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

118 factor that remains common to most fungal centromeres is the presence of histone
119 H3 variant CENP-A^{Cse4} except for in *Mucor circinelloides* (10). Most of the
120 kinetochore proteins evolved from pre-eukaryotic lineages and remained conserved
121 within closely-related species complex or expanded through gene duplication (11-
122 13). It remains a long-standing paradox that the underlying centromere DNA
123 sequences keep evolving so fast while the kinetochore structure remains relatively
124 well conserved (14). Therefore, an understanding of the evolutionary processes
125 driving species-specific changes in centromere DNA sequences is essential for a
126 better understanding of the centromere biology.
127
128 The first centromere was discovered in *S. cerevisiae,* which carries conserved
129 genetic elements capable of activating *de novo* centromere function when cloned
130 into a yeast replicative plasmid (8). DNA sequence-dependent regulation of
131 centromere function is also identified in *Schizosaccharomyces pombe,* where the
132 centromeres inverted repeat-associated structures of 40-100 kb (15). Other closely
133 related *Saccharomyces* and *Schizosaccharomyces* species were also identified to
134 harbor a DNA sequence-dependent regulation of centromere function (16-18).
135 Although the DNA sequence-dependent mechanism for centromere function is
136 present in certain organisms, the advantage of having such regulation is not well
137 understood. In fact, the majority of the species with known centromeres are thought
138 to be regulated in a non-DNA sequence-dependent mechanism (14). The first
139 epigenetically-regulated fungal centromere carrying 3-5 kb long CENP-A^{Cse4}-bound
140 unique DNA sequences were identified in *C. albicans* (19), a CUG-Ser1 clade
141 species in the fungal phylum of Ascomycota. Subsequently, unique centromeres
142 were identified in closely related *Candida dubliniensis* (20) and *Candida lusitaniae*
143 (21). Strikingly, all seven centromeres of another CUG-Ser1 clade species *C.*
144 *tropicalis,* carry 3-4 kb long inverted repeats (IR) flanking ~3 kb long CENP-A^{Cse4} rich
145 central core (CC) and their DNA sequences are highly identical to each other.
146 Intriguingly, the centromere DNA of *C. tropicalis* can facilitate *de novo* recruitment of
147 CENP-A^{Cse4} (22). In contrast, the centromeres of *C. albicans* lack such a DNA
148 sequence-dependent mechanism facilitating *de novo* CENP-A^{Cse4} recruitment (23).
149 Such a rapid transition in the structural and functional properties of the centromeres
150 within two closely related species offers a unique opportunity to study the process of
151 centromere-type transition.

152

153 Our previous analysis suggested that centromeres of *C. tropicalis* are located
154 near inter chromosomal synteny breakpoints (ICSBs), which are relics of ancient
155 translocations in the common ancestor of *C. tropicalis* and *C. albicans* (22).
156 Additionally, the subcellular localization of the kinetochore proteins as a single
157 punctum per nucleus indicated the clustering of centromeres in *C. tropicalis* (22).
158 However, due to the nature of the then-available fragmented genome assembly, the
159 genome-wide distribution of the ICSBs and the spatial organization of the genome in
160 *C. tropicalis* remained unknown. Therefore, the influence of the spatial proximity on
161 the outcome of the translocations near the centromeres guiding the karyotype
162 evolution in the CUG-Ser1 clade remains as a hypothesis to be tested.
163

164 In this study, we constructed a chromosome-level gapless genome assembly
165 of the *C. tropicalis* type strain MYA-3404 by combining information from previously
166 available contigs, NGS reads and high-throughput 3C-seq data. Using this assembly
167 and 3C-seq data, we studied the spatial genome organization in *C. tropicalis.* Next,
168 we mapped the ICSBs in *C. tropicalis* genome with reference to *C. albicans*
169 (ASM18296v3) to ask if the frequency of translocations correlates with the spatial
170 genome organization. In addition, we performed Oxford Nanopore and Illumina
171 sequencing and assembled the genome of *Candida sojae,* a sister species of *C.*
172 *tropicalis* in the CUG-Ser1 clade (24). Finally, we used our genome assembly of *C.*
173 *sojae* and publicly available genome assembly of *C. viswanathii* (ASM332773v1) and
174 identified the putative centromeres of these two species as HIR-associated loci
175 syntenic to the centromeres of *C. tropicalis*. Based on our results, we propose a
176 model suggesting homology and proximity guided centromere proximal
177 translocations facilitated karyotype evolution and possibly aided in rapid transition
178 from HIR-associated to unique centromere types in the members of CUG-Ser1
179 clade.
180

181 **Results**

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

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

186 assemble the nuclear genome of *C. tropicalis* in seven chromosomes, we combined
187 short-read Illumina sequencing, and long-read single molecule real-time sequencing
188 (SMRT-seq) approaches together with high-throughput 3C-seq (simplified Hi-C)
189 experiment (**Figure 1A, S1A-D**). We started from the publicly available genome
190 assembly of *C. tropicalis* strain MYA-3404 in 23 nuclear contigs (ASM633v3,
191 Assembly A) (25) and used Illumina sequence reads to scaffold them into 16 contigs
192 to get Assembly B (**Figure 1A**). Next, we used the SMRT-seq long reads to join
193 these contigs, which resulted in an assembly of 12 contigs (Assembly C, **Table S1**).
194 Based on the contour clamped homogenized electric field (CHEF)-gel karyotyping
195 (**Figure 1B**) and 3C-seq data (**Figure S1E-G**), we joined two contigs and rectified a
196 misjoin in Assembly C to produce an assembly of seven chromosomes and five short
197 orphan haplotigs (OHs). Based on our analysis of the *de novo* contigs (**Figure S1H,**
198 Methods), sequence coverage data (**Figure S2A-B**), and Southern blot analysis of
199 the engineered aneuploid strains, we demonstrate that the small orphan contigs fall
200 in heterozygous regions of the genome (**Figure S2C-G,** Methods). Next, we used the
201 *de novo* contigs to fill pre-existing 104 N-gaps and scaffold 14 sub-telomeres (**Figure**
202 **S3A-C, Table S2**). Finally, we used the 3C-seq reads to polish the complete genome
203 assembly of *C. tropicalis* constituting of 14,609,527 bp in seven telomere-to-telomere
204 long gapless chromosomes (**Figure 1B**). We call this new assembly as
205 Assembly2020.
206
207 We then named the chromosomes in the order of their length from
208 chromosome 1 (Chr1) through chromosome 6 (Chr6), and the chromosome
209 containing rDNA locus is named as chromosome R (ChrR) (**Figure 1C**). Accordingly,
210 the centromere on each chromosome is named after the respective chromosome
211 number. Additionally, we assembled the genome sequence of each chromosome in
212 a way to consistently maintain the short arm of chromosomes at the 5′ end. The
213 statistics of the intermediate and final genome assemblies are summarized in **Table**
214 **S3**. In Assembly2020, 1278 out of 1315, Ascomycota-specific BUSCO gene sets
215 could be identified compared to 1255 identified using Assembly A (**Table S4**,
216 Methods). Inclusion of 23 additional BUSCO gene sets as compared to the
217 Assembly A suggests improved contiguity and completeness of Assembly2020.
218

219 Previously, using centromere-proximal probes, we could distinctly identify five
220 chromosomes (Chr1, Chr2, Chr3, Chr5, Chr6) in chromoblot analysis (22). However,
221 the length of Chr4 as well as ChrR remaind unknown. To validate the correct
222 assembly of these two chromosomes (Chr4 and ChrR), we performed chromoblot
223 analysis. We observed that the Chr4 homologs differ in size (**Figure S4A**). Analysis
224 of the sequence coverage across Chr4 identified an internal duplication of ~235 kb
225 region, which explains the size difference between the homologs Chr4A and Chr4B
226 (**Figure 1C, S4B**). We named this duplicated locus as *DUP4.* Subsequently, we
227 scanned the entire genome for the presence of copy number variations (CNVs),
228 which led to the identification of two additional large duplication events: one each on
229 Chr5 (*DUP5*, ~23 kb) and ChrR (*DUPR*, ~80 kb) (**Figure 1C, S4B**). Additionally, we
230 detected a balanced heterozygous translocation between Chr1 and Chr4 (**Figure**
231 **S4C**) through analyses of 3C-seq data and the *de novo* contigs (**Figure S4D**). This
232 translocation was validated using chromoblot analysis (**Figure S4E**), Illumina and
233 SMRT-seq read mapping (**Figure S4F**). A chromoblot analysis for ChrR revealed
234 that the actual length of ChrR is ~2.8 Mb, while the assembled length is 2.1 Mb
235 (**Figure 1C, S4G**). Considering the length of rDNA locus is ~700 kb in *C. albicans*
236 (26), we reason that the difference between the assembled length and actual length
237 (derived from the chromoblot analysis) of ChrR in *C. tropicalis* can be attributed due
238 to the presence of the repetitive rDNA of ~700 kb, which is not completely
239 assembled in Assembly2020.
240
241 Next, we performed phasing of the diploid genome of *C. tropicalis* using our
242 SMRT-seq, and 3C-seq data to identify the homolog-specific variations (Methods).
243 This analysis produced 16 nuclear contigs, which were colinear with the
244 chromosomes of Assembly2020, except for the previously validated heterozygous
245 translocation between Chr1 and Chr4 (**Figure S4H**). In order to characterize the
246 sequence variations in the diploid genome of *C. tropicalis,* we identified the single
247 nucleotide polymorphisms (SNPs) and insertions-deletions (indels) (Methods).
248 Intriguingly, we detected a long chromosomal region depleted of SNPs and indels on
249 the left arm of ChrR (**Figure 1D**). We refer to this region with loss of heterozygosity
250 on ChrR as LOH^R. Strikingly, we found parts of the syntenic regions of LOH^R to be
251 SNP and indel depleted in the *Candida sojae* strain NCYC-2607, a closely related
252 species of *C. tropicalis*, as well as in *C. albicans* reference strain SC5314 (**Figure**

253 **S5**). We also identified the genome-wide distribution of transposons and simple
254 repeats but could not detect preferential enrichment of these sequence elements at
255 any specific genomic location in *C. tropicalis* (**Figure 1D**). Together, we identified
256 multiple long CNVs, long-track LOH, and heterozygous translocation events in the
257 diploid genome of *C. tropicalis.* Possible implications of these events in virulence and
258 drug resistance of this successful human fungal pathogen need to be explored.
259
260 **Conserved principle of spatial genome organization in *C. tropicalis* and *C.***
261 ***albicans***
262 Indirect immunofluorescence imaging of *C. tropicalis* strain expressing
263 protein-A tagged Cse4 suggests the clustering of the centromere-kinetochore
264 complex, which is localized at the periphery of the DAPI-stained nuclear DNA mass
265 as a single punctum (**Figure 2A-B**). We re-aligned 3C-seq data to the
266 Assembly2020 to generate the genome-wide chromatin contact map of *C. tropicalis*.
267 The resultant heatmap shows high signal intensity along the diagonal indicating that
268 the intra-chromosomal interactions are generally stronger than interchromosomal
269 interactions (**Figure 2C**). However, the most striking feature of the heatmap is the
270 presence of conspicuous puncta in the interchromosomal areas, which signify strong
271 spatial proximity between centromeres (**Figure 2C-D**). Aggregate signal analysis
272 further reiterates the enrichment of centromere-centromere interactions (**Figure 2E**).
273 All these observations suggest the clustering of centromeres and conservation of the
274 Rabl configuration in *C. tropicalis*, a well-known feature of a higher-order genome
275 organization in yeasts (27-29). Strikingly, we also noted enrichment of interactions
276 between telomeres of different chromosomes (**Figure 2E**). These interchromosomal
277 telomeric interactions were significantly greater than the average interchromosomal
278 interaction (Mann-Whitney U test *P* value = 6.547×10^{-7}) (**Figure S6A**). We also
279 observed enhanced *cis* interaction between the two telomeres of an individual
280 chromosome compared to average intra-chromosomal long-range (≥100 kb)
281 interaction (Mann-Whitney U test *P* value = 1.091×10^{-9}) (**Figure S6B**).
282
283 Previously, the genomic contacts in *C. albicans* were analyzed by Hi-C, which
284 showed physical interaction among the centromeres (27, 30, 31). Together, our
285 analysis reveals a conserved pattern of centromere clustering in two closely related

286 fungal species with completely different centromere DNA sequences and structural
287 features. This observation suggests a DNA sequence-independent mechanism for
288 centromere clustering in yeasts. Moreover, our analysis demonstrates the conserved
289 principles of chromosomal organization in two human pathogenic ascomycetes, *C.*
290 *albicans*, and *C. tropicalis*, despite substantial karyotypic changes during the
291 speciation.
292
293 **Centromere and telomere proximal loci are hotspots for complex**
294 **translocations**
295 Using the chromosome-level Assembly2020 of *C. tropicalis* and publicly
296 available chromosome-level assembly of the *C. albicans* reference genome of
297 SC5314 strain (ASM18296v3), we performed a detailed genome-wide synteny
298 analysis following four different approaches. We used two published analysis tools,
299 Symap (32) and Satsuma synteny (33), and a custom approach to identify the ICSBs
300 based on the synteny of the conserved orthologs (**Figure 3A**). Next, we compared
301 and validated the results obtained from our custom approach of analysis with
302 another published tool Synchro (**Figure S7A-B**) (34). All four methods of analysis
303 detected that six out of seven centromeres (except *CEN6*) of *C. tropicalis* are located
304 proximal to multiple ICSBs, which are rare at the chromosomal arms (**Figure 3A**).
305 The ORF-level synteny analysis detected four out of seven centromeres (*CEN2*,
306 *CEN3*, *CEN5*, *CENR*) in *C. tropicalis* to be precisely located at the ICSBs, while
307 multiple ICSBs are located within ~100 kb of other two centromeres (**Figure 3B**).
308 However, no ICSB could be identified on Chr6. Additionally, we found a convergence
309 of orthoblocks from as many as four different chromosomes within 100 kb of
310 centromeres (**Figure 3B**).
311
312 To correlate the frequency of translocations with the spatial genome
313 organization, we quantified ICSB density (the number of ICSBs per 100 kb of the
314 genome) at different zones across the chromosome for all chromosomes except for
315 Chr6 (**Figure 3C**). Since no ICSBs were mapped on Chr6, it was excluded from the
316 analysis. This analysis revealed that the ICSB density is the highest at the
317 centromere proximal zones for all six chromosomes, but dropped sharply at the
318 chromosomal arms. However, the ICSB density near the telomere proximal zone for
319 Chr2, Chr4, and ChrR showed an increase over the chromosomal arms, albeit at a

320 lower magnitude than the centromeres. We also compared the length of the
321 orthoblocks across three different genomic zones- the centromere proximal (within
322 300 kb from the centromere on both sides), centromere distal (beyond 300 kb from
323 the centromere to 200 kb from the telomeres), and telomere proximal (within 200 kb
324 from the telomeres) zones. This analysis revealed that the length of the orthoblocks
325 located proximal to centromeres and telomeres are significantly smaller compared to
326 the orthoblocks located at the centromere distal zone (**Figure 3D**).
327
328 Does this mean there were inter-centromeric translocations in the common
329 ancestor of *C. albicans* and *C. tropicalis*? If such inter-centromeric translocations
330 occurred, then the ORFs present near different centromeres in *C. tropicalis* should
331 converge together on the *C. albicans* genome. Indeed, we found 10 loci where such
332 convergence is observed (**Figure S7C**). Intriguingly, four such loci are proximal to
333 the centromeres (CEN3, CEN4, CEN7, and CENR) in *C. albicans* (**Figure 3E-F,**
334 **S7D-E**). This observation supports the possibility of inter-centromeric translocation
335 events in the common ancestor of *C. albicans* and *C. tropicalis.* Additionally, the
336 other four centromeres are located proximal to ORFs, homologs of which are also
337 proximal to the centromeres in *C. tropicalis* (**Figure S7C**). Together, these
338 observations posit that the ancestral HIR-associated centromeres are lost in *C.*
339 *albicans* and evolutionary new centromeres (ENCs) formed proximal to the ancestral
340 centromere loci on unique and different DNA sequences (19).
341
342 **Rapid transition in the centromere type within the members of the CUG-Ser1**
343 **clade**
344 Based on the identification of multiple translocation events concentrated near
345 the centromeric regions of the *C. tropicalis* genome, we hypothesize that complex
346 translocations between HIR-associated centromeres in the common ancestor of *C.*
347 *albicans* and *C. tropicalis* led to the loss of HIR and evolution of unique centromere
348 types observed in *C. albicans* and *C. dubliniensis.* However, the genomic
349 rearrangements are rare events, even at the evolutionary time scale. Therefore, if
350 the HIR-associated centromeres are the ancestral state, from which the unique
351 centromeres would have derived during a rare chromothripsis-like event, then the
352 other closely related species should have retained HIR-associated centromeres. To
353 gain further insights into the centromere type of the common ancestor of *C. albicans*

354 and *C. tropicalis,* we scanned for the presence of HIR-like structures in the genomes
355 of *C. parapsilosis,* an early-diverging members of the CUG-Ser1 clade. Indeed, we
356 identified eight HIR-associated structures (**Figure S8A**), present once in each of the
357 eight chromosomes of *C. parapsilosis.* Identification of the HIR-associated structures
358 present at the intergenic and transcription poor regions, once each on all eight
359 chromosomes, suggests that these loci are the putative centromeres of *C.*
360 *parapsilosis.* This observation indicates that the common ancestor of *C. albicans* and
361 *C. tropicalis* possibly carried HIR-associated centromeres. Next, we performed a
362 genome-wide synteny analysis between *C. orthopsilosis* and *C. parapsilosis* and
363 found evidence of translocations at seven out of eight HIR-associated loci, five of
364 which are ICSB associated (**Figure S8B**). This result indicates the involvement of
365 HIR-associated structures in translocation events, similar to those translocation
366 events involving *C. tropicalis* centromeres.
367
368 Such structure-defined HIR-associated centromeres have only been identified
369 in *C. tropicalis* in the CUG-Ser1 clade species (22). Although IRs are present in
370 *CEN4*, *CEN5*, and *CENR* of *C. albicans,* these sequences are not homogenized like
371 the HIR-associated centromeres in *C. tropicalis* (**Figure 4A**). In order to study the
372 presence or absence of HIRs in *C. sojae,* a sister species of *C. tropicalis* (24), we
373 assembled its genome into 42 contigs, including seven chromosome-length contigs
374 (Methods). Using this assembly, we identified seven putative centromeres in *C.*
375 *sojae* as intergenic and HIR-associated loci syntenic to the centromeres in *C.*
376 *tropicalis* (**Figure S9A-B, D**). Each of these seven centromeres in *C. sojae* consists
377 of a ~2 kb long central core (CC) region flanked by 3-12 kb long inverted repeats
378 (**Table S5**). Using a similar approach, we identified six HIR-associated centromeres
379 in the publicly available genome assembly (ASM332773v1) of *Candida viswanathii*,
380 another species closely related to *C. tropicalis* (**Figure S9C, E, Table S6**) (35). A
381 dot-plot analysis found extensive homology shared across the IRs but not among the
382 CC elements (**Figure 4A**) of the HIR-associated centromeres present in *C. tropicalis*
383 and the putative centromeres of *C. sojae* and *C. viswanathii* (**Table S7**). Moreover,
384 We detected extensive structural conservation in *CEN* DNA-elements, especially
385 among IRs within an individual species (**Figure S10A**). This structural feature of IRs
386 is also significantly conserved across the three species, *C. tropicalis*, *C. sojae,* and
387 *C. viswanathii,* with HIR-associated centromeres (**Figure S10B**).

388

389 Cloning of a full-length centromere of *C. tropicalis* in a replicative plasmid
390 facilitates *de novo* CENP-A deposition but fails when the IRs are replaced with
391 Ca*CEN5* IRs (22). This result indicated the presence of a genetic element
392 specifically on the IRs of *C. tropicalis* but absent in Ca*CEN5* IR. To identify the
393 putative genetic element, we analyzed the *CEN* DNA sequences of all three HIR-
394 associated centromeres and the unique centromeres of *C. albicans* for the presence
395 of conserved motifs. This analysis identified a highly conserved 12 bp motif (dubbed
396 as IR-motif) (**Figure 4B**) clustered specifically at the centromeres but not anywhere
397 else in the entire genome of *C. tropicalis, C. sojae* and *C. viswanathii* (**Figure 4C-D,**
398 **S10C**). On the contrary, the IR-motif density at the centromeres in *C. albicans*
399 remains approximately an order of magnitude lower than that of *C. tropicalis* (**Figure**
400 **4C**). This observation indicates a potential function of the IR-motif in the regulation of
401 *de novo* CENP-A loading in *C. tropicalis*. Moreover, the *CEN*-enriched motif was
402 found to be specifically concentrated on the IRs but not at the mid-core region in
403 HIR-associated centromeres present in *C. tropicalis* (**Figure 4E**) and at the putative
404 centromeres in *C. soaje* and *C. viswanathii* (**S10D**). Additionally, we detected that
405 the direction of the IR-motif is diverging away from the central core of the
406 centromeres in *C. tropicalis* (**Figure S10E**), and this pattern remained conserved in
407 the other two species as well (**Figure S10F**). The conserved structure and
408 organization of the IR-motif sequences in the HIR-associated centromeres of three
409 *Candida* species suggest an inter-species conserved function of the IR DNA
410 sequence among these three species, although the clusters of IR-motifs are located
411 at a variable distance from the CC in these three species (**Figure S10G**). The
412 importance of this 12-bp conserved motif on the centromere function is yet to be
413 determined.

414

415 **Discussion**

416

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

422 indels. We constructed genome-wide contact maps and identified centromere-
423 centromere as well as telomere-telomere spatial interactions. A comparative genome
424 analysis between *C. albicans* and *C. tropicalis* revealed that six out of seven
425 centromeres of *C. tropicalis* are mapped precisely at or proximal to
426 interchromosomal synteny breakpoints. Strikingly, ORFs proximal to the centromeres
427 of *C. tropicalis* are converged into specific regions on the *C. albicans* genome in
428 some occasions, suggesting possibilities of inter-centromeric translocations in their
429 common ancestor. Moreover, the presence of homogenized inverted repeat
430 associated putative centromeres in *C. sojae* and *C. viswanathii,* like in *C. tropicalis*,
431 suggest that such a centromere structure is plausibly the ancestral form in the CUG-
432 Ser1 clade species complex but lost in *C. albicans* and *C. dubliniensis.* We propose
433 that loss of such a centromere structure possibly happened during translocation
434 events involving centromeres in the common ancestor might have given rise to
435 evolutionary new centromeres on unique DNA sequences and facilitated speciation.
436
437 The availability of the chromosome-level genome assembly, and improved
438 annotations of genomic variants and genes absent in the publicly available
439 fragmented genome assembly of *C. tropicalis* should greatly facilitate genome-wide
440 association studies to understand the pathobiology of the organism including the
441 cause of antifungal drug resistance. In addition, this study sheds lights on how
442 primordial mechanisms of *de novo* centromere establishment present in an ancestral
443 species become dispensable in the derived lineages.
444
445 *C. tropicalis* is a human pathogenic ascomycete, closely related to the well-
446 studied model fungal pathogen *C. albicans* (36). These two species diverged from
447 their common ancestor ~39 million years ago (37) and evolved into two distinct
448 karyotypes (22), having different phenotypic traits (38), and ecological niches (39).
449 While *C. albicans* remains the primary cause of candidiasis worldwide, systemic
450 ICU-acquired candidiasis is primarily (30.5-41.6%) caused by *C. tropicalis* in tropical
451 countries including India (40), Pakistan (41), and Brazil (42). Moreover, the
452 occurrence of drug resistance, particularly multidrug resistance, in *C. tropicalis* is on
453 the rise (40, 43, 44). Therefore, relatively less-studied *C. tropicalis* is emerging as a
454 major threat for nosocomial candidemia with 29-72% broad spectrum mortality rate
455 (45). Fluconazole resistance in *C. albicans* can be gained due to segmental

456 aneuploidy of chromosome 5 containing long IRs at the centromere, by the formation
457 of isochromosomes (46), which is also identified in chromosome 4 with IRs at its
458 centromere (47). All seven centromeres in *C. tropicalis* are associated with long IRs,
459 hence it is possible that each of them can form isochromosomes. Now, with the
460 availability of the chromosome-level assembly of the *C. tropicalis* genome, it should
461 be possible to initiate genome-wide association studies to understand the genomic
462 causes of pathogenicity and the rapid emergence of drug resistance in *C. tropicalis.*
463
464 Since the mechanism of homology search during HR is positively influenced
465 by spatial proximity and the extent of DNA sequence homology (4, 48), at least in the
466 engineered model systems, it is expected that spatially clustered homologous DNA
467 sequences experience more translocations than other loci. Although these factors
468 were not shown to be involved in karyotypic rearrangements during speciation, a
469 retrospective survey in light of spatial proximity and homology now offers a better
470 explanation. For example, bipolar to tetrapolar transition of the mating type locus in
471 the *Cryptococcus* species complex was associated with inter-centromeric
472 recombination following pericentric inversion (49). Similar inter-centromeric
473 recombination has been reported in the common ancestor of two fission yeast
474 species, *Schizosaccharomyces cryophilus* and *Schizosaccharomyces octosporus*
475 (18). These examples raise an intriguing notion that centromeres serve as sites of
476 recombination, which may lead to centromere loss and/or emergence of evolutionary
477 new centromeres. This notion is supported by the fact that DNA breaks at the
478 centromere following fusion of the acentric fragments to other chromosomes led to
479 chromosome number reduction in *Ashbya* species (16) and *Malassezia* species (50).
480 Genomic instability at the centromere can also lead to fluconazole resistance, as in
481 the case of isochromosome formation on Chr5 of *C. albicans* (46). Additionally,
482 breaks at the centromeres are reported to be associated with cancers (51).
483
484 What would be the conseqence of spatial proximity of chromosomal regions
485 with high DNA sequence homology, observed in fungal systems, in other domains of
486 life? Chromoplexy, a type of chromothripsis, where a series of translocations occur
487 among multiple chromosomes, is associated with cancers (1). Although fine mapping
488 of translocation events at the repetitive regions in human cancer cells becomes
489 difficult, the growing evidence that such events are associated with the formation of

490 micronuclei (52) supports the idea that spatial genome organization may influence
491 such events (53). With the availability of Hi-C and other techniques to probe genomic
492 contacts in high-resolution, it may now be possible to test whether chromoplexies
493 occur due to close physical proximity of homologous DNA sequences.
494
495 The identification of HIR-associated putative centromeres in *C. parapsilosis*,
496 *C. sojae,* and *C. viswanathii* supports the idea that the unique centromeres might
497 have evolved from an ancestral HIR-associated centromere (54) (**Figure 5A**). Such
498 a rapid transition in the structural and functional properties of centromeres is
499 unprecedented. While HIR-associated centromeres of *C. tropicalis, C. sojae,* and *C.*
500 *viswanathii* form on different DNA sequences, a well-conserved IR-motif was
501 identified in this study that is present in multiple copies on the centromeric IR
502 sequences across these three species. Some centromeres in *C. albicans* carry
503 chromosome-specific IRs, which lack IR-motifs. In addition, *CaCEN5* IRs could not
504 functionally complement the centromere function in *C. tropicalis* for the *de novo*
505 CENP-A^{Cse4} recruitment. This indicates a possible role of the conserved IR-motifs on
506 species-specifc centromere function (22). Therefore, the loss of HIR-associated
507 centromere in *C. albicans* that are only epigenetically propagated (23) clearly shows
508 how ability of *de novo* establishment of kinetochore assembly in an ancestral lineage
509 can be lost in an derived lineage. However, the details of the mechanism through
510 which IR-motifs may regulate centromere identity remains to be explored.
511
512 Loss of HIR-associated centromeres during inter-centromeric translocations
513 or MIR must have been catastrophic for the cell, and the survivor needed to activate
514 another centromere at an alternative locus. How is such a location determined?
515 Artificial removal of a native centromere in *C. albicans* leads to the activation of a
516 neocentromere (55, 56), which then becomes part of the centromere cluster (27).
517 This evidence supports the existence of a spatial determinant, known as the CENP-
518 A cloud or CENP-A-rich zone (55, 57), influencing preferential formation of
519 neocentromere at loci proximal to the native centromere (55, 58). We found that the
520 unique and different centromeres of *C. albicans* are located proximal to the ORFs,
521 which are also proximal to the centromeres in *C. tropicalis*. This observation
522 indicates that the formation of the new centromeres in *C. albicans* may have been
523 influenced by spatial proximity to the ancestral *CEN* cluster. However, the new

524 centromeres of *C. albicans* are formed on loci with completely unique and different
525 DNA sequences. Because of these reasons, it may be logical to consider the
526 centromeres of *C. albicans* as ENCs (**Figure 5B**). Intriguingly, even after the
527 catastrophic chromosomal rearrangements, the ENCs in *C. albicans* remain
528 clustered similar to *C. tropicalis* (**Figure 5C**). This observation identifies spatial
529 clustering of centromeres as a matter of cardinal importance for the fungal genome
530 organization.
531
532 **Materials and Methods**
533 The strains, primers, and plasmids used in this study are listed in SI Appendix,
534 Tables S8, S9, and S10, respectively. Details of all of the experimental procedures
535 and sequence analysis are given in SI Materials and Methods. All sequencing data
536 used in the study and the genome assembly of *C. tropicalis* and *C. sojae* have been
537 submitted to NCBI under the BioProject accession number PRJNA596050.
538
539 **Acknowledgments**
540 We thank all the members of KS laboratory and AS laboratory for critical reading of
541 the manuscript and inputs. We acknowledge Dr. Sheng Sun and Dr. Joseph Heitman
542 for helping in SMRT-seq of *C. tropicalis* at the PacBio sequencing facility at Duke
543 University. Illumina sequencing experiment for *C. sojae* genome was performed at
544 Clevergene Biocorp, Bangalore, India. We also thank B. Suma for confocal
545 microscopy, JNCASR. K.G. acknowledges Shyama Prasad Mukherjee Fellowship
546 from Council of Scientific and Industrial Research (CSIR), Govt. of India
547 [07/733(0181)/2013-EMR-I and financial assistance from JNCASR. K.S.
548 acknowledges TATA innovation fellowship (BT/HRD/35/01/03/2017), Department of
549 Biotechnology (DBT), Govt. of India. KS laboratory is supported by funding from
550 DBT, Science and Engineering Research Board (SERB), Indian Council of Medical
551 Research (ICMR), and Indo-French Centre for the Promotion of Advanced Research
552 (CEFIPRA). Intramural funding from JNCASR is acknowledged. This work is also
553 supported by Nanyang Technological University's Nanyang Assistant Professorship
554 grant and Singapore Ministry of Education Academic Research Fund Tier 1 grant
555 [RG39/18 (S)] to A.S.
556
557 **Author contributions**

558 Author contributions: K.S., and A.S. designed research; K.G. and Y.C. performed
559 research; K.G., Y.C., R.M., S.R.B.M.M., C.B., and G.B. contributed new
560 reagents/analytic tools; K.G., Y.C., B.C.T., C.B., and G.B. analyzed data; and K.G.,
561 K.S., A.S., and Y.C. wrote the paper.
562
563 **References**
564

565 1. Zhang CZ, Leibowitz ML, & Pellman D (2013) Chromothripsis and beyond: rapid
566 genome evolution from complex chromosomal rearrangements. *Genes Dev*
567 27(23):2513-2530.
568 2. Searle JB (1998) Speciation, chromosomes, and genomes. *Genome Research* 8(1):1-3.
569 3. Lee CS*, et al.* (2016) Chromosome position determines the success of double-strand
570 break repair. *Proceedings of the National Academy of Sciences of the United States of*
571 *America* 113(2):E146-154.
572 4. Agmon N, Liefshitz B, Zimmer C, Fabre E, & Kupiec M (2013) Effect of nuclear
573 architecture on the efficiency of double-strand break repair. *Nat Cell Biol* 15(6):694-
574 699.
575 5. Burgess SM & Kleckner N (1999) Collisions between yeast chromosomal loci in vivo
576 are governed by three layers of organization. *Genes Dev* 13(14):1871-1883.
577 6. Piazza A, Wright WD, & Heyer WD (2017) Multi-invasions Are Recombination
578 Byproducts that Induce Chromosomal Rearrangements. *Cell* 170(4):760-773 e715.
579 7. Muller H, Gil J, Jr., & Drinnenberg IA (2019) The Impact of Centromeres on Spatial
580 Genome Architecture. *Trends Genet* 35(8):565-578.
581 8. Clarke L & Carbon J (1980) Isolation of a yeast centromere and construction of
582 functional small circular chromosomes. *Nature* 287(5782):504-509.
583 9. Mahtani MM & Willard HF (1990) Pulsed-field gel analysis of α-satellite DNA at the
584 human X chromosome centromere: high-frequency polymorphisms and array size
585 estimate. *Genomics* 7(4):607-613.
586 10. Navarro-Mendoza MI*, et al.* (2019) Early diverging fungus *Mucor circinelloides* lacks
587 centromeric histone CENP-A and displays a mosaic of point and regional
588 centromeres. *Curr Biol* 29(22):3791-3802 e3796.
589 11. Meraldi P, McAinsh AD, Rheinbay E, & Sorger PK (2006) Phylogenetic and structural
590 analysis of centromeric DNA and kinetochore proteins. *Genome Biol* 7(3):R23.
591 12. van Hooff JJ, Tromer E, van Wijk LM, Snel B, & Kops GJ (2017) Evolutionary dynamics
592 of the kinetochore network in eukaryotes as revealed by comparative genomics.
593 *EMBO Reports* 18(9):1559-1571.
594 13. Tromer EC, van Hooff JJE, Kops G, & Snel B (2019) Mosaic origin of the eukaryotic
595 kinetochore. *Proceedings of the National Academy of Sciences of the United States of*
596 *America* 116(26):12873-12882.
597 14. Ekwall K (2007) Epigenetic control of centromere behavior. *Annu Rev Genet* 41(1):63-
598 81.
599 15. Clarke L & Baum MP (1990) Functional analysis of a centromere from fission yeast: a
600 role for centromere-specific repeated DNA sequences. *Molecular and Cellular*
601 *Biology* 10(5):1863-1872.

602 16. Gordon JL, Byrne KP, & Wolfe KH (2011) Mechanisms of chromosome number
603 evolution in yeast. *PLoS Genet* 7(7):e1002190.
604 17. Kobayashi N*, et al.* (2015) Discovery of an unconventional centromere in budding
605 yeast redefines evolution of point centromeres. *Curr Biol* 25(15):2026-2033.
606 18. Tong P*, et al.* (2019) Interspecies conservation of organisation and function between
607 nonhomologous regional centromeres. *Nature Communications* 10(1):2343.
608 19. Sanyal K, Baum M, & Carbon J (2004) Centromeric DNA sequences in the pathogenic
609 yeast *Candida albicans* are all different and unique. *Proceedings of the National*
610 *Academy of Sciences of the United States of America* 101(31):11374-11379.
611 20. Padmanabhan S, Thakur J, Siddharthan R, & Sanyal K (2008) Rapid evolution of
612 Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida*
613 *albicans* and *Candida dubliniensis*. *Proceedings of the National Academy of Sciences*
614 *of the United States of America* 105(50):19797-19802.
615 21. Kapoor S, Zhu L, Froyd C, Liu T, & Rusche LN (2015) Regional centromeres in the
616 yeast *Candida lusitaniae* lack pericentromeric heterochromatin. *Proceedings of the*
617 *National Academy of Sciences of the United States of America* 112(39):12139-12144.
618 22. Chatterjee G*, et al.* (2016) Repeat-associated fission yeast-like regional centromeres
619 in the ascomycetous budding yeast *Candida tropicalis*. *PLoS Genet* 12(2):e1005839.
620 23. Baum M SK, Mishra PK, Thaler N, Carbon J. (2006) Formation of functional
621 centromeric chromatin is specified epigenetically in *Candida albicans*. *Proceedings of*
622 *the National Academy of Sciences* 103(40)(Oct 3):14877-14882.
623 24. Shen XX*, et al.* (2018) Tempo and mode of genome evolution in the budding yeast
624 subphylum. *Cell* 175(6):1533-1545 e1520.
625 25. Butler G*, et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight
626 *Candida* genomes. *Nature* 459(7247):657-662.
627 26. Jones T*, et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proceedings*
628 *of the National Academy of Sciences of the United States of America* 101(19):7329-
629 7334.
630 27. Burrack LS*, et al.* (2016) Neocentromeres provide chromosome segregation accuracy
631 and centromere clustering to multiple loci along a *Candida albicans* chromosome.
632 *PLoS Genet* 12(9):e1006317.
633 28. Descorps-Declere S*, et al.* (2015) Genome-wide replication landscape of *Candida*
634 *glabrata*. *BMC Biol* 13:69.
635 29. Duan Z*, et al.* (2010) A three-dimensional model of the yeast genome. *Nature*
636 465(7296):363-367.
637 30. Sreekumar L*, et al.* (2019) Cis- and trans-chromosomal interactions define pericentric
638 boundaries in the absence of conventional heterochromatin. *Genetics* 212(4):1121-
639 1132.
640 31. Sreekumar L*, et al.* (2019) Orc4 spatiotemporally stabilizes centromeric chromatin.
641 *bioRxiv*:465880, DOI: 465810.461101/465880.
642 32. Soderlund C, Bomhoff M, & Nelson WM (2011) SyMAP v3.4: a turnkey synteny
643 system with application to plant genomes. *Nucleic Acids Res* 39(10):e68.
644 33. Grabherr MG RP, Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. (2010)
645 Genome-wide synteny through highly sensitive sequence alignment: Satsuma.
646 *Bioinformatics.* 26(9):1145-1151.
647 34. Drillon G, Carbone A, & Fischer G (2014) SynChro: a fast and easy tool to reconstruct
648 and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* 9(3):e92621.

649 35. Tsui CK, Daniel HM, Robert V, & Meyer W (2008) Re-examining the phylogeny of
650 clinically relevant *Candida* species and allied genera based on multigene analyses.
651 *FEMS Yeast Res* 8(4):651-659.
652 36. Legrand M, Jaitly P, Feri A, d'Enfert C, & Sanyal K (2019) *Candida albicans*: An
653 emerging yeast model to study eukaryotic genome plasticity. *Trends Genet*
654 35(4):292-307.
655 37. Kumar S, Stecher G, Suleski M, & Hedges SB (2017) TimeTree: a resource for
656 timelines, timetrees, and divergence times. *Mol Biol Evol* 34(7):1812-1819.
657 38. Cavalheiro M & Teixeira MC (2018) *Candida* Biofilms: threats, challenges, and
658 promising strategies. *Front Med (Lausanne)* 5:28.
659 39. Pappas PG, Lionakis MS, Arendrup MC, Ostrosky-Zeichner L, & Kullberg BJ (2018)
660 Invasive candidiasis. *Nat Rev Dis Primers* 4:18026.
661 40. Chakrabarti A*, et al.* (2015) Incidence, characteristics and outcome of ICU-acquired
662 candidemia in India. *Intensive Care Med* 41(2):285-295.
663 41. Farooqi JQ*, et al.* (2013) Invasive candidiasis in Pakistan: clinical characteristics,
664 species distribution and antifungal susceptibility. *J Med Microbiol* 62(Pt 2):259-268.
665 42. da Costa VG, Quesada RM, Abe AT, Furlaneto-Maia L, & Furlaneto MC (2014)
666 Nosocomial bloodstream *Candida* infections in a tertiary-care hospital in South
667 Brazil: a 4-year survey. *Mycopathologia* 178(3-4):243-250.
668 43. Xiao M*, et al.* (2015) Antifungal susceptibilities of *Candida glabrata* species complex,
669 *Candida krusei*, *Candida parapsilosis* species complex and *Candida tropicalis* causing
670 invasive candidiasis in China: 3 year national surveillance. *J Antimicrob Chemother*
671 70(3):802-810.
672 44. Goncalves SS, Souza ACR, Chowdhary A, Meis JF, & Colombo AL (2016) Epidemiology
673 and molecular mechanisms of antifungal resistance in *Candida* and *Aspergillus*.
674 *Mycoses* 59(4):198-219.
675 45. Lamoth F, Lockhart SR, Berkow EL, & Calandra T (2018) Changes in the
676 epidemiological landscape of invasive candidiasis. *J Antimicrob Chemother*
677 73(suppl_1):i4-i13.
678 46. Selmecki A, Forche A, & Berman J (2006) Aneuploidy and isochromosome formation
679 in drug-resistant *Candida albicans*. *Science* 313(5785):367-370.
680 47. Todd RT, Wikoff TD, Forche A, & Selmecki A (2019) Genome plasticity in *Candida*
681 *albicans* is driven by long repeat sequences. *Elife* 8:e45954.
682 48. Seeber A, Hauer MH, & Gasser SM (2018) Chromosome dynamics in response to
683 DNA damage. *Annu Rev Genet* 52(1):295-319.
684 49. Wolfe K*, et al.* (2017) Fungal genome and mating system transitions facilitated by
685 chromosomal translocations involving intercentromeric recombination. *PLOS Biology*
686 15(8):e2002527.
687 50. Sankaranarayanan SR*, et al.* (2020) Loss of centromere function drives karyotype
688 evolution in closely related Malassezia species. *eLife* 9:e53944.
689 51. Barra V & Fachinetti D (2018) The dark side of centromeres: types, causes and
690 consequences of structural abnormalities implicating centromeric DNA. *Nat*
691 *Commun* 9(1):4340.
692 52. Crasta K*, et al.* (2012) DNA breaks and chromosome pulverization from errors in
693 mitosis. *Nature* 482(7383):53-58.

694 53. Meaburn KJ, Misteli T, & Soutoglou E (2007) Spatial genome organization in the
695 formation of chromosomal translocations. *Seminars in cancer biology*, (Elsevier), pp
696 80-90.
697 54. Coughlan AY, Hanson SJ, Byrne KP, & Wolfe KH (2016) Centromeres of the yeast
698 *Komagataella phaffii* (*Pichia pastoris*) have a simple inverted-repeat structure.
699 *Genome Biology and Evolution* 8(8):2482-2492.
700 55. Thakur J & Sanyal K (2013) Efficient neocentromere formation is suppressed by gene
701 conversion to maintain centromere function at native physical chromosomal loci in
702 *Candida albicans*. *Genome Res* 23(4):638-652.
703 56. Ketel C*, et al.* (2009) Neocentromeres form efficiently at multiple possible loci in
704 *Candida albicans*. *PLoS Genet* 5(3):e1000400.
705 57. Fukagawa T & Earnshaw WC (2014) The centromere: chromatin foundation for the
706 kinetochore machinery. *Dev Cell* 30(5):496-508.
707 58. Scott KC & Sullivan BA (2014) Neocentromeres: a place for everything and everything
708 in its place. *Trends Genet* 30(2):66-74.
709
710

711 **Figure Legends**

712

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

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

732

733 **Figure 2. Spatial genome organization reveals centromeric and telomeric *trans***
734 **contacts in *C. tropicalis.***
735 A. A representative field image of indirect immuno-fluorescence microscopy of
736 Protein-A tagged CENP-A^{Cse4} (red) and DAPI-stained nuclear mass (blue). The
737 images were acquired using a DeltaVision imaging system (GE) and processed
738 using FIJI software. Scale, 2 μm. B. A 3D reconstruction of colocalization of DAPI
739 stained genome (blue) and CENP-A^{Cse4} (red) using Imaris software (Oxford
740 Instruments). Scale, 2 μm. C. A genome-wide contact probability heatmap (bin size =
741 10 kb) generated using 3C-seq data. Chromosome labels and their corresponding
742 ideograms are shown on the heatmap. Color-bar represents the contact probability in
743 log2 scale. D. Zoomed-in heatmap of chr4 and chr5 from panel C (blue box). E.
744 Average signal strength of aggregate interactions (bin size = 2 kb) between
745 centromeres (left) or telomeres (right) of different chromosomes. Left, genomic loci
746 containing mid-points of centromeres are aligned at the center (red bar); right,
747 genomic loci from 5′ or 3′-ends of chromosomes are aligned at the top right corner
748 (arrow).
749
750 **Figure 3. Genome-wide mapping of interchromosomal synteny breakpoints in**
751 ***C. tropicalis* identifies a spatial cue for karyotype evolution.**
752 A. A scaled representation of the color coded orthoblocks (relative to *C. albicans*
753 chromosomes) and interchromosomal synteny breakpoints (ICSBs) (white lines) on
754 *C. tropicalis* (Methods). Orthoblocks are defined as stretches of the target genome
755 (*C. tropicalis*) carrying more than two syntenic ORFs from the same chromosome of
756 the reference genome (*C. albicans*). The centromeres are represented with red
757 arrowheads. B. Zoomed view of the centromere-specific ICSBs on *CEN2*, *CEN3*,
758 *CEN5* and *CENR* showing the color-coded (relative to *C. albicans* chromosomes)
759 ORFs flanking each centromere. *C. tropicalis*-specific unique ORFs proximal to
760 *CEN3* and *CEN5* are shown in red. C. A smooth-line connected scatter-pot of the
761 chromosome-wise ICSB density, calculated as number of ICSBs per 100 kb of the C.
762 *tropicalis* genome (*y*-axis) as a function of the linear distance from the centromere in
763 nine bins, which are a) within 100 kb of centromere (bin I), b) 100-200 kb (bin II), c)
764 200-300 kb (bin III), d) 300-400 kb (bin IV), e) 400-500 kb (bin V), f) 500-600 kb (bin
765 VI), g) 600-700 kb (bin VII), h) >700 kb to telomere proximal 200 kb (bin VIII), and i)
766 200 kb from the telomeres (bin IX). Chr6 was excluded from the analysis, as it does

767 not have any ICSBs. E. A violin plot comparing the distribution of the orthoblock
768 lengths (*y*-axis) at three different genomic zones, which are a) the centromere
769 proximal zone (CP, within 300 kb from the centromere on both sides), b) the
770 centromere distal zone (CD, beyond 300 kb from the centromere to telomere
771 proximal 200 kb), and c) telomere-proximal zone (TP: within 200 kb from the
772 telomeres). Orthoblocks, which span over more than one zone, were assigned to the
773 zone with maximum overlap. The centromere-distal dataset was compared with the
774 other two groups using the Mann Whitney test and the respective *P* values are
775 presented. E - F. Circos representation showing the convergence of centromere
776 proximal ORFs of *C. tropicalis* chromosomes near the centromeres on *C. albicans*
777 Chr4 or ChrR. Chromosomes of *C. tropicalis* and *C. albicans* are marked with black
778 and purple filled circles at the beginning of each chromosome, respectively.
779
780 **Figure 4. Genome-wide analysis of centromere DNA sequences across the**
781 **CUG-Ser1 clade reveals divergence of unique centromeres from an ancestral**
782 **homogenized inverted repeat-associated centromere type.**
783 A. A dot-plot matrix representing the sequence and structural homology among
784 species of the CUG-Ser1 clade was generated using Gepard (Methods). B. A logo
785 plot showing the 12 bp long inter-species conserved motif (IR-motif), identified using
786 MEME-suit (Methods). C. The density of the IR-motif on centromere DNA and across
787 the entire genome of each species was calculated as the number of motifs per kb of
788 DNA (Methods). Note that *C. albicans* and *C. dubliniensis* centromeres that form on
789 unique and different DNA sequence do not contain the IR-motif. D. IGV track images
790 showing the IR-motif density across seven chromosomes of *C. tropicalis.* Location of
791 the centromere on each chromosome is marked with a red arrowhead. E. IGV track
792 images showing IR-motif distribution across seven HIR-associated centromeres of *C.*
793 *tropicalis.*
794
795 **Figure 5. Conservation of the spatial genome organization after inter-**
796 **centromeric translocation facilitated the centromere-type transition in the**
797 **CUG-Ser1 clade.**
798 A. A maximum likelihood based phylogenetic tree of closely related CUG-Ser1
799 species analyzed in this study. The centromere structure of each species is shown
800 and drawn to scale. B. A model showing possible events during the loss of

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

Figure 1

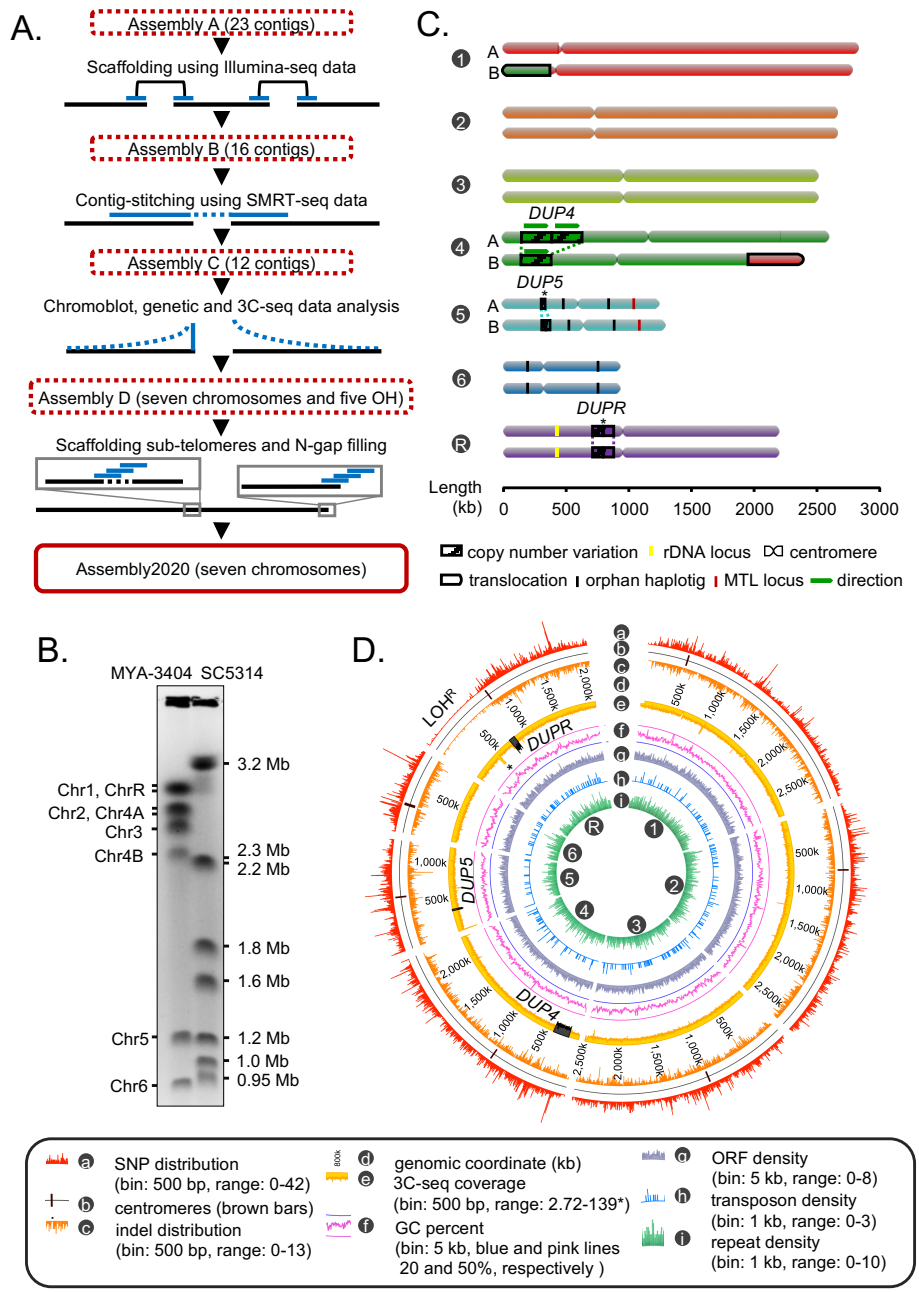

A.
Assembly A (23 contigs)
Scaffolding using Illumina-seq data
Assembly B (16 contigs)
Contig-stitching using SMRT-seq data
Assembly C (12 contigs)
Chromoblot, genetic and 3C-seq data analysis
Assembly D (seven chromosomes and five OH)
Scaffolding sub-telomeres and N-gap filling
Assembly2020 (seven chromosomes)
C.
A
B
DUP4
DUP5
DUPR
Length (kb)
0
500
1000
1500
2000
2500
3000
copy number variation
rDNA locus
centromere
translocation
orphan haplotig
MTL locus
direction
B.
MYA-3404
SC5314
Chr1, ChrR
Chr2, Chr4A
Chr3
Chr4B
Chr5
Chr6
3.2 Mb
2.3 Mb
2.2 Mb
1.8 Mb
1.6 Mb
1.2 Mb
1.0 Mb
0.95 Mb
D.
LOHR
DUPR
DUP5
DUP4
SNP distribution (bin: 500 bp, range: 0-42)
centromeres (brown bars)
indel distribution (bin: 500 bp, range: 0-13)
genomic coordinate (kb)
3C-seq coverage (bin: 500 bp, range: 2.72-139*)
GC percent (bin: 5 kb, blue and pink lines 20 and 50%, respectively)
ORF density (bin: 5 kb, range: 0-8)
transposon density (bin: 1 kb, range: 0-3)
repeat density (bin: 1 kb, range: 0-10)

Figure 2

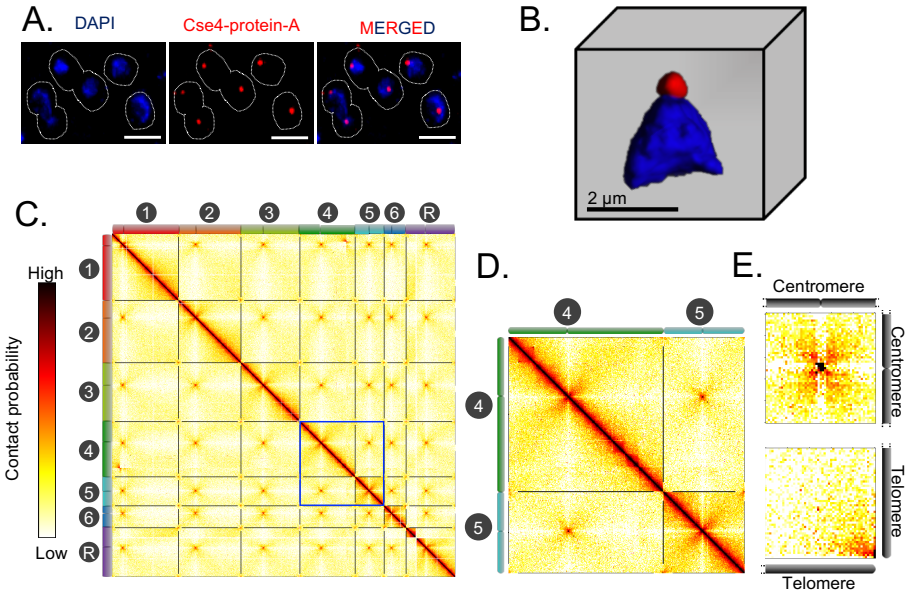

A.
DAPI
Cse4-protein-A
MERGED
B.
2 µm
C.
1
2
3
4
5
6
R
High
Low
Contact probability
D.
E.
Centromere
Telomere
Centromere
Telomere

Figure 3

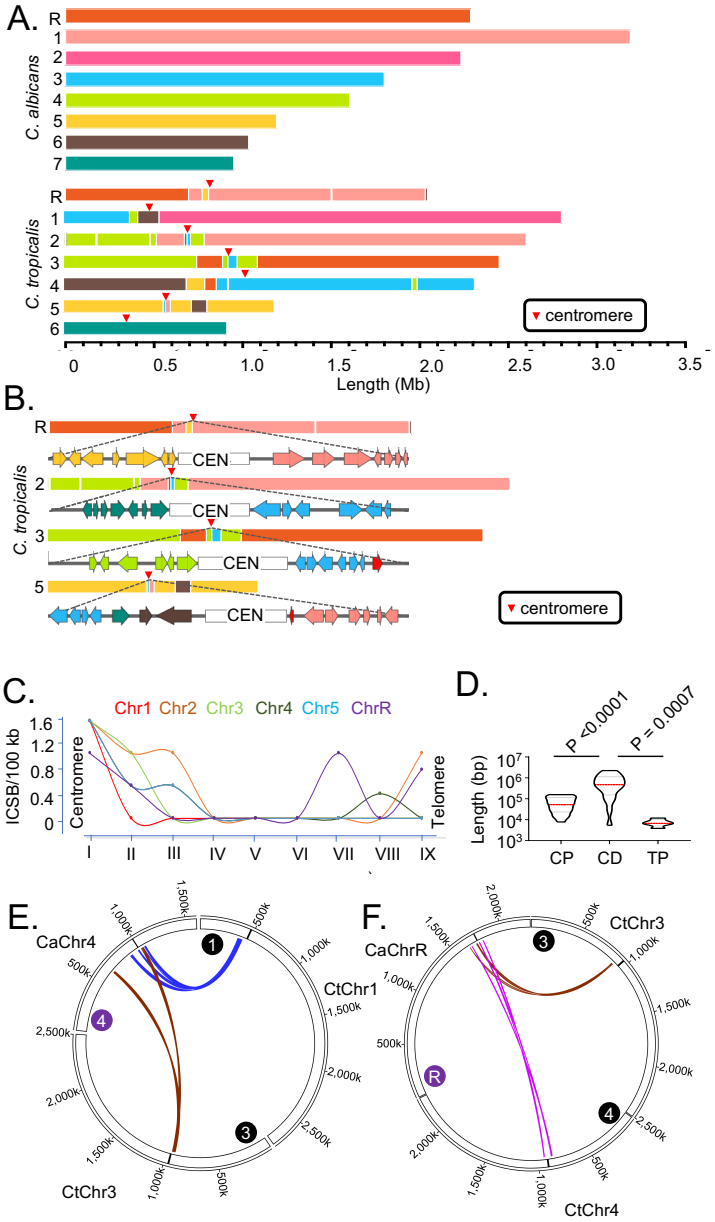

A.
C. albicans
R
1
2
3
4
5
6
7
C. tropicalis
R
1
2
3
4
5
6
centromere
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Length (Mb)
B.
C. tropicalis
R
2
3
5
CEN
centromere
C.
Chr1 Chr2 Chr3 Chr4 Chr5 ChrR
ICSB/100 kb
1.6
1.2
0.8
0.4
0
Centromere
Telomere
I II III IV V VI VII VIII IX
D.
P <0.0001
P = 0.0007
Length (bp)
CP CD TP
E.
CaChr4
CtChr1
CtChr3
F.
CaChrR
CtChr3
CtChr4

Figure 4

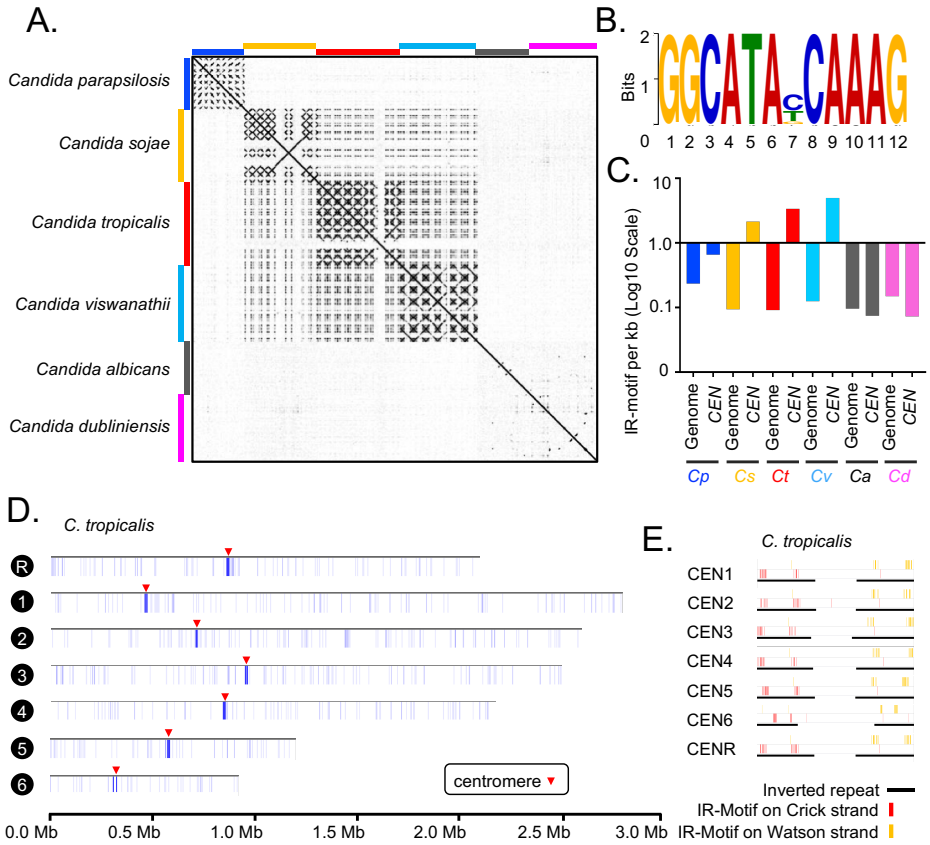
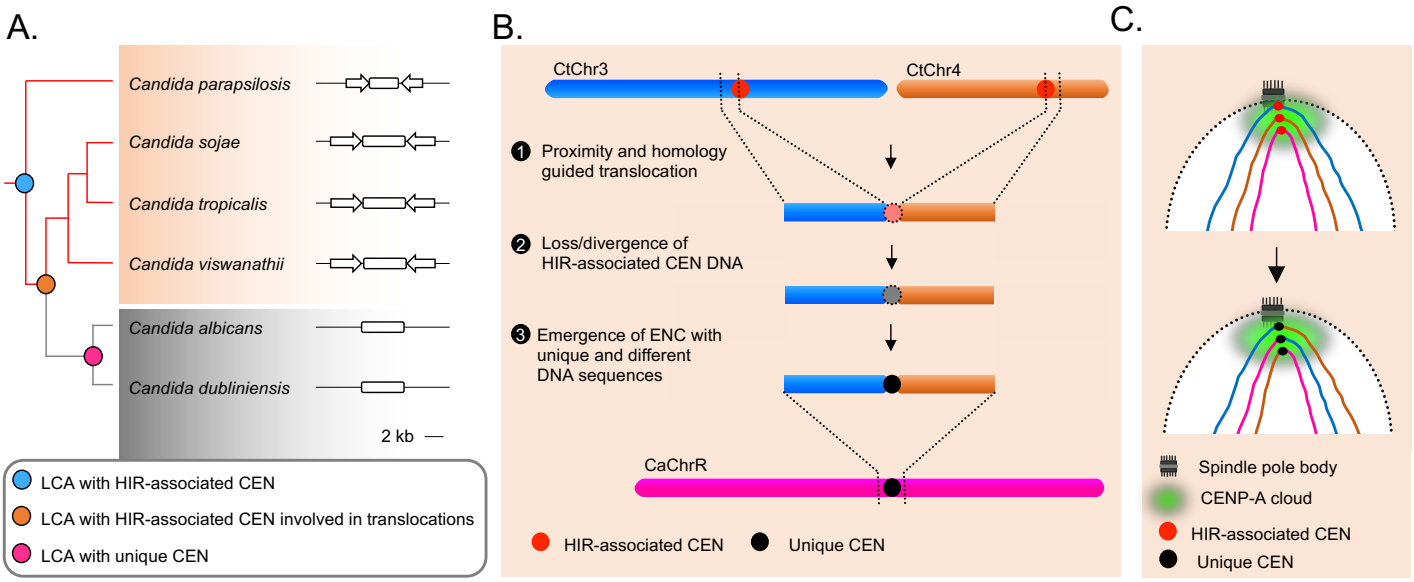

A.
Candida parapsilosis
Candida sojae
Candida tropicalis
Candida viswanathii
Candida albicans
Candida dubliniensis
B.
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12
C.
IR-motif per kb (Log10 Scale)
10
1.0
0.1
0
Genome
CEN
Cp
Cs
Ct
Cv
Ca
Cd
D.
C. tropicalis
R
1
2
3
4
5
6
centromere
0.0 Mb
0.5 Mb
1.0 Mb
1.5 Mb
2.0 Mb
2.5 Mb
3.0 Mb
E.
C. tropicalis
CEN1
CEN2
CEN3
CEN4
CEN5
CEN6
CENR
Inverted repeat
IR-Motif on Crick strand
IR-Motif on Watson strand

Figure 5


A.
Candida parapsilosis
Candida sojae
Candida tropicalis
Candida viswanathii
Candida albicans
Candida dubliniensis
2 kb
LCA with HIR-associated CEN
LCA with HIR-associated CEN involved in translocations
LCA with unique CEN
B.
CtChr3
CtChr4
1 Proximity and homology guided translocation
2 Loss/divergence of HIR-associated CEN DNA
3 Emergence of ENC with unique and different DNA sequences
CaChrR
HIR-associated CEN
Unique CEN
C.
Spindle pole body
CENP-A cloud
HIR-associated CEN
Unique CEN