O'Brien[3], Geraldine Butler[3], Amartya Sanyal[2*], Kaustuv Sanyal[1*]

[1] Molecular Mycology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore-560064, India; [2] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551; [3] The Butler Laboratory, School Of Biomolecular & Biomed Science, Conway Institute of Biomolecular and Biomedical Research, University of Dublin, Belfield, Dublin 4, Ireland.

*corresponding author

Kaustuv Sanyal, Ph.D.
Molecular Biology & Genetics Unit
Jawaharlal Nehru Centre for Advanced Scientific Research
Jakkur, Bangalore – 560064
India
Email: sanyal@jncasr.ac.in
Telephone : +91 80 2208 2878
Fax : +91 80 2208 2766

Amartya Sanyal, Ph.D.
Nanyang Assistant Professor
School of Biological Sciences
Nanyang Technological University
60 Nanyang Drive, SBS-05n-22
Singapore 637551
Email: asanyal@ntu.edu.sg
Telephone: (+65) 6513-8270

Present address:
Bhagya C. Thimmappa, Department of Biochemistry, Robert-Cedergren Centre of Bioinformatics and Genomics, University of Montreal, Montreal, Canada.
Radha Mishra, Department of Cellular & Molecular Medicine, University of Ottawa, ON, Canada, K1H 8M5

Classification:
Biological Science, Genetics

Keywords:
Genome assembly, 3D-genome, 3C-seq, CUG-Ser1 clade, Evolutionarily new centromere, Chromosome segregation

Abstract

Aneuploidy is associated with drug resistance in fungal pathogens. In tropical countries, *Candida tropicalis* is the most frequently isolated *Candida* species from patients. To facilitate the study of genomic rearrangements in *C. tropicalis*, we assembled its genome in seven gapless chromosomes by combining next-generation sequencing (NGS) technologies with chromosome conformation capture sequencing (3C-seq). Our 3C-seq data revealed interchromosomal centromeric and telomeric interactions in *C. tropicalis*, similar to a closely related fungal pathogen *Candida albicans*. By performing a genome-wide synteny analysis between *C. tropicalis* and *C. albicans*, we identified 39 interchromosomal synteny breakpoints (ICSBs), which are relics of ancient translocations. Majority of ICSBs are mapped within 100 kb of homogenized inverted repeat-associated (HIR) centromeres (17/39) or telomere-proximal regions (7/39) in *C. tropicalis*. Further, we developed a genome assembly of *Candida sojae* and used the available genome assembly of *Candida viswanathii,* two closely related species of *C. tropicalis,* to identify the putative centromeres. In both species, we identified the putative centromeres as HIR-associated loci, syntenic to the centromeres of *C. tropicalis*. Strikingly, a centromere-specific motif is conserved in these three species. Presence of similar HIR-associated putative centromeres in early-diverging *Candida parapsilosis* indicated that the ancestral CUG-Ser1 clade species possessed HIR-associated centromeres. We propose that homology and spatial proximity-aided translocations among the ancestral centromeres and loss of HIR-associated centromere DNA sequences led to the emergence of evolutionary new centromeres (ENCs) on unique DNA sequences. These events might have facilitated karyotype evolution and centromere-type transition in closely-related CUG-Ser1 clade species.

Significance Statement

We assembled the genome of *Candida tropicalis*, a frequently isolated fungal pathogen from patients in tropical countries, in seven complete chromosomes. Comparative analysis of the CUG-Ser1 clade members suggests chromosomal rearrangements are mediated by homogenized inverted repeat (HIR)-associated centromeres present in close proximity in the nucleus as revealed by chromosome conformation capture. These translocation events facilitated loss of ancestral HIR-

associated centromeres and seeding of evolutionary new centromeres on unique DNA sequences. Such karyotypic rearrangements can be a major source of genetic variability in the otherwise majorly clonally propagated human fungal pathogens of the CUG-Ser1 clade. The improved genome assembly will facilitate studies related to aneuploidy-induced drug resistance in *C. tropicalis*.

Introduction

The efficient maintenance of the genetic material and its propagation to subsequent generations determine the fitness of an organism. Genomic rearrangements are often associated with the development of multiple diseases including cancer. Multiple classes of clustered genomic rearrangements, collectively referred to as chromothripsis, are associated with cancer (1). Similarly, structural rearrangements in the genome are often observed during speciation (2). Such structural changes begin with the formation of at least one DNA double-stranded break (DSB), which is generally repaired by homologous recombination (HR) or non-homologous end joining (NHEJ) *in vivo*. Studies using engineered *in vivo* model systems showed that the success of the DSB repair through the HR pathway depends upon efficient identification of the template donor. This process of 'homology search' is facilitated by the physical proximity and the extent of DNA sequence homology (3-5). Multi-invasion-induced rearrangements (MIR) involving more than one template donor has recently been shown to be influenced by physical proximity and homology (6). Therefore, the outcome of the genomic rearrangements is largely dependent on the nature of the spatial genome organization. In yeasts, apicomplexans, and certain plants, centromeres cluster inside the nucleus (7), which may facilitate translocations between two chromosomes through their pericentromeric loci.

The centromere, one of the guardians of the genome stability, assembles a large DNA-protein complex to form the kinetochore, which ensures fidelity of chromosome segregation by correctly attaching every chromosome to the spindle machinery. Paradoxically, this conserved process of centromere function is carried out by highly diverged species-specific centromere DNA sequences. For example, the length of the functional centromere DNA is ~125 bp in budding yeast *S. cerevisiae* (8), but it can be as long as a few megabases in humans (9). The only

factor that remains common to most fungal centromeres is the presence of histone H3 variant CENP-A^{Cse4} except for in *Mucor circinelloides* (10). Most of the kinetochore proteins evolved from pre-eukaryotic lineages and remained conserved within closely-related species complex or expanded through gene duplication (11-13). It remains a long-standing paradox that the underlying centromere DNA sequences keep evolving so fast while the kinetochore structure remains relatively well conserved (14). Therefore, an understanding of the evolutionary processes driving species-specific changes in centromere DNA sequences is essential for a better understanding of the centromere biology.

The first centromere was discovered in *S. cerevisiae,* which carries conserved genetic elements capable of activating *de novo* centromere function when cloned into a yeast replicative plasmid (8). DNA sequence-dependent regulation of centromere function is also identified in *Schizosaccharomyces pombe,* where the centromeres inverted repeat-associated structures of 40-100 kb (15). Other closely related *Saccharomyces* and *Schizosaccharomyces* species were also identified to harbor a DNA sequence-dependent regulation of centromere function (16-18). Although the DNA sequence-dependent mechanism for centromere function is present in certain organisms, the advantage of having such regulation is not well understood. In fact, the majority of the species with known centromeres are thought to be regulated in a non-DNA sequence-dependent mechanism (14). The first epigenetically-regulated fungal centromere carrying 3-5 kb long CENP-A^{Cse4}-bound unique DNA sequences were identified in *C. albicans* (19), a CUG-Ser1 clade species in the fungal phylum of Ascomycota. Subsequently, unique centromeres were identified in closely related *Candida dubliniensis* (20) and *Candida lusitaniae* (21). Strikingly, all seven centromeres of another CUG-Ser1 clade species *C. tropicalis,* carry 3-4 kb long inverted repeats (IR) flanking ~3 kb long CENP-A^{Cse4} rich central core (CC) and their DNA sequences are highly identical to each other. Intriguingly, the centromere DNA of *C. tropicalis* can facilitate *de novo* recruitment of CENP-A^{Cse4} (22). In contrast, the centromeres of *C. albicans* lack such a DNA sequence-dependent mechanism facilitating *de novo* CENP-A^{Cse4} recruitment (23). Such a rapid transition in the structural and functional properties of the centromeres within two closely related species offers a unique opportunity to study the process of centromere-type transition.

Our previous analysis suggested that centromeres of *C. tropicalis* are located near inter chromosomal synteny breakpoints (ICSBs), which are relics of ancient translocations in the common ancestor of *C. tropicalis* and *C. albicans* (22). Additionally, the subcellular localization of the kinetochore proteins as a single punctum per nucleus indicated the clustering of centromeres in *C. tropicalis* (22). However, due to the nature of the then-available fragmented genome assembly, the genome-wide distribution of the ICSBs and the spatial organization of the genome in *C. tropicalis* remained unknown. Therefore, the influence of the spatial proximity on the outcome of the translocations near the centromeres guiding the karyotype evolution in the CUG-Ser1 clade remains as a hypothesis to be tested.

In this study, we constructed a chromosome-level gapless genome assembly of the *C. tropicalis* type strain MYA-3404 by combining information from previously available contigs, NGS reads and high-throughput 3C-seq data. Using this assembly and 3C-seq data, we studied the spatial genome organization in *C. tropicalis.* Next, we mapped the ICSBs in *C. tropicalis* genome with reference to *C. albicans* (ASM18296v3) to ask if the frequency of translocations correlates with the spatial genome organization. In addition, we performed Oxford Nanopore and Illumina sequencing and assembled the genome of *Candida sojae,* a sister species of *C. tropicalis* in the CUG-Ser1 clade (24). Finally, we used our genome assembly of *C. sojae* and publicly available genome assembly of *C. viswanathii* (ASM332773v1) and identified the putative centromeres of these two species as HIR-associated loci syntenic to the centromeres of *C. tropicalis*. Based on our results, we propose a model suggesting homology and proximity guided centromere proximal translocations facilitated karyotype evolution and possibly aided in rapid transition from HIR-associated to unique centromere types in the members of CUG-Ser1 clade.

Results

A chromosome-level gapless assembly of *C. tropicalis* genome in seven chromosomes

C. tropicalis has seven pairs of chromosomes (22, 25), but the current publicly available genome assembly (ASM633v3) has 23 nuclear contigs. To completely

assemble the nuclear genome of *C. tropicalis* in seven chromosomes, we combined short-read Illumina sequencing, and long-read single molecule real-time sequencing (SMRT-seq) approaches together with high-throughput 3C-seq (simplified Hi-C) experiment (**Figure 1A, S1A-D**). We started from the publicly available genome assembly of *C. tropicalis* strain MYA-3404 in 23 nuclear contigs (ASM633v3, Assembly A) (25) and used Illumina sequence reads to scaffold them into 16 contigs to get Assembly B (**Figure 1A**). Next, we used the SMRT-seq long reads to join these contigs, which resulted in an assembly of 12 contigs (Assembly C, **Table S1**). Based on the contour clamped homogenized electric field (CHEF)-gel karyotyping (**Figure 1B**) and 3C-seq data (**Figure S1E-G**), we joined two contigs and rectified a misjoin in Assembly C to produce an assembly of seven chromosomes and five short orphan haplotigs (OHs). Based on our analysis of the *de novo* contigs (**Figure S1H,** Methods), sequence coverage data (**Figure S2A-B**), and Southern blot analysis of the engineered aneuploid strains, we demonstrate that the small orphan contigs fall in heterozygous regions of the genome (**Figure S2C-G,** Methods). Next, we used the *de novo* contigs to fill pre-existing 104 N-gaps and scaffold 14 sub-telomeres (**Figure S3A-C, Table S2**). Finally, we used the 3C-seq reads to polish the complete genome assembly of *C. tropicalis* constituting of 14,609,527 bp in seven telomere-to-telomere long gapless chromosomes (**Figure 1B**). We call this new assembly as Assembly2020.

We then named the chromosomes in the order of their length from chromosome 1 (Chr1) through chromosome 6 (Chr6), and the chromosome containing rDNA locus is named as chromosome R (ChrR) (**Figure 1C**). Accordingly, the centromere on each chromosome is named after the respective chromosome number. Additionally, we assembled the genome sequence of each chromosome in a way to consistently maintain the short arm of chromosomes at the 5′ end. The statistics of the intermediate and final genome assemblies are summarized in **Table S3**. In Assembly2020, 1278 out of 1315, Ascomycota-specific BUSCO gene sets could be identified compared to 1255 identified using Assembly A (**Table S4**, Methods). Inclusion of 23 additional BUSCO gene sets as compared to the Assembly A suggests improved contiguity and completeness of Assembly2020.

Previously, using centromere-proximal probes, we could distinctly identify five chromosomes (Chr1, Chr2, Chr3, Chr5, Chr6) in chromoblot analysis (22). However, the length of Chr4 as well as ChrR remaind unknown. To validate the correct assembly of these two chromosomes (Chr4 and ChrR), we performed chromoblot analysis. We observed that the Chr4 homologs differ in size (**Figure S4A**). Analysis of the sequence coverage across Chr4 identified an internal duplication of ~235 kb region, which explains the size difference between the homologs Chr4A and Chr4B (**Figure 1C, S4B**). We named this duplicated locus as *DUP4.* Subsequently, we scanned the entire genome for the presence of copy number variations (CNVs), which led to the identification of two additional large duplication events: one each on Chr5 (*DUP5*, ~23 kb) and ChrR (*DUPR*, ~80 kb) (**Figure 1C, S4B**). Additionally, we detected a balanced heterozygous translocation between Chr1 and Chr4 (**Figure S4C**) through analyses of 3C-seq data and the *de novo* contigs (**Figure S4D**). This translocation was validated using chromoblot analysis (**Figure S4E**), Illumina and SMRT-seq read mapping (**Figure S4F**). A chromoblot analysis for ChrR revealed that the actual length of ChrR is ~2.8 Mb, while the assembled length is 2.1 Mb (**Figure 1C, S4G**). Considering the length of rDNA locus is ~700 kb in *C. albicans* (26), we reason that the difference between the assembled length and actual length (derived from the chromoblot analysis) of ChrR in *C. tropicalis* can be attributed due to the presence of the repetitive rDNA of ~700 kb, which is not completely assembled in Assembly2020.

Next, we performed phasing of the diploid genome of *C. tropicalis* using our SMRT-seq, and 3C-seq data to identify the homolog-specific variations (Methods). This analysis produced 16 nuclear contigs, which were colinear with the chromosomes of Assembly2020, except for the previously validated heterozygous translocation between Chr1 and Chr4 (**Figure S4H**). In order to characterize the sequence variations in the diploid genome of *C. tropicalis,* we identified the single nucleotide polymorphisms (SNPs) and insertions-deletions (indels) (Methods). Intriguingly, we detected a long chromosomal region depleted of SNPs and indels on the left arm of ChrR (**Figure 1D**). We refer to this region with loss of heterozygosity on ChrR as LOH^R. Strikingly, we found parts of the syntenic regions of LOH^R to be SNP and indel depleted in the *Candida sojae* strain NCYC-2607, a closely related species of *C. tropicalis*, as well as in *C. albicans* reference strain SC5314 (**Figure**

S5). We also identified the genome-wide distribution of transposons and simple repeats but could not detect preferential enrichment of these sequence elements at any specific genomic location in *C. tropicalis* (**Figure 1D**). Together, we identified multiple long CNVs, long-track LOH, and heterozygous translocation events in the diploid genome of *C. tropicalis.* Possible implications of these events in virulence and drug resistance of this successful human fungal pathogen need to be explored.

Conserved principle of spatial genome organization in *C. tropicalis* and *C. albicans*

Indirect immunofluorescence imaging of *C. tropicalis* strain expressing protein-A tagged Cse4 suggests the clustering of the centromere-kinetochore complex, which is localized at the periphery of the DAPI-stained nuclear DNA mass as a single punctum (**Figure 2A-B**). We re-aligned 3C-seq data to the Assembly2020 to generate the genome-wide chromatin contact map of *C. tropicalis*. The resultant heatmap shows high signal intensity along the diagonal indicating that the intra-chromosomal interactions are generally stronger than interchromosomal interactions (**Figure 2C**). However, the most striking feature of the heatmap is the presence of conspicuous puncta in the interchromosomal areas, which signify strong spatial proximity between centromeres (**Figure 2C-D**). Aggregate signal analysis further reiterates the enrichment of centromere-centromere interactions (**Figure 2E**). All these observations suggest the clustering of centromeres and conservation of the Rabl configuration in *C. tropicalis*, a well-known feature of a higher-order genome organization in yeasts (27-29). Strikingly, we also noted enrichment of interactions between telomeres of different chromosomes (**Figure 2E**). These interchromosomal telomeric interactions were significantly greater than the average interchromosomal interaction (Mann-Whitney U test *P* value = 6.547×10^{-7}) (**Figure S6A**). We also observed enhanced *cis* interaction between the two telomeres of an individual chromosome compared to average intra-chromosomal long-range (≥100 kb) interaction (Mann-Whitney U test *P* value = 1.091×10^{-9}) (**Figure S6B**).

Previously, the genomic contacts in *C. albicans* were analyzed by Hi-C, which showed physical interaction among the centromeres (27, 30, 31). Together, our analysis reveals a conserved pattern of centromere clustering in two closely related

fungal species with completely different centromere DNA sequences and structural features. This observation suggests a DNA sequence-independent mechanism for centromere clustering in yeasts. Moreover, our analysis demonstrates the conserved principles of chromosomal organization in two human pathogenic ascomycetes, *C. albicans*, and *C. tropicalis*, despite substantial karyotypic changes during the speciation.

Centromere and telomere proximal loci are hotspots for complex translocations

Using the chromosome-level Assembly2020 of *C. tropicalis* and publicly available chromosome-level assembly of the *C. albicans* reference genome of SC5314 strain (ASM18296v3), we performed a detailed genome-wide synteny analysis following four different approaches. We used two published analysis tools, Symap (32) and Satsuma synteny (33), and a custom approach to identify the ICSBs based on the synteny of the conserved orthologs (**Figure 3A**). Next, we compared and validated the results obtained from our custom approach of analysis with another published tool Synchro (**Figure S7A-B**) (34). All four methods of analysis detected that six out of seven centromeres (except *CEN6*) of *C. tropicalis* are located proximal to multiple ICSBs, which are rare at the chromosomal arms (**Figure 3A**). The ORF-level synteny analysis detected four out of seven centromeres (*CEN2*, *CEN3*, *CEN5*, *CENR*) in *C. tropicalis* to be precisely located at the ICSBs, while multiple ICSBs are located within ~100 kb of other two centromeres (**Figure 3B**). However, no ICSB could be identified on Chr6. Additionally, we found a convergence of orthoblocks from as many as four different chromosomes within 100 kb of centromeres (**Figure 3B**).

To correlate the frequency of translocations with the spatial genome organization, we quantified ICSB density (the number of ICSBs per 100 kb of the genome) at different zones across the chromosome for all chromosomes except for Chr6 (**Figure 3C**). Since no ICSBs were mapped on Chr6, it was excluded from the analysis. This analysis revealed that the ICSB density is the highest at the centromere proximal zones for all six chromosomes, but dropped sharply at the chromosomal arms. However, the ICSB density near the telomere proximal zone for Chr2, Chr4, and ChrR showed an increase over the chromosomal arms, albeit at a

lower magnitude than the centromeres. We also compared the length of the orthoblocks across three different genomic zones- the centromere proximal (within 300 kb from the centromere on both sides), centromere distal (beyond 300 kb from the centromere to 200 kb from the telomeres), and telomere proximal (within 200 kb from the telomeres) zones. This analysis revealed that the length of the orthoblocks located proximal to centromeres and telomeres are significantly smaller compared to the orthoblocks located at the centromere distal zone (**Figure 3D**).

Does this mean there were inter-centromeric translocations in the common ancestor of *C. albicans* and *C. tropicalis*? If such inter-centromeric translocations occurred, then the ORFs present near different centromeres in *C. tropicalis* should converge together on the *C. albicans* genome. Indeed, we found 10 loci where such convergence is observed (**Figure S7C**). Intriguingly, four such loci are proximal to the centromeres (CEN3, CEN4, CEN7, and CENR) in *C. albicans* (**Figure 3E-F, S7D-E**). This observation supports the possibility of inter-centromeric translocation events in the common ancestor of *C. albicans* and *C. tropicalis.* Additionally, the other four centromeres are located proximal to ORFs, homologs of which are also proximal to the centromeres in *C. tropicalis* (**Figure S7C**). Together, these observations posit that the ancestral HIR-associated centromeres are lost in *C. albicans* and evolutionary new centromeres (ENCs) formed proximal to the ancestral centromere loci on unique and different DNA sequences (19).

Rapid transition in the centromere type within the members of the CUG-Ser1 clade

Based on the identification of multiple translocation events concentrated near the centromeric regions of the *C. tropicalis* genome, we hypothesize that complex translocations between HIR-associated centromeres in the common ancestor of *C. albicans* and *C. tropicalis* led to the loss of HIR and evolution of unique centromere types observed in *C. albicans* and *C. dubliniensis.* However, the genomic rearrangements are rare events, even at the evolutionary time scale. Therefore, if the HIR-associated centromeres are the ancestral state, from which the unique centromeres would have derived during a rare chromothripsis-like event, then the other closely related species should have retained HIR-associated centromeres. To gain further insights into the centromere type of the common ancestor of *C. albicans*

and *C. tropicalis,* we scanned for the presence of HIR-like structures in the genomes of *C. parapsilosis,* an early-diverging members of the CUG-Ser1 clade. Indeed, we identified eight HIR-associated structures (**Figure S8A**), present once in each of the eight chromosomes of *C. parapsilosis.* Identification of the HIR-associated structures present at the intergenic and transcription poor regions, once each on all eight chromosomes, suggests that these loci are the putative centromeres of *C. parapsilosis.* This observation indicates that the common ancestor of *C. albicans* and *C. tropicalis* possibly carried HIR-associated centromeres. Next, we performed a genome-wide synteny analysis between *C. orthopsilosis* and *C. parapsilosis* and found evidence of translocations at seven out of eight HIR-associated loci, five of which are ICSB associated (**Figure S8B**). This result indicates the involvement of HIR-associated structures in translocation events, similar to those translocation events involving *C. tropicalis* centromeres.

Such structure-defined HIR-associated centromeres have only been identified in *C. tropicalis* in the CUG-Ser1 clade species (22). Although IRs are present in *CEN4*, *CEN5*, and *CENR* of *C. albicans,* these sequences are not homogenized like the HIR-associated centromeres in *C. tropicalis* (**Figure 4A**). In order to study the presence or absence of HIRs in *C. sojae,* a sister species of *C. tropicalis* (24), we assembled its genome into 42 contigs, including seven chromosome-length contigs (Methods). Using this assembly, we identified seven putative centromeres in *C. sojae* as intergenic and HIR-associated loci syntenic to the centromeres in *C. tropicalis* (**Figure S9A-B, D**). Each of these seven centromeres in *C. sojae* consists of a ~2 kb long central core (CC) region flanked by 3-12 kb long inverted repeats (**Table S5**). Using a similar approach, we identified six HIR-associated centromeres in the publicly available genome assembly (ASM332773v1) of *Candida viswanathii*, another species closely related to *C. tropicalis* (**Figure S9C, E, Table S6**) (35). A dot-plot analysis found extensive homology shared across the IRs but not among the CC elements (**Figure 4A**) of the HIR-associated centromeres present in *C. tropicalis* and the putative centromeres of *C. sojae* and *C. viswanathii* (**Table S7**). Moreover, We detected extensive structural conservation in *CEN* DNA-elements, especially among IRs within an individual species (**Figure S10A**). This structural feature of IRs is also significantly conserved across the three species, *C. tropicalis*, *C. sojae,* and *C. viswanathii,* with HIR-associated centromeres (**Figure S10B**).

Cloning of a full-length centromere of *C. tropicalis* in a replicative plasmid facilitates *de novo* CENP-A deposition but fails when the IRs are replaced with Ca*CEN5* IRs (22). This result indicated the presence of a genetic element specifically on the IRs of *C. tropicalis* but absent in Ca*CEN5* IR. To identify the putative genetic element, we analyzed the *CEN* DNA sequences of all three HIR-associated centromeres and the unique centromeres of *C. albicans* for the presence of conserved motifs. This analysis identified a highly conserved 12 bp motif (dubbed as IR-motif) (**Figure 4B**) clustered specifically at the centromeres but not anywhere else in the entire genome of *C. tropicalis, C. sojae* and *C. viswanathii* (**Figure 4C-D, S10C**). On the contrary, the IR-motif density at the centromeres in *C. albicans* remains approximately an order of magnitude lower than that of *C. tropicalis* (**Figure 4C**). This observation indicates a potential function of the IR-motif in the regulation of *de novo* CENP-A loading in *C. tropicalis*. Moreover, the *CEN*-enriched motif was found to be specifically concentrated on the IRs but not at the mid-core region in HIR-associated centromeres present in *C. tropicalis* (**Figure 4E**) and at the putative centromeres in *C. soaje* and *C. viswanathii* (**S10D**). Additionally, we detected that the direction of the IR-motif is diverging away from the central core of the centromeres in *C. tropicalis* (**Figure S10E**), and this pattern remained conserved in the other two species as well (**Figure S10F**). The conserved structure and organization of the IR-motif sequences in the HIR-associated centromeres of three *Candida* species suggest an inter-species conserved function of the IR DNA sequence among these three species, although the clusters of IR-motifs are located at a variable distance from the CC in these three species (**Figure S10G**). The importance of this 12-bp conserved motif on the centromere function is yet to be determined.

Discussion

In this study, we improved the current genome assembly of the human fungal pathogen *C. tropicalis* by employing SMRT-seq, 3C-seq, and CHEF-chromoblot experiments, and present Assembly2020, the first chromosome-level gapless genome assembly of this organism. We identified three long duplication events in its genome, phased the diploid genome of *C. tropicalis* and mapped the SNPs and

indels. We constructed genome-wide contact maps and identified centromere-centromere as well as telomere-telomere spatial interactions. A comparative genome analysis between *C. albicans* and *C. tropicalis* revealed that six out of seven centromeres of *C. tropicalis* are mapped precisely at or proximal to interchromosomal synteny breakpoints. Strikingly, ORFs proximal to the centromeres of *C. tropicalis* are converged into specific regions on the *C. albicans* genome in some occasions, suggesting possibilities of inter-centromeric translocations in their common ancestor. Moreover, the presence of homogenized inverted repeat associated putative centromeres in *C. sojae* and *C. viswanathii,* like in *C. tropicalis*, suggest that such a centromere structure is plausibly the ancestral form in the CUG-Ser1 clade species complex but lost in *C. albicans* and *C. dubliniensis.* We propose that loss of such a centromere structure possibly happened during translocation events involving centromeres in the common ancestor might have given rise to evolutionary new centromeres on unique DNA sequences and facilitated speciation.

The availability of the chromosome-level genome assembly, and improved annotations of genomic variants and genes absent in the publicly available fragmented genome assembly of *C. tropicalis* should greatly facilitate genome-wide association studies to understand the pathobiology of the organism including the cause of antifungal drug resistance. In addition, this study sheds lights on how primordial mechanisms of *de novo* centromere establishment present in an ancestral species become dispensable in the derived lineages.

C. tropicalis is a human pathogenic ascomycete, closely related to the well-studied model fungal pathogen *C. albicans* (36). These two species diverged from their common ancestor ~39 million years ago (37) and evolved into two distinct karyotypes (22), having different phenotypic traits (38), and ecological niches (39). While *C. albicans* remains the primary cause of candidiasis worldwide, systemic ICU-acquired candidiasis is primarily (30.5-41.6%) caused by *C. tropicalis* in tropical countries including India (40), Pakistan (41), and Brazil (42). Moreover, the occurrence of drug resistance, particularly multidrug resistance, in *C. tropicalis* is on the rise (40, 43, 44). Therefore, relatively less-studied *C. tropicalis* is emerging as a major threat for nosocomial candidemia with 29-72% broad spectrum mortality rate (45). Fluconazole resistance in *C. albicans* can be gained due to segmental

aneuploidy of chromosome 5 containing long IRs at the centromere, by the formation of isochromosomes (46), which is also identified in chromosome 4 with IRs at its centromere (47). All seven centromeres in *C. tropicalis* are associated with long IRs, hence it is possible that each of them can form isochromosomes. Now, with the availability of the chromosome-level assembly of the *C. tropicalis* genome, it should be possible to initiate genome-wide association studies to understand the genomic causes of pathogenicity and the rapid emergence of drug resistance in *C. tropicalis.*

Since the mechanism of homology search during HR is positively influenced by spatial proximity and the extent of DNA sequence homology (4, 48), at least in the engineered model systems, it is expected that spatially clustered homologous DNA sequences experience more translocations than other loci. Although these factors were not shown to be involved in karyotypic rearrangements during speciation, a retrospective survey in light of spatial proximity and homology now offers a better explanation. For example, bipolar to tetrapolar transition of the mating type locus in the *Cryptococcus* species complex was associated with inter-centromeric recombination following pericentric inversion (49). Similar inter-centromeric recombination has been reported in the common ancestor of two fission yeast species, *Schizosaccharomyces cryophilus* and *Schizosaccharomyces octosporus* (18). These examples raise an intriguing notion that centromeres serve as sites of recombination, which may lead to centromere loss and/or emergence of evolutionary new centromeres. This notion is supported by the fact that DNA breaks at the centromere following fusion of the acentric fragments to other chromosomes led to chromosome number reduction in *Ashbya* species (16) and *Malassezia* species (50). Genomic instability at the centromere can also lead to fluconazole resistance, as in the case of isochromosome formation on Chr5 of *C. albicans* (46). Additionally, breaks at the centromeres are reported to be associated with cancers (51).

What would be the conseqence of spatial proximity of chromosomal regions with high DNA sequence homology, observed in fungal systems, in other domains of life? Chromoplexy, a type of chromothripsis, where a series of translocations occur among multiple chromosomes, is associated with cancers (1). Although fine mapping of translocation events at the repetitive regions in human cancer cells becomes difficult, the growing evidence that such events are associated with the formation of

micronuclei (52) supports the idea that spatial genome organization may influence such events (53). With the availability of Hi-C and other techniques to probe genomic contacts in high-resolution, it may now be possible to test whether chromoplexies occur due to close physical proximity of homologous DNA sequences.

The identification of HIR-associated putative centromeres in *C. parapsilosis*, *C. sojae,* and *C. viswanathii* supports the idea that the unique centromeres might have evolved from an ancestral HIR-associated centromere (54) (**Figure 5A**). Such a rapid transition in the structural and functional properties of centromeres is unprecedented. While HIR-associated centromeres of *C. tropicalis, C. sojae,* and *C. viswanathii* form on different DNA sequences, a well-conserved IR-motif was identified in this study that is present in multiple copies on the centromeric IR sequences across these three species. Some centromeres in *C. albicans* carry chromosome-specific IRs, which lack IR-motifs. In addition, *CaCEN5* IRs could not functionally complement the centromere function in *C. tropicalis* for the *de novo* CENP-A^{Cse4} recruitment. This indicates a possible role of the conserved IR-motifs on species-specifc centromere function (22). Therefore, the loss of HIR-associated centromere in *C. albicans* that are only epigenetically propagated (23) clearly shows how ability of *de novo* establishment of kinetochore assembly in an ancestral lineage can be lost in an derived lineage. However, the details of the mechanism through which IR-motifs may regulate centromere identity remains to be explored.

Loss of HIR-associated centromeres during inter-centromeric translocations or MIR must have been catastrophic for the cell, and the survivor needed to activate another centromere at an alternative locus. How is such a location determined? Artificial removal of a native centromere in *C. albicans* leads to the activation of a neocentromere (55, 56), which then becomes part of the centromere cluster (27). This evidence supports the existence of a spatial determinant, known as the CENP-A cloud or CENP-A-rich zone (55, 57), influencing preferential formation of neocentromere at loci proximal to the native centromere (55, 58). We found that the unique and different centromeres of *C. albicans* are located proximal to the ORFs, which are also proximal to the centromeres in *C. tropicalis*. This observation indicates that the formation of the new centromeres in *C. albicans* may have been influenced by spatial proximity to the ancestral *CEN* cluster. However, the new

centromeres of *C. albicans* are formed on loci with completely unique and different DNA sequences. Because of these reasons, it may be logical to consider the centromeres of *C. albicans* as ENCs (**Figure 5B**). Intriguingly, even after the catastrophic chromosomal rearrangements, the ENCs in *C. albicans* remain clustered similar to *C. tropicalis* (**Figure 5C**). This observation identifies spatial clustering of centromeres as a matter of cardinal importance for the fungal genome organization.

Materials and Methods

The strains, primers, and plasmids used in this study are listed in SI Appendix, Tables S8, S9, and S10, respectively. Details of all of the experimental procedures and sequence analysis are given in SI Materials and Methods. All sequencing data used in the study and the genome assembly of *C. tropicalis* and *C. sojae* have been submitted to NCBI under the BioProject accession number PRJNA596050.

Acknowledgments

We thank all the members of KS laboratory and AS laboratory for critical reading of the manuscript and inputs. We acknowledge Dr. Sheng Sun and Dr. Joseph Heitman for helping in SMRT-seq of *C. tropicalis* at the PacBio sequencing facility at Duke University. Illumina sequencing experiment for *C. sojae* genome was performed at Clevergene Biocorp, Bangalore, India. We also thank B. Suma for confocal microscopy, JNCASR. K.G. acknowledges Shyama Prasad Mukherjee Fellowship from Council of Scientific and Industrial Research (CSIR), Govt. of India [07/733(0181)/2013-EMR-I and financial assistance from JNCASR. K.S. acknowledges TATA innovation fellowship (BT/HRD/35/01/03/2017), Department of Biotechnology (DBT), Govt. of India. KS laboratory is supported by funding from DBT, Science and Engineering Research Board (SERB), Indian Council of Medical Research (ICMR), and Indo-French Centre for the Promotion of Advanced Research (CEFIPRA). Intramural funding from JNCASR is acknowledged. This work is also supported by Nanyang Technological University's Nanyang Assistant Professorship grant and Singapore Ministry of Education Academic Research Fund Tier 1 grant [RG39/18 (S)] to A.S.

Author contributions

Author contributions: K.S., and A.S. designed research; K.G. and Y.C. performed research; K.G., Y.C., R.M., S.R.B.M.M., C.B., and G.B. contributed new reagents/analytic tools; K.G., Y.C., B.C.T., C.B., and G.B. analyzed data; and K.G., K.S., A.S., and Y.C. wrote the paper.

References

1. Zhang CZ, Leibowitz ML, & Pellman D (2013) Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev* 27(23):2513-2530.
2. Searle JB (1998) Speciation, chromosomes, and genomes. *Genome Research* 8(1):1-3.
3. Lee CS*, et al.* (2016) Chromosome position determines the success of double-strand break repair. *Proceedings of the National Academy of Sciences of the United States of America* 113(2):E146-154.
4. Agmon N, Liefshitz B, Zimmer C, Fabre E, & Kupiec M (2013) Effect of nuclear architecture on the efficiency of double-strand break repair. *Nat Cell Biol* 15(6):694-699.
5. Burgess SM & Kleckner N (1999) Collisions between yeast chromosomal loci in vivo are governed by three layers of organization. *Genes Dev* 13(14):1871-1883.
6. Piazza A, Wright WD, & Heyer WD (2017) Multi-invasions Are Recombination Byproducts that Induce Chromosomal Rearrangements. *Cell* 170(4):760-773 e715.
7. Muller H, Gil J, Jr., & Drinnenberg IA (2019) The Impact of Centromeres on Spatial Genome Architecture. *Trends Genet* 35(8):565-578.
8. Clarke L & Carbon J (1980) Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 287(5782):504-509.
9. Mahtani MM & Willard HF (1990) Pulsed-field gel analysis of α-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics* 7(4):607-613.
10. Navarro-Mendoza MI*, et al.* (2019) Early diverging fungus *Mucor circinelloides* lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. *Curr Biol* 29(22):3791-3802 e3796.
11. Meraldi P, McAinsh AD, Rheinbay E, & Sorger PK (2006) Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol* 7(3):R23.
12. van Hooff JJ, Tromer E, van Wijk LM, Snel B, & Kops GJ (2017) Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Reports* 18(9):1559-1571.
13. Tromer EC, van Hooff JJE, Kops G, & Snel B (2019) Mosaic origin of the eukaryotic kinetochore. *Proceedings of the National Academy of Sciences of the United States of America* 116(26):12873-12882.
14. Ekwall K (2007) Epigenetic control of centromere behavior. *Annu Rev Genet* 41(1):63-81.
15. Clarke L & Baum MP (1990) Functional analysis of a centromere from fission yeast: a role for centromere-specific repeated DNA sequences. *Molecular and Cellular Biology* 10(5):1863-1872.

16. Gordon JL, Byrne KP, & Wolfe KH (2011) Mechanisms of chromosome number evolution in yeast. *PLoS Genet* 7(7):e1002190.
17. Kobayashi N*, et al.* (2015) Discovery of an unconventional centromere in budding yeast redefines evolution of point centromeres. *Curr Biol* 25(15):2026-2033.
18. Tong P*, et al.* (2019) Interspecies conservation of organisation and function between nonhomologous regional centromeres. *Nature Communications* 10(1):2343.
19. Sanyal K, Baum M, & Carbon J (2004) Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. *Proceedings of the National Academy of Sciences of the United States of America* 101(31):11374-11379.
20. Padmanabhan S, Thakur J, Siddharthan R, & Sanyal K (2008) Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida albicans* and *Candida dubliniensis*. *Proceedings of the National Academy of Sciences of the United States of America* 105(50):19797-19802.
21. Kapoor S, Zhu L, Froyd C, Liu T, & Rusche LN (2015) Regional centromeres in the yeast *Candida lusitaniae* lack pericentromeric heterochromatin. *Proceedings of the National Academy of Sciences of the United States of America* 112(39):12139-12144.
22. Chatterjee G*, et al.* (2016) Repeat-associated fission yeast-like regional centromeres in the ascomycetous budding yeast *Candida tropicalis*. *PLoS Genet* 12(2):e1005839.
23. Baum M SK, Mishra PK, Thaler N, Carbon J. (2006) Formation of functional centromeric chromatin is specified epigenetically in *Candida albicans*. *Proceedings of the National Academy of Sciences* 103(40)(Oct 3):14877-14882.
24. Shen XX*, et al.* (2018) Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 175(6):1533-1545 e1520.
25. Butler G*, et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459(7247):657-662.
26. Jones T*, et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America* 101(19):7329-7334.
27. Burrack LS*, et al.* (2016) Neocentromeres provide chromosome segregation accuracy and centromere clustering to multiple loci along a *Candida albicans* chromosome. *PLoS Genet* 12(9):e1006317.
28. Descorps-Declere S*, et al.* (2015) Genome-wide replication landscape of *Candida glabrata*. *BMC Biol* 13:69.
29. Duan Z*, et al.* (2010) A three-dimensional model of the yeast genome. *Nature* 465(7296):363-367.
30. Sreekumar L*, et al.* (2019) Cis- and trans-chromosomal interactions define pericentric boundaries in the absence of conventional heterochromatin. *Genetics* 212(4):1121-1132.
31. Sreekumar L*, et al.* (2019) Orc4 spatiotemporally stabilizes centromeric chromatin. *bioRxiv*:465880, DOI: 465810.461101/465880.
32. Soderlund C, Bomhoff M, & Nelson WM (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* 39(10):e68.
33. Grabherr MG RP, Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics.* 26(9):1145-1151.
34. Drillon G, Carbone A, & Fischer G (2014) SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* 9(3):e92621.

35. Tsui CK, Daniel HM, Robert V, & Meyer W (2008) Re-examining the phylogeny of clinically relevant *Candida* species and allied genera based on multigene analyses. *FEMS Yeast Res* 8(4):651-659.
36. Legrand M, Jaitly P, Feri A, d'Enfert C, & Sanyal K (2019) *Candida albicans*: An emerging yeast model to study eukaryotic genome plasticity. *Trends Genet* 35(4):292-307.
37. Kumar S, Stecher G, Suleski M, & Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 34(7):1812-1819.
38. Cavalheiro M & Teixeira MC (2018) *Candida* Biofilms: threats, challenges, and promising strategies. *Front Med (Lausanne)* 5:28.
39. Pappas PG, Lionakis MS, Arendrup MC, Ostrosky-Zeichner L, & Kullberg BJ (2018) Invasive candidiasis. *Nat Rev Dis Primers* 4:18026.
40. Chakrabarti A*, et al.* (2015) Incidence, characteristics and outcome of ICU-acquired candidemia in India. *Intensive Care Med* 41(2):285-295.
41. Farooqi JQ*, et al.* (2013) Invasive candidiasis in Pakistan: clinical characteristics, species distribution and antifungal susceptibility. *J Med Microbiol* 62(Pt 2):259-268.
42. da Costa VG, Quesada RM, Abe AT, Furlaneto-Maia L, & Furlaneto MC (2014) Nosocomial bloodstream *Candida* infections in a tertiary-care hospital in South Brazil: a 4-year survey. *Mycopathologia* 178(3-4):243-250.
43. Xiao M*, et al.* (2015) Antifungal susceptibilities of *Candida glabrata* species complex, *Candida krusei*, *Candida parapsilosis* species complex and *Candida tropicalis* causing invasive candidiasis in China: 3 year national surveillance. *J Antimicrob Chemother* 70(3):802-810.
44. Goncalves SS, Souza ACR, Chowdhary A, Meis JF, & Colombo AL (2016) Epidemiology and molecular mechanisms of antifungal resistance in *Candida* and *Aspergillus*. *Mycoses* 59(4):198-219.
45. Lamoth F, Lockhart SR, Berkow EL, & Calandra T (2018) Changes in the epidemiological landscape of invasive candidiasis. *J Antimicrob Chemother* 73(suppl_1):i4-i13.
46. Selmecki A, Forche A, & Berman J (2006) Aneuploidy and isochromosome formation in drug-resistant *Candida albicans*. *Science* 313(5785):367-370.
47. Todd RT, Wikoff TD, Forche A, & Selmecki A (2019) Genome plasticity in *Candida albicans* is driven by long repeat sequences. *Elife* 8:e45954.
48. Seeber A, Hauer MH, & Gasser SM (2018) Chromosome dynamics in response to DNA damage. *Annu Rev Genet* 52(1):295-319.
49. Wolfe K*, et al.* (2017) Fungal genome and mating system transitions facilitated by chromosomal translocations involving intercentromeric recombination. *PLOS Biology* 15(8):e2002527.
50. Sankaranarayanan SR*, et al.* (2020) Loss of centromere function drives karyotype evolution in closely related Malassezia species. *eLife* 9:e53944.
51. Barra V & Fachinetti D (2018) The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat Commun* 9(1):4340.
52. Crasta K*, et al.* (2012) DNA breaks and chromosome pulverization from errors in mitosis. *Nature* 482(7383):53-58.

53. Meaburn KJ, Misteli T, & Soutoglou E (2007) Spatial genome organization in the formation of chromosomal translocations. *Seminars in cancer biology*, (Elsevier), pp 80-90.
54. Coughlan AY, Hanson SJ, Byrne KP, & Wolfe KH (2016) Centromeres of the yeast *Komagataella phaffii* (*Pichia pastoris*) have a simple inverted-repeat structure. *Genome Biology and Evolution* 8(8):2482-2492.
55. Thakur J & Sanyal K (2013) Efficient neocentromere formation is suppressed by gene conversion to maintain centromere function at native physical chromosomal loci in *Candida albicans*. *Genome Res* 23(4):638-652.
56. Ketel C*, et al.* (2009) Neocentromeres form efficiently at multiple possible loci in *Candida albicans*. *PLoS Genet* 5(3):e1000400.
57. Fukagawa T & Earnshaw WC (2014) The centromere: chromatin foundation for the kinetochore machinery. *Dev Cell* 30(5):496-508.
58. Scott KC & Sullivan BA (2014) Neocentromeres: a place for everything and everything in its place. *Trends Genet* 30(2):66-74.

Figure Legends

Figure 1. Construction of the gapless assembly of *C. tropicalis* type strain MYA-3404 in seven chromosomes.

A. Schematic showing the stepwise construction of the gapless chromosome-level assembly (Assembly2020) of *C. tropicalis* (also see Figure S1 and S2). B. An ideogram of seven chromosomes of *C. tropicalis* as deduced from Assembly2020 and drawn to scale. The genomic location of the three loci showing copy number variations (CNVs): *DUP4, DUP5* and *DUPR* located on Chr4, Chr5 and ChrR respectively are marked and shown as using black mesh. The CNVs for which the correct homolog-wise distribution of the duplicated copy is unknown are marked with asterisks. Homolog-specific differences for Chr1 and Chr4, occurred due to an exchange of chromosomal parts in a balanced heterozygous translocation between Chr1B and Chr4B, is highlighted with black borders (also see Figure S4C). C. An ethidium bromide (EtBr)-stained CHEF gel picture where the chromosomes of the *C. tropicalis* strain MYA-3404 and *C. albicans* strain SC5314 were separated (Methods). The known sizes of *C. albicans* chromosomes are presented for size estimation and validation of the chromosomes of *C. tropicalis* in the newly constructed Assembly2020. D. A circos plot showing genome-wide distribution of various sequence features. Very high sequence coverage at rDNA locus is clipped for clearer representation and marked with an asterisk.

Figure 2. Spatial genome organization reveals centromeric and telomeric *trans* contacts in *C. tropicalis.*

A. A representative field image of indirect immuno-fluorescence microscopy of Protein-A tagged CENP-A^{Cse4} (red) and DAPI-stained nuclear mass (blue). The images were acquired using a DeltaVision imaging system (GE) and processed using FIJI software. Scale, 2 μm. B. A 3D reconstruction of colocalization of DAPI stained genome (blue) and CENP-A^{Cse4} (red) using Imaris software (Oxford Instruments). Scale, 2 μm. C. A genome-wide contact probability heatmap (bin size = 10 kb) generated using 3C-seq data. Chromosome labels and their corresponding ideograms are shown on the heatmap. Color-bar represents the contact probability in log2 scale. D. Zoomed-in heatmap of chr4 and chr5 from panel C (blue box). E. Average signal strength of aggregate interactions (bin size = 2 kb) between centromeres (left) or telomeres (right) of different chromosomes. Left, genomic loci containing mid-points of centromeres are aligned at the center (red bar); right, genomic loci from 5′ or 3′-ends of chromosomes are aligned at the top right corner (arrow).

Figure 3. Genome-wide mapping of interchromosomal synteny breakpoints in *C. tropicalis* identifies a spatial cue for karyotype evolution.

A. A scaled representation of the color coded orthoblocks (relative to *C. albicans* chromosomes) and interchromosomal synteny breakpoints (ICSBs) (white lines) on *C. tropicalis* (Methods). Orthoblocks are defined as stretches of the target genome (*C. tropicalis*) carrying more than two syntenic ORFs from the same chromosome of the reference genome (*C. albicans*). The centromeres are represented with red arrowheads. B. Zoomed view of the centromere-specific ICSBs on *CEN2*, *CEN3*, *CEN5* and *CENR* showing the color-coded (relative to *C. albicans* chromosomes) ORFs flanking each centromere. *C. tropicalis*-specific unique ORFs proximal to *CEN3* and *CEN5* are shown in red. C. A smooth-line connected scatter-pot of the chromosome-wise ICSB density, calculated as number of ICSBs per 100 kb of the C. *tropicalis* genome (*y*-axis) as a function of the linear distance from the centromere in nine bins, which are a) within 100 kb of centromere (bin I), b) 100-200 kb (bin II), c) 200-300 kb (bin III), d) 300-400 kb (bin IV), e) 400-500 kb (bin V), f) 500-600 kb (bin VI), g) 600-700 kb (bin VII), h) >700 kb to telomere proximal 200 kb (bin VIII), and i) 200 kb from the telomeres (bin IX). Chr6 was excluded from the analysis, as it does

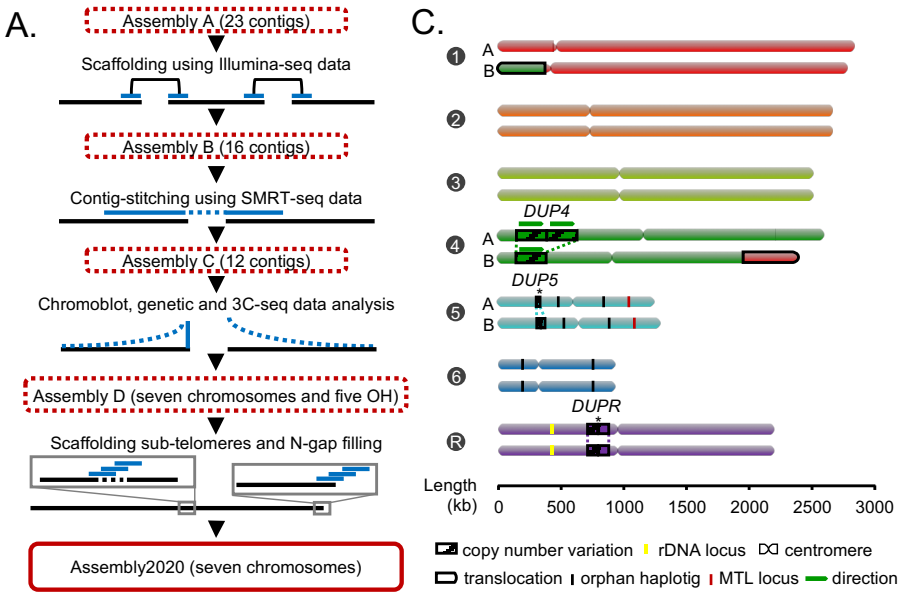
not have any ICSBs. E. A violin plot comparing the distribution of the orthoblock lengths (*y*-axis) at three different genomic zones, which are a) the centromere proximal zone (CP, within 300 kb from the centromere on both sides), b) the centromere distal zone (CD, beyond 300 kb from the centromere to telomere proximal 200 kb), and c) telomere-proximal zone (TP: within 200 kb from the telomeres). Orthoblocks, which span over more than one zone, were assigned to the zone with maximum overlap. The centromere-distal dataset was compared with the other two groups using the Mann Whitney test and the respective *P* values are presented. E - F. Circos representation showing the convergence of centromere proximal ORFs of *C. tropicalis* chromosomes near the centromeres on *C. albicans* Chr4 or ChrR. Chromosomes of *C. tropicalis* and *C. albicans* are marked with black and purple filled circles at the beginning of each chromosome, respectively.

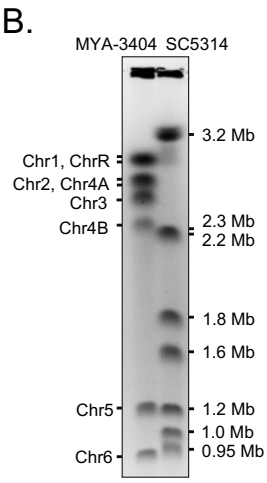
Figure 4. Genome-wide analysis of centromere DNA sequences across the CUG-Ser1 clade reveals divergence of unique centromeres from an ancestral homogenized inverted repeat-associated centromere type.
A. A dot-plot matrix representing the sequence and structural homology among species of the CUG-Ser1 clade was generated using Gepard (Methods). B. A logo plot showing the 12 bp long inter-species conserved motif (IR-motif), identified using MEME-suit (Methods). C. The density of the IR-motif on centromere DNA and across the entire genome of each species was calculated as the number of motifs per kb of DNA (Methods). Note that *C. albicans* and *C. dubliniensis* centromeres that form on unique and different DNA sequence do not contain the IR-motif. D. IGV track images showing the IR-motif density across seven chromosomes of *C. tropicalis.* Location of the centromere on each chromosome is marked with a red arrowhead. E. IGV track images showing IR-motif distribution across seven HIR-associated centromeres of *C. tropicalis.*

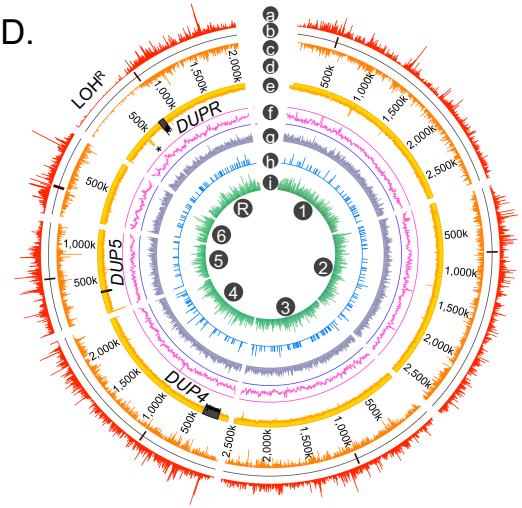
Figure 5. Conservation of the spatial genome organization after inter-centromeric translocation facilitated the centromere-type transition in the CUG-Ser1 clade.
A. A maximum likelihood based phylogenetic tree of closely related CUG-Ser1 species analyzed in this study. The centromere structure of each species is shown and drawn to scale. B. A model showing possible events during the loss of

homogenized repeat associated-associated centromeres and emergence of the unique centromere type through inter-centromeric translocations in the common ancestor of *C. tropicalis* and *C. albicans.* The model is drawn to show translocation between CtChr3 and CtChr4, as representative chromosomes, which can be mapped proximal to the centromere on CaChrR (as shown in Figure 3F). C. A cartoon representing the conservation of spatial genomic organization during inter-centromeric translocation that mediated centromere-type transition.

Figure 1







a SNP distribution (bin: 500 bp, range: 0-42)
b centromeres (brown bars)
c indel distribution (bin: 500 bp, range: 0-13)
d genomic coordinate (kb)
e 3C-seq coverage (bin: 500 bp, range: 2.72-139*)
f GC percent (bin: 5 kb, blue and pink lines 20 and 50%, respectively)
g ORF density (bin: 5 kb, range: 0-8)
h transposon density (bin: 1 kb, range: 0-3)
i repeat density (bin: 1 kb, range: 0-10)

Figure 2

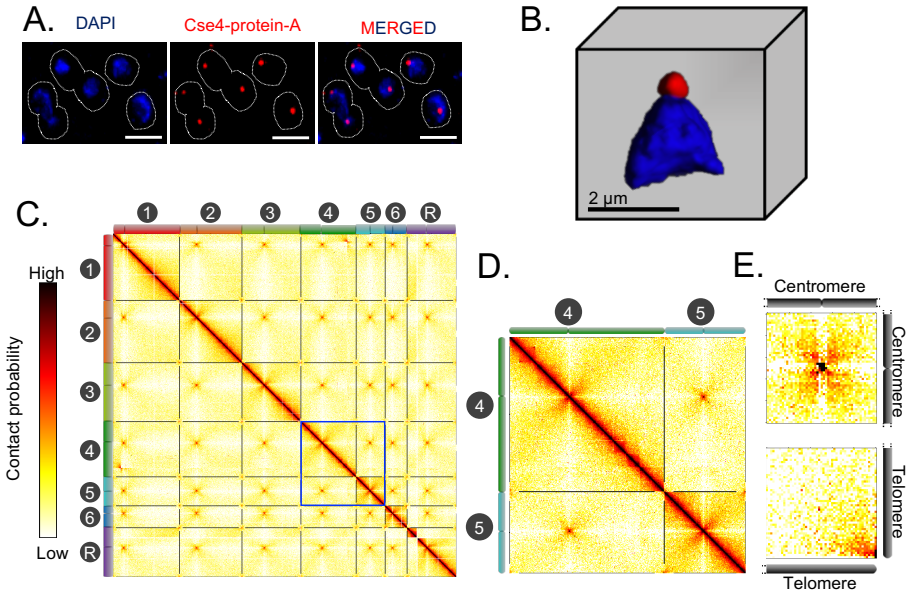


Figure 3

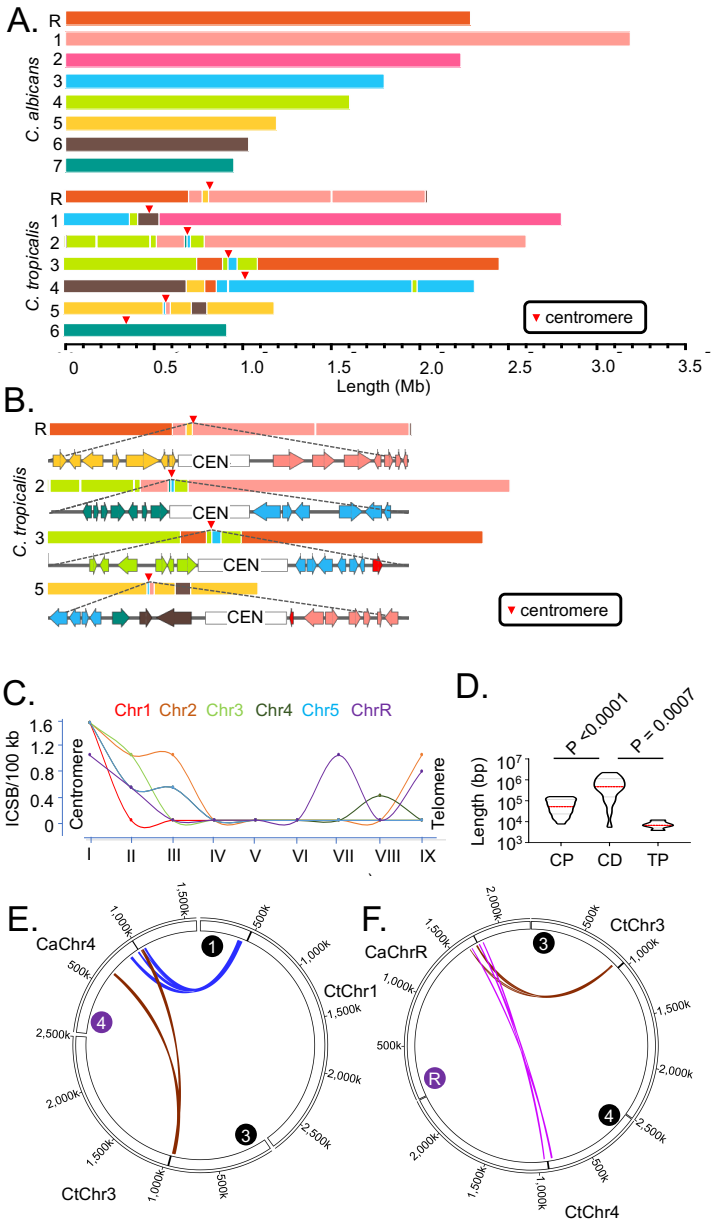


Figure 4

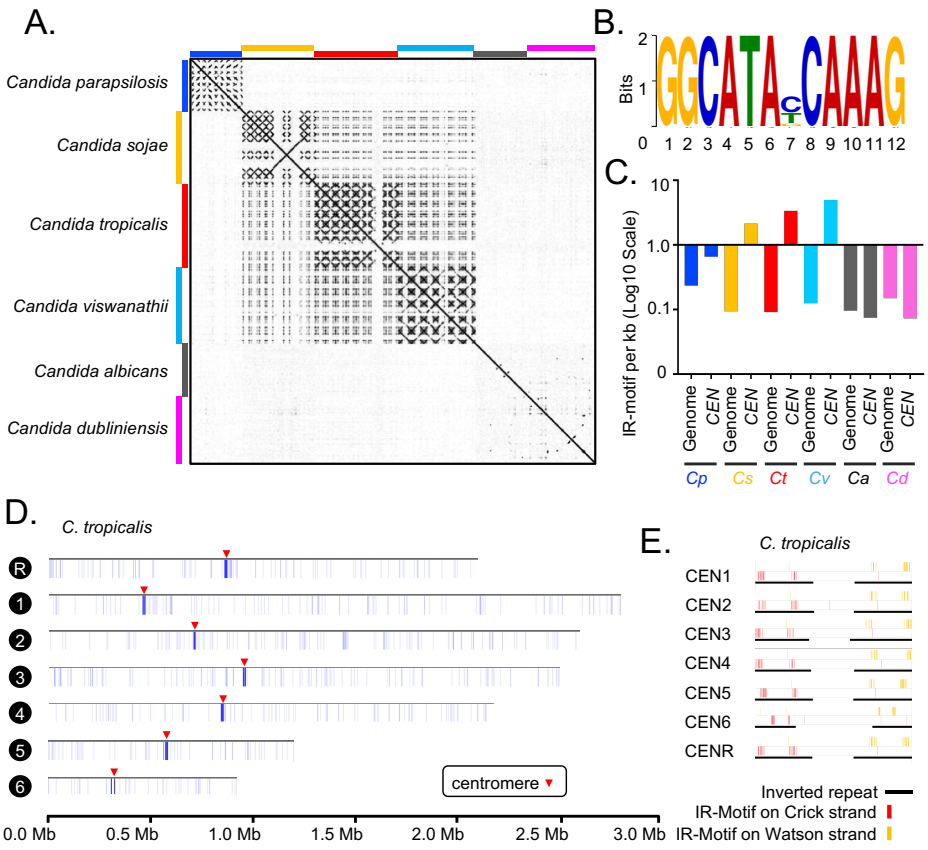


Figure 5