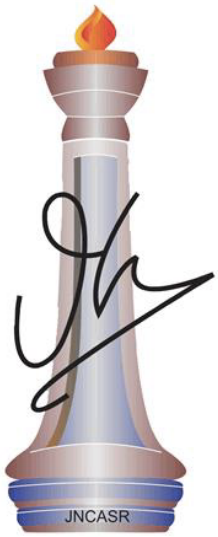
Chromosome-level Genome Assembly of the Human Fungal Pathogen *Candida tropicalis*

A thesis submitted for the degree of

Doctor of Philosophy

by

Krishnendu Guin



Molecular Biology and Genetics Unit

Jawaharlal Nehru Centre for Advanced Scientific Research

Jakkur, Bangalore-560064

May 2020

Dedicated to my parents

Declaration

I hereby declare that this thesis entitled "Chromosome-level genome assembly of the human fungal pathogen *Candida tropicalis*" is an authentic record of the research work carried out by myself under the supervision of Prof. Kaustuv Sanyal, Professor, Molecular Biology and Genetics Unit (MBGU), Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bangalore, India and that this work has not been submitted elsewhere for the award of any other degree. In keeping with the norm of reporting the scientific observations, due acknowledgements have been made whenever the work described was carried out in collaboration with other researchers. Any omission, which might have occurred by oversight or misjudgment, is regretted.

Krishnendu Guin
Place: Jakra
Date: 15th May 2020

Jawaharlal Nehru Center for Advanced Scientific Research

Kaustuv Sanyal *PhD, FAAM (USA), FNA, FASc, FNASc*
Professor & Tata Innovation Fellow

Visiting Professor
Graduate School of Frontier Biosciences
Osaka University
Suita, Osaka 565 0871, Japan

15 May 2020

Certificate

This is to certify that the work described in this thesis entitled "Chromosome-level genome assembly of the human fungal pathogen *Candida tropicalis*" is the result of investigations carried out by **Mr. Krishnendu Guin** in the Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India, under my supervision and guidance. The results presented here have not previously formed the complete basis for the award of any other diploma or degree.

Kaustuv Sanyal, Ph.D
Professor
Molecular Mycology Laboratory
Molecular Biology and Genetics Unit
Jawaharlal Nehru Centre for
Advanced Scientific Research
Jakkur Post, Bangalore - 560 064, India.
Ph :+91-80-22082878, Email: sanyal@jncasr.ac.in

Place: Bangalore

Acknowledgements

I take this opportunity to acknowledge the individuals whose pertinent contribution(s) enabled me to present this thesis.

I am grateful to my Ph.D. supervisor Prof. Kaustuv Sanyal who has believed in my abilities and provided me with the opportunities to work on different projects. I cannot thank him enough for the patience, constant guidance, and enthusiasm with which he provided me the formal training for scientific research. During these years, I found inspiration from him to develop a curiosity-driven attitude to take up and actively pursue any scientific problem.

I thank all the members of MBGU faculty, Professors MRS Rao, Uday Kumar Ranga, Anuranjan Anand, Maneesha Inamdar, Tapas Kumar Kundu, Hemalatha Balaram, Namita Surolia, Dr. Ravi Manjithaya, Dr. James Chelliah, and Dr. Kushagra Bansal for the stimulating discussions during the annual departmental work presentations. I also acknowledge Prof. Hemalatha Balaram and Dr. Ravi Manjithaya for teaching various aspects of protein chemistry and membrane trafficking, respectively, during the formal course works, which helped me think broadly. I want to convey my gratitude to Prof. Uday Kumar Ranga for supporting the Oxford Nanopore sequencing experiment with reagents and instruments, Prof. Namita Surolia, for allowing us to run the CHEF-gels in her lab, and Dr. Ravi Manjithaya for extending his advice as well as his lab facilities for my project. I convey my regards to Dr. Meher Prakash from Theoretical Science Unit, JNCASR, for providing computer facility to perform 3C-seq data analysis initially.

This work is the outcome of a fruitful collaborative effort with Prof. Amartya Sanyal's laboratory from Nanyang Technological University, Singapore, and Prof. Geraldine Butler's laboratory in University College Dublin, Ireland. Siti Rawaidah B. M. Muzaki and Yao Chen from Prof. Amartya Sanyal's laboratory generated the 3C-seq library and performed data analysis, respectively. Caoimhe E. O'Brien from Prof. Geraldine Butler's laboratory performed the initial scaffolding of *C. tropicalis* genome. It was a great experience to be part of the engaging scientific discussion, which often straightened out technical problems and taught me the value of teamwork.

I sincerely acknowledge the funding support from CSIR, Govt. of India, in the form of Shyama Prasad Mukherjee fellowship [07/733(0181)/2013-EMR-I] and financial assistance from JNCASR. I also acknowledge the funding support from JNCASR and EMBO travel grant, enabling me to attend a conference on genome evolution, which improved my understanding of the subject to a great extent. While attending several meetings, I enjoyed the lively discussion with Prof. Wolf-Dietrich Heyer, Prof. Akira Shinohara, Prof. Christophe d'Enfert, Prof. Joseph Heitman, Prof. Scott Keeney, Prof. Tatsuo Fukagawa, Prof. Anna Selmecki, which always taught me something new. I wholeheartedly thank Prof. Ganesh Nagaraju and Prof. Utpal Nath, for their continuous support and critical inputs on the progress of my Ph.D. work during the interim evaluation process.

The in-house facilities at JNCASR have been beneficial in my work. Clevergene Biocorp has helped us with Illumina sequencing experiments. Many thanks to Mr. Tony Jose, Dr. Reddy, and Amrita. I thank Suma ma'am from the confocal facility, Dr. Narendra, from the flow cytometry facility for their continuous support. Thanks to JNCASR admin staff, academic staff, and MBGU office for their efficiency and making things so easy for us. I would also like to express my gratitude to the JNCASR administration and gardeners for maintaining a picturesque campus, which made my stay at JNCASR enjoyable. I acknowledge the doctors and health staff at Dhanvanthri, security staff, mess workers, cleaning persons, utility storekeeper Raju and his family, among others whose continuous efforts during everyday chores go unnoticed.

While looking back to the start of this eventful journey, I feel like being part of a family and cherish the memories I have shared with my senior and junior lab mates. I sincerely acknowledge my seniors Gautam, Sreyoshi, Laxmi Sir, Tanmay, Abhishek, Sreedevi, Arti, Neha, Vikas, and Lakshmi, for teaching me how to design and perform experiments, analyze the experimental data when it works and most importantly, troubleshoot when it fails. I express my heartfelt gratitude for getting a place among these people, who are not only excellent researchers but also admirable human beings. I was lucky to spend quality time with three of my labmate-batchmates, Sreyas, Sundar, and Rima, whom I revere as terrific researchers. I acknowledge Shweta, Aswathy, and Hasim for engaging discussions on science and beyond. I thank my juniors in the lab, including Priya J., Jigyasa, Aditi, Radha, Bhagya, Priya B., Satyadev, Kuldeep, Rashi, Tejas, Padmalaya, and Ankita for their help and support. There is no better way to learn than to teach. I cherish the fond memories of teaching the summer and POBE trainees, including Dipashree, Bhavya, Ninadini, Kajal, Chitra, Shakur, Medha, among others, who brought a breeze of fresh air in the lab with their cheerful personality. Last but not least, I convey my sincere regards to Nagraj, whose relentless efforts ensuring a steady supply of reusable glassware, plasticware, and other reagents to a lab of almost 20 enthusiastic researchers, deserve a BIG shout out.

I am indebted to my friend Anand, Samarth, Sunanda, Anshul, Pawan, Adma, Manjeet, among others, for either forgetting or celebrating my birthday on a wrong day! I would like to specifically mention Jiyada, Sunitadi, and Parvin, with whom I could only spend very little but quality time. I cherish the dinner table discussion with these people, Niloyendu, Manodeep, Brijesh, Navneet, Badri, Aditya, Vijaya, Avijit, Pallavi, and Shashank, often blended with large doses of a good laugh or two. If not for them, my Ph.D. would have lacked much of the philosophy part. I am grateful to my departmental seniors and friends, including Shashank bhaiya, Vijay, Shweta, Shalini, Deboshree, Arnab, Lakshmisha, Amrutha, Arpit, Suchismita, Aparna, Palak, Arun, Ananya, Harshit, Shuvangini, Veena, Gaurav who always found a moment to help me out with discussion or assistance during experiments. Special thanks to Harshit for helping me with the Oxford Nanopore sequencing experiment. I wish to acknowledge my friend Anjali who has been supportive of both academic and non-academic matters. I am grateful to Amrita, who have been very enthusiastic about life and ready to talk about literally anything under the sun. I wish to thank my friends Rumesh, Swarup, Somnath, Srijani, Ananya, Satbeer for their years of cheerful company and support.

I express my respect and gratitude to my teachers, including Professors Kanti Pradhan, Somnath Bhattacharya, R. K. Jain, Sunita Jain, S. K. Sethi, and Dhiraj Singh, among others, who were instrumental in guiding me to the path of academic research. I acknowledge Satyajit Ray, Arthur Conan Doyle, Sunil Ganguli, Frédéric Chopin, Richard Dawkins, Steven Spielberg, A. P. J. Abdul Kalam, J. K. Rowling, Frank Darabont, Wolfgang Amadeus Mozart, George R. R. Martin, Albert L. Lehninger, Rabindranath Thakur, and many other great personalities whose work have shaped up my emotional world and inspired to live life to the fullest.

Last but not least, I am grateful to my ma (Santwana Guin) and baba (Ranjan Kumar Guin), for half of their genomes reside in each of my body cells, literally making me who I am. I acknowledge the loving company of my elder sister, didi (Mahua Guin Nandi), who imparted the habit of reading from a very young age and taught me to strive for knowledge. I owe my gratitude to my brother-in-law (Shuvendu Nandi) and nieces (Sanchari and Samriddhi), for their unconditional love and support, without which I would never have managed to reach this point.

-Krishnendu Guin

Table of contents

Chapter 1: Introduction

- *Candida* species complex 2
- *Candida tropicalis,* an emerging human pathogenic budding yeast species 5
- Evolution of DNA sequencing techniques 7
- Long-read/ 3rd generation sequencing techniques 8
- Development of computational algorithms for analysis of genome sequencing data 10
- Higher-order genome organization in eukaryotes 15
- Available methods for studying genome organization 16
- Structure and function of topologically associated domains (TADs) 21
- Higher-order chromatin structure in yeasts 23
- Centromere 24
- Point centromere 29
- Short regional centromere 31
- Long regional centromere 32
- Mosaic centromere 33
- Establishment and propagation of centromere identity 33
- Structure and properties of centromeric chromatin 37
- Regulation of centromere identity in space and time 39
- Cause and consequence of centromere mediated genome rearrangements 42
- Rationale of the present study 44
- Available information and plan for construction of chromosome-level genome assembly of *C. tropicalis* 44
- Summary of the present work 47

Chapter 2: A gapless assembly of *Candida tropicalis* genome in seven chromosomes

- Construction of chromosome-level assembly of *C. tropicalis* 49
- Validation of the genome assembly with CHEF gel and chromoblot analysis 59
- Large CNVs lead to copy-number dependent fluconazole resistance in *C. tropicalis* 61
- Identification of SNPs and indels in *C. tropicalis* strain MYA-3404 69
- Haplotype phasing of MYA-3404 genome 71

Chapter 3: Higher-order genome organization in *Candida tropicalis*

- Conserved principle of genome organization in *C. tropicalis* 77
- Co-localization of CENP-A^{Cse4} with DAPI stained genome in *C. tropicalis* 78

- Localization of inner and outer kinetochore proteins in *C. tropicalis* 78
- Analysis of chromosome conformation capture sequencing (3C-seq) data reveals a conserved Rabl conformation of chromosomes in *C. tropicalis* 79
- Evidence of topologically associated domains (TADs) in *C. tropicalis* 82

Chapter 4: Genomic rearrangements and centromere-type transition in the CUG-Ser1 clade species

- Rapid evolution of centromere-types in closely related members of the CUG-Ser1 clade 88
- Mitotic stability assay of the centromeric constructs 89
- Structure and sequence dependent *de novo* CENP-A^{Cse4} recruitment on the *C. tropicalis* CEN-ARS plasmid 90
- Identification of homogenized inverted repeat (HIR)-associated centromeres in closely related CUG-Ser1 clade species 92
- Analysis of the interchromosomal synteny breaks (ICSBs) in *C. tropicalis* genome 104

Chapter 5: Discussion 112

Chapter 6: Materials and methods

- Media, growth conditions and transformation 120
- Construction and confirmation of strains and plasmids 120
  - Construction and confirmation of *C. tropicalis* strains (CtKS101 and CtKS102) expressing Protein-A tagged CENP-A^{Cse4} 120
  - Construction of *C. tropicalis* strains expressing Nuf2-GFP (CtKG500) and CENP-C^{Mif2}-GFP (CtKG501) 122
  - Construction of *dup4* and *dupR* mutant strains 122
  - Construction of pARS2-λ, pCEN501, pCEN502 and pCaCEN5 plasmids 123
- Cell lysate preparation and western blot 124
- Indirect immunofluorescence microscopy 125
- Pulsed-field gel electrophoresis 125
- Preparation of high molecular weight genomic DNA 126
- SMRT sequencing of *C. tropicalis* strain MYA-3404 on PacBio sequel system 126
- Construction of the *de novo* SMRT assembly and contig stitching using SMIS 127
- Filling N-gaps 127
- Assembly of sub-telomeric regions 127
- Mapping of the orphan haplotigs using the de novo SMRT assembly 127
- Pilon polishing of the genome assembly 128
- Construction of aneuploids for confirmation of 128

heterozygosity of the OHs
- Deletion of *SCH9* in *C. tropicalis* 128
- Construction of a reporter strain by integration of *URA3* on Chr5 for isolation of Chr5 monosomic isolates 129
- Isolation and confirmation of the 2n-1 aneuploids for Chr5 129

▪ Library preparation and sequencing of the library DNA for chromosome conformation capture (3C-seq) 129
▪ 3C-seq data analysis 130
- Mapping of 3C-seq data, contact probability matrix generation, and visualization 130
- Contig scaffolding 131

▪ Identification of SNPs, indels and CNVs 131
▪ Haplotype analysis 132
▪ Assessment of the genome assembly completeness using BUSCO 132
▪ Oxford Nanopore sequencing of *C. sojae* strain NCYC-2607 132
▪ Illumina sequencing of *C. sojae* strain NCYC-2607 133
▪ De novo genome assembly of *C. sojae* strain NCYC-2607 133
▪ Synteny analysis 134
▪ Identification of the putative centromeres in the members of the CUG-Ser1 clade 134
▪ Construction of the phylogenetic tree 135
▪ Total RNA isolation 135
▪ Transcriptome analysis 136
▪ Mitotic stability assay 137
▪ Chromatin immunoprecipitation 137
▪ Data access 138

Chapter 7: References 139

❖ Appendix-I: List of strains used in this study 160
❖ Appendix-II List of primers used in this study 161
❖ Appendix-III List of plasmids used in this study 166
❖ Appendix-IV Script used for analysis of 3C-seq data using Homer 167
❖ Appendix-V Script used in SNP/indel analysis using GATK software 168
❖ Appendix-VI Script used in haplotype analysis using FALCON, FALCON-Unzip and FALCON-Phase software 170
❖ Appendix-VII Script used for RNA-seq data analysis 173

List of publications 174

List of tables and figures

Introduction

- Figure 1.1 Phylogram of Saccharomycotina, a subphylum of Ascomycota 2
- Figure 1.2 Parasexual mating cycle of *Candida* species 3
- Figure 1.3 Structural and numerical genomic alterations 4
- Figure 1.4 Cellular targets of various classes of antifungal drugs and key mechanisms of drug resistance in *Candida* species 6
- Figure 1.5 Euler's solution to the 'seven bridges of Königsberg' problem and the origin of the graph theory 11
- Figure 1.6 Schematic depicting key steps in variant detection using GATK software 13
- Figure 1.7 Landmark events of genome sequencing in the last 45 years 15
- Figure 1.8 Schematic of the key steps in a typical Hi-C experiment 16
- Figure 1.9 Schematic of the key steps in 3C-seq experiment 18
- Figure 1.10 A comparison of different variants of 'C-techniques' 19
- Figure 1.11 Hierarchical chromatin organization inside the cell nucleus 20
- Figure 1.12 Integration of contact probability data and the chromatin binding of cohesin and CTCF reveals TADs and sub-TAD structures 22
- Figure 1.13 Phylogenetic distribution of fungal species with known centromeres in Ascomycota, Basidiomycota, and Mucoromycota 28
- Figure 1.14 DNA sequence, structural, and chromatin properties of seven major fungal centromere types 30
- Figure 1.15 Molecular determinants of centromere formation in fungi 34
- Figure 1.16 Centromere mediated karyotype evolution in fungal species 41
- Figure 1.17 Identification of seven inverted repeat-associated centromeres in contig-level genome assembly of *C. tropicalis* 45
- Figure 1.18 Homogenized inverted repeat-associated centromere loci in *C. tropicalis* increases the possibility of mis-assembly 46
- Table 1.1 A comprehensive list of fungal species belonging to Ascomycota, Basidiomycota and Mucoromycota with known or predicted centromeres and their features. 25

Results

- Figure 2.1 Schematic showing the stepwise construction of the gapless chromosome-level assembly (Assembly2020) of *C. tropicalis* 49
- Figure 2.2 CHEF-karyotyping, analysis of *de novo* contigs and 3C-seq contact probability data indicate contig5 and contig6 are parts of the same chromosome 51
- Figure 2.3 Use of contact probability data from 3C-seq 52

experiment for correction of mis-assembly of contig13 in Assembly C
- Figure 2.4 Analyses of the sequence coverage data and *de novo* contigs suggest orphan haplotigs are heterozygous loci in *C. tropicalis* genome 54
- Figure 2.5 Genetic analysis of engineered monosomic strains to confirm heterozygosity of the orphan haplotigs 55
- Figure 2.6 Schematic of the strategy followed for N-gap filling and scaffolding of sub-telomeres 58
- Figure 2.7 CHEF-chromoblot analysis for validation of the assembly of Chr4 and ChrR 61
- Figure 2.8 Copy number variation in *C. tropicalis* strain MYA-3404 is correlated to increased gene expression 62
- Figure 2.9 The *dupR* mutants shows fluconazole sensitivity 64
- Figure 2.10 Dilution spotting assay to test fluconazole sensitivity of the *dup4* mutant strains 65
- Figure 2.11 The *dupR* mutants show compromised membrane function 66
- Figure 2.12 Genome-wide mapping of SNP/indels and CNVs in *C. tropicalis* 68
- Figure 2.13 Phasing of diploid genome of *C. tropicalis* using FALCON 72
- Figure 2.14 Analysis of chromoblots, 3C-seq contact probability data, *de novo* contigs and sequence coverage validates a balanced heterozygous translocation between Chr1 and Chr4 73
- Figure 2.15 Chromosomal features of *C. tropicalis* strain MYA-3404 as revealed in Assembly2020 74
- Table 2.1 Details of the genome assembly with 13 contigs after SMIS run 50
- Table 2.2 Assembly of sub-telomeres and filling up N-gaps in the genome assembly of *C. tropicalis* using *de novo* assembled contigs 57
- Table 2.3 Statistics for intermediate and final version of genome assemblies of *C. tropicalis* (MYA-3404) 59
- Table 2.4 Improvements of *C. tropicalis* genome assembly 60
- Table 2.5 Type of effects due to the SNPs 69
- Table 2.6 Type of effects due to the indels 70
- Figure 3.1 Nuclear localization of CENP-A^{Cse4} in *C. tropicalis* 77
- Figure 3.2 Localization of inner and outer kinetochore proteins in *C. tropicalis* 79
- Figure 3.3 The contact probability matrix of the *C. tropicalis* genome reveals significant *CEN-CEN* and *TEL-TEL trans* interactions 80
- Figure 3.4 The contact probability matrix of the *C. tropicalis* genome obtained from analysis of 3C-seq data using Homer 81
- Figure 3.5 Identification of putative TADs in the *C. tropicalis* genome 86
- Figure 4.1 DNA sequence dependent regulation of centromere identity in *C. tropicalis* 89
- Figure 4.2 Identification of SNPs and indels in the *C. sojae* 94

strain NCYC-2607
- Figure 4.3 Partial conservation of a LOH block in each of the *C. albicans, C. tropicalis* and *C. sojae* genome 95
- Figure 4.4 Identification of HIR-associated centromeres in the CUG-Ser1 clade 96
- Figure 4.5 Putative centromeres of *C. parapsilosis* are ORF-free and transcription poor loci similar to the centromeres of *C. albicans* 99
- Figure 4.6 Intra- and inter-species conservation in homogenized inverted repeat-associated centromeres in the CUG-Ser1 clade 101
- Figure 4.7 Identification of an inter-species conserved and centromere-enriched DNA sequence motif in CUG-Ser1 clade species 102
- Figure 4.8 Organization of IR-motifs on homogenized inverted repeat-associated centromeres of *C. tropicalis, C. sojae* and *C. viswanathii* 103
- Figure 4.9 Genome-wide mapping of ICSBs on *C. tropicalis* genome reveals spatial regulation of centromere-proximal translocations in their common ancestor 105
- Figure 4.10 ORF-level synteny of the centromere proximal loci in *C. tropicalis* with respect to *C. albicans* genome 107
- Figure 4.11 Genome-wide synteny analysis between *C. albicans* and *C. tropicalis* finds evidence of inter-centromere translocations in the last common ancestor 109
- Figure 4.12 Centromere-proximal interchromosomal rearrangements leads to loss of inosine-uridine nucleoside N-ribohydrolase homolog 110
- Table 4.1 Statistics for the *C. sojae* genome assembly 93
- Table 4.2 Genomic coordinates of putative HIR associated centromeres in *C. sojae, C. viswanathii,* and *C. parapsilosis* 98
- Table 4.3 Length of the centromere DNA elements in *C. sojae* 100
- Table 4.4 Length of the centromere DNA elements in *C. viswanathii* 100

Discussion
- Figure 5.1 The spatial genome organization remained conserved in the CUG-Ser1 clade despite centromere type diversity 116

Materials and methods
- Figure 6.1 Construction and confirmation of CENP-A^{Cse4}-Protein-A tagged strain (CtKS101) of *C. tropicalis* 121
- Figure 6.2 Double homologous recombination mediated deletion of CNV loci and PCR confirmation 123
- Figure 6.3 Schematic of the steps followed during isolation of DNA-free total RNA from *C. tropicalis* cells 136

Abbreviations

%	Percent
ºC	Degree Celsius
3C	Chromosome conformation capture
3D	Three dimensional
4C	Circular chromosome conformation capture
5-FOA	5-Fluoroorotic acid
5C	Chromosome conformation capture carbon copy
ABC	The ATP-binding cassette
ANOVA	Analysis of variance
ARS	Autonomously replicating sequence
BAC	Bacterial artificial chromosome
BAM	Binary alignment map
BED	Browser extensible data
BLAST	Basic local alignment search tool
bp	Base pair
CDD	Conserved domain database
CDE	Centromere DNA element
CEN	Centromere
CENP-A	Centromere protein-A
CGH	Comparative genome hybridization
CHEF	Contour clamped homogenized electric field
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP sequencing
Chr	Chromosome
cm	Centimetre
CNV	Copy number variation
CPU	Central processing unit
CTAB	Cetyltrimethyl ammonium bromide
CTCF	CCCTC-binding factor
DAPI	4, 6-Diamino-2- phenylindole
DBG	de Burjin graph
DI	Directionality index
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide

DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
ENC	Evolutionary new centromere
EtBr	Ethidium bromide
FISH	Fluorescence in situ hybridization
FRET	Fluorescence resonance energy transfer
g	gram
GATK	Genome analysis tool-kit
GFP	Green fluorescent protein
h	Hour
HIR	Homogenized inverted repeat
HR	Homologous recombination
ICE	Iterative correction and eigenvector decomposition
ICSB	Interchromosomal synteny breakpoint
ICU	Intensive care unit
IGV	Integrated genomics viewer
IR	Inverted repeat
IS	Insulation score
kb	kilo base
kDa	kilodalton
KT	Kinetochore
M	Molar
Mb	Megabase
MD	Molecular dynamics
MFS	Major facilitator superfamily
mg	milligram
min	minute
ml	millilitre
mM	milli molar
MNase	Micrococcal nuclease
MTL	Mating type locus
MYA	Million years ago
N	Normal
NCAC	non-*Candida albicans Candida*
NCBI	National centre for biotechnology information
NGS	Next generation sequencing

NMR	Nuclear magnetic resonance
NTC	Nourseothricin
OD_{600}	Optical density at 600nm
OH	Orphan haplotig
OLC	Overlap-layout-consensus
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
psi	Pounds per square inch
RE	Restriction enzyme
RIN	RNA integrity number
RIP	Repeat-induced point mutation
RITS	RNA-induced transcriptional silencing
RNA	Ribonucleic acid
RNAi	RNA interference
rpm	Revolutions per minute
RT	Robertsonian translocation
s	Second
SAM	Sequence alignment map
SAXS	Small-angle X-ray scattering
SDS	Sodium dodecyl sulphate
SMRT	Single Molecule Real-Time
SNP	Single nucleotide polymorphism
SPB	Spindle pole body
TAD	Topologically associated domain
TAP	Tandem affinity purification
TCA	Trichloroacetic acid
ZMW	Zero-mode waveguides
µg	Microgram
µl	Microlitre
µm	Micrometre
µM	Micromolar

Chapter 1

Introduction

Candida species complex

Candida species are the most common cause of local or systemic fungal infection in immunocompromised humans. These species are members of the CUG-Ser1 clade (Figure 1.1), of the fungal phylum of Ascomycota. The majority of *Candida* infections are caused by five species: *Candida albicans, Candida glabrata, Candida tropicalis, Candida parapsilosis*, and *Candida krusei* (1, 2). In addition, cases of *Candida auris* infection are rapidly emerging

Figure 1.1 Phylogram of Saccharomycotina, a subphylum of Ascomycota.
Major clades of fungal subphylum Saccharomycotina are shown in color-coded branches. Collapsed nodes bearing multiple species are represented with triangles. The phylogram was generated using Evolview (3) from the phylogenetic tree data presented in reference (4).

worldwide (5). Except for *C. auris,* these species are majorly clonally propagated and contain a diploid genome (6). Due to the clonal nature of reproduction, multiple clinical isolates bear identical patterns of loss of heterozygosity (LOH) at certain portions of their genomes. This phenomenon of LOH is well documented in the case of *C. albicans* clinical isolates (7, 8). Absence of meiosis in *C. albicans* is puzzling because certain *Candida* species, including a haploid species *Candida lusitaniae* undergoes sexual cycle and carries a similar set of meiosis-specific genes (9). However, it is worth noting that Ime1, the master regulator of the meiotic program in *Saccharomyces cerevisiae,* is absent in *Candida* species (6). Genome sequencing of several clinical isolates revealed rare genomic changes possibly generated through para-sexual reproduction (Figure 1.2) (7, 8). In line with this genomic evidence, a cryptic parasexual cycle has been first reported in an experimentally manipulated laboratory strain of *C. albicans* (10) and subsequently studied in both *C. albicans* and *C. tropicalis* (11-14). Moreover, the fusants obtained from artificially induced parasexual mating show fitness advantage (15). Recently, it was found that drug-induced mating competence and parasexual recombination led to the evolution of fluconazole-resistant *C. albicans* strains (16). However, direct evidence of the parasexual cycle in a natural cell population and its contribution towards the pathobiology of *Candida* species remain elusive.

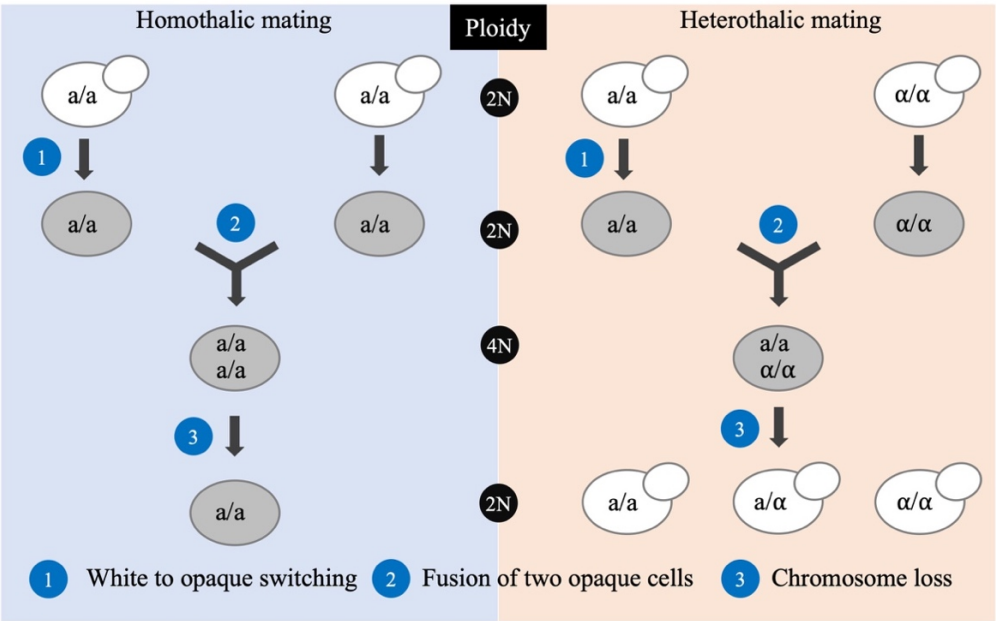


Figure 1.2 Parasexual mating cycle of *Candida* species.
Schematic showing three key steps during the homothallic and heterothallic mating cycle in *Candida* species. Both homothallic and heterothallic mating requires fusion of two diploid opaque cells, which produce an unstable tetraploid intermediate. The tetraploid intermediate undergoes random loss of chromosomes and achieves a stable diploid state (also termed as the parasexual cycle). Compared to the white cells, 'opaque cells' are larger in size, more

elongated, and form darker and flat colonies on solid agar (17). Later, it was identified that the opaque cells mate 10^6 times more efficiently than the white cells (18). The figure was adapted from reference (19).

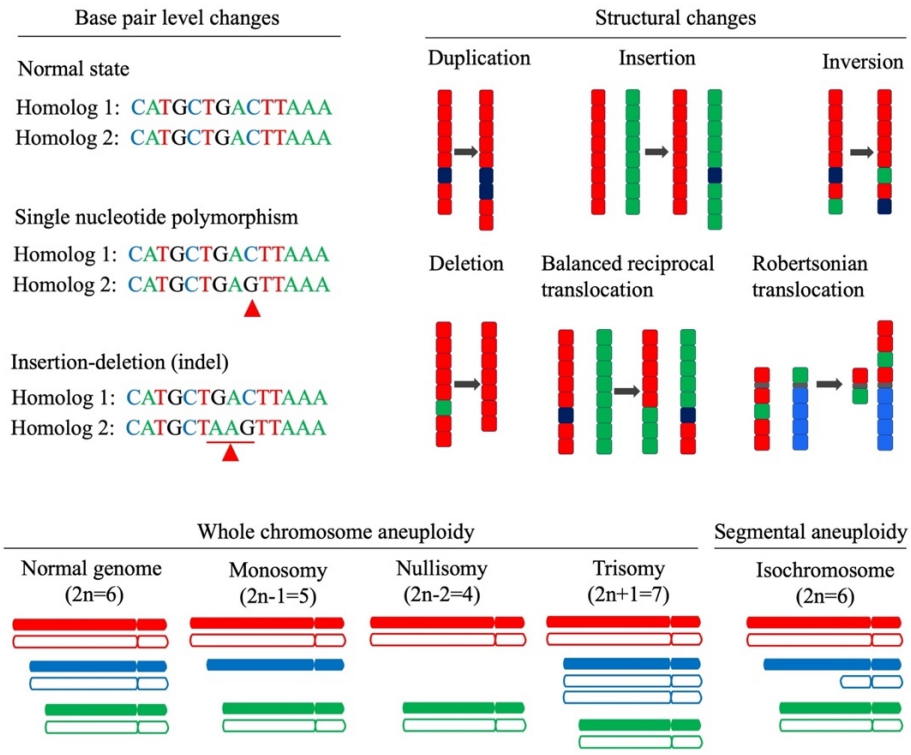


Figure 1.3 Structural and numerical genomic alterations.
Schematic shows examples of base substitutions, chromosomal structural changes, whole chromosome aneuploidy and segmental aneuploidy in a diploid genome.

The absence of true meiosis and crossovers limits the emergence of new alleles and compromise the fitness of a pathogenic species. However, strict selection under challenging host environments, especially the presence of antifungal drugs, often led to the emergence of isolates carrying large-scale structural and numerical changes (Figure 1.3) in their genomes (20). Among these large-scale genomic changes, the specific contribution of segmental aneuploidy and trisomy to the evolution of drug resistance has been studied in *C. albicans* (21). Similarly, whole chromosome aneuploidy of Chr1 and Chr4 is also associated with the development of fluconazole resistance in the human fungal pathogen *Cryptococcus neoformans* (22). Further, identification of isochromosome formation in drug resistant isolates of *C. albicans* revealed a correlation between fluconazole resistance and increased copy number of genes involved in the biosynthesis of ergosterol (23-25), which is a component of the cell membrane in fungi. Many of these genome-wide studies could be

conducted in *C. albicans* due to the availability of a chromosome-level genome assembly (26-30). Although these studies demonstrated an association of genomic alterations with drug resistance in *C. albicans,* the validity of this notion remains to be explored in *C. tropicalis* in the absence of a chromosome-level genome assembly.

Candida tropicalis, **an emerging human pathogenic budding yeast species**

In the last 30 years, there has been a significant increase in the incidence of fungal infections in humans (31). Among different fungi, the members of the genus *Candida* are the most frequently recovered species from human fungal infections. Until recently*, C. albicans* was considered to be the major *Candida* species involved in human fungal infections worldwide. However, in parallel with the overall increase of fungal infections, it has been observed that infections caused by non-*Candida albicans Candida* (NCAC) species are increasing (32-34). It was found that *C. tropicalis* is the most frequently isolated NCAC species from the bloodstream and urinary tract infections worldwide and particularly in India (35). Additionally, *C. tropicalis* is often found in patients admitted to ICUs, especially in patients with cancer, requiring prolonged catheterization, and receiving broad-spectrum antibiotics (36, 37). *C. tropicalis* possesses a diversity of virulence factors that induce serious damage to patients and increases the mortality risk (38). For example, this species appears to possess a higher potential for dissemination in the neutropenic host than *C. albicans* and other NCAC species. This propensity for dissemination in some way may explain the reported relatively high mortality associated with *C. tropicalis* (39, 40). However, more work is necessary to get deeper insights into the strategies used by *C. tropicalis* to change from a harmless commensal microorganism to become a human pathogen of serious clinical concern.

Globally, the infections due to *C. tropicalis* have been increasing steadily, proclaiming this organism to be an emerging pathogenic yeast (41). In the Asia-Pacific region, NCAC species constitute 50 – 80% of all isolates. The incidences of candidemia vary from 1-12 per 1000 ICU admissions in India. Another worrying trend has been the emergence of triazole resistance in some of the centers of India, with 10-14% of *C. tropicalis* isolates demonstrating resistance to antifungals (42). The reasons for this organism's prevalence and its resistance to fluconazole have been difficult to elucidate. In a recent report published in 2015, the authors studied 1400 ICU acquired candidemia cases out of which 918 *Candida*

strains were collected. Among these isolates, *C. tropicalis* (n = 382; 41.6%) was the most prevalent species followed by *C. albicans* (n = 192; 20.9%) and

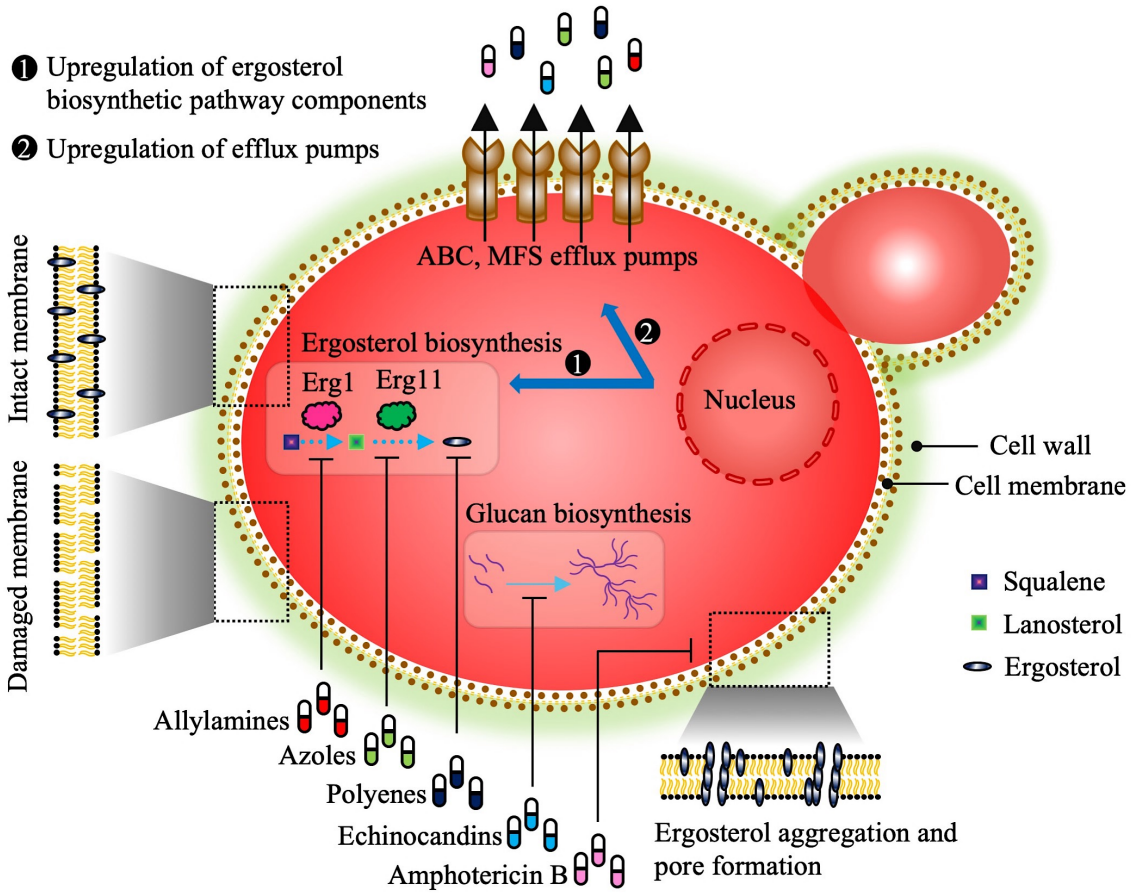


Figure 1.4 Cellular targets of various classes of antifungal drugs and key mechanisms of drug resistance in *Candida* species.
Various classes of antifungal drugs, including allylamines, azoles, polyenes, echinocandins target glucan and ergosterol biosynthetic pathway components (*e.g.,* Erg1 and Erg11) or destabilize ergosterol. In the absence of ergosterol, both the structure and function of the cell membrane are compromised. Similarly, amphotericin B leads to ergosterol aggregation on the cell membrane, which causes pore formation and damages membrane integrity. Cells that show resistance against these drugs often show upregulation of genes implicated in ergosterol biosynthetic pathway (1), and genes encoding efflux pump proteins (2), which can compensate the membrane damage, and reduce the intracellular drug concentration, respectively.

C. parapsilosis (n = 100; 10.9%). Drug resistance, as well as multidrug resistance, is also being reported in *C. tropicalis* clinical isolates (43). Among these 918 isolates multi-drug resistance was noted in *C. tropicalis* (n = 4; 23.5%), *C. auris* (n=4; 23.5%) and *C. krusei* (n=3; 17.6%) (43). Apart from India, *C. tropicalis* was also reported to cause the majority of ICU-acquired candidemia in Pakistan (44). In *C. albicans*, the most common mechanisms of drug resistance (Figure 1.4) involve mutations in a target gene (*e.g., ERG11*) of the drugs or

over-expression of efflux pumps like ABC transporters (*CDR1* and *CDR2*), and the major facilitator super-family efflux pump (MFS-MDR) (45). Another commonly found mechanism of drug resistance in *C. albicans,* is mediated through genome rearrangements such as whole chromosome and segmental aneuploidy (25, 45). The possibility of such genome rearrangements as a mechanism for drug resistance in *C. tropicalis* has not been studied in detail and needs to be explored. However, fragmented genome assembly remains the key limitation to perform such studies in *C. tropicalis.*

In addition to drug resistance, the mechanism of this organism's pathogenicity and the consequent immune response in the host remain to be determined. In order to understand the effects of genome rearrangements on the pathobiology of *C. tropicalis*, a properly annotated gapless chromosome-level genome assembly is essential. Though the genome sequencing of *C. tropicalis* has been done at Broad Institute, MIT (*Candida* Sequencing Project, Broad Institute of Harvard and MIT; http://www.broadinstitute.org), the supercontigs have not been assigned to individual chromosomes (6). Moreover, identification of ORF present in the *C. tropicalis* genome but absent in the fragmented genome assembly indicates the scope of further improvement in the genome assembly of *C. tropicalis* (46). However, the construction of a reliable genome assembly of *C. tropicalis* chromosomes with homogenized and inverted repeat-associated centromeres can be a challenging task to accomplish. Nevertheless, recent improvements in the sequencing techniques offer a critical advantage in this area.

Evolution of DNA sequencing techniques

The famous discovery of the double-helical structure of the DNA by J. D. Watson and F. H. C. Crick opened up a new area of molecular biology research (47). Subsequently, it was demonstrated how elegant structural properties of complementary base pairing between two antiparallel strands in a DNA double helix facilitates a semi-conservative model of DNA replication (48). These fundamental studies formed the bedrock for recent advances in sequencing technology. Determination of the exact order of nucleotides in a DNA polymer was first demonstrated independently by Sanger *et al.* (49) using a chain termination reaction mediated by incorporation of dideoxynucleotides and Maxam and Gilbert (50) using a chemical cleavage procedure. Sanger's approach of DNA sequencing was commercialized by Applied Biosystem Instruments (ABI), which constitute the first-generation sequencers.

Even though the Sanger and Maxam-Gilbert sequencing methods offer a high level of accuracy, both lack the speed and throughput required for the sequencing of larger genomes in shorter times. Therefore, alternative techniques have been developed. Among these methods, sequencing by hybridization (51), ligation-cleavage (52), and pyrosequencing (53) were widely used. The method of 'pyrosequencing' works by detecting the incorporation of a nucleotide through the luminescence of a specific fluorophore attached to each of the four nucleotides. This method quickly gained attention because of its scalability and amenability to automation, which are characteristic features of the second-generation sequencers. The first commercially available pyrosequencing was developed by Roche/454 on the GS-FLX platform. Parallelly, the first Solexa sequencing platform 'Genome Analyzer' was developed that utilized four unique-dyes for labeling of nucleotides (54). Ligase-based chemistry of sequencing was applied in another popular approach developed by Applied Biosciences termed as 'SOLiD'. All the next generation sequencing (NGS) techniques described above, works through an imaging-based detection method. An alternative approach to determine the change in current by measuring the change in pH due to a single proton release was implemented in the 'Ion-torent' technique developed by Thermo Fisher.

These second-generation sequencing platforms were developed and employed during the human genome sequencing project (HGP) and led to rapid development in NGS technologies. Consequently, the cost of sequencing was reduced dramatically and facilitated the sequencing of several eukaryotic genomes. However, the quality of the genome assemblies produced using these methods can vary between the contig level to chromosome-level, depending on the complexity of the genome and the coverage of NGS data. Another confounding factor is the read length of the NGS data. Despite the advantage of speed and scalability, the length of the reads obtained from the second generation NGS techniques remained short, in the range of 28-700 bp. This poses a challenge to assemble the repetitive loci, often spanning over a few Mbs. The development of long-read sequencing techniques offered a possible solution to this problem.

Long-read/ 3rd generation sequencing techniques

At the beginning of 2011, the PacBio RS platform was launched by Pacific Biosciences (PacBio). In this technique of Single Molecule Real-Time (SMRT) sequencing, the template molecules are size selected to enrich the long reads. The size selected DNA

fragments are then used to generate libraries, which are single-stranded circular hairpins known as SMRTbell. When the library is loaded in a SMRT cell, the SMRTbells diffuse into smaller cavities known as Zero-mode waveguides (ZMW). The SMRTbell then binds to the DNA-polymerase immobilized at the bottom of each ZMW. As the synthesis proceeds, during the incorporation of nucleotides, a distinct fluorescence signal is produced from each of the four nucleotides labeled with a specific fluorescence dye. Since the radius of the ZMWs is shorter than the wavelength of the light produced, it cannot escape the ZMW well and detected by a sensor located bellow the ZMWs.

The entire process of sequencing is recorded in a movie that provides the time information along with the fluorescence pattern. Together, these two data can be used to find the rate of nucleotide incorporation at a given base-pair, allowing identification of certain chemical modifications of the base present on the template DNA. The difference between a modified and an unmodified base is expressed as the IPD ratio. This approach can detect 4-methylcytosine, 5-methylcytosine, 5-hydroxymethylcytosine, 6-methyladenine, and 8-oxoguanine (www.pacb.com/basemod). Due to the size selection and the improved sequence chemistry, the PacBio SMRT sequencing approach can produce up to ~80 kb long sequence reads. SMRT-sequencing does not involve PCR amplification during the library preparation step, and hence the reads are free from PCR-duplicates, a common artifact found in Illumina and other NGS platforms. However, SMRT-sequencing suffers from an error rate as high as 15% in single-pass mode, which can be corrected further by increasing the coverage (55). Another limitation of SMRT-sequencing is that the read length is dependent on the processivity of the polymerase used for sequencing.

A recently developed approach, the Oxford Nanopore sequencing technique, was first commercialized in 2012. Oxford Nanopore sequencing is also available on three different scales: MinION, GridION, and PromethION. The extreme portability offered by the MinION system allows real-time sequencing of the samples at the site of collection. The currently available Oxford Nanopore MinION system can deliver reads more than 150 kilobases (56), but the read length can be extended even up to 1 Mb (57). Since the longer reads can provide larger scaffolds, one of the main applications of the Oxford Nanopore sequencing technique is to generate *de novo* genome assembly.

In this approach of sequencing, a processive enzyme attaches to the 5′ end of the template strand and guides the template strand through the nanopore. This process generates ionic current due to the differences in the shifting nucleotide sequences through the pore. This change in ionic current can be detected as separable events using a suitable sensor. This process of template strand transfer through the nanopore is known as '1D read'. Alternatively, with the help of a hairpin adapter ligated to one end of the DNA duplex, transfer of the complementary strand results in the sequencing of both the strands in a process known as '2D read'. Oxford Nanopore sequencing generates long high quality reads. For example, 1D reads over 300 kb, and 2D reads up to 60 kb have been achieved using the *Escherichia coli* genomic DNA template (58). Longer reads generated in this approach are extremely helpful in assembling the repetitive elements in a genome assembly. For example, 36-kb + MinION reads were used to resolve a putative 50-kb gap in the human Xq24 reference sequence (59). More recently, the same group could, for the first time, generate a linear assembly of centromere DNA of the human Y chromosome composed of a few mega base pairs of alpha-satellite repeats (57). These lines of work demonstrate the potential use of the Oxford Nanopore sequencing approach to assembly repetitive DNA sequences of the genome. Other advantages of this approach, similar to the PacBio SMRT sequencing technique, include the identification of modified bases and the absence of PCR duplicates. In addition, the Oxford Nanopore sequencing technique offers portability to remote areas and direct sequencing of freshly collected samples and generate sequencing data in real-time on the spot. However, this technique also suffers from an error rate as high as 15% in 1D mode (55), which can be compensated using higher coverage.

Development of computational algorithms for analysis of genome sequencing data

The development of sequencing techniques started producing a vast amount of sequence data, which needed efficient computational approaches for assembling a genome and subsequent analyses. During the sequencing of the *Drosophila* genome using the shotgun method, the Celera assembler was developed, which could perform *de novo* assembly from ~3 million reads in less than a week on an eight-core machine with 32 Gb memory (60). The Celera assembler follows an overlap-layout-consensus (OLC) approach. In the first step, the reads are used in a matrix-based all-versus-all pair-wise read comparison. Next, the overlap-maps are used to generate an approximate read layout to define the order and orientation of the reads, following which the multiple sequence alignment (61) based consensus is

generated and contigs are assembled. A similar OLC-based strategy was employed in other contemporary assemblers, including Arachne (62, 63), CAP, and PCAP (64) for shotgun sequence data and Newbler (65) for the Roche 454 platform. An improved OLC-based software EDNA (66) was designed for unpaired short reads of uniform length obtained from the SOLiD and Solexa platform. This assembler discards duplicate reads and generate error-free overlaps (67). However, extensive time requirement for the analysis was a major limitation of these methods. Therefore, more efficient algorithms were needed.

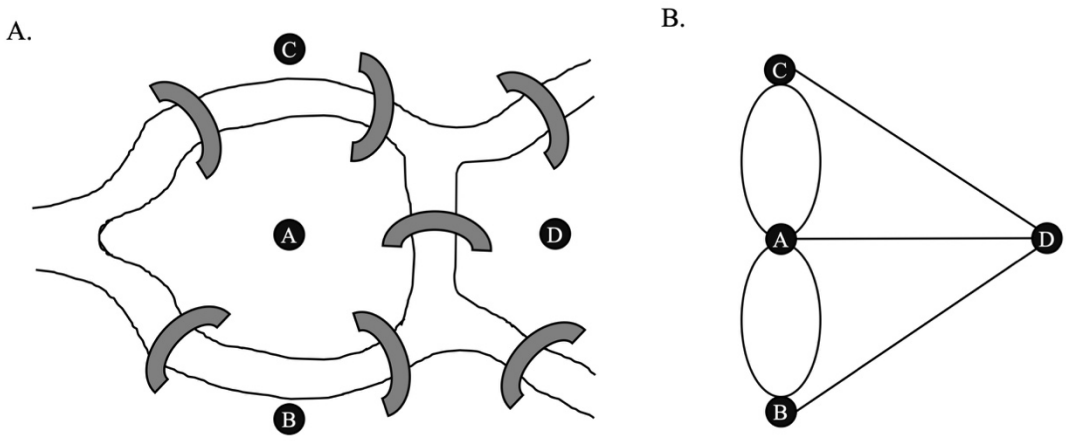


Figure 1.5 Euler's solution to the 'seven bridges of Königsberg' problem and the origin of the graph theory.
A. Schematic showing the seven bridges (gray arches) over the Pregel river that separated the four connected landmasses (black filled circles marked with A, B, C, and D) of Königsberg, Prussia. The challenge was to travel to all four parts of the city using each of these seven bridges only once. B. Graphical representation of seven edges (the seven bridges shown as black lines) connecting four vertices (connected landmasses). This graphical representation for 'seven bridges of Königsberg' problem allowed Euler to devise a general model to predict if a solution to such problems exists or not. One can find a solution to this problem if they write all possible paths. However, that would take a long time for problems involving more bridges and landmasses. The application of the graph theory can significantly reduce the time required to solve such problems.

An elegant solution for a computationally efficient approach of *de novo* genome assembly came from a 300 years old puzzle: the 'Bridges of Königsberg problem' (Figure 1.5). This problem was solved by Leonhard Euler in 1735, which opened up a new branch of mathematics: the graph theory. Euler's solution was later adapted by Nicolas de Burjin to find a cyclic sequence of letters taken from a given alphabet for which every possible word of a certain length (k) appears as a string of consecutive characters in the cyclic sequence exactly once (68). This strategy of construction of de Burjin graphs (DBG) was employed in SOAPdenovo (69) and Velvet (70) assembler. Although both OLC and DBG approaches

offer robust assembly, by avoiding the computationally exhaustive all-versus-all search, DBG offers better efficiency for a large volume of short-read sequence data over the OLC-based assemblers. Markedly different from these two approaches, a greedy-graph based algorithm is followed in SHARCGS (71) and SSAKE (72) assembler, where the basic operation is to add a read or a contig for a given read or contig. This operation is repeated until no more addition is possible. A comparative analysis of different assembly strategies employed in a set of 24 academically available *de novo* genome assemblers suggests DBG-based assemblers perform better for the large genomes using large datasets, while the OLC-based assemblers are preferred for assembling smaller genomes and smaller data sets (73).

Both DBG- and OLC-based assemblers facilitated the construction of *de novo* genome assembly of a large number of prokaryotic and eukaryotic genomes using short reads. However, the short-read assembly often breaks at the repetitive regions of the genome. Initially, two solutions were proposed to address this problem: (a) end sequencing of a large piece of cloned DNA (*e.g.,* BACs) (74) and (b) use of paired-end sequence data from longer-fragment libraries. However, it was found that none of these two approaches can effectively assemble all types of repetitive regions (75). In such cases, long reads obtained from PacBio-SMRT-seq, Oxford Nanopore, or other platforms can be used to improve the contiguity of the genome and resolve the repetitive regions accurately. One major caveat of the long-read sequencing techniques is the high error rate. Therefore, a solution was proposed in the form of a hybrid assembly technique combining the accuracy of short-read sequences and the contiguity of the multi-kilobase long reads. In this method, high-quality short reads are used to correct the PacBio long reads, which achieves >99.99% base-call accuracy. These corrected reads are then used for the construction of *de novo* genome assemblies following the OLC approach. The use of this method also demonstrated a significant improvement in contiguity over the second-generation assemblies (76). Later along with the development of the Oxford Nanopore sequencing technique, Canu (77) was developed, which could integrate single- or paired-end sequence data from various platforms. In addition, Canu generates graph-based assembly outputs for integration with complementary phasing and scaffolding techniques (77). This advantage of long-read-based assembly for phasing the haplotype difference in the diploid genomes was further improved in FALCON-Unzip (78), Supernova (79), and Purge-Haplotigs (80). The phasing of haplotype and the contiguity of the genome assembly was shown to be further improved when the long-read sequencing data was combined with the contact probability information obtained from chromatin conformation

capture sequencing (3C-seq) or Hi-C based experiments (81). These tools offer a promising solution for haplotype phasing for species where sexual reproduction is rare or absent. However, when available, a combination of long read sequence data from parental and offspring genome can be used to develop high quality phased genome assembly (82).

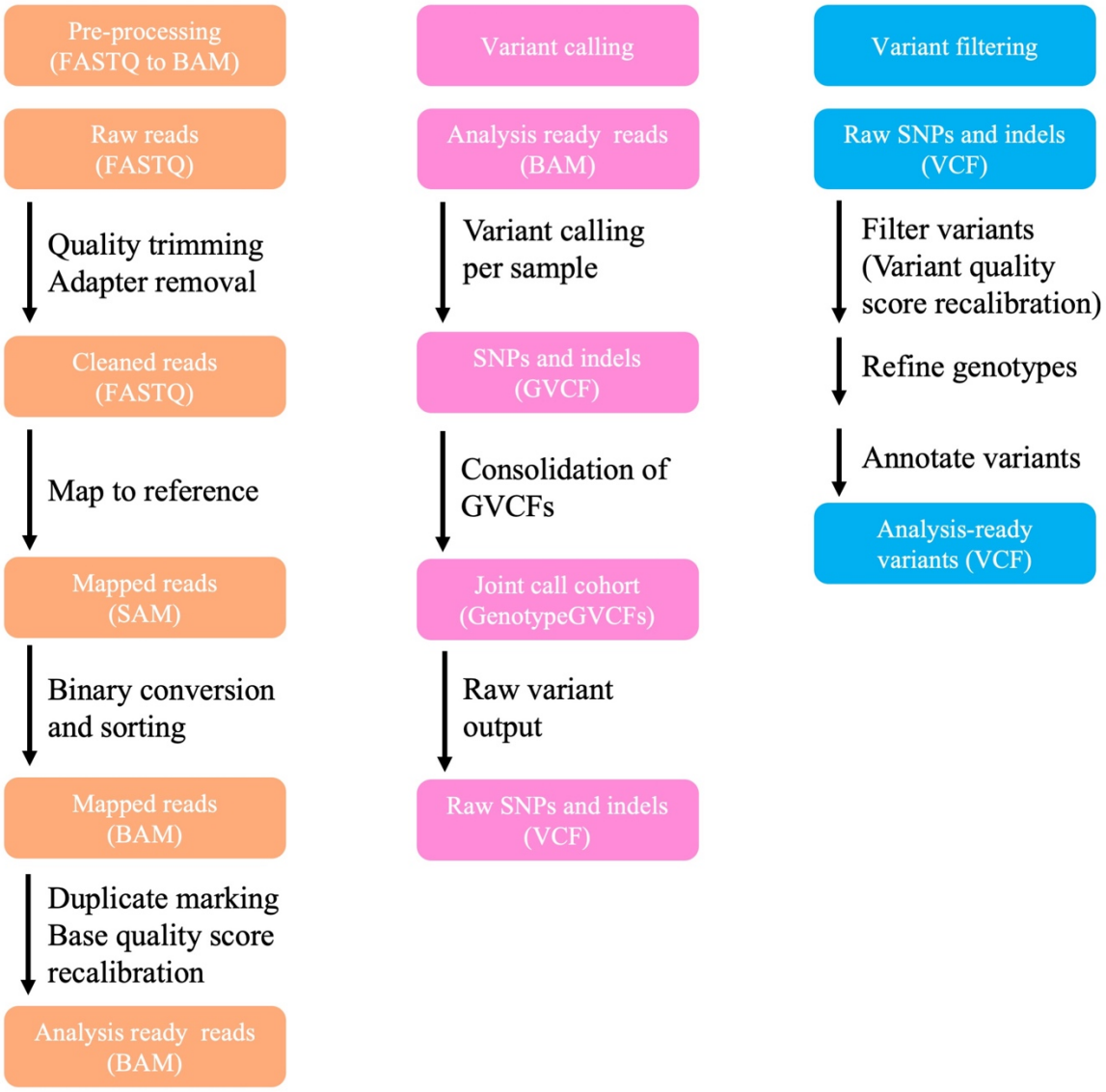


Figure 1.6 Schematic depicting key steps in variant detection using GATK software.

In a diploid genome, two alleles of the same genes are distributed on the homologous chromosomes. A given locus is called 'heterozygous' when the two alleles are not identical to each other. The heterozygosity can arise due to three major ways. First, the two alleles can differ by one single nucleotide, which is termed as single nucleotide polymorphism (SNP). Second, insertion or deletion of one or a few bases known as indels. The third type of heterozygosity arises due to a deletion or a duplication of a region spanning across multiple

kb of a genome and termed as copy number variation (CNV). Another type of structural variation might arise between the two homologs of a chromosome due to a balanced translocation. During the early days of cytogenetics, structural genomic variations were studied using staining of metaphase chromosome spread and fluorescence in situ hybridization (FISH). A major limitation of these assays was the lack of resolution. Later, the development of hybridization-based techniques such as comparative genome hybridization (CGH) facilitated the detection of smaller variations, including CNVs, SNPs, and indels, but failed to detect chromosomal translocation events (83).

Initial genome assemblies developed using the shotgun method for inbred lines of mouse and selfing hermaphroditic species *Caenorhabditis briggsae* carried little heterozygosity as expected (84). Later, the development of improved assembly algorithms allowed the construction of genome assemblies, even for the heterozygous genomes. The availability of whole-genome sequence data of different isolates of a single species helped to understand the polymorphism present in a population. Genome Analysis Tool-kit (GATK) is one of the most popular and computationally robust pipelines developed during the 1000 Genomes Project and the Cancer Genome Atlas (85). GATK is developed using the programming principle of MapReduce (85). By doing so, GATK can distribute large datasets into clusters and allow efficient use of parallel computation. GATK can also be used for the identification of structural variations in the diploid genome by implementing the base quality recalibration and indel realignment in the HaplotypeCaller (86). A flow chart describing the key steps in the variant calling procedure in GATK is presented in figure 1.6.

In the last 45 years, the rapid development of sequencing techniques and improved assembly algorithms initially allowed genome sequencing of the well-studied model organisms as well as other closely related species across various domains of life (Figure 1.7). Huge resources of genome sequence data eased reverse genetics and molecular dissection of various aspects of evolution, disease, development, drug resistance, among others. With the availability of more genome-wide data sets, the influence of spatial genome organization on the regulation of DNA replication, chromosome segregation, genome maintenance, and expression of gene-sets became clearer. Although the linear one-dimensional DNA sequence of the chromosomes provides little or no information about its spatial organization inside the nucleus, a gapless genome assembly empowered further analysis of the higher-order genome organization.

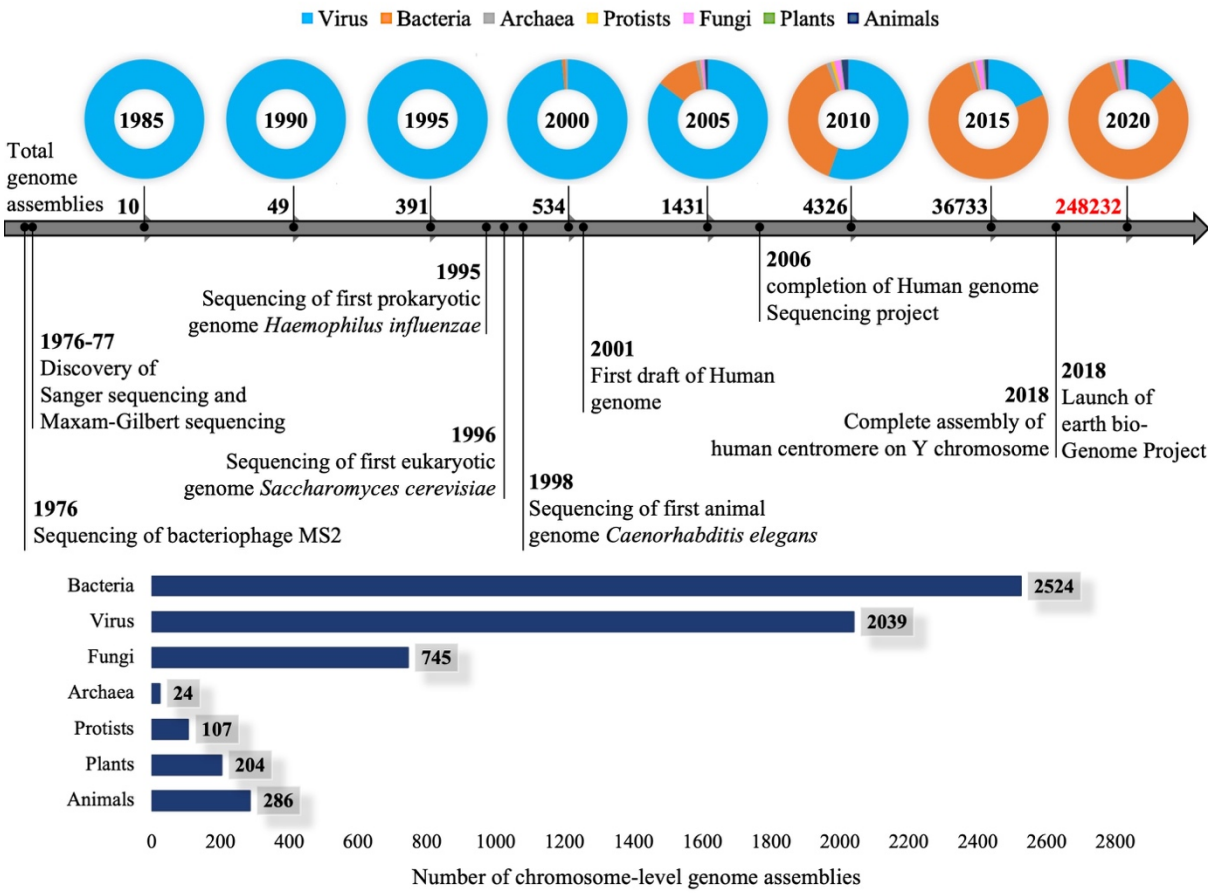


Figure 1.7 Landmark events of genome sequencing in the last 45 years.
Color-coded doughnut plots showing the relative proportions of genome assemblies across different domains of life available at different time intervals in the NCBI Assembly database (87) during the last 45 years. The bottom panel shows a bar chart depicting the total number of chromosome-level genome assemblies available for species belonging to different domains of life till April 2020.

Higher-order genome organization in eukaryotes

One of the most significant changes during the evolution of Eukarya is the partitioning of the nuclear genome into linear chromosomes. Each chromosome contains a polymer of DNA organized into a compact three-dimensional structure with the help of several structural proteins. The fundamental unit of chromatin packaging is a nucleosome, which was initially identified as 'beads on a string' structure (88-91). In general, a single nucleosome core particle comprises of histone octamer: two molecules of each of histone H2A, H2B, H3 and H4 and wrapped with ~147 bp DNA double helix in about 1.75 left-handed super helical turns providing ~7 fold compaction into a disk-like structure of about 11 nm in diameter and 5.5 nm in height (92, 93). This level of compaction is not enough to fit

the entire genome inside the nucleus. Therefore, a better understanding of the higher-order chromatin structure is the need of the hour.

Available methods for studying genome organization

The fundamental unit of the chromatin is a nucleosome. The structure of a single nucleosome can be studied using various structural techniques such as fluorescence resonance energy transfer (FRET), small-angle X-ray scattering (SAXS), NMR, hydrogen/deuterium exchange followed by mass spectrometry (HDX–MS) coupled with molecular dynamics (MD) simulation (94). However, precise arrangement of the nucleosomes facilitating higher-order chromatin compaction remained a debatable topic among different research groups. Based on *in vitro* data, two alternative models, namely (a) solenoid and (b) zigzag models, were proposed to explain the secondary structure of chromatin into a ~30 nm fiber (95).

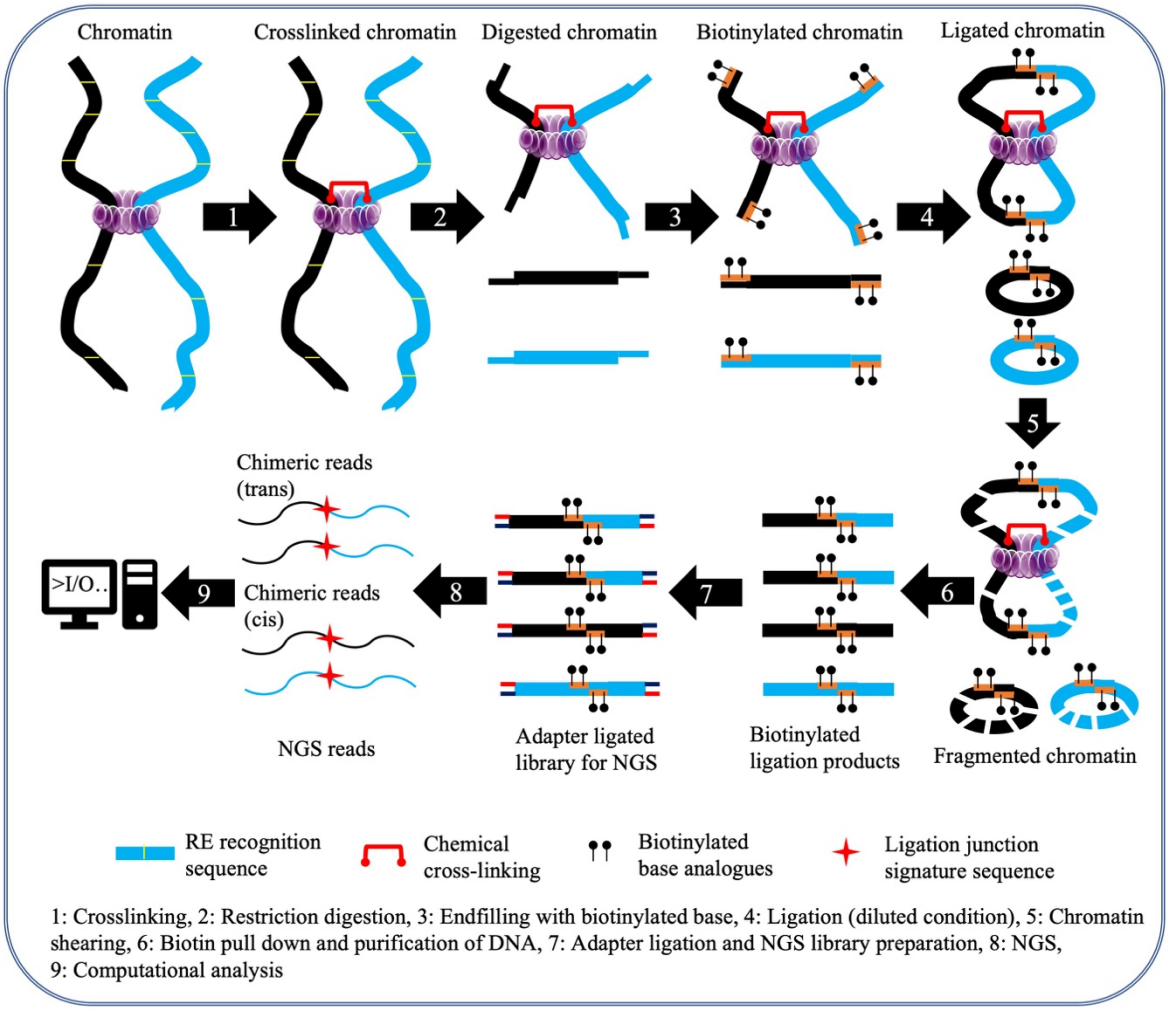


Figure 1.8 Schematic of the key steps in a typical Hi-C experiment.

A typical Hi-C experiment starts with the chemical crosslinking of the chromatin to preserve the spatial genomic interactions between (*trans*) and within (*cis*) the chromosomes. Second, the crosslinked chromatin sample is digested with a given restriction enzyme. Biotinylated base analogs are incorporated into the staggered cut ends by an end-filling reaction in the third step. Subsequently, the digested and biotinylated chromatin fraction is allowed to ligate in a diluted condition, which reduces the chances of random ligation events by increasing the chances of ligation of ends that are in closer proximities *in vivo*. At this step, chimeric ligation junctions are created either between two different DNA molecules held together in proximity or between two ends of the same DNA molecule. The later produces self-ligated circles. In the fifth step, the ligated chromatin fraction is fragmented by a mechanical force such as sonication. Next, the chromatin fragments are subjected to streptavidin pull-down, to enrich the biotin-labeled ligation junctions specifically. In the seventh step, the DNA fraction of the streptavidin-precipitated chromatin fragments are purified, and adapter-ligated to prepare a library for Illumina sequencing. This library is then sequenced which produces *cis* or *trans* chimeric reads spanning over a ligation junction signature sequence. Finally, the Illumina reads are analyzed to identify the spatial genomic interactions.

However, *in vivo* evidence for a ~30 nm chromatin fiber structure remained elusive (96). Alternatively, FISH allows the visualization of large genomic regions. With the development of chromosome conformation capture (3C) (97), which offers a significant improvement in spatial resolution of two genome regions over FISH (98), it was possible to study chromatin structure in kilo base-pair resolution. The principle of the 3C experiment is 'proximity ligation', which preferentially generates chimera of two spatially proximal loci over any other combination of two random genomic loci. In this experiment, crosslinked chromatin is digested with a restriction enzyme (RE) and allowed to ligate in a diluted condition. A dilution reaction condition reduces the chances of random ligation and increases the proportion of ligated products between DNA loci, which are spatially proximal to each other. The presence of chimeric molecules, which serve as evidence of interactions between the two given loci, can then be tested by polymerase chain reaction (PCR) (99).

Although 3C offers a unique way to study the chromatin structure *in vivo,* the use of this method is limited to test only specific spatial contacts in the genome at a time. Subsequent improvements in the 3C method addressed this issue to allow detection of multiple spatial interactions in a genome through circular chromosome conformation capture (4C) (100) and chromosome conformation capture carbon copy (5C) (101). In another variant technique termed as Hi-C, NGS following a modified 3C protocol, allows the identification of all versus all interactions in the genome. In this modified protocol, digested DNA ends are biotin labeled. Pulldown of the chimeric molecules by streptavidin allows enrichment of

ligation junctions (102, 103). The schematic of the key steps in a typical Hi-C experiment has been shown (Figure 1.8). Later, modified Hi-C or genome conformation capture or 3C-sequencing was developed, in which the biotin ligation step is avoided (104, 105) (Figure 1.9). However, the majority of 3C-sequencing reads originate from the genomic regions, and fewer reads are generated from chimeric DNA fragments. A comparison of different variants of 'C-techniques' for their ability to detect spatial interactions is presented (Figure 1.10).

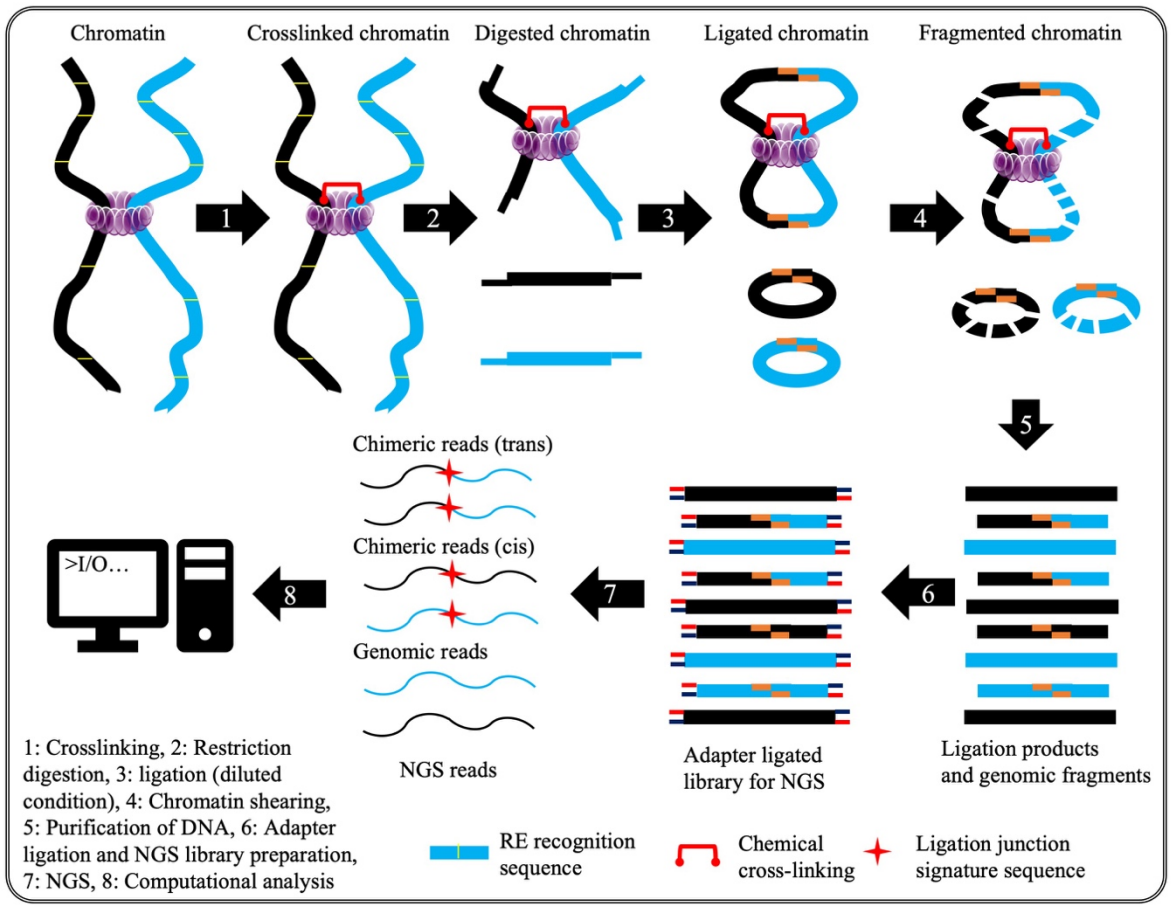


Figure 1.9 Schematic of the key steps in 3C-seq experiment.
The key steps in a 3C-seq experiment are identical to that of a typical Hi-C experiment (as described in figure 1.8), with one major exception. Unlike Hi-C, incorporation of the biotinylated base analogs in the end-filling reaction to specifically label and enrich the ligation products is not performed in a 3C-seq experiment. As a result, the percentage of chimeric reads obtained in 3C-seq can be significantly lower than that of a Hi-C experiment.

Analysis of 3C-seq or Hi-C data involves the identification of chimera between the two DNA loci joined at a hybrid-restriction site sequence while eliminating products of self-ligation. Moreover, this analysis requires data normalization steps to avoid the effect of CNVs and maintain uniformity across the entire genome. Therefore, accurate mapping of the spatial contacts in a relatively large eukaryotic genome can be computationally intensive.

Newer algorithms were developed to address these issues, although certain tools were already available for quality filtering, trimming, and mapping of reads (106). In addition, the steps followed during the Hi-C experiment, such as crosslinking, chromatin fragmentation, biotin-

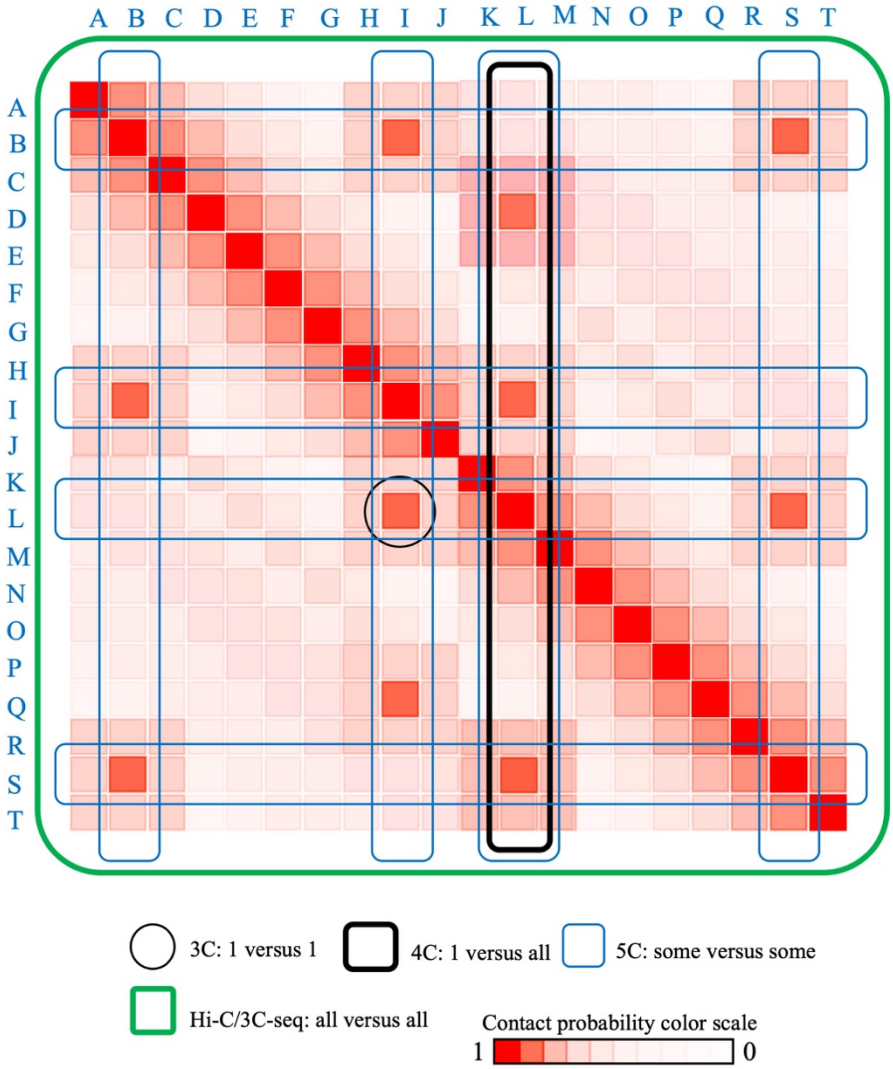


Figure 1.10 A comparison of different variants of 'C-techniques'
Schematic of contact probability matrix indicating the types of data generated by 3C and related techniques. A 3C experiment queries interactions between one locus to another (black circle), 4C identifies interactions of a single locus with the rest of the genome (black rectangle), 5C generates an interaction matrix of a set of loci (blue rectangles), whereas by using Hi-C, one finds out genome-wide chromatin interactions (green square). This figure was adapted from the reference (107).

---

labeling, and re-ligation, can induce biases and complicate the interpretation of sequence data (108-110). Thus, one of the major challenges in the analysis of Hi-C data is the application of the correct normalization method to eliminate biases in the raw data, if any.

Methods for normalization of Hi-C data follows two broad approaches: (a) explicit approach, and (b) implicit approach. In the explicit approach, it is assumed that all sources of systematic biases are known based on biases determined empirically from the observed data. For example, the variability in restriction enzyme fragment lengths, GC content, and sequence mappability are identified as three major sources of experimental bias in Hi-C data (111). On the contrary, an implicit model assumes that each locus should receive equal sequence coverage after biases are removed (108). Two such widely used models are (a) iterative correction and eigenvector decomposition (ICE) (112) and (b) Knight and Ruiz algorithm (113). The basic assumption of both these models is that all parts of the genome interact equally with each other. Due to this assumption, both models fail to avoid the effect of CNVs present in the genome. This issue was addressed in the newly developed models, such as 'LOIC/CAIC' and 'OneD', which can be used to accurately interpret Hi-C data, especially in cases like cancer cells carrying multiple CNVs in their genome (114, 115).

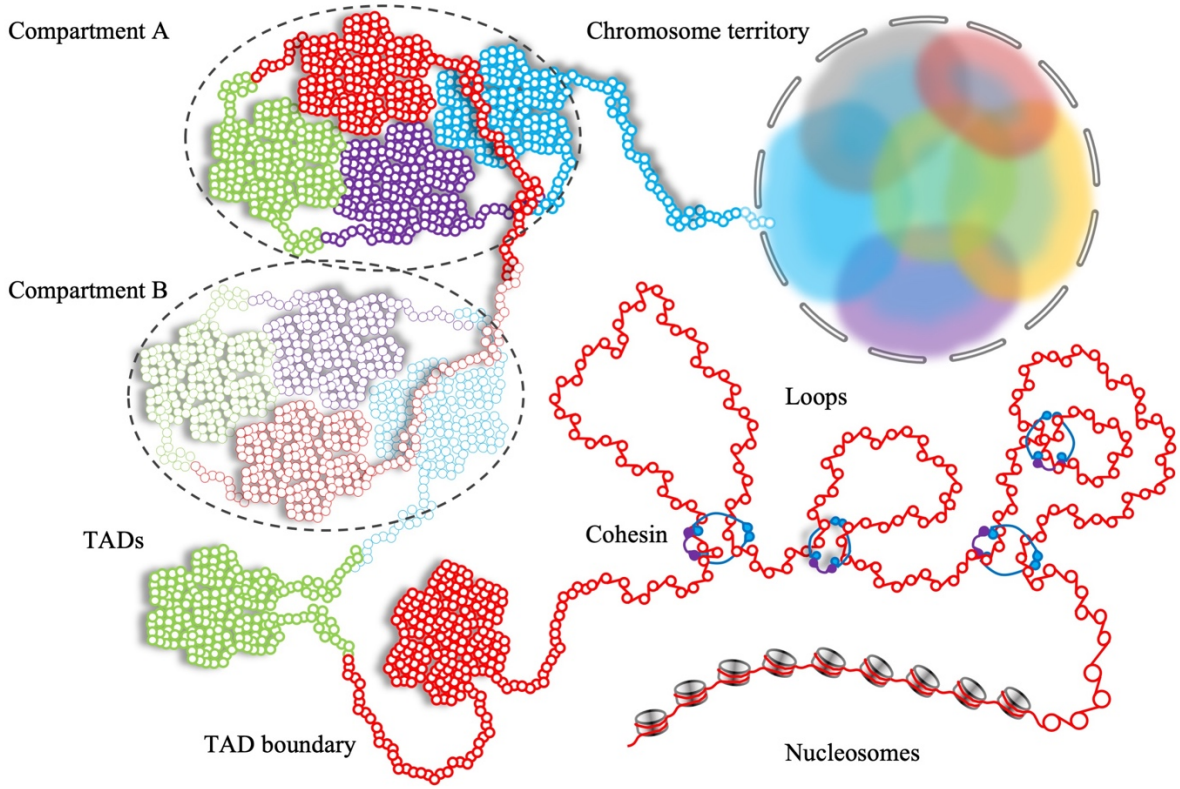


Figure 1.11 Hierarchical chromatin organization inside the cell nucleus.
A simplified scheme showing hierarchical chromatin organization. The DNA wraps around histone octamers to form a typical canonical nucleosome. This 10 nm chromatin fiber can form loops that are held together by cohesin or other chromatin-associated proteins. This chromatin fiber can be further compacted into compact chromatin domains, which can interact among themselves to form higher-order chromatin structures known as chromatin compartments. Compartment A is identified as transcriptionally active parts of the genome, while the transcriptionally inert regions are organized into Compartment B. In the interphase

nuclei, several compartments are associated with a chromosome and form the so called 'chromosome territory'. The concept of this drawing is adapted from reference (116).

Analysis of Hi-C data led to the identification of genomic domains that are folded into compact local structures or topologically associated domains (TAD) (Figure 1.11). Newer algorithms were developed for statistical analysis of the contact probability data to study local chromatin compaction and identify these self-associating genomic domains. Some of the widely used algorithms for TAD calling are Directionality Index (DI) (117), Armatus (118), TADtree (119), insulation score (IS) (120), IC-finder (121). These algorithms assess the self-association properties of chromatin for genome-wide identification of TADs. For example, DI is a statistical tool that measures the imbalance between the upstream and downstream contacts of a genomic locus. The DI statistic is then used to call domains with respect to the inherent bias state of each locus conferred from the Hidden Markov model (HMM) analysis (117). Similarly, IS for a given bin is defined as an average number of interactions across that bin in close vicinity. The local minima of IS lie at the TAD borders (120). Additionally, principal component analysis of the genome-wide contact probability matrix identified two major compartments in the human genome named Compartment A and Compartment B (103). Two genomic loci belonging to the same compartment are found to be spatially proximal to each other than to another pair of loci belonging to separate compartments. Moreover, genomic loci belonging to Compartment A were found to be less compact and poised for gene expression than the loci belonging to Compartment B (103).

Structure and function of topologically associated domains (TADs)

Statistical analysis of the spatial contacts revealed TADs as one of the conserved features of genome organization in a wide range of species. TADs were initially identified in human and mouse cells as megabase-size local chromatin interaction domains (117, 122). Subsequently, similar higher-order chromatin structures were identified in flies (104, 123), worms (124), plants (125), yeasts (126-128), and even in bacteria (129). Evidence for the existence of these chromatin domains has also been obtained from high-resolution imaging studies (130) and super-resolution chromatin tracking at single-cell resolution (131). These observations indicate that TAD-like structures are fundamental units of the chromatin organization *in vivo*. However, the length of DNA folded into one domain varies both within and across the species. For example, in *Caenorhabditis elegans,* TADs of ~1 Mb length are

present on the X chromosome (120). Similarly, hundreds of kilobases long TADs have been identified in mammals (117). On the contrary, smaller TADs of ~100 kb length have been observed in *Drosophila melanogaster* (104) and *Schizosaccharomyces pombe* (126). It is also important to mention that a large TAD can be composed of several smaller sub-TADs. Improvement in data resolution may lead to the identification of additional sub-TADs, which are part of a larger TAD (Figure 1.12). For example, an improved resolution of Hi-C data led to the identification of >4000 TADs in *D. melanogaster* genome (132), while an earlier study reported approximately 1300 TADs (133). The use of Hi-C data generated from a population of cells may also influence the identification of TADs due to the inherent heterogeneity and cell to cell variability in TAD structures (134). Such heterogeneity is also evident from the observation that two alleles can independently interact with different parts of the genome (134, 135).

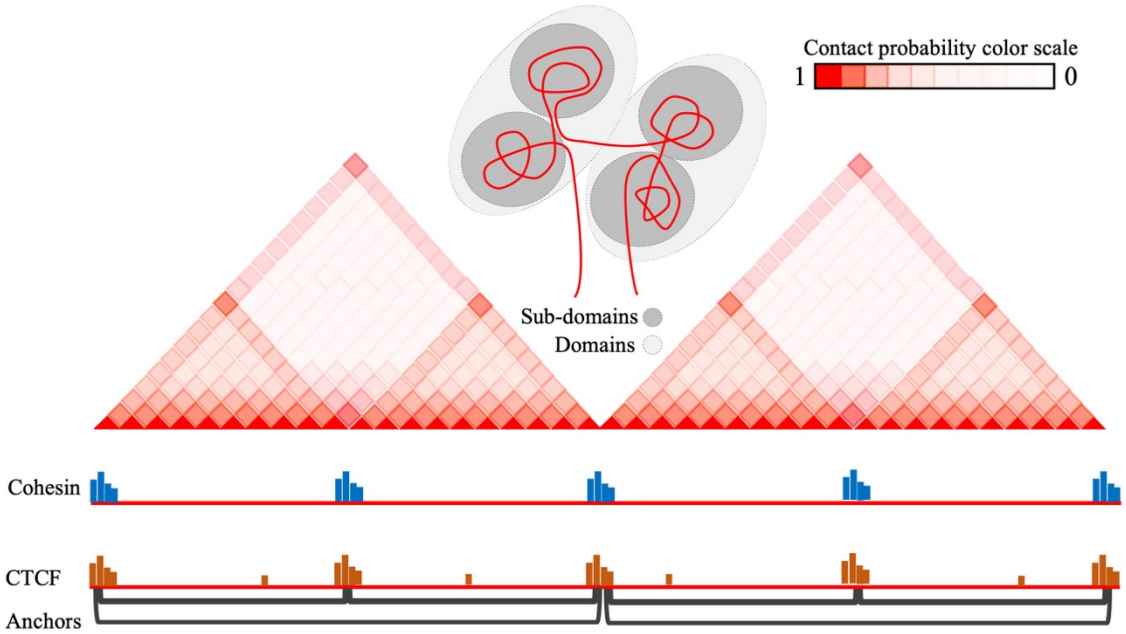


Figure 1.12 Integration of contact probability data and the chromatin binding of cohesin and CTCF reveals TADs and sub-TAD structures.
Integration of the contact probability heatmap along with the ChIP-seq data of a genomic locus confirms spatial contacts between the anchor points bound to cohesin and CTCF, which facilitates the formation of topologically associated domains (TADs) in mammals. Based on these data, this given locus appears to carry two TADs, each of which is composed of two smaller subdomains. A model representing the approximate spatial organization of the genomic locus is shown on the top.

Identification of TAD-like structures across species raised questions about their function, which revealed a wide range of biological functions of TADs ranging from

development to cognition. Earlier studies reported transition in chromatin state in response to the environmental and developmental cues (136). Subsequent studies uncovered the biological role of TADs in the regulation of gene expression (137-139), deciding cell fate during differentiation (140-142), and cognition (143, 144). Importantly, dysregulation in domain structures can be associated with certain diseases (135, 145, 146), and therefore, alteration in chromosomal domains can act as potential biomarkers for disease diagnostics (147). In addition, probing the higher-order chromatin organization provided a structural basis of long-range interactions between gene promoters and other distal regulatory elements such as enhancers (148). The spatial genome organization in TADs also correlates with replication timing (149). Recently it was found out that the deletion of early replicating control elements can affect the domain-wide replication timing, A/B compartmentalization, and TAD structures (150).

What regulates the dynamic nature of the TADs? Studying the structure of TADs in multiple model systems identified both underlying DNA sequence (151) and multiple conserved architectural proteins such as cohesins, CCCTC-binding factor (CTCF), polycomb repressor complex (PRC) are involved in maintaining the TAD structure (152-154). CTCF and cohesins facilitate close contact between the boundary loci flanking a TAD and thereby forming a loop-like structure. In metazoan species, including *Homo sapiens*, *C. elegans,* and *D. melanogaster* boundaries of TADs are enriched with insulator elements and active genes, but the interiors generally contain a relatively uniform chromatin state (155). Mechanisms facilitating the dynamics of TAD structures remain poorly understood until a 'loop extrusion' mechanism of a new TAD formation was demonstrated (156). Recently, high-resolution imaging produced compelling evidence of loop extrusion, which is facilitated by condensins and cohesins (157, 158).

Higher-order chromatin structure in yeasts

A structural homolog of mammalian CTCF has not been identified in the budding yeast genome. However, Hi-C (127) and Micro-C (128) based studied reported presence of TADs in *S. cerevisiae*. Intriguingly, the length of TADs reported in these two studies are significantly different from each other. While the majority of TADs identified from the analysis of the Micro-C data were smaller than 10 kb, TADs as long as ~400 kb were identified from Hi-C analysis. These studies revealed fundamental properties of the TAD

boundaries as nucleosome-depleted and transcriptionally active loci. A clear homolog of CTCF remains unidentified in *S. cerevisiae,* but the presence of chromatin remodeler Sth1, cohesin loader Scc2 and Forkhead transcription factor Fkh1 and Fkh2 at the boundary regions indicate a putative role of these proteins in the maintenance of TAD structures (127, 128). Similarly, higher-order chromatin structures were studied in fission yeast *S. pombe* using Hi-C analysis, which identified cohesion dependent globules (126). Dynamic reorganization of TADs takes place during different stages of the cell cycle in *S. pombe,* as the cohesion-dependent large domains of 300 kb - 1 Mb length formed during mitosis are gradually restructured to smaller domains of 30 - 40 kb (159). Similarly, observation of cohesion- and condensin-dependent chromatin reorganization during different stages of the cell cycle in *S. cerevisiae* indicates the existence of a conserved mechanism (160).

The adoption of non-random positioning of chromosomes in interphase nuclei in preferred conformations (161-163) allows disparate DNA elements to engage in replication (164) or transcription 'factories'(165). Genetic and cell biological analysis in *S. cerevisiae* revealed Rabl-like conformation in yeast nuclei, in which centromeres, as well as telomeres of all chromosomes are located in a close physical proximity (166-168). However, high-resolution contact probability data obtained from Hi-C experiments allowed the construction of a structural model revealing spatial contacts within and between all the chromosomes present in *S. cerevisiae* (169). Similar to the budding yeast species *S. cerevisiae*, spatial genome organization facilitating trans-interactions among the centromeres of different chromosomes is also observed in fungi, apicomplexans, flies, plants, mice and humans (170, 171) carrying highly divergent underlying centromere DNA sequences. Especially, clustering of the centromeres has been identified to be a conserved feature of the fungal genome organization. However, the significance of such a non-random genome organization at the centromere is not well understood.

Centromere

The primary constrictions were first described by Walther Flemming (172). Later these structures were identified as centromeres, which serve as the chromosomal binding sites of spindle microtubules. In most organisms, centromeres are localized chromosomal domains, present only once in every chromosome. The centromere-kinetochore complex ensures timely and accurate attachment of the spindle microtubules to facilitate the faithful

Table 1.1: A comprehensive list of fungal species belonging to Ascomycota, Basidiomycota and Mucoromycota with known or predicted centromeres and their features.

Sl. no.*	Species	Method used to identify centromere	Genome size (Mb)	No. of chromosomes	AT richness	RNAi/HP1 mediated heterochromatinization	Active or truncated transposons at the centromere	Length of centromere (based on plasmid stability function, contact probability data)	Length of CENP-A/KT domain	References (Identification, annotation)
1	*Saccharomyces cerevisiae*	Plasmid stability assay, CDE	11.87	16	Yes	No	No	~120-125 bp	~125 bp	(173)
2	*Saccharomyces paradoxus*	CDE, Hi-C	11.94	16	Yes	No	No	~120-125 bp	NA	(174, 175)
3	*Saccharomyces uvarum*	CDE, plasmid stability assay	11.60	NA	Yes	No	No	~120-125 bp	NA	(176)
4	*Saccharomyces bayanus*	CDE, plasmid stability assay	11.86	NA	Yes	No	No	118-121 bp	NA	(177, 178)
5	*Zygosaccharomyces rouxii*	CDE, plasmid stability assay	9.76	7	Yes	No	No	167-169 bp	NA	(179, 180)
6	*Vanderwaltozyma polyspora*	CDE	14.67	NA	Yes	No	No	109-120 bp	NA	(178)
7	*Saccharomyces mikatae*	Hi-C	11.47	NA	Yes	No	No	~125 bp	NA	(175)
8	*Saccharomyces kudriavzevii*	Hi-C	11.85	16	Yes	No	No	~125 bp	NA	(175)
9	*Candida glabrata*	Synteny, conserved elements, plasmid stability assay	12.47	13	Yes	No	No	107-113 bp	NA	(181, 182)
10	*Naumovozyma castellii*	Ndc10, Ndc80 ChIP-seq	11.23	10	Yes	No	No	~110 bp	~110 bp	(183)
11	*Naumovozyma direnensis*	Synteny, conserved sequence	13.75	11	Yes	No	No	~110 bp	NA	
12	*Lachancea meyersii*	CDE	11.26	8	Yes	No	No	125-138 bp	NA	(178, 180, 184)
13	*Lachancea dasiensis*	CDE	10.70	8	Yes	No	No	116-128 bp	NA	
14	*Lachancea nothofagi*	CDE	11.31	8	Yes	No	No	126-128 bp	NA	
15	*Lachancea thermotolerans*	CDE	10.39	8	Yes	No	No	127-138 bp	NA	

16	*Lachancea waltii*	CDE	10.91	NA	Yes	No	No	127-139 bp	NA	
17	*Lachancea mirantina*	CDE	10.11	8	Yes	No	No	115-118 bp	NA	
18	*Lachancea fermentati*	CDE	10.26	8	Yes	No	No	186-197 bp	NA	
19	*Lachancea cidri*	CDE	NA		Yes	No	No	187-217 bp	NA	
20	*Lachancea kluyveri*	CDE	11.50	8	Yes	No	No	187-200 bp	NA	
21	*Lachancea fantastica*	CDE	11.33	7	Yes	No	No	124-136 bp	NA	
22	*Eremothecium gossypii*	CDE-like elements	9.13	7	Yes	No	No	199-202 bp	NA	(178, 185)
23	*Eremothecium cymbalariae*	CDE-like elements	9.66	8	Yes	No	No	~200 bp	NA	(186, 187)
24	*Eremothecium coryli*	CDE-like elements	9.09	NA	Yes	No	No	~200 bp	NA	(186)
25	*Kluyveromyces lactis*	CDE-like elements (KlCDE)	10.70	6	Yes	No	No	~200 bp	NA	(182, 188)
26	*Kluyveromyces marxianus*	Plasmid stability assay, conserved KlCDE elements	10.90	8	Yes	No	No	~260 bp	NA	(189)
27	*Kluyveromyces wickerhamii*	Hi-C	9.80	NA	NA	No	No	NA	NA	(175)
28	*Candida lusitaniae*	CENP-A ChIP-seq	12.11	8	Yes	No	No	4-4.5 kb	4-4.5 kb	(190)
29	*Candida dubliniensis*	CENP-A ChIP-seq	14.04	8	NA	No	No	3-5 kb	3-5 kb	(191)
30	*Candida albicans*	CENP-A ChIP-PCR	14.67	8	No	No	No	2.9-4 kb	2.9-4 kb	(192)
31	*Candida tropicalis*	CENP-A ChIP-seq	14.63	7	No	No	No	9-22.4 kb	3-5 kb	(193)
32	*Scheffersomyces stipitis*	Plasmid stability assay, Hi-C	15.48	8	NA	No	Yes	NA	NA	(175)
33	*Debaryomyces hansenii*	Hi-C	12.06	7	NA	No	No	1186-4402 bp	NA	(194)
34	*Kuraishia capsulata*	Hi-C	11.37	7	NA	NA	Yes	851-6741 bp	NA	
35	*Ogataea polymorpha*	Bioinformatic prediction	8.97	7	NA	NA	Yes	10-20 kb	NA	(195)
36	*Komagataella pastoris*	CENP-A ChIP-seq	9.35	4	Yes	No	No	5,354-6,655 bp	5,354-6,655 bp	(196)
37	*Blastobotrys adeninivoran*	Bioinformatic prediction	NA	NA	NA		No	~1 kb	NA	(197)
38	*Yarrowia lipolytica*	Plasmid stability assay	20.55	6	NA	No	No	~1 kb	NA	(182, 198)
39	*Zymoseptoria tritici*	CENP-A ChIP-seq	37.68	21[#]	No	Yes	Yes	5.57-13.55 kb	5.57-13.55 kb	(199)

40	*Malassezia sympodialis*	Bioinformatic prediction, Mtw1 ChIP-seq	7.72	8	Yes	No	No	81-1152 bp	2-5 kb	(200, 201)
41	*Malassezia furfur*	CENP-A ChIP-qPCR	7.79	NA	Yes	No	No	168-721 bp	NA	(201)
42	*Malassezia globosa*	H3/H4 ChIP-PCR	8.93	NA	Yes	No	No	107-455 bp	NA	
43	*Malassezia restricta*	Bioinformatic prediction	7.29	9	Yes	No	No	594-687 bp	NA	
44	*Malassezia sloffiae*	H3/H4 ChIP-PCR	8.42	NA	Yes	No	No	477-583 bp	NA	
45	*Malassezia nana*	Bioinformatic prediction	7.57	NA	Yes	No	No	512-794 bp	NA	
46	*Malassezia dermatis*	Bioinformatic prediction	7.54	NA	Yes	NA	No	499-848 bp	NA	
47	*Malassezia vespertilionis*	Bioinformatic prediction	7.58	NA	Yes	NA	No	521-1200 bp	NA	
48	*Malassezia japonica*	Bioinformatic prediction	8.35	NA	Yes	NA	No	497-598 bp	NA	
49	*Epichole festucae*	Hi-C	35.04	7	NA	NA	No	~10 kb	NA	(202)
50	*Neurospora crassa*	Plasmid stability assay	41.10	7	Yes	Yes	Yes	175–300 kb	NA	(203)
51	*Schizosaccharomyces pombe*	Plasmid stability assay	12.81	3	No	Yes	Yes	40-100 kb	10-15 kb	(204)
52	*Schizosaccharomyces octosporus*	CENP-A ChIP-seq	11.63	3	No	Yes	Yes	66-77 kb	10-15 kb	(205)
53	*Schizosaccharomyces cryophilus*	CENP-A ChIP-seq	11.55	3	No	Yes	Yes	73-85 kb	10-15 kb	
54	*Cryptococcus amylolentus*	CENP-A ChIP-seq	20.26	14	NA	NA	yes	22-48 kb	10-15 kb	(206)
55	*Cryptococcus neoformans*	CENP-A ChIP-seq	18.59	14	NA	Yes	Yes	27-64 kb	27-64 kb	(207)
56	*Cryptococcus deneoformans*	CENP-A ChIP-seq	19.05	14	NA	Yes	Yes	29-110 kb	29-110 kb	
57	*Cryptococcus deuterogattii*	CENP-A ChIP-seq	17.47	14	NA	No	Yes	8-21 kb	8-21 kb	
58	*Ustilago maydis*	Bioinformatic prediction	20.06	23	NA	No	Yes	14.5 kb	NA	
59	*Ustilago bromivora*	Bioinformatic prediction	20.70	23	NA	Yes	Yes	27.8 kb	NA	
60	*Ustilago hordei*	Bioinformatic prediction	24.63	NA	NA	Yes	Yes	39.3 kb	NA	
61	*Magnaporthe oryzae*	CENP-A ChIP-seq	38.75	7	Yes	Yes	Yes	57-109 kb	57-109 kb	(208)
62	*Trichoderma reesei*	3C-seq	32.68	7	Yes	NA	Yes	30-43 kb	NA	(209)
63	*Fusarium graminearum*	CenH3-ChIP-seq	36.66	4	Yes	Yes	No	150-300 kb	174-287 kb	(210)

64	*Mucor circinelloides*	Mis12 and Dsn1 ChIP-seq	36.56	9	Yes	Yes	Yes	15-73 kb	744-1132 bp	(211)

*Centromere type: 1-27: point centromeres, 28-49: short regional centromeres, 50-63: long regional centromeres, 64: mosaic centromere; NA: Data not available; #*Z. tritici* contains 13 core and 8 accessory chromosomes.

segregation of sister chromatids. Initial studies on centromere biology were facilitated by fungal model systems due to the ease of laboratory manipulation and availability of genome assembly. To date, the identity of centromeres from over 60 fungal species has been predicted by DNA sequence analysis; a majority of them are validated by genetic and/or biochemical experiments (Table 1.1). Initially, cloning of centromere DNA on a replicative

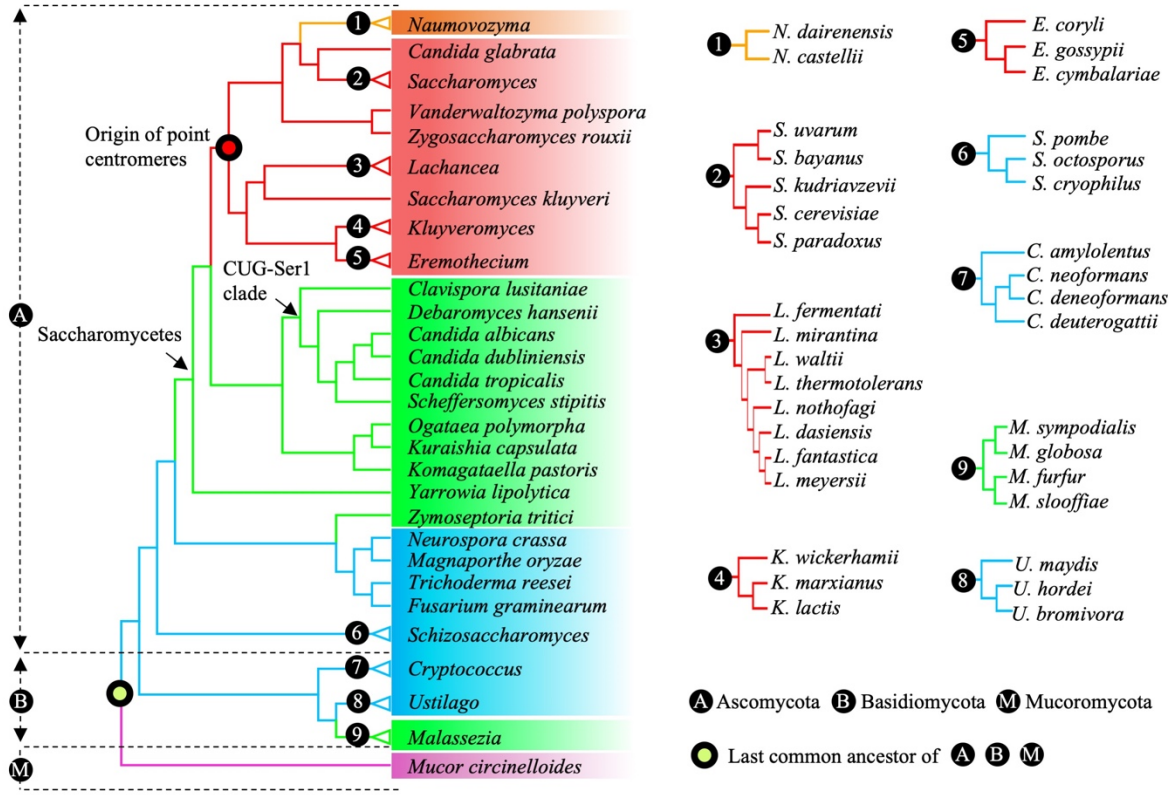


Figure 1.13 Phylogenetic distribution of fungal species with known centromeres in Ascomycota, Basidiomycota, and Mucoromycota.
Left, a maximum-likelihood based tree of 55 fungal species with known centromeres was generated from the conserved orthologs identified in these species using OrthoFinder (212), MAFFT (213), FastTree (214). *Blastobotrys adeninivorans, Epichloe festucae* and some species belonging to the genera *Lachancea* and *Malassezia* with known centromere loci were not included in this phylogenetic tree as the complete annotation of ORFs are not available. The branches and the groups of species are color-coded based on the centromere type. Orange and red, unconventional and conventional point centromeres respectively; green, short regional; blue, long regional and pink, mosaic-type centromere. Nine nodes, marked with black circles numbered from one to nine, containing species with similar centromere type, were collapsed and represented as triangles. *Right,* the internal species-level tree

topology of the collapsed nodes is expanded, and branches are color-coded as described in the left panel.

---

plasmid and scoring for its mitotic stability allowed genetic dissection of functional DNA elements. Later, the use of chromatin immunoprecipitation (ChIP) assays led to the identification of kinetochore-bound domains on these centromeres. Based on the length of the kinetochore bound domain centromeres identified in three fungal phyla Ascomycota, Basidiomycota, and Mycoromycota (Figure 1.13) can be categorized into four major classes. These are a. point centromere, b. short regional centromere, c. long regional centromere, and d. mosaic centromere. Comparative analyses of the structural and functional properties of the fungal centromeres facilitated the detection of the evolutionary patterns within a group of closely related fungi. One of the major findings from these studies is the non-universality of the factors that define and regulate centromere structure and function.

Point centromere

Molecular understanding of centromere DNA was initiated by the cloning of centromeres in *S. cerevisiae* that led to the construction of the first artificial minichromosome (173). The 125-bp long 'point' centromere of *S. cerevisiae*, roughly the same length of DNA wrapped around a single nucleosome, consists of conserved DNA elements (CDEs): CDEI, CDEII, and CDEIII (215). CDEI and CDEIII share conserved but degenerate motifs of 8 and 26 nucleotides, respectively (216). Although the highly AT-rich CDEII (78-86 bp) (217) is not conserved, its length is important for centromere function (218). A single base-pair mutation in the CCG-motif in CDEIII is sufficient to abolish the centromere function. Centromeric nucleosomes contain centromere-specific histone H3 variant CENP-A^{Cse4} (219). Binding of kinetochore proteins facilitates bending of the DNA flanking CDEII, which has an intrinsic ability to form curves (220, 221). These physical properties and DNA sequence recognition by the point centromere-specific protein complexes contribute to the genetic identity of centromere DNA, enabling these sequences to *de novo* assemble kinetochore components. Approximately 25 closely related Saccharomycetes in the fungal phylum of Ascomycota have been found to contain conventional CDE-like elements at their centromeres (178). In these organisms, the length of CDEII varies from 93 bp in *Lachancea waltii* to 161 bp in *Kluyveromyces lactis* (178, 188, 222). These conserved structural features of centromere DNA shared by organisms in the subphylum Saccharomycotina indicate a single

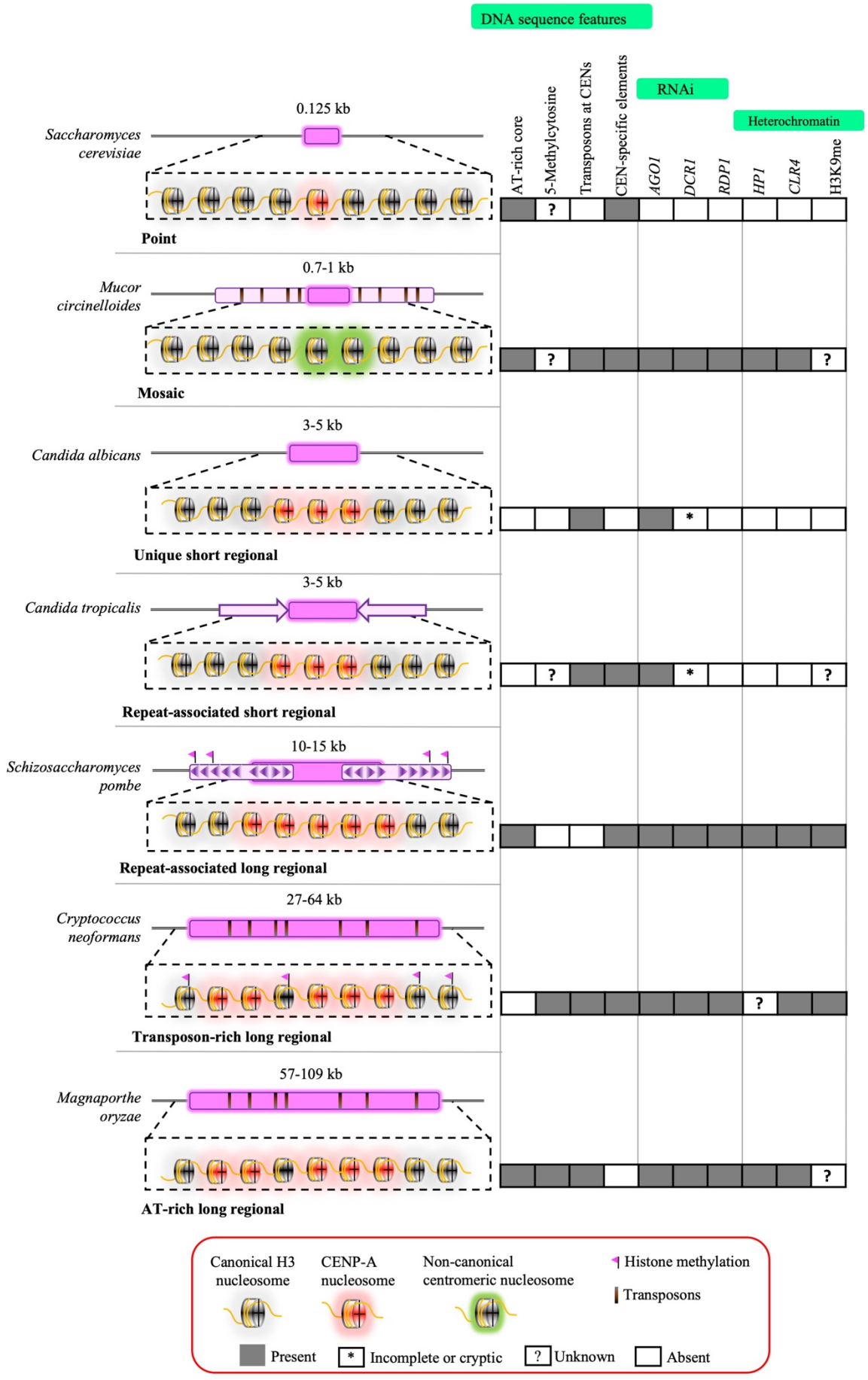

DNA sequence features
RNAi
Heterochromatin
AT-rich core
5-Methylcytosine
Transposons at CENs
CEN-specific elements
AGO1
DCR1
RDP1
HP1
CLR4
H3K9me
0.125 kb
Saccharomyces cerevisiae
Point
0.7-1 kb
Mucor circinelloides
Mosaic
3-5 kb
Candida albicans
Unique short regional
3-5 kb
Candida tropicalis
Repeat-associated short regional
10-15 kb
Schizosaccharomyces pombe
Repeat-associated long regional
27-64 kb
Cryptococcus neoformans
Transposon-rich long regional
57-109 kb
Magnaporthe oryzae
AT-rich long regional
Canonical H3 nucleosome
CENP-A nucleosome
Non-canonical centromeric nucleosome
Histone methylation
Transposons
Present
Incomplete or cryptic
Unknown
Absent

Figure 1.14 DNA sequence, structural, and chromatin properties of seven major fungal centromere types.
Left, schematic of the centromere organization highlighting the centromeric chromatin domain (purple) and flanking pericentric region (pink) in each representative type of fungal centromeres. Line diagrams are not drawn to scale. A representative arrangement of nucleosomes in each type of centromere is shown in a dotted box. *Right*, the presence or absence of various determinants of centromere structure and function is shown. Transposon refers to the presence of either full-length or truncated version of it in at least one centromere of a given species. Centromere-specific DNA sequence elements include conserved DNA sequences present exclusively at the centromeres, but not necessarily common to all centromeres. Centromere-specific elements include CDEs (*S. cerevisiae*), an AT-rich motif (*M. circinelloides*), pericentric repeats (*C. tropicalis* and *S. pombe*), and full length Tcn retrotransposons (*C. neoformans*).

---

origin of point centromere. More recently, the unconventional point centromeres that harbor CDEs, different from those of *S. cerevisiae*, have been reported in *Naumovozyma castellii* and *Naumovozyma dairenensis* (183). The genetic identity of these unconventional point centromeres also revealed a rapid co-evolution of the CBF3 complex components Ndc10 and Cep3, which recognize diverged point centromere DNA sequences (183).

Short regional centromere

Most other fungal species have regional centromeres spanning beyond a single nucleosome and are not strictly defined by the underlying DNA sequence (Figure 1.14). The short regional centromere (< 20 kb) was first identified in a CUG-Ser1 clade species *C. albicans* that contains unique DNA sequence spanning 3-5 kb long CENP-A^{Cse4}-bound centromeric chromatin (192, 223). Lack of sequence conservation and the inability of centromere DNA to stabilize a centromeric plasmid carrying an autonomously replicating sequence (ARS) suggested a DNA sequence-independent inheritance of centromere function (224). Centromeres of *Candida dubliniensis* also share similar features, containing unique DNA sequences that are remarkably diverged from their *C. albicans* counterparts (191). AT-rich short regional centromeres on unique DNA sequences were identified in another CUG-Ser1 clade species *Candida lusitaniae* (190). Using various genetic, genomic and biochemical approaches, short regional centromeres were identified in other Saccharomycetes, including *Kuraishia capsulata* (194), *Ogataea polymorpha* (195), *Blastobotrys adeninivorans* (197), and *Yarrowia lipolytica* (225). Unusual short regional centromeres of *Y. lipolytica* carry conserved blocks of 9-14 bp regions with dyad symmetry (198).

Inverted repeat (IR)-associated short regional centromeres were identified in the CUG-Ser1 clade species *C. tropicalis* (193), which diverged ~39 million years ago from *C. albicans*. Unlike the unique centromeres in *C. albicans*, all seven centromeres of *C. tropicalis* are highly homogeneous (226), containing 2-3 kb long CENP-A^{Cse4}-bound mid-core flanked by 3-5 kb long IRs. Intriguingly, the entire mid-core flanked by IRs present on a plasmid can facilitate the *de novo* recruitment of CENP-A^{Cse4} and improve its mitotic stability, albeit at a lower frequency than that of *S. cerevisiae* (193). Similar IR-associated centromeres were identified in *Komagataella phaffi* that consist of ~2 kb IRs flanking ~1 kb *mid* regions (196). *Zymoseptoria tritici*, a filamentous ascomycete, contains 5.5-13.5 kb CENP-A^{CenH3} enriched centromeric chromatin (199). Apart from these ascomycetes described above, organisms of the *Malassezia* species complex of the fungal phylum Basidiomycota also possess short regional centromeres that are highly AT-rich with 2-5 kb long centromeric chromatin (201).

Long regional centromere

A class of DNA sequence-dependent long regional centromeres (>20 kb) were identified in the fission yeast *S. pombe* (227-229). The length of fission yeast centromeres ranges from 40-110 kb encompassing the kinetochore-bound central core (CC) region flanked by various types of repeats (204) (Figure 1.14). The central regions of *CEN1* and *CEN2* of *S. pombe* share homology, whereas the central region of *CEN2* is unique (204). The pericentric region consists of *dg* and *dh* class of repeats (229). However, a part of CC and one arm of pericentric chromatin proved to be sufficient for the establishment of centromere identity and proper segregation of minichromosomes (230). Similar repeat-associated long regional centromeres were identified in closely related *Schizosaccharomyces* species: *Schizosaccharomyces cryophilus* and *Schizosaccharomyces octosporus* (205, 231).

Long regional centromeres, which are rich in transposons, have been reported in both Ascomycota and Basidiomycota (Figure 1.14). Centromeres of *Neurospora crassa, Magnaporthe oryzae*, and *C. neoformans* are highly repetitive and harbor both active and/or truncated transposon elements (207, 208, 232). The length of centromeres ranges from 150-300 kb of heterochromatic DNA in *N. crassa* (232). The repeats at the centromeres of *N. crassa* introduce numerous C:T and G:A transitions by repeat-induced point mutation (RIP) (97) randomly through recurring cycles of an unknown mechanism leading to centromere DNA sequence divergence (233, 234). AT-rich centromeres of *M. oryzae* contain 57-109 kb

centromeric chromatin (235). Analysis of 3C-seq data revealed putative centromeric regions containing clusters of Tdh5 retrotransposon spanning 18-27 kb regions on all chromosomes of *Debaryomyces hansenii* (194). The RNAi-proficient species of the *Cryptococcus* species complex harbor 20-110 kb long centromeric chromatin. RNAi seems to help maintain full-length retrotransposons at centromeres in these organisms (207). A correlation between accumulation or loss of retrotransposons with alteration in the centromere length has been reported in the *Cryptococcus* as well as the *Ustilago* species complex (207).

Mosaic centromere

Most fungal centromeres studied to date are enriched with CENP-A (236). The loss of CENP-A has been described in kinetoplastid kinetochores present in Trypanosomes (237). In addition, certain insect lineages lacking CENP-A (238) harbor holocentric chromosomes, implying an independent transition to holocentricity (diffuse centromeres along the entire length of a chromosome) upon CENP-A loss in these lineages (237, 238). Among fungi, CENP-A loss has been recently reported in an early diverging sub-phylum Mucoromycotina (211). Strikingly, *Mucor circinelloides* has monocentric chromosomes in spite of lacking CENP-A. The average kinetochore binding length is 941 bp with a conserved AT-rich motif in this organism. These centromeres are mosaic-type given their point centromere-like kinetochore binding domain and unusually long pericentric regions (Figure 1.14). These pericentric regions range between 15-75 kb interspersed with Grem-LINE1 elements, which are repeats of LINE1-like non-LTR retrotransposable elements. The diversity in both the length and the structure of fungal centromeres hint that additional factors beyond centromere DNA play crucial role in the establishment and propagation of centromeric chromatin.

Establishment and propagation of centromere identity

The establishment of centromeric chromatin involves interactions between kinetochore proteins and centromere DNA that can either be at the level of primary DNA sequence, chromatin architecture, and/or three-dimensional conformation of the genome. Factors required for the maintenance of centromeric chromatin include heterochromatin components, transcriptional status, replication timing, and spatial chromosomal interactions. The establishment of centromeric chromatin on naked DNA sequences was first demonstrated by improved mitotic stability of minichromosomes in *S. cerevisiae* (173). However, in many fungal species, the mode of centromere establishment is independent of

the underlying DNA sequence. In *S. pombe*, a heterochromatic environment facilitated by the HP1 homolog Swi6 and RNAi-mediated machinery helps in the efficient recruitment of CENP-A^{Cnp1} to the central regions (Figure 1.15A) (239, 240). On the other hand, the epigenetic nature of centromeres in *C. albicans* that lacks RNAi and conventional heterochromatin does not permit the stabilization of a kinetochore on an externally introduced centromeric plasmid (224). This raises possibilities that species-specific factors are involved in centromere establishment. This plasticity of centromeric chromatin has been exemplified in experiments carried out in fungal species when studied in neocentromere formation, transgene silencing at the centromere, artificial centromere construction, and dicentric inactivation.

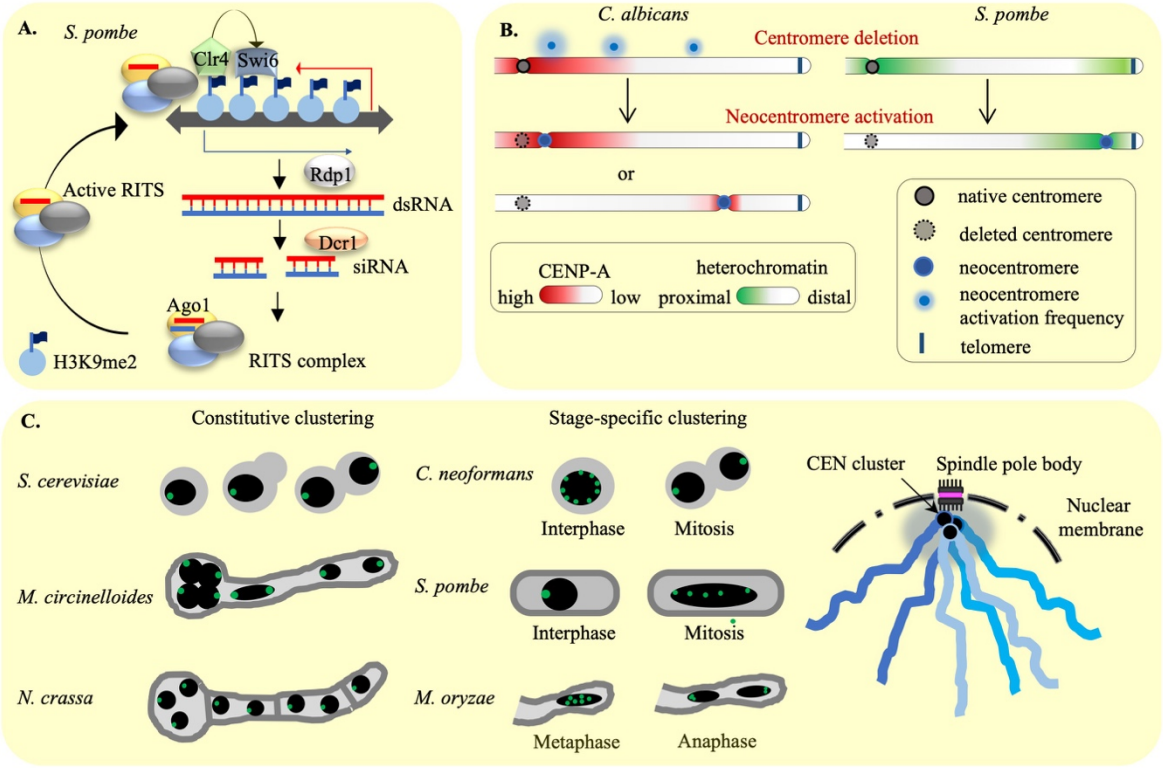


Figure 1.15 Molecular determinants of centromere formation in fungi.
A. Maintenance of centromeric heterochromatin at the outer repeats in *S. pombe* is mediated by RNAi-dependent machinery, where both strands of the outer repeats are transcribed by RNA polymerase II. Double-stranded RNA molecules are generated with the help of RNA-dependent RNA polymerase I (Rdp1) and processed by Dicer (Dcr1) to yield small interfering RNAs (siRNAs). The resulting duplex siRNAs are loaded onto the Argonaute (Ago1) complex and converted into single-stranded siRNAs after cleavage and release as the passenger strand in the RNA-induced transcriptional silencing (RITS) complex. The RITS complex also recruits H3K9 methyltransferase, Clr4. H3K9 methylation stabilizes the association of RITS with centromeric chromatin and also provides binding sites for Swi6. B. Experimental deletion of native centromere leads to formation of neocentromeres at a site proximal or distal to the native centromere in *C. albicans*. Note that the frequency of neocentromere activation, proportional to the area of halo around the blue circles, is higher at the centromere-proximal location than the centromere-distal sites in *C. albicans*. In *S. pombe*,

a heterochromatin-mediated mechanism guides the activation of neocentromeres at sub-telomeric regions. C. The spatial clustering of centromeres, either constitutive or cell cycle stage-specific, is a unique feature across fungal species. (*Left* and *middle* panels) Clustering patterns for representative fungal species have been depicted by kinetochores (green) arranged at the periphery of the nuclear mass (black). *Right*, the microscopic observations of spatial clustering have been supported by 3C-seq and derived techniques. Evolutionarily conserved clustering of centromeres near the spindle pole body at the nuclear periphery in fungi is shown as a schematic drawing.

Neocentromeres, which are sites acquiring centromeric properties in the event of native centromere inactivation, serve as an excellent tool to study factors contributing to centromere establishment. Systematic deletion of CENP-A^{Cse4} binding and the flanking DNA sequences in *C. albicans* resulted in the formation of neocentromeres in close proximity to the deleted region (241) (Figure 1.15B). An independent study reported the activation of both proximal and distal neocentromeres in *C. albicans* (242). Strikingly, Hi-C studies indicated that even the distal neocentromere clusters with other native centromeres of various chromosomes. This indicates that proximity to the CENP-A–rich zone or CENP-A cloud where endogenous centromeres cluster together at the nuclear periphery is a stronger determinant than the DNA sequence itself for neocentromere establishment in this organism (241). This was seen to be consistent in *C. dubliniensis* as well (241). On the other hand, the conditional deletion of a centromere in *S. pombe* produces survivors in which chromosomes are largely rescued by telomeric fusions with another chromosome or, in rare cases, activate a neocentromere at a sub-telomeric region (243) (Figure 1.15B). The similarities in the heterochromatin environment at both these loci and the presence of sequences homologous to the *dg* and *dh* elements identified in the sub-telomeric regions explain the preferential activation of neocentromeres at these loci (244).

Reversible transgene-silencing is a unique feature of centromeric chromatin. When a transgene, *URA4* or *ADE2,* is integrated at the central region of centromeres in *S. pombe*, the transgene undergoes reversible transcriptional silencing, rendering variable expression patterns (245-247). However, the expression of the same transgene integrated at the outer repeats was efficiently turned off due to the highly heterochromatic nature of these repeats (248). The boundary of centromeric heterochromatin that retains the property of reversible silencing was determined in *C. albicans* by integrating *URA3* as a transgene at the core and centromere-flanking regions. This study suggested that flexible positioning of CENP-A^{Cse4} within a domain that permits neocentromere activation when the native centromeric DNA

sequence was deleted (249). In *S. pombe,* however, the tRNA genes were identified as the boundary elements between CENP-A^{Cnp1} chromatin and flanking heterochromatin regions (250). These studies indicate that low levels of CENP-A^{Cse4} can be present beyond the 3-5 kb region of centromeric chromatin in the absence of any boundary elements in *C. albicans*, while CENP-A^{Cnp1} is restricted by defined boundary elements in *S. pombe*. Structural boundary elements are not identified in other classes of regional centromeres, and thus, it is not well understood what restricts the length of the functional centromeric chromatin that seeds kinetochore assembly.

New insights into factors required for centromere function could be gained by studying the fate of dicentric chromosomes. In *S. cerevisiae*, dicentric chromosomes are unstable but are stabilized exclusively by DNA rearrangements when one of the two centromeres becomes inactivated (251). The artificial dicentric chromosome generated in *S. pombe* using site-directed recombination led to a cell cycle arrest at the interphase stage. Less than 1% of the survivors were shown to inactivate one of the centromeres either by DNA sequence rearrangement or by heterochromatinization of centromere DNA sequence leading to epigenetic inactivation (252). The fact that in spite of the presence of several potential neocentromere sites, the native centromere always serves as the sole functional centromere indicates the existence of an active suppression mechanism to keep neocentromeres dormant. Maintenance of centromeric chromatin involves the efficient propagation of already established centromeric chromatin marks. Even the genetically determined point centromere in *S. cerevisiae* displays an epigenetic mode of maintenance. Chl4 is a non-essential kinetochore protein in *S. cerevisiae*. A centromeric plasmid introduced into *chl4* mutants display reduced mitotic stability. Whereas if the same mutation is introduced after the centromere is allowed to establish on the plasmid centromere, 50% of the cells show high mitotic stability, indicative of the semi-essential role of Chl4 in centromere maintenance (253). In *S. pombe*, when various centromeric plasmids with incomplete centromere DNA sequences were transformed, the mitotically unstable plasmid switched to a stable state by epigenetic means. Strikingly this stable state was efficiently propagated in subsequent cell divisions (254). These observations evoked interest in understanding the changes in the structure and composition of the chromatin environment of the centromeric plasmids.

Structure and properties of centromeric chromatin

Despite the divergence of the centromere DNA sequences, the centromeric histone H3 variant CENP-A is a unifying feature that specifies the centromere location in most fungal species. CENP-A is considered as the molecular marker for centromere specification, as it is excluded from the rest of bulk chromatin. CENP-A targeting domain (CATD) of CENP-A is required for its efficient targeting to the centromere (255).

Most of the biophysical and structural studies on centromeric chromatin have been performed in *S. cerevisiae* and *S. pombe*. Initial methods to dissect the differences between CENP-A nucleosomes and H3 nucleosomes utilized *in vitro* reconstitution assays. There are ongoing debates regarding the stoichiometry and DNA binding properties of CENP-A chromatin, which differs significantly from H3 chromatin. First, the budding yeast centromeric nucleosome forms a stable octamer containing two molecules of CENP-A^{Cse4} that wrap the DNA in a left-handed manner both *in vitro* and *in vivo* (256). Second, partial micrococcal nuclease digestion of centromeric chromatin generates a 125-bp fragment, shorter than the 147-bp fragment protected in an H3 nucleosome (257). The tight association of Cbf1 and the CBF3 complex to the tetramer nucleosome has been proposed to help in the formation of the bigger centromeric complex. Later it was shown that a dimer of Ndc10 binds to ~50 bp fragment of centromere DNA, leaving ~80 bp DNA to wrap around a single nucleosome (258). Finally, CENP-A nucleosomes induce an alternative topology, which is positive supercoiling, consistent with right-handed wrapping (259). This altered handedness has been suggested for kinetochore accessibility and binding of non-histone proteins that can regulate the centromeric activity. Later, the measurement of DNA linking number difference (ΔLk) revealed that the positively supercoiled centromere DNA achieves a ΔLk value of +0.6 in the CEN-ARS plasmid context. Induction of positive supercoiling of the centromeric nucleosome remains unaffected by alterations of CDEII length and Cbf1 binding to CDEI, but lost when the CDEIII sequence was mutated (260).

In vitro reconstitution experiments revealed the presence of a single octameric nucleosome at the budding yeast point centromere, which comprises of a dimer of CENP-A^{Cse4}, two molecules each of H2A, H2B, and H4 (261). It has also been proposed that the chaperone Scm3 evicts H2A, H2B leaving two copies each of Scm3, CENP-A^{Cse4}, and H4 to form a hexameric nucleosome at the centromere (262). The hexameric model is difficult to

reconcile as previous reports show that H2A is important to maintain centromere function (263). Another proposed model for centromeric chromatin is the hemisome model, which was only observed in the interphase cells of *Drosophila* (264). It was later shown that CDEII is a good substrate for hemisome formation under high salt concentrations in *S. cerevisiae* (265). This also provides structural evidence for the importance of AT-rich regions at the centromeres. AT-rich regions are stiff to prevent octamer formation, and hence favor hemisomes.

Regional centromeres of *S. pombe* have multiple nucleosomes of CENP-A, which have distinct chromatin properties than bulk H3 chromatin, as seen by MNase digestion assays (266). The array of CENP-A nucleosomes is orderly positioned in the unique mid-core sequences of *S. pombe*. The smeary pattern obtained upon MNase digestion of centromere DNA was largely attributed to the protection rendered by the kinetochore complex. Similar smeary patterns were observed in fission yeast centromeres during meiosis (61) and also in *C. albicans* centromeres (224). A contrasting report emerged where the *in vivo* chemical mapping of nucleosome positioning in *S. pombe* revealed a fuzzy positioning of CENP-A nucleosomes at the central core (267). The reason for such an irregular positioning was attributed to the AT richness of the central core. On the other hand, centromeres in *C. albicans* do not show any AT-richness, even though they display the smeary MNase digestion pattern similar to *S. pombe*. High-resolution mapping of centromeric nucleosomes revealed that H3 nucleosomes are nearly absent from the central core in *S. pombe* (268). The central core is also enriched with two H4 molecules per nucleosome that are unpositioned and widely spaced than the flanking H3 nucleosomes. The same study revealed the enrichment of CENP-T and Scm3 throughout the central core, with CENP-T linking adjacent CENP-A nucleosomes. Scm3 has shown to copurify with CENP-A^{Cse4} *in vitro* (269). *In vivo*, Scm3 dissociates from the centromere transiently during early mitosis, whereas CENP-A^{Cse4} is constitutively present at the centromeric nucleosome (269). Hence, it is favorable to consider Scm3 to be a stable incorporator of CENP-A rather than a constitutively present structural component at the centromeric nucleosome.

Intriguingly, dyad symmetry at the centromere DNA of *S. pombe* was found to be correlated with enrichment of non-B-form DNA, which remains conserved in other domains of life (270). Centromere DNAs are found to be covalently modified in certain fungal species. Cytosine methylation of DNA (5mC) has been implicated in the formation of

heterochromatin in *N. crassa* (271, 272), which is also present in *Cryptococcus* species (207). On the contrary, the centromeres in *C. albicans* are devoid of DNA CpG methylation marks (273). Therefore, the requirement of these covalent modifications for centromere function is not conserved across all fungal models. Together, the higher-order centromeric and pericentromeric chromatin structure, the unusual secondary structure of centromere DNA and its covalent modification might influence the complex regulation of centromere identity.

The structural organization and topology of the centromeric chromatin are studied in detail in *S. cerevisiae.* The three-dimensional structure of the centromeric and pericentromeric chromatin was dissected using microscopic studies of fluorescently labeled spindle pole body proteins, centromeric markers, and cohesin molecules (274-276). These studies revealed the presence of pericentric loops with the centromeres located at the apex of each loop (276). These loops display coordinated motion and stretching in metaphase (277). Higher-order organization of pericentric chromatin through loops is proposed to function as nonlinear spring, which helps in mitotic force balance (274) and facilitates correct orientation and stabilization of the microtubule attachment sites (277).

Regulation of centromere identity in space and time

The positioning of centromeres at the nuclear periphery near SPBs in a transcription-poor zone facilitates spindle attachment and shields the centromere from pervasive transcription (278). Centromeres are clustered throughout the cell cycle in *S. cerevisiae* (279) and *C. albicans* (223) and the existence of a CENP-A-rich zone or CENP-A cloud at a perinuclear space has been proposed (241). In *S. cerevisiae*, a locally enriched population of CENP-A^{Cse4} molecules at pericentromeres serves as a reservoir for the rapid incorporation of CENP-A^{Cse4} in an event when they are prematurely evicted from the centromeres (280). The CENP-A cloud hypothesis stems from the activation of native centromere-proximal neocentromeres in *C. albicans* (241, 281). Unlike budding yeast, centromeres in fission yeast cluster during interphase and uncluster for a brief period during mitosis (282). These clustered centromeres are attached to the nuclear envelope near the site of SPBs during interphase (283). In *C. neoformans*, unclustered centromeres in interphase eventually cluster at the mitotic onset in a microtubule-dependent manner (284). Apart from unicellular yeasts, centromere clustering has also been observed in filamentous fungi like *Fusarium graminearum, N. crassa, M. oryzae*, wherein with the exception of *M. oryzae*, all centromeres

were found to constitutively cluster as a single punctum by fluorescence microscopic analyses (Figure 1.15C) (235, 285). Despite the differences in the centromere clustering patterns across fungal species examined to date, it has been consistently shown that centromere clustering is important for proper kinetochore-microtubule attachment during mitosis (283, 286, 287).

Recent progress in microscopic imaging and sequencing techniques has enabled the successful mapping of distinct compartments within the nucleus to address fundamental questions regarding the structure and functional states of chromosomes. 3C-sequencing in *S. cerevisiae* revealed that the clustered centromeres are present in close spatial proximity, leading to physical interactions between different chromosomes (Figure 1.15C) (169). In *S. pombe*, where heterochromatin is a major determinant of centromere organization, centromere-proximal regions interact with a higher contact frequency, as revealed by Hi-C analysis. A similar correlation supported by Hi-C analysis in *N. crassa* revealed predominant interactions across constitutively heterochromatic regions enriched with H3K9me3 and HP1. Due to the conserved clustering features of fungal centromeres, Hi-C and related techniques have been used to accurately predict centromere loci in fungal genomes (175).
In the absence of heterochromatic marks and well-defined DNA sequences, what determines the clustering of centromeres remains an enigma. Clustering of *C. albicans* centromeres, which are devoid of conventional heterochromatin, indicates the involvement of additional factors facilitating this process (249). As discussed previously, centromere clustering favors the site of centromere formation in subsequent cell cycles, possibly by CENP-A nucleation. Surprisingly, even in the CENP-A-deficient species *M. circinelloides*, centromeres are constitutively clustered both in the spore and the germinating tube (Figure 1.15C) (211).

Although the biological implications for the conserved spatial organization of centromeres in fungi remain to be explored, its impact on the replication program of the genome is being revealed recently (160, 288). Centromeres are spatially and temporally distinguishable from the rest of the genome owing to their distinct clustering patterns and replication timing, respectively. Centromeres are replicated in the earliest part of the S phase in certain *Saccharomyces* species (289), *C. albicans* (290), and *S. pombe* (291). What is the significance of fungal centromeres being early replicating? Early replication timing ensures proper kinetochore assembly at the centromeres (292) and helps to maintain the viability of cells in the face of any replication stress in *S. cerevisiae* (293). Early replication of

centromeres due to the early firing of the centromere-proximal origins can be attributed to their characteristic clustering and nuclear sub-positioning (294). The relocation of a centromere to a late firing region resets the replication timing of the latter, reinstating that the mere presence of a centromeric sequence can modulate replication timing (289). DNA replication fork pause at centromeres helps in centromere DNA loop formation, which is essential for sister centromere separation and kinetochore assembly in *S. cerevisiae* (295, 296). In *S. pombe*, centromeres and the sub-telomeric regions have similar heterochromatin environment but differ in their replication timings. The heterochromatin protein Swi6 helps in early replication of centromeres (297, 298), exhibiting the prominent role played by heterochromatin in influencing replication timing and the consequent effect on centromere function.

The temporal effect on the DNA replication origin firing has also been studied in *C. albicans,* in which deletion of a native centromere gives rise to a neocentromere with the activation of an early firing neo-origin (290). This clearly states that centromeric location positively influences the replication timing of the adjacent regions. In *Y. lipolytica,* a centromere-linked replication origin, helps to maintain plasmid stability (225). Hence, centromere-proximal origins seem to have a role more than just acting as initiator sites for DNA replication.

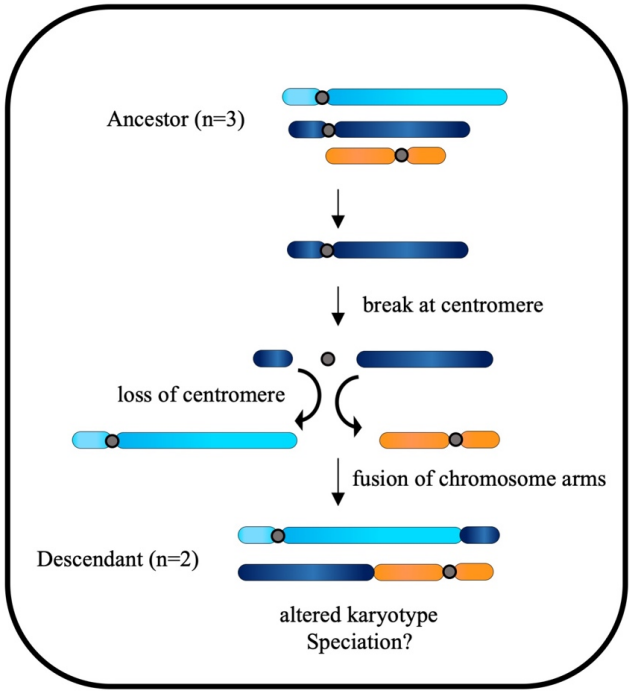


Figure 1.16 Centromere mediated karyotype evolution in fungal species.
A possible consequence of chromosome breakage at centromere. The resulting acentric fragments can be stabilized by fusion with other chromosomes, eventually leading to an

altered karyotype as observed in the species complexes belonging to Ascomycota and Basidiomycota.

Cause and consequence of centromere mediated genome rearrangements

The mechanisms contributing to the rapid evolution of centromere DNA, especially in asexual fungi, remain unclear. Genome synteny analyses in *C. albicans,* and *C. tropicalis* helped identify genomic rearrangements near centromeres suggesting inter-centromeric translocations in the last common ancestor (193). A similar inter-centromeric translocation has been observed in the common ancestor of *S. cryophilus* and *S. octosporus* (205). These examples indicate that centromeres can mediate karyotypic rearrangements. In fungi, centromere mediated rearrangements can lead to a change in chromosome number. One such example is found in *Eremothecium gossypii,* where a break at the centromere following fusion of the broken arms with two other chromosomes in the ancestor led to chromosome number reduction (178). A similar mechanism of centromere breaks resulting in chromosome number alteration has been reported recently in the *Malassezia* species complex (201) (Figure 1.16). In mammals, centromere-proximal translocation events are popularly known as Robertsonian translocation, which causes sterility in humans (299) and often linked with the heterogeneity of carcinomas (300). Robertsonian translocations are also proposed to be a driver of speciation in mice (301). Since chromosomal rearrangements mediated large karyotypic differences creating incompatible homolog pairing, reproductive isolation, and speciation (302), centromere mediated karyotype reshuffling can be one of the mechanisms driving speciation. Striking evidence supporting this notion was obtained from experimental evolution based study, which showed karyotype shuffling after CRISPR guided centromere scission led to reproductive isolation in *C. neoformans* (303).

Although the number chromosomes present in different species can be different, the biological significance of this factor remains elusive. To address this question, two independent groups adopted an elegant approach chromosome engineering, which involves CRISPR mediated centromere deletion following fusion of the acentric chromosomes through the telomeres (132, 304). Applying this technique, each of these two groups constructed a strains of *S. cerevisiae* with either one or two chromosomes. Then they have used this strain to assay for changes in phenotypic traits, compaction of chromosomal domains, and tested for reproductive isolation between the engineered strain and the parental

strain. These studies showed chromosome number change leads to reproductive isolation. Although the reduction in chromosome number led to a reduction in the spore production and compromised fitness across environments, the overall transcriptome and properties of local chromatin compaction remained unperturbed (132, 304).

Karyotypic rearrangements often lead to the repositioning of centromeres or the formation of neocentromeres. For example, the fusion between two chromosomes can create a dicentric chromosome with two active centromeres, the catastrophic consequence of which was originally identified in maize and described as 'breakage-fusion-bridge' cycle leading to genomic instability (305). Since then, the occurrence of the 'breakage-fusion-bridge' cycle has been reported in cancers (306), tetraploid mouse cells (307), *D. melanogaster* (308), *C. elegans* (309) and fungal pathogen *Zymoseptoria tritici* (310). However, inactivation of one centromere can rescue the fused chromosome and stabilize the karyotype (306, 311, 312). On the contrary, acentric chromosomal fragments originating from chromosome breakage can be stabilized by the formation of neocentromeres, which was first identified in humans (313). Other examples of naturally occurring neocentromeres are reported in plants (314). Apart from these two cases, centromere repositioning events on an evolutionary time scale may lead to the formation of evolutionary new centromeres (ENCs), which are often associated with speciation in mammals (315, 316). It was found that the location of one centromere in horse varies across individuals (317, 318). The driving force facilitating centromere relocation was proposed to be associated with chromosomal inversion and translocation in certain cases (319).

However, the altered karyotype should also offer a fitness advantage for it to be selected over the ancestral karyotype. Since it is difficult to predict the factors driving speciation, the fitness advantages conferred by species-specific rearrangements are not well understood. One of the ways to achieve karyotypic alteration is through centromere-mediated chromosomal rearrangements. An example of such a translocation includes the bipolar to tetrapolar mating-type transition in the *Cryptococcus* species complex involving a pericentric inversion, thereby rewiring the regulation of the mating-type locus (206). Another instance where karyotype alteration provides specific fitness advantage involves the generation of an isochromosome of chromosome 5L in *C. albicans,* which confers fluconazole resistance (25). Thus, centromere DNA, one of the guardians of genome stability, may contribute towards chromosomal rearrangements and possibly speciation. 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 (320-322). In *C. tropicalis* the centromeres are formed on highly homogenized inverted repeat (HIR)-associated DNA sequences. Therefore, it is possible that genomic rearrangements through spatially proximal and HIR-associated centromeres in the last common ancestor led to the loss of HIRs and the emergence of ENCs in *C. albicans.*

Rationale of the present study

Our previous analysis suggested that centromeres of *C. tropicalis* are located near interchromosomal synteny breakpoints (ICSBs), which are relics of ancient translocations in the common ancestor of *C. tropicalis* and *C. albicans* (193). Additionally, the subcellular localization of the kinetochore proteins as a single punctum per nucleus indicated the clustering of centromeres in *C. tropicalis* (193). These observations suggest an intriguing possibility that ICSBs are specifically enriched near the centromere cluster. 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.

Available information and plan for construction of chromosome-level genome assembly of *C. tropicalis*

The currently available nuclear genome assembly of *C. tropicalis* type strain MYA-3404 contains 23 contigs for the nuclear genome and one contig for the mitochondrial genome, comprising of 14.58 Mb, with a contig N50 of 221,103 bp (ASM633V3) (6). However, 14 unique sequences associated with the telomeric repeats have been identified in *C. tropicalis,* which indicate the presence of seven pairs of telomeres and, therefore, seven chromosomes (6). Later, the identification of seven centromeres in this species supported the idea that this species contains seven pairs of chromosomes in its nuclear genome (193) (Figure 1.17). Therefore, the currently available assembly of the *C. tropicalis* nuclear genome in 23 scaffolds contains 16 gaps, which need to be closed to construct a complete chromosome-level genome assembly.

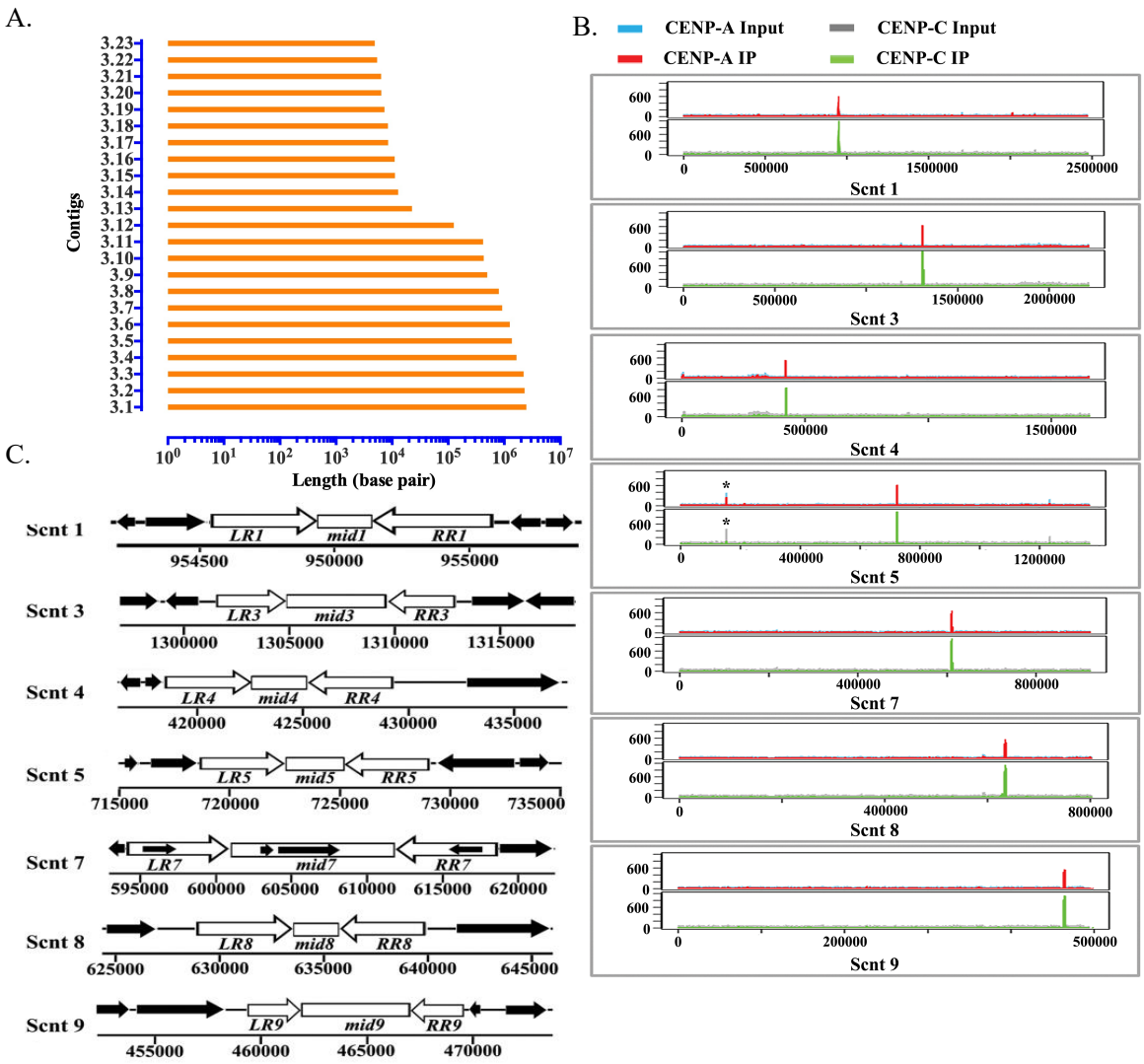


Figure 1.17 Identification of seven inverted repeat-associated centromeres in contig-level genome assembly of *C. tropicalis.*
A. Assembly of *C. tropicalis* nuclear genome in 23 contigs is represented in the order of increasing length (*x*-axis) (6). B. Genome-wide binding profile of conserved kinetochore proteins CENP-A^{Cse4} and CENP-C revealed by ChIP-seq analysis identified seven centromeres in *C. tropicalis*. Plots showing enrichment of CENP-A^{Cse4} (red), and CENP-C (green) ChIP-seq reads (*y*-axis) across the genomic coordinates (*x*-axis) at seven distinct locations on *C. tropicalis* genome of seven different contigs (193). C. Analysis of the centromere DNA sequences revealed the presence of 2 - 3 kb long inverted repeat flanking the CENP-A- and CENP-C-bound region on all seven centromeres (323).

Identification of HIR-associated centromere in this species raises another concern about the contiguity of the assembly across the centromeres. As the available contigs were generated using short Illumina sequence reads, there is a chance of mis-assembly creating a chimeric chromosome in which the chromosome arms are

A.

	N_{tot}	$N_{aligned}$	$N_{conserved}$	Percent aligned	Percent conserved
mid, avg	31437	15711	12465	50.0	79.34
IR, avg (other)	51332	34047	30571	66.3	89.80
IR, avg (self)	51332	48478	46248	94.4	95.30
IR, Scnt 1	7753	7394	6854	95.3	92.90
IR, Scnt 3	4941	4654	4330	94.2	93.00
IR, Scnt 4	7464	7348	7108	98.4	96.70
IR, Scnt 5	7571	7222	7002	95.4	97.00
IR, Scnt 7	11569	11394	11278	98.5	99.00
IR, Scnt 8	7663	7500	7164	97.9	95.50
IR, Scnt 9	4371	4276	4088	97.8	95.60

B.

Figure 1.18 Homogenized inverted repeat-associated centromere loci in *C. tropicalis* increases the possibility of mis-assembly.
A. Total number of bases (N_{tot}), number of aligned bases ($N_{aligned}$), number of conserved bases ($N_{conserved}$), percent aligned, and percent conserved bases among the IRs, and *mid* elements of seven centromeres in *C. tropicalis* as reported previously (323). B. *Left,* schematic drawing of four chromosomes showing the actual karyotype and *right,* possible mis-assembly due to swapping of chromosomal arms at the homogenized centromeres.

---

swapped at the centromere (Figure 1.18). In addition, the current genome assembly has multiple N-gaps. Therefore, we chose to leverage the recent development in the long-read sequencing chemistry and performed SMRT sequencing of the already sequenced *C. tropicalis* strain MYA-3404. Additionally, we used 3C-seq data to validate the contiguity across the chromosomes. There have been additional improvements over the 24-contig assembly (Assembly A) by Dr. Geraldine Butler's group at School of Biomolecular and Biomedical Science, University College Dublin, Ireland. This improved genome assembly consists of 16 contigs (Assembly B). Therefore, our effort of construction of chromosome-level genome assembly of *C. tropicalis* started with these contigs (Assembly B).

Summary of the present work

Centromeres of *C. albicans* form on unique and different DNA sequences, but a closely related species, *C. tropicalis*, possesses homogenized inverted repeat-associated centromeres. In addition, centromere DNA of *C. tropicalis* can initiate *de novo* CENP-A^{Cse4} recruitment on a plasmid and improves its mitotic stability. On the contrary, DNA sequence-dependent *de novo* centromere establishment is absent in *C. albicans.* To investigate the mechanism of centromere-type transition, we improved the fragmented genome assembly and constructed a chromosome-level genome assembly of *C. tropicalis* by employing PacBio sequencing, chromosome conformation capture sequencing (3C-seq), chromoblot, and genetic analysis of engineered aneuploid strains. Further, we analyzed the 3D genome organization using 3C-seq data, which revealed spatial proximity among the centromeres as well as telomeres of seven chromosomes in *C. tropicalis*. Intriguingly, we observed evidence of inter-centromeric translocations in the common ancestor of *C. albicans* and *C. tropicalis*. Identification of putative centromeres in closely related *Candida sojae*, *Candida viswanathii* and *C. parapsilosis* indicate loss of ancestral HIR-associated centromeres and establishment of evolutionary new centromeres in *C. albicans*. Based on these results, we propose that spatial proximity of the homologous centromere DNA sequences facilitated karyotype rearrangements and centromere type transitions in human pathogenic yeasts of the CUG-Ser1 clade.

Chapter 2

Results

A gapless assembly of *Candida tropicalis* genome in seven chromosomes

Construction of chromosome-level assembly of *C. tropicalis*

Our efforts to construct a chromosome-level assembly of *C. tropicalis* type strain MYA-3404 comprises of five key steps (Figure 2.1). First, the SMRT-seq long reads were used for scaffolding the Assembly B contigs using Single Molecular Integrative Scaffolding (SMIS) software (https://github.com/fg6/smis) producing Assembly C. Second, we integrated our contour clamped homogenized electric field (CHEF) gel data, *de novo* assembled contigs generated using the SMRT-seq data, and contact probability data to join two contigs and rectify a mis-join in Assembly C, which produces assembly D. Third, based on the analysis of the genome sequence data and genetic analysis, we identified the smaller contigs present in assembly D as heterozygous loci in the diploid genome of *C. tropicalis* strain MYA-3404.

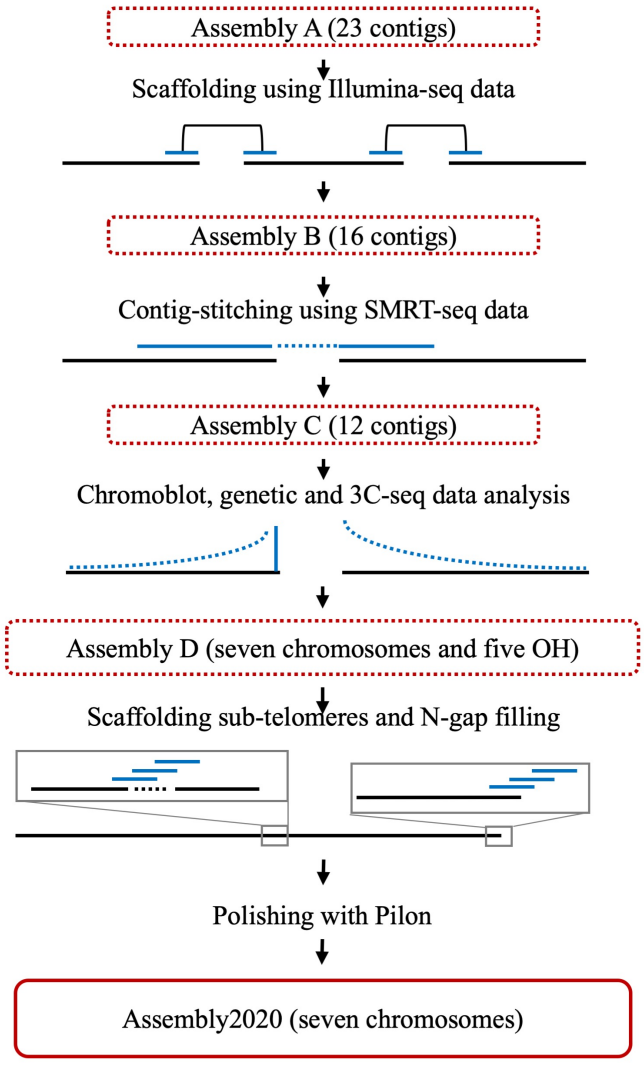


Figure 2.1 Schematic showing the stepwise construction of the gapless chromosome-level assembly (Assembly2020) of *C. tropicalis*.

Fourth, using the *de novo* contigs generated from the SMRT-seq data, the preexisting N-gaps were filled, and the sub-telomeres were scaffolded. The fifth and final step was polishing of the chromosomes with Illumina sequence data using Pilon (324). The final assembly of *C. tropicalis* strain MYA-3404 in seven complete chromosomes was named as Assembly2020. A schematic of this procedure is detailed in figure 2.1. Genome sequencing of *C. tropicalis* strain MYA-3404 was performed on PacBio Sequel platform with version V2 chemistry. The genomic DNA molecules were size selected to enrich ~20 kb long fragments and used for library preparation (Materials and methods). This run generated 996041

Table 2.1: Details of the genome assembly with 13 contigs after SMIS run.
*As reported by Chatterjee G. *et al.,* 2016.

Sl. No.	Contig No.	Length (bp)	CEN*	No. of N-gaps	Feature
1	2	2797331	9	20	NA
2	1	2474493	1	15	NA
3	10	2342205	3	22	NA
4	11	2098823	4	15	rDNA locus
5	5	1379441	5	17	Not observed in CHEF gel
6	6	1255791	None		
7	8	1225766	8	7	MTLa
8	7	921083	7	8	NA
9	12	12816	None	-	MTLalpha
10	13	22703	None	-	NA
11	14	8448	None	-	NA
12	15	6389	None	-	NA
13	16	5408	None	-	NA
Total	13	14652842	7	104	

reads with an average read length of 5.8 kb. This data was used to scaffold the contigs present in Assembly B using SMIS (https://github.com/fg6/smis) with the default parameters (Materials and methods). This analysis led to the joining of four more contigs and produced Assembly C with 12 contigs comprising of 14652842 bp (Table 2.1). Additionally, the

SMRT-seq reads were used to run Canu (77) and FALCON (78, 325) to generate *de novo* assemblies of *C. tropicalis*.

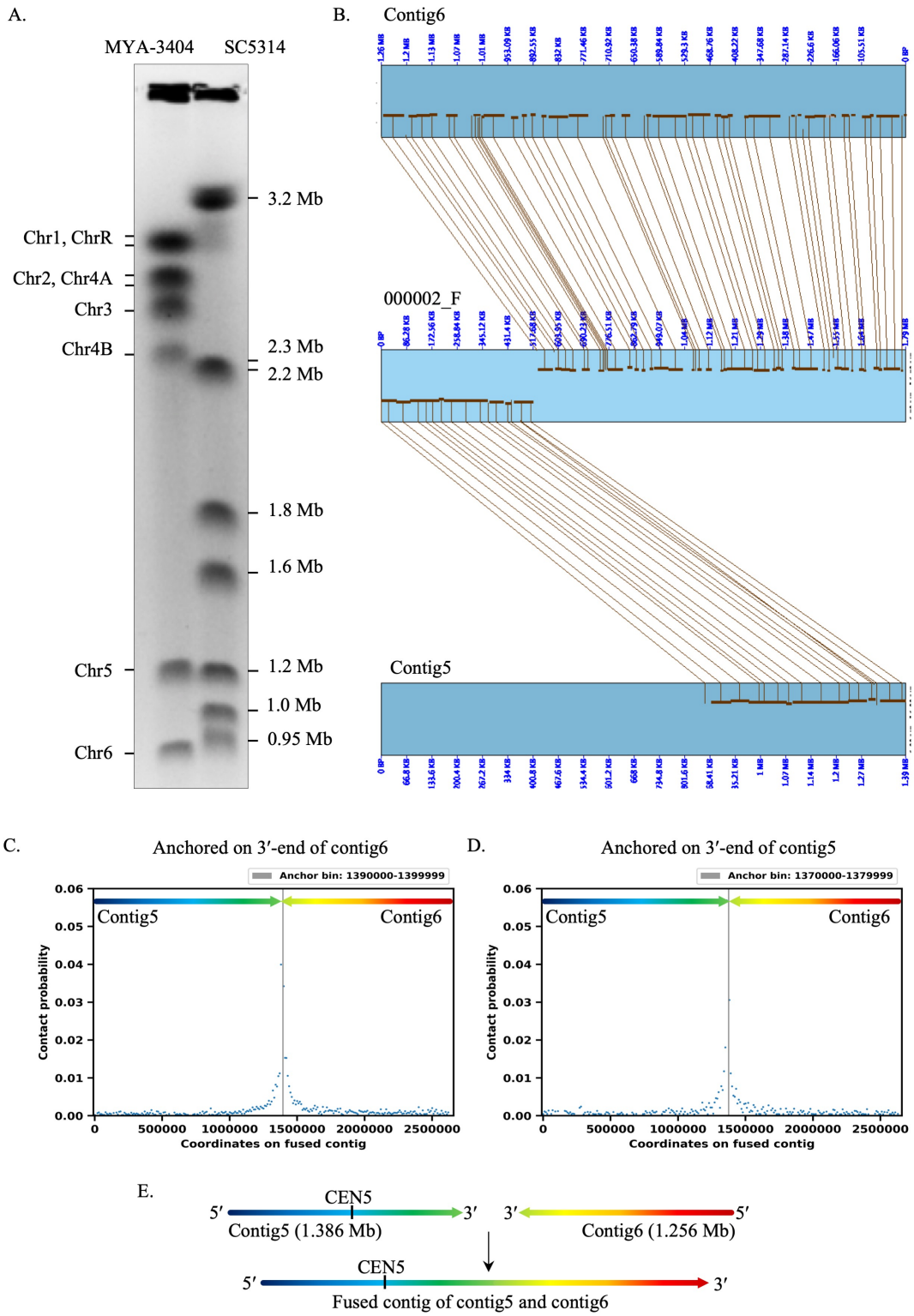


Figure 2.2 CHEF-karyotyping, analysis of *de novo* contigs and 3C-seq contact probability data indicate contig5 and contig6 are parts of the same chromosome.
A. An ethidium bromide (EtBr)-stained CHEF gel image where the chromosomes of the *C. tropicalis* strain MYA-3404 and *C. albicans* strain SC5314 were separated (Materials and

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. B. A synteny map of *de novo* FALCON generated contig 000002_F with respect to contig5 and contig6 indicate that these two contigs should join in a 3' to 3' orientation. The synteny map was generated using Symap (326) (Materials and methods). C. The 3C profile (bin size = 10 kb) of 3′-terminal bin of contig6 (anchor; gray vertical line) showing its contact probabilities (blue dots) with bins on contig5 and contig6. D. The 3C profile (bin size = 10 kb) of 3′-terminal bin of contig5 (anchor; gray vertical line) showing its contact probabilities (blue dots) with bins on contig5 and contig6. E. The cartoon representation of chromosome 2 assembly by fusing contig5 and contig6 in a tail-to-tail (3′ to 3′) orientation based on the 3C profile results.

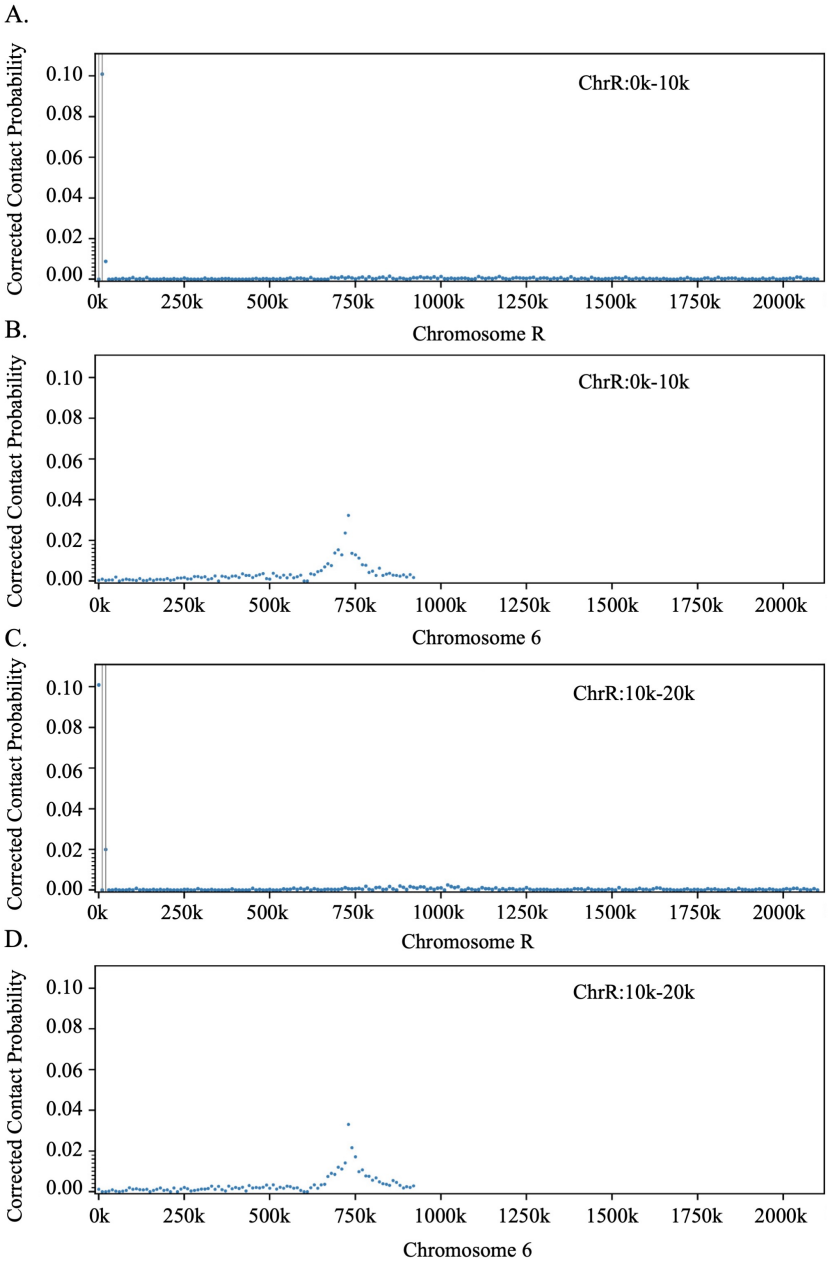


Figure 2.3 Use of contact probability data from 3C-seq experiment for correction of mis-assembly of contig13 in Assembly C.

A. and B. 4C-like plot showing contact probability profiles (blue dots) of ChrR and Chr6, respectively, with respect to the anchor bin ChrR:0k-10k. C. and D. 4C-like plot similar to A and B, with respect to the anchor bin at ChrR:10k-20k. Both anchor bins are located on the contig13 fused to the 5′ end of ChrR. In all the plots the *y*-axis represents corrected contact probability and *x*-axis represents the chromosomal coordinates.

In Assembly C, the length of contig5 is 1379441 bp. However, we could not detect any chromosome of this size in our CHEF gel analysis (Figure 2.2A). It was also found that 1255791 bp long contig6 does not carry any centromere, suggesting it is part of another chromosome. We performed a dot plot analysis between Assembly C and the *de novo* assembled contigs by FALCON (78, 325) to prove that contig5 and contig6 are indeed part of the same chromosome. This analysis suggested that the contig5 and contig6 are joined together through their 3' end in a single chromosome (Figure 2.2B). Based on this data, we joined contig5 and contig6. In addition, the contact probability decay pattern also supported that these two contigs are part of the same chromosome (Figure 2.2C - E). The sequence coverage analysis and contact probability decay pattern suggested that the contig13 of Assembly B is wrongly assembled in Assembly C (Figure 2.3A - D). We corrected the mis-assembly. Therefore, after joining of contig5 with contig6 and correction of mis-assembly of contig13 resulted in Assembly D, which contain seven chromosome-length scaffolds, and five small contigs. We referred these five small contigs as orphan haplotigs (OH).

One of these five OHs, contig12 carries mating type locus (*MTLα*), which is present as a heterozygous locus in the diploid genome of *C. tropicalis* (10). Therefore, this observation indicated a possibility that the other smaller contigs present in Assembly D are also heterozygous loci in the diploid genome of *C. tropicalis*. To test this possibility, sequence coverage on these contigs were analyzed. Quantification of the sequence coverage suggested that all five OHs are either partly or completely heterozygous regions (Figure 2.4A - B). Further, analysis of the *de novo* contigs generated by diploid-aware assembler Canu (77) suggested that these OHs are heterozygous loci, the chromosomal coordinates for which were also mapped (Figure 2.4C). In order to experimentally validate that the OHs are indeed heterozygous loci, we performed genetic analysis.

We constructed *C. tropicalis* strains monosomic for Chr5 and used them to demonstrate that loss of one homolog of Chr5 leads to loss of one of the two alleles of the orphan contigs: contig14 and contig16, that are mapped on Chr5. Since the *sch9* mutants in

C. albicans were viable but lost chromosomes at a significantly higher rate than the wild-type (327), we adopted the same strategy to delete both copies of *SCH9* homolog in *C. tropicalis*. Next, a reporter strain was created in this *sch9* mutant strain background of *C. tropicalis* to assay for loss of a Chr5 homolog. These strains (2n-1) that lacked one homolog of Chr5 were used to confirm the heterozygosity of orphan haplotigs (OHs) of CtChr5 (Figure 2.5A).

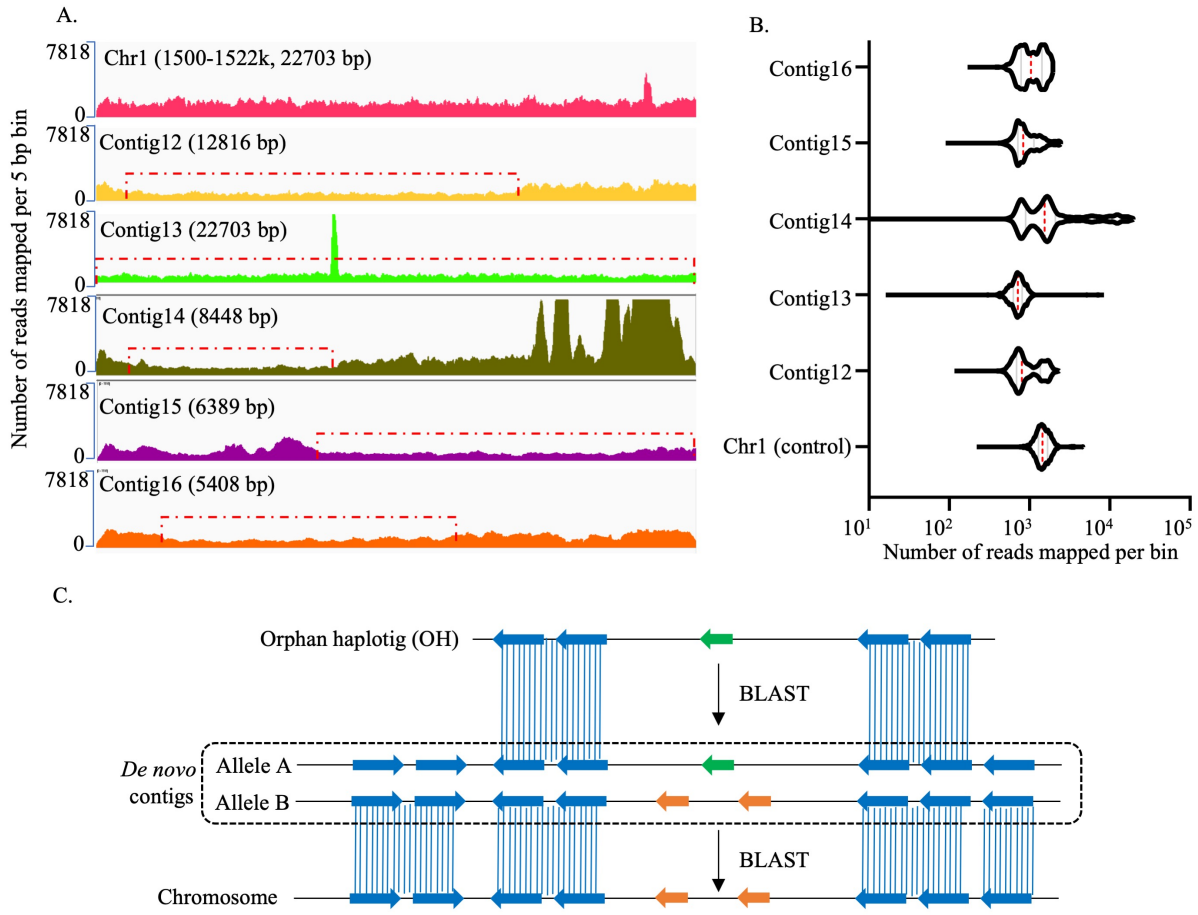


Figure 2.4 Analyses of the sequence coverage data and *de novo* contigs suggest orphan haplotigs are heterozygous loci in *C. tropicalis* genome.
A. IGV track images showing the coverage of 3C-seq data on the *y*-axis (number of reads mapped per bin for each million of the total reads) over the orphan contigs and a control locus from Chr1. B. A violin plot showing the distribution of 3C-seq read coverage across the OHs (bin size = 5 bp) and a control region on Chr1 was generated using deepTools2 bamCoverage script. C. OHs were mapped to the chromosomes by performing two step BLAST analysis. First, the *de novo* contig bearing OH locus was identified. Next, 10 ORFs present on each side OH homology region located on the *de novo* contig were used as query sequences against the chromosomes of *C. tropicalis* Assmebly2020. The allelic difference between the OH and its chromosomal homolog is depicted by color-coded ORFs (orange and green arrows).

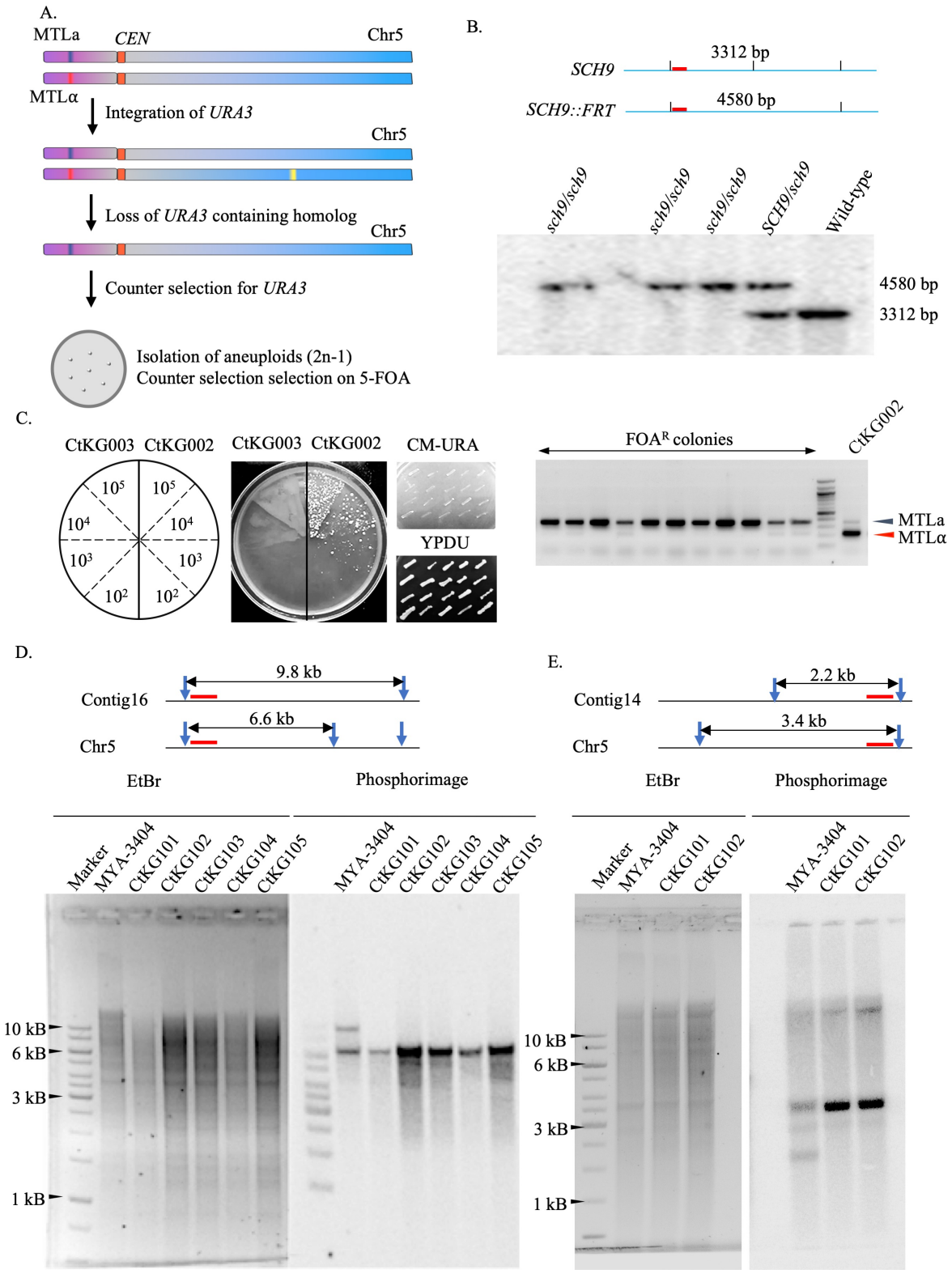


Figure 2.5 Genetic analysis of engineered monosomic strains to confirm heterozygosity of the orphan haplotigs.

A. Schematic for construction of monosomic strain with only one homolog of Chr5. B. Schematic showing the positions of HindIII sites (vertical black lines) and the length of the expected bands detectable by the probe used (red bars) in Southern hybridization to confirm *sch9* deletion strain (CtKG001). Phosphorimage of the blot showing result of the southern

hybridization experiment for confirmation of transformants of *sch9Δ/sch9Δ* mutant strain (CtKG001). C. Different number of cells (10^5, 10^4, 10^3, and 10^2) of CtKG002 and CtKG003 were plated on CM+FOA plate along with wild-type controls, as shown in schematic. Isolated FOA^R colonies were picked up, patched on CM-URA, and YPDU plates, incubated at 30°C for 48 h and imaged. Loss of *MTL* alleles in each of individual FOA^R colonies were tested using multiplex PCR experiment with the parental control strain CtKG002. The EtBr stained gel picture for the multiplex PCR experiment is shown with the bands for *MTL***a** and *MTL*α marked with blue and red arrowheads, respectively. D. and E. Experimental validation of the allelic nature of contig16 and contig14, respectively, using Southern blot analysis. The length of restriction fragments polymorphism between the alleles after digesting with ClaI and EcoRI (restriction enzyme sites are indicated using blue arrowheads) are graphically represented for contig16 and contig14, respectively. The lanes in ethidium bromide stained gels (*left*) and corresponding phosphorimages (*right*) represent the wild-type MYA-3404 (2^{nd} lane) and the monosomic aneuploid strains. The location of the probes used in this experiment are marked using red bars. The genotype of each strain used in this experiment, and the primers used to amplify probes are mentioned in Appendix-I and Appendix-II, respectively.

The *SCH9* homolog in *C. tropicalis* was identified in a BLAST search using *CaSCH9* as the query sequence against the *C. tropicalis* proteome. A putative homolog of *SCH9* was located on Chr1:1994521-1996662 and encoded by the Crick strand. A deletion cassette (pKG1) for double homologous recombination-mediated deletion of *SCH9* ORF was constructed by cloning upstream and downstream homology regions in pSFS2a plasmid (328). This construct was transformed into CtKS102 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP* (*CaHIS1*)) for the deletion of both copies of *SCH9* ORF by recycling the *CaSAT1* marker after the deletion of the first copy of *SCH9* gene. Independent colonies of the *sch9*/*sch9* null mutant strain CtKG001 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT*) were confirmed using Southern hybridization (Figure 2.5B). Primers used in this study are mentioned in Appendix-II.

Upstream and downstream homology regions of the target intergenic locus (Chr5_497_kb) in Chr5 were amplified, and cloned into pBSCaURA3 plasmid (193) to construct pKG2 (Appendix-III). This cassette was released by restriction digestion with BamHI and ApaI and transformed into the *sch9* mutant strain CtKG001 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT*) to construct the reporter strain CtKG002 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT Chr5-497kb/ Chr5-497kb::CaURA3*). Similarly, we integrated

CaURA3 into the target intergenic locus (Chr5_497_kb) of CtKS102 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP* (CaHIS1)) to create a control strain CtKG003 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) Chr5-497kb/ Chr5-497kb::CaURA3*). In both CtKG002 and CtKG003 the short arm (5′ end) of one of the two homologs is marked with *CaURA3* marker and the long arm (3′ end) carries the heterozygous *MTL* locus (*MTL***a** or *MTL*α) with two distinct alleles present on two homologs. Concomitant loss of one of the *MTL* alleles together with *CaURA3* marker would indicate loss of one homolog of Chr5.

Table 2.2: Assembly of sub-telomeres and filling up N-gaps in the genome assembly of *C. tropicalis* using *de novo* assembled contigs.

Chr	5′ tel	3′ tel	Total N-gaps	Filled N-gap using strategy-I	Filled N-gaps using strategy-II
1	Tig4	Tig28	20	14	6
2	Tig224	Tig10	17	14	3
3	Tig236	Tig2	15	11	4
4	Tig4	Tig238 5′	22	14	8
5	Tig251	Tig1909	7	7	0
6	Tig110 5′	Tig110 3′	8	4	4
R	Tig254	Tig244	15	14	1
Total			104	78	26

Different cell numbers (10^5, 10^4, 10^3, and 10^2) of the reporter strain CtKG002 and the control strain CtKG003 were plated on complete media (CM) + 5-FOA and incubated for 48-72 h at 30°C. Multiple FOA^R colonies appeared for CtKG002 strain but no colonies appeared for the control strain CtKG003. The colonies were then patched on YPDU and CM-URA plates to confirm the loss of the *CaURA3* marker. Next, PCR was performed to confirm the loss of one of the *MTL* loci (*MTL***a** or *MTL*α) in these colonies using a multiplex PCR strategy described previously (Figure 2.5C) (13). We devised a Southern strategy to distinguish between two alleles for each of contig14 and contig16. In this strategy, using the monosomic strains of Chr5, CtKG101 – CtKG105 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy), as controls we demonstrate that the contig14

(Figure 2.5D) and contig16 (Figure 2.5E) are indeed heterozygous loci in the diploid genome of *C. tropicalis* strain MYA-3404.

Next, the *de novo* contigs generated using Canu (77) and FALCON (78, 325) were used for scaffolding the sub-telomeres and filling the N-gaps present in the chromosomes present in assembly D. Using this approach, all 14 sub-telomeres on seven chromosomes were scaffolded (Table 2.2, Figure 2.6A). Two different strategies were employed to fill the N-gaps flanked by either unique or repetitive sequences (Figure 2.6B - C; Materials and methods). All 104 N-gaps were filled using these two strategies (Table 2.2). Next, to examine if the gaps were closed correctly, 3C-seq data and SMRT-seq data were mapped on the chromosomes, and the read mapping on these coordinates was checked by manual inspection using IGV (329). Finally, the chromosomes were polished with the 3C-seq data using Pilon (324) to rectify base-pair-level errors. After following all the five steps, the final chromosome-level genome assembly of *C. tropicalis* type-strain MYA-3404 contain 14609527 bp in seven telomere-to-telomere long gapless chromosomes. We named this chromosome-level genome assembly of *C. tropicalis* as Assembly2020.

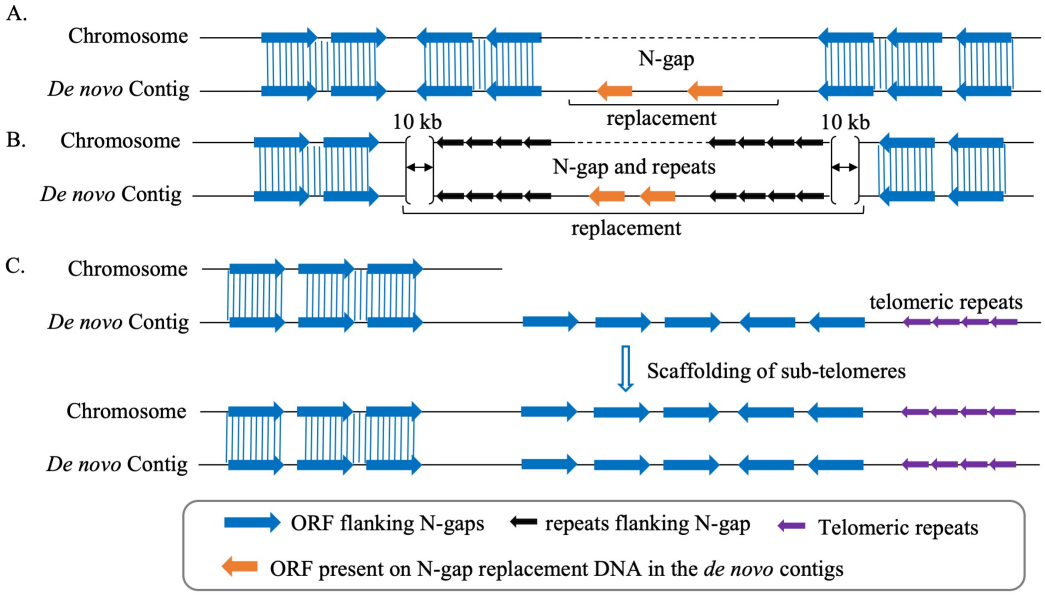


Figure 2.6 Schematic of the strategy followed for N-gap filling and scaffolding of sub-telomeres.
A - B. Strategy-I and strategy-II (Materials and methods) for filling N-gaps without flanking repeats or with flanking repeats, respectively. Repeats are presented as black arrows. C. Schematic for scaffolding of sub-telomeres using the *de novo* assembled contigs. Telomeric repeats are presented as purple arrows.

Table 2.3: Statistics for intermediate and final version of genome assemblies of *C. tropicalis* (MYA-3404)

Parameters	Assembly B	Assembly C	*De novo* (Canu)	*De novo* (FALCON)	Assembly2020
asm_contigs	16	12	135	17	7
asm_esize	1948827	2042279	1074242	1651545	2305488
asm_max	2797331	2797331	2853807	2860637	2835428
asm_mean	907756	1210666	138616	874119	2087075
asm_median	419327	1225766	35218	784244	2347188
asm_min	5408	5408	16806	30345	927416
asm_n50	2216334	2342205	755598	1541447	2504192
asm_n90	921083	1225766	35218	483039	1239682
asm_n95	802573	921083	29731	464065	927416
asm_total_bp	14524098	14527994	18713202	14860026	14609527

The chromosomes were named 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). Accordingly, the centromeres on each chromosome are named after the respective chromosome number. Additionally, we assembled the genome sequence of each chromosome in a way to consistently maintain the small arm of chromosomes at the 5′ end. The statistics of the intermediate and final genome assemblies are summarized (Table 2.3). In this chromosome-level assembly, 1278 out of 1315 Ascomycota-specific BUSCO (330) gene sets could be identified compared to 1255 identified using assembly A (Materials and methods). Inclusion of 23 additional gene sets as compared to Assembly A suggests improved contiguity and completeness of Assembly2020 (Table 2.4).

Validation of the genome assembly with CHEF gel and chromoblot analysis

Comparison between the chromosome bands observed for *C. tropicalis* strain MYA-3404 with reference to the *C. albicans* reference strain SC5314 in CHEF-gel suggested that the length of the two smallest chromosomes are in perfect agreement with the assembly (Figure 2.2A). However, the other five chromosomes appear as four bands at the top portion of the gel (Figure 2.2A). Previously, the identity of five chromosomes carrying centromeres has been determined using chromoblot analysis (193), where a centromere adjacent unique locus

from each of Chr1, Chr2, Chr3, Chr5, and Chr6 were used as probes to identify the corresponding chromosomes in the CHEF-gel by chromoblot experiments. Therefore, to confirm the correct assembly of the remaining two chromosomes ChrR and Chr4, chromoblot analysis was performed. First, a centromere-proximal locus upstream of *CEN4* was used as the probe (Probe A). This probe detected two bands in the chromoblot experiment. To confirm the observation, and eliminate the possibility of one of the bands being an artifact, a second probe (Probe B) was selected from the telomere-proximal region of the opposite arm of Probe A. However, the same two bands were detected in this experiment. This result confirmed the validity of the initial result and implied that the two homologs of Chr4 in *C. tropicalis* strain MYA-3404 are of different sizes (Figure 2.7A). Comparison with the band lengths with the chromosome-bands observed for *C. albicans* (SC5314) suggests the difference between the two homologs is ~250 kb. Since the smaller homolog (Chr4B) matched with the assembled length, it was suspected that the larger homolog Chr4A carries a ~250 kb long duplicated region.

Table 2.4: Improvements of *C. tropicalis* genome assembly

	Assembly A	**Assembly2020**
Contigs/Chromosomes	24 (23 and mtDNA)	8 (7 and mtDNA)
Protein coding genes	6254	6136
tRNA genes	187	190
SNP density	1 in 576 bp	1 in 388 bp
Indel density	NA	1 in 1340 bp
long heterozygous loci	0	5
Large CNVs	0	3
Telomeres	0	14
Total length	14630139 bp	14609527 bp
N50	1654078 bp	2504192 bp
N90	498422 bp	1239682 bp
Completeness (BUSCO)	1255 out of 1315	1278 out of 1315
N-gaps	104	0

Similarly, using an analogous approach, a centromere-proximal locus of ChrR was used as a probe to examine the correct length of ChrR. In this experiment, the probe detected the topmost band migrating ~2.8 Mb (Figure 2.7B). However, the assembly length of Chr1 is

2.8 Mb. To confirm if the topmost band observed in the CHEF-gel contains both Chr1 and ChrR, a centromere-proximal probe from Chr1 was used. In this experiment, both the probes from Chr1 and ChrR detected the same band. This experiment suggests that the length of ChrR is ~700 kb longer than the assembled length of ChrR. This difference in the length of ChrR can be explained by the presence of rDNA locus on it. Although we have used SMRT-seq reads with an average length of ~6 kb, the rDNA locus is not completely assembled. Similarly, the length of the rDNA locus in *C. albicans* is ~700 kb (27). Based on this fact and the experimental evidence, we speculate that the length of the rDNA locus in *C. tropicalis* is ~700 kb.

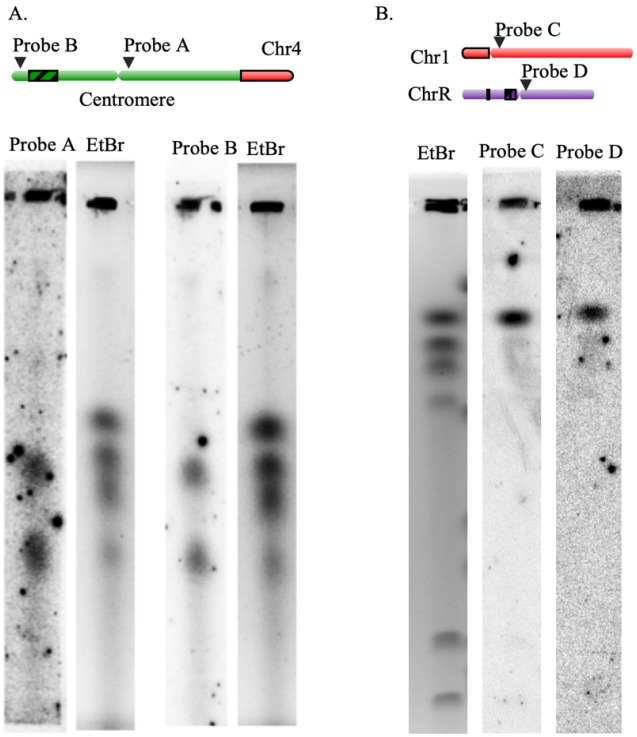


Figure 2.7 CHEF-chromoblot analysis for validation of the assembly of Chr4 and ChrR
A. *Top,* a schematic showing genomic location of probes (black triangle) used for Southern hybridization experiment on Chr4. *Bottom*, EtBr stained gel images and corresponding phosphoriamges obtained from Southern hybridization experiments using a centromere-proximal (Probe A) and a centromere-distal probe (Probe B). B. *Top,* a schematic showing genomic location of Probe C and Probe D located on Chr1 and ChrR. *Bottom*, an ethidium bromide stained gel picture and phosphoriamges obtained from Southern hybridization experiments using centromere-proximal probes from Chr1 (Probe C) and ChrR (Probe D).

Large CNVs lead to copy-number dependent fluconazole resistance in *C. tropicalis*

To test if the difference in the length of two homologs of Chr4 is due to copy number variation (CNV) of a region, a sequence coverage analysis was performed. The simple rationale behind this analysis is that the number of reads obtained from any given genomic

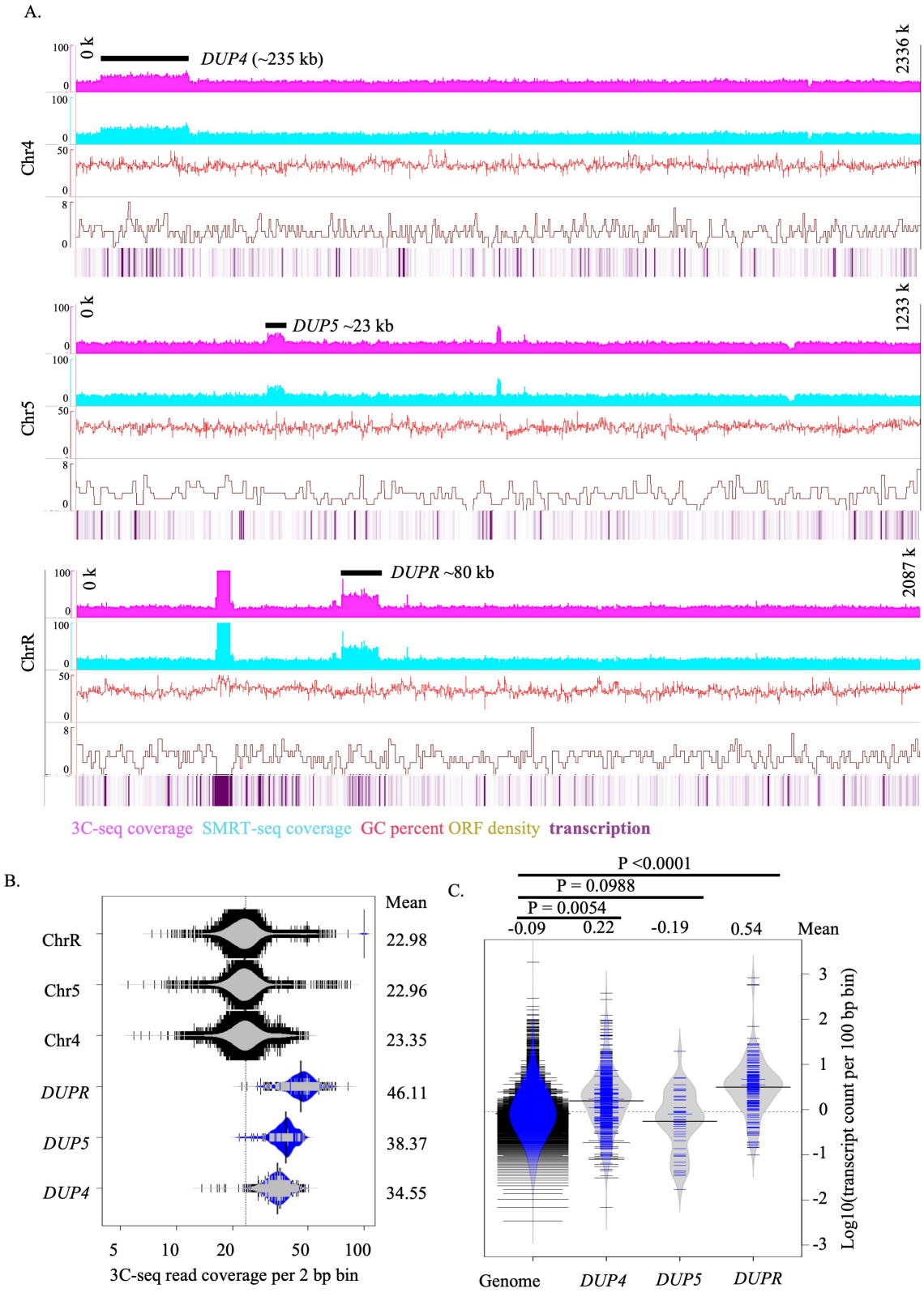


Figure 2.8 Copy number variation in *C. tropicalis* strain MYA-3404 is correlated to increased gene expression.

A. IGV track images showing coverage of 3C-seq data (pink), SMRT-seq data (cyan), GC percent profile (red), ORF density (maroon), and the mRNA-seq heatmap profile (purple) for Chr4, Chr5, and ChrR (Materials and methods). Each of the CNV loci *DUP4, DUP5,* and *DUPR* is marked (black bar) and labelled with the respective length mentioned within the

parenthesis. B. Bean plots showing quantification of 3C-seq data coverage over *DUP4, DUP5, DUPR,* and chromosome averages for Chr4, Chr5, and ChrR (chromosome average for ChrR was calculated excluding the reads mapped on rDNA locus). Mean value of sequence depth is mentioned for each locus or chromosome. C. Bean plots showing quantification of the average gene expression level of *DUP4, DUP5,* and *DUPR* loci along with the genome average control. Unpaired t-test was performed to test the statistical significance of difference between genome average and each of the CNV loci. The *x*-axis of B and *y*-axis of C represents $\log_{10}$(read count per million mapped reads) per 2 bp and 1000 bp bins, respectively and obtained using bamCoverage utility of deepTools2.

---

locus is proportional to their copy number. Therefore, the copy number of a given genomic locus was estimated as follows:

$$Copy\ number = \frac{number\ of\ reads\ mapped\ per\ kb\ of\ a\ given\ locus}{average\ number\ of\ reads\ mapped\ per\ kb\ across\ the\ genome}$$

To detect the CNVs across the genome, both SMRT-seq reads and 3C-seq reads were mapped on the chromosomes of Assembly2020 using BLASR (331) and Bowtie2 (332), respectively (Materials and methods). Next, the number of aligned reads were counted using deepTools2 (333) and the sequence coverage across all seven chromosomes was visualized using IGV (329). In this analysis, three long regions were identified, which are present in more than two copies in the genome of *C. tropicalis* strain MYA-3404 (Figure 2.8A). These three long CNV loci present on Chr4, Chr5 and ChrR were named as *DUP4, DUP5,* and *DUPR,* respectively. Estimation of copy number suggests that *DUP4* and *DUP5* are present in three copies in the genome, while the *DUPR* locus is present in four copies (Figure 2.8B). Quantification of read mapping in high-resolution facilitated an accurate estimation of the length of these CNV loci. The largest among the three CNVs, *DUP4,* spans over a ~235 kb long region and present on the larger homolog of Chr4. The other two CNV loci, *DUPR* and *DUP5,* span across DNA regions of ~80 kb and ~23 kb, respectively. The identification of CNV of the *DUP4* locus explains the previously observed length difference between the homologs of Chr4.

To examine whether presence of CNVs influences the expression level of a given locus, an mRNA sequencing experiment was performed in *C. tropicalis* isolate MYA-3404. Illumina RNA-seq reads were mapped on the chromosomes of Assembly2020 using STAR (334). Global quantification of genomic transcripts was performed using deepTools2 (333)

and visualized using IGV (329). Examination of the transcription status revealed a noticeable increase at *DUP4* and *DUPR,* but not at *DUP5* (Figure 2.8A). However, we found that the transcriptional bias at the *DUP4* and *DUPR* locus is independent of the ORF density and the overall base composition of the underlying DNA (Figure 2.8A). Precise quantification of the transcription level suggested that increased copy number of *DUP4* and *DUPR* loci is directly correlated with an increase in the average gene expression level compared to the genome average (Figure 2.8C).

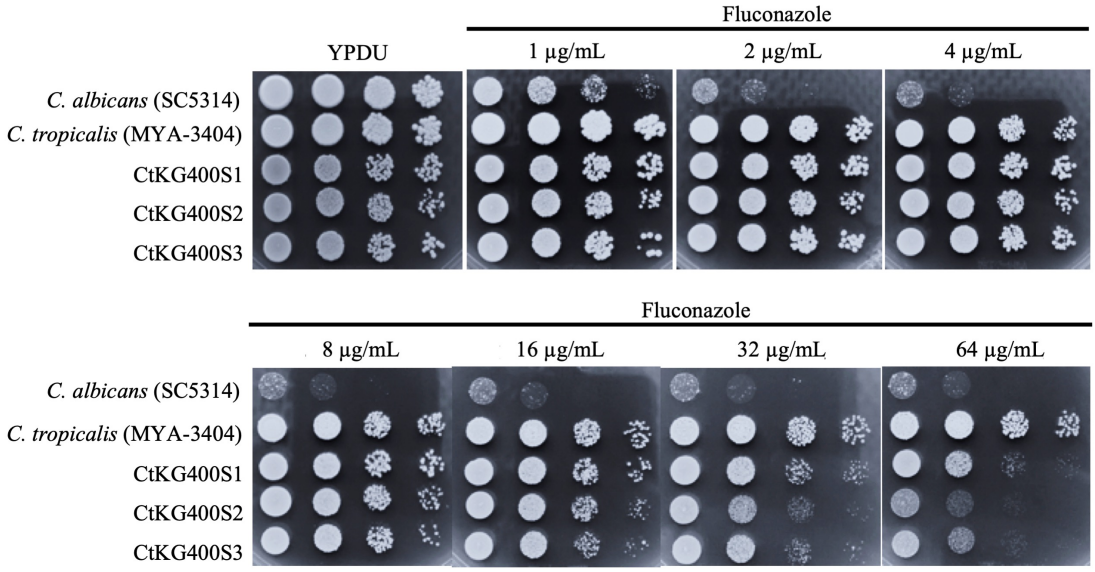


Figure 2.9 The *dupR* mutants shows fluconazole sensitivity.
Various dilutions of cells ($2x10^4$, $2x10^3$, $2x10^2$, and $2x10^1$ from left to right) for each of the *dupR* deletion mutant strains (genotypes are mentioned in Appendix-I), the wild-type strains of *C. tropicalis* (MYA-3404) and *C. albicans* (SC5314) were spotted on YPDU plates without or with different concentrations of fluconazole and the plates were incubated at 30°C for 36 h before images were taken.

Our result of the genome-wide profiling of mRNA sequencing data in *C. tropicalis*, suggests that the increased copy number of the genomic DNA sequence can positively influence the overall expression level of a given locus. However, at this point, it was not clear if the increased expression profile of *DUP4* and *DUPR* loci translates into a phenotype. Therefore, we deleted at least one copy of the entire length of both the *DUP4* (~235 kb) and *DUPR* (~80 kb) loci in MYA-3404 strain background and generated *dup4* (*DUP4/DUP4/DUP4::CaSAT1*; three independent transformants CtKG300S1, CtKG300S2, and CtKG300S3) and *dupR* (*DUPR/DUPR/DUPR/DUPR::CaSAT1*; three independent transformants CtKG400S1, CtKG400S2, and CtKG400S3) mutant strains (Materials and

methods). We used these mutant strains to test the contribution of these CNVs on the fluconazole-resistance phenotype in *C. tropicalis* by performing a dilution spotting assay.

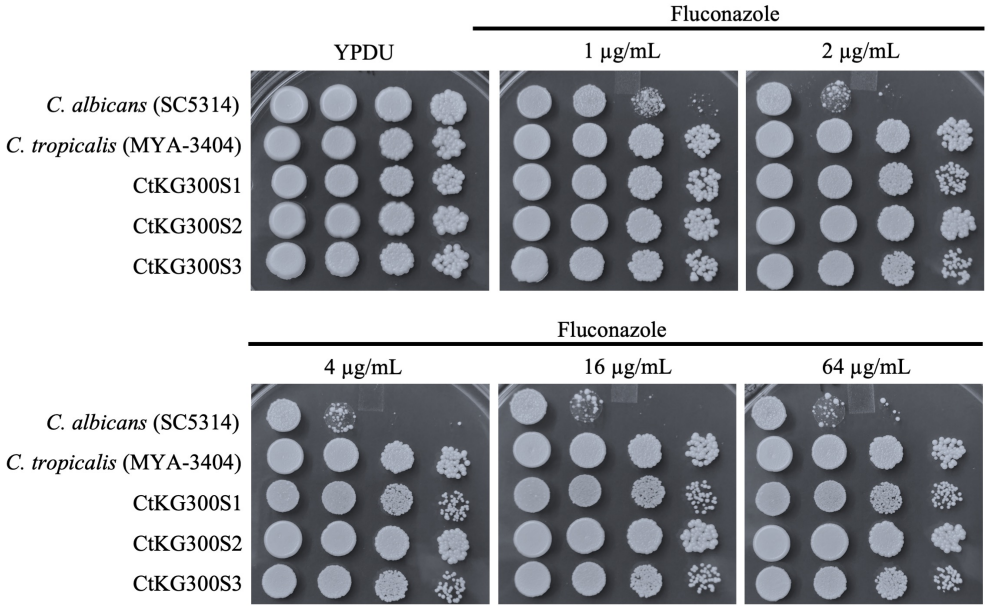


Figure 2.10 Dilution spotting assay to test fluconazole sensitivity of the *dup4* mutant strains.
Various dilutions of cells ($2x10^4$, $2x10^3$, $2x10^2$, and $2x10^1$ from left to right) for each of the *dup4* deletion mutant strains (genotypes are mentioned in Appendix-I), the wild-type strains of *C. tropicalis* (MYA-3404) and *C. albicans* (SC5314) were spotted on YPDU plates without or with different concentrations of fluconazole and the plates were incubated at 30°C for 36 h before images were taken.

In this assay, various dilutions of cells for each of the deletion mutants, the wild-type strains of *C. tropicalis* (MYA-3404) and *C. albicans* (SC5314) were spotted on YPDU plates without or with different concentrations of fluconazole. We observed that the *dupR* mutants show a dose-dependent reduction in growth compared to the parental control strain in the presence of fluconazole. This result suggests that the *DUPR* locus is implicated in fluconazole resistance of MYA-3404. However, all three *dupR* mutant strains show a significantly higher level of fluconazole resistance than the wild-type strain of *C. albicans* (SC5314). This observation indicates that additional genomic factors contribute to fluconazole resistance in *C. tropicalis* (Figure 2.9). On the contrary, the *dup4* mutants do not show a significant growth disadvantage when challenged with fluconazole in a similar dilution spotting assay (Figure 2.10).

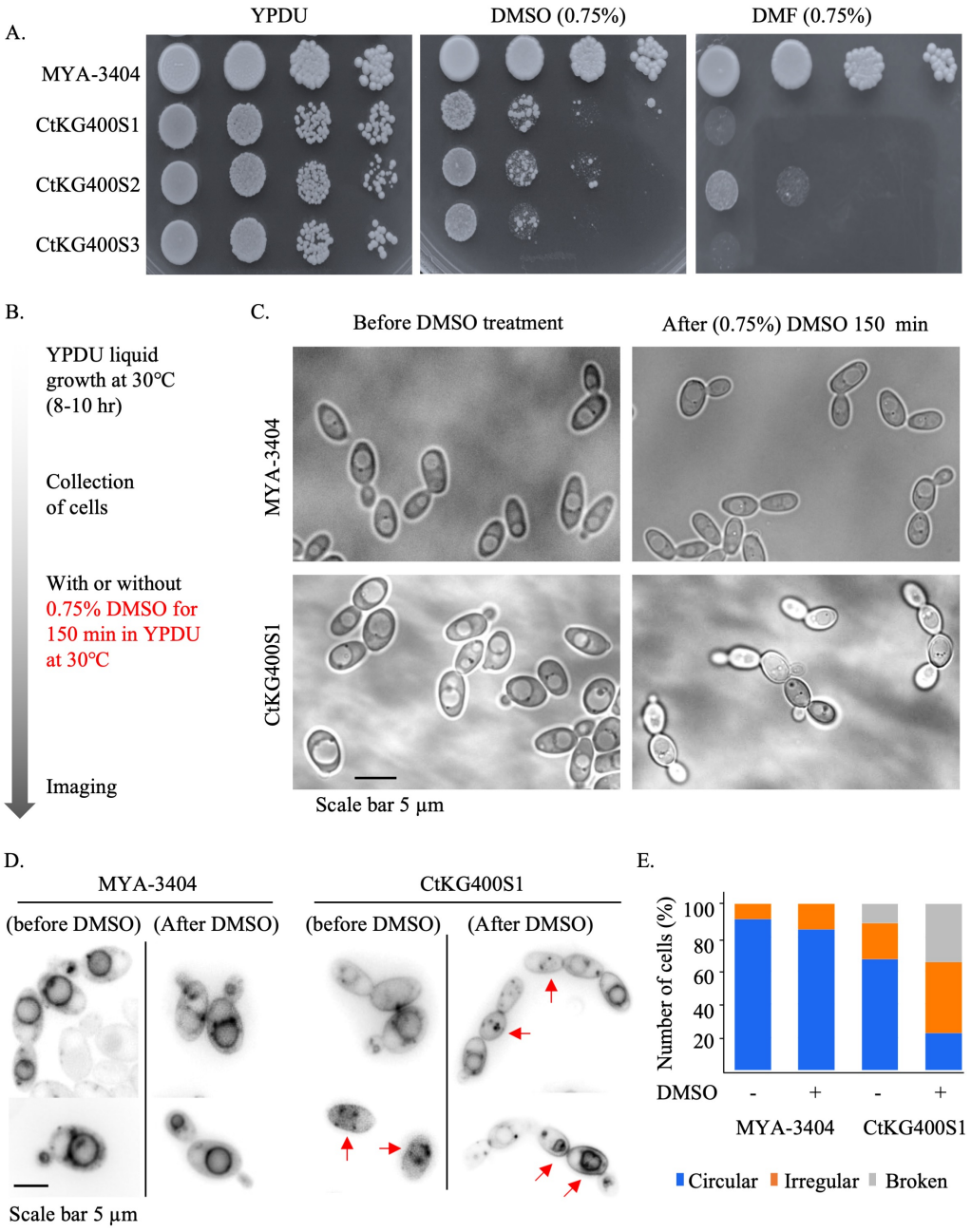


Figure 2.11 The *dupR* mutants show compromised membrane function.
A. Various dilutions of cells ($2x10^4$, $2x10^3$, $2x10^2$, and $2x10^1$ from left to right) for each of the *dupR* deletion mutant strains (genotypes are mentioned in Appendix-I) and the wild-type strains of *C. tropicalis* (MYA-3404) were spotted on YPDU, YPDU with 0.75% DMSO and YPDU with 0.75% DMF plates and incubated at 30°C for 36 h before images were taken. B. Flow-chart of the experimental outline to test the effect of DMSO and DMF treatment on the cell morphology of wild-type and *dupR* mutant strain (CtKG400S1). C. Representative microscopic images of cells, as observed using 100x magnification in a bright-field microscope. D. Representative field view of wild-type (MYA-3404) and *dupR* mutant (CtKG400S1) cells stained using FM 4-64 dye before and after DMSO treatment. The red arrows points at the unusual morphology of the vacuolar membrane, which are not observed in wild-type cells. E. A bar chart showing proportion of wild-type (MYA-3404) and *dupR* mutant (CtKG400S1) cells (n>100) with circular (blue), irregular (orange), and broken (gray) morphology of vacuolar membrane before and after DMSO treatment.

To understand why the *dupR* mutants display a fluconazole sensitive phenotype, we annotated the ORFs present in the *dupR* locus by performing BLAST search using the amino acid sequences for each of the 32 ORFs as query against the *C. albicans* genome. Using this approach, orthologs of 26 out of 32 ORFs present on *DUPR* locus could be annotated. On the other hand, five of these ORFs were uncharacterized in *C. albicans*, and the ortholog for one ORF could not be detected. One of these 26 annotated ORFs is *UPC2. UPC2* encodes for a transcription factor, which regulates the ergosterol biosynthetic pathway in *C. albicans,* and transcriptionally induced by antifungal drugs and anaerobicity (335, 336). It was also known to autoregulate its expression (337). Therefore, we suspected that the increased copy number of *UPC2* might lead to further upregulation of its own expression and increased fluconazole resistance. To test, if the fluconazole sensitivity of the *dupR* mutants arises due to compromised membrane function, a dilution spotting assay was performed. In this assay, the *dupR* mutants and the parental strain MYA-3404 were challenged with organic solvents dimethyl sulfoxide (DMSO) and dimethylformamide (DMF), which can affect membrane integrity and thereby enhance the permeability (338). In this dilution spotting assay, it was found that the *dupR* mutant strains are sensitive to both the organic solvents at a concentration, in which the parental strain MYA-3404 did not show any growth defect (Figure 2.11A).

Next, we performed cell biological experiments to know if the DMSO and DMF sensitivity of the *dupR* mutants are associated with compromised membrane morphology. In this experiment, the parental strain and the *dupR* mutant strain (CtKG400S1) were treated with 0.75% DMSO for 150 min in YPDU at 30℃ and the cells were observed under the microscope (Figure 2.11B). While the parental strain (MYA-3404) grew as yeast cells both before and after DMSO treatment, only after DMSO treatment the yeast cells of *dupR* mutant strain (CtKG400S1) became pseudohyphal, which can be a general manifestation of cellular stress (Figure 2.11C). We then studied the morphology of the cell membrane by performing fluorescence microscopy of cells stained with FM 4-64 dye. Both the mutant and the wild-type cells were treated with DMSO for 150 min, recovered and stained with FM 4-64 dye following microscopic observation (Figure 2.11B). We noted that upon DMSO treatment, the integrity of vacuolar membranes is compromised in the *dupR* mutant strain CtKG400S1 but not in MYA-3404 (Figure 2.11D). The vacuolar membranes in *dupR* mutant cells accumulated unusual aggregates, which are rarely observed in MYA-3404. In addition, the vacuolar membrane appeared to be disrupted/broken in certain cells. This phenotype was

aggravated upon DMSO treatment in *dupR* mutant, while the MYA-3404 cells remained unperturbed. This observation suggests that the vacuolar membrane organization or maintenance function in *dupR* mutant cells is compromised. This observation may explain the reduced viability of these cells upon exposure to fluconazole, which targets biosynthesis of ergosterol, one of the key membrane lipids in yeasts (339).

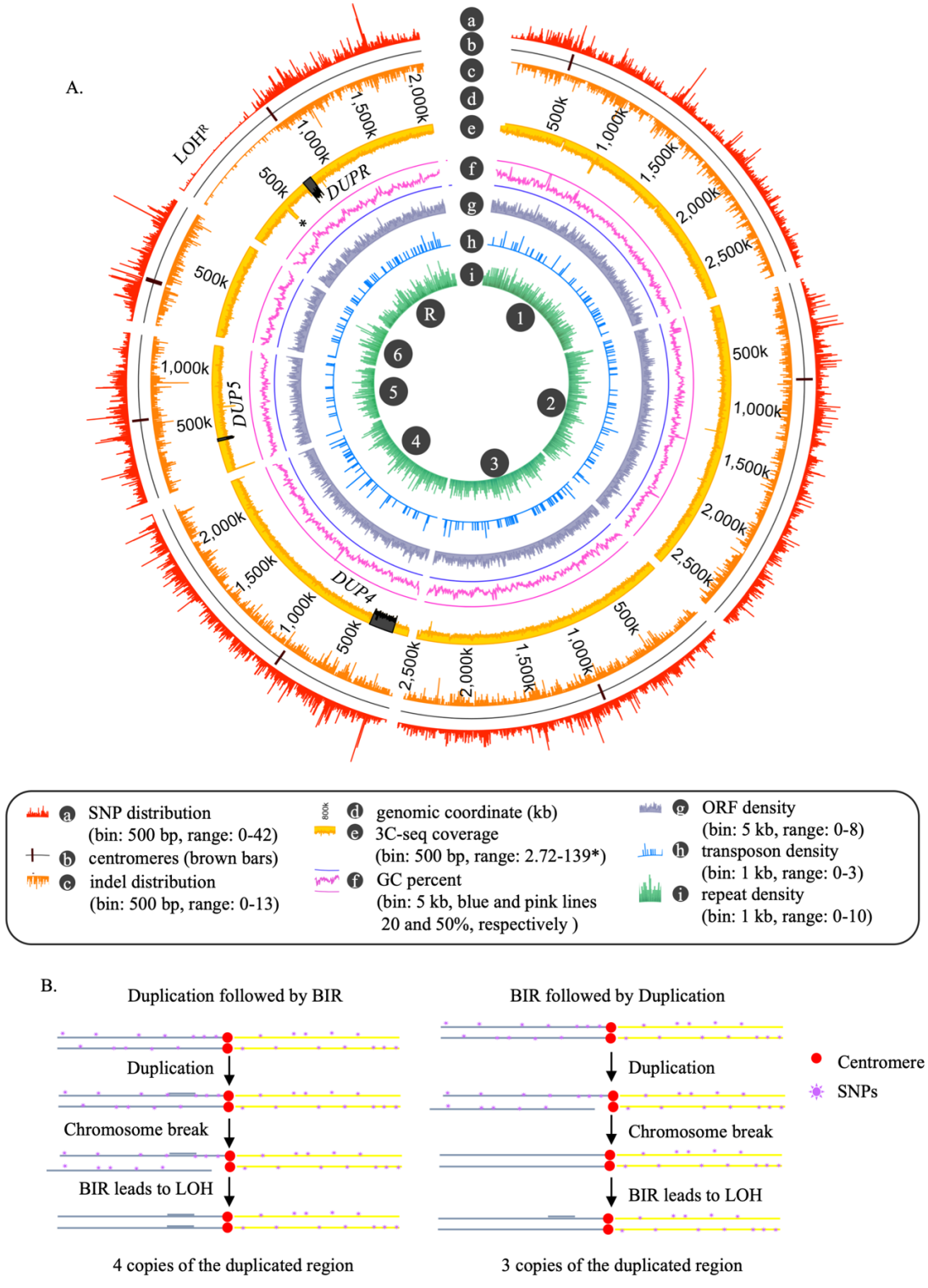


Figure 2.12 Genome-wide mapping of SNP/indels and CNVs in *C. tropicalis*.

A. 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. B. Models showing possible sequence of events shaping present-day configuration of ChrR in *C. tropicalis* strain MYA-3404.

Identification of SNPs and indels in *C. tropicalis* strain MYA-3404

The diploid genome of *C. tropicalis* carries two homologs for each of the seven chromosomes. Ideally, the two homologs should harbor exactly the same DNA sequences. However, as it was observed for Chr4 of MYA-3404, often, the two homologs can be different. The difference between the homologs can be in the form of large duplications or deletion, giving rise to CNVs. However, changes at the base-pair level in the DNA sequences of the two homologs are also possible, which generates Single Nucleotide Polymorphisms (SNPs) or insertion-deletions (indels). Previously, the SNPs in the genome of *C. tropicalis* have been identified by Butler *et al.* 2008 (6). However, the improved chromosome-level assembly of *C. tropicalis* strain MYA-3404 now allows a high-precision identification and

Table 2.5 Type of effects due to the SNPs.

Type (alphabetical order)	Count	Percent
downstream_gene_variant	85,877	39.08%
intergenic_region	23,377	10.638%
intron_variant	333	0.152%
missense_variant	7,150	3.254%
splice_acceptor_variant	1	0%
splice_region_variant	46	0.021%
start_lost	6	0.003%
stop_gained	37	0.017%
stop_lost	8	0.004%
stop_retained_variant	21	0.01%
synonymous_variant	11,577	5.268%
upstream_gene_variant	91,315	41.55%

mapping of SNPs and indels. To identify the SNPs and indels, the paired-ended 3C-seq data was used. However, to avoid any spurious hits due to chimeric reads obtained after 3C, only

reads that mapped completely were taken forward for this analysis using the genome analysis tool kit (GATK) software (Materials and methods) (85). Based on this analysis, a total of 37668 high confidence SNPs and 10900 Indels distributed over14609527 bp total genome were identified, which represents a genome average of one SNP per 388 bases and one indel per 1340 bp. Next, these SNPs and indels were mapped on the chromosomes and their relative abundance across the chromosomes was visualized using IGV. Primary inspection of the SNP and indel distribution across the chromosomes revealed a long SNP and indel depleted region on the left arm (5′ end) of the ChrR. This locus (LOH^R) spanning ~800 kb from the 5′ end of ChrR till the centromere-proximal region also encompasses the *DUPR* locus (Figure 2.12A). Present-day configuration of ChrR in MYA-3404 genome carrying four copies of *DUPR* locus within the LOH^R locus might have arisen due to a duplication event following a break-induced replication (BIR)-like event (Figure 2.12B).

Table 2.6 Type of effects due to the indels.

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	108	0.18%
conservative_inframe_insertion	157	0.25%
disruptive_inframe_deletion	229	0.37%
disruptive_inframe_insertion	217	0.35%
downstream_gene_variant	23,560	38.09%
frameshift_variant	349	0.56%
intergenic_region	10,082	16.30%
intron_variant	103	0.17%
splice_acceptor_variant	1	0.00%
splice_region_variant	8	0.01%
start_lost	8	0.01%
stop_gained	5	0.01%
stop_lost	3	0.01%
upstream_gene_variant	27,029	43.70%

The SNPs and indels may lead to mutations in coding sequences in the genome. However, to understand the impact of a particular SNP/indel, annotation of the ORFs is required. Multiple tools are available for *ab initio* gene prediction from the genome FASTA files. One such widely used tool is Augustus (340). Augustus is a generalized hidden Markov model-based tool for *ab initio* gene prediction, and it has been already trained for gene prediction in *C. tropicalis* (340). Therefore, Augustus was used for gene prediction in the

chromosome-level genome assembly of *C. tropicalis.* In this analysis, a total of 6129 ORFs were identified in the haploid genome of *C. tropicalis* type-strain MYA-3404. Next, to predict the nature of the mutation caused by a particular SNP or indel was analyzed using SNPeff (341). In this analysis, we detected certain cases where start codon is lost, stop codon is gained or lost due to the SNPs (Table 2.5) and indels (Table 2.6). However, most of the SNPs and indels are located away from the coding sequences.

Haplotype phasing of MYA-3404 genome

Analysis of SNPs, indels, and CNVs characterizes the genetic variation present in the diploid genome of in the *C. tropicalis* strain MYA-3404. However, it does not reveal the linkage information among the alleles. Conventionally the haplotyping of the offspring can be performed if the parental linkage information is available. However, this approach cannot be applied to organisms that do not undergo true meiosis. Therefore, the haplotyping of *C. albicans* was performed by comparative genomic hybridization (CGH) analysis (29) and later Illumina sequencing (26) of a set of monosomic isolates in which one of the eight chromosomes is present in one copy. Although this approach is the best available one to date, it may generate a biased result. In a scenario where the only functional allele of two unlinked essential genes are located on two different homologs, an aneuploid strain lacking either of the homologs will never be recovered. The only way this aneuploid can be recovered if the homologous chromosomes undergo recombination to retain the functional alleles of both the genes. However, this recombination event will remain undetected and portray an incorrect haplotype that is not present in the parental strain. Therefore, in spite of extensive efforts to generate all the aneuploid strains and their genome sequencing, there are chances of error in the results.

Recently, a new computational tool-set FALCON and FALCON-Unzip was developed for haplotyping, which uses the long read SMRT-seq data together with the contact probability information from Hi-C experiment to reliably phase the haplotypes (78, 325) (Figure 2.13A). Therefore, we used FALCON to perform haplotyping of the *C. tropicalis* strain MYA-3404 using the long-read SMRT-seq data and the 3C-seq data. Details of the scripts and parameters used are described in the methods section. This analysis generated 16 phased pseudo haplotigs comprising of total 14860026 bases for phase0 contigs and 14886827 bases for the phase1 contigs. In order to identify and map the genomic

variation between the phased haplotigs, phase0 and phase1 contigs were aligned against each other using MUMmer (342), and the output delta file was analyzed using Assemblytics (343) (Figure 2.13B).

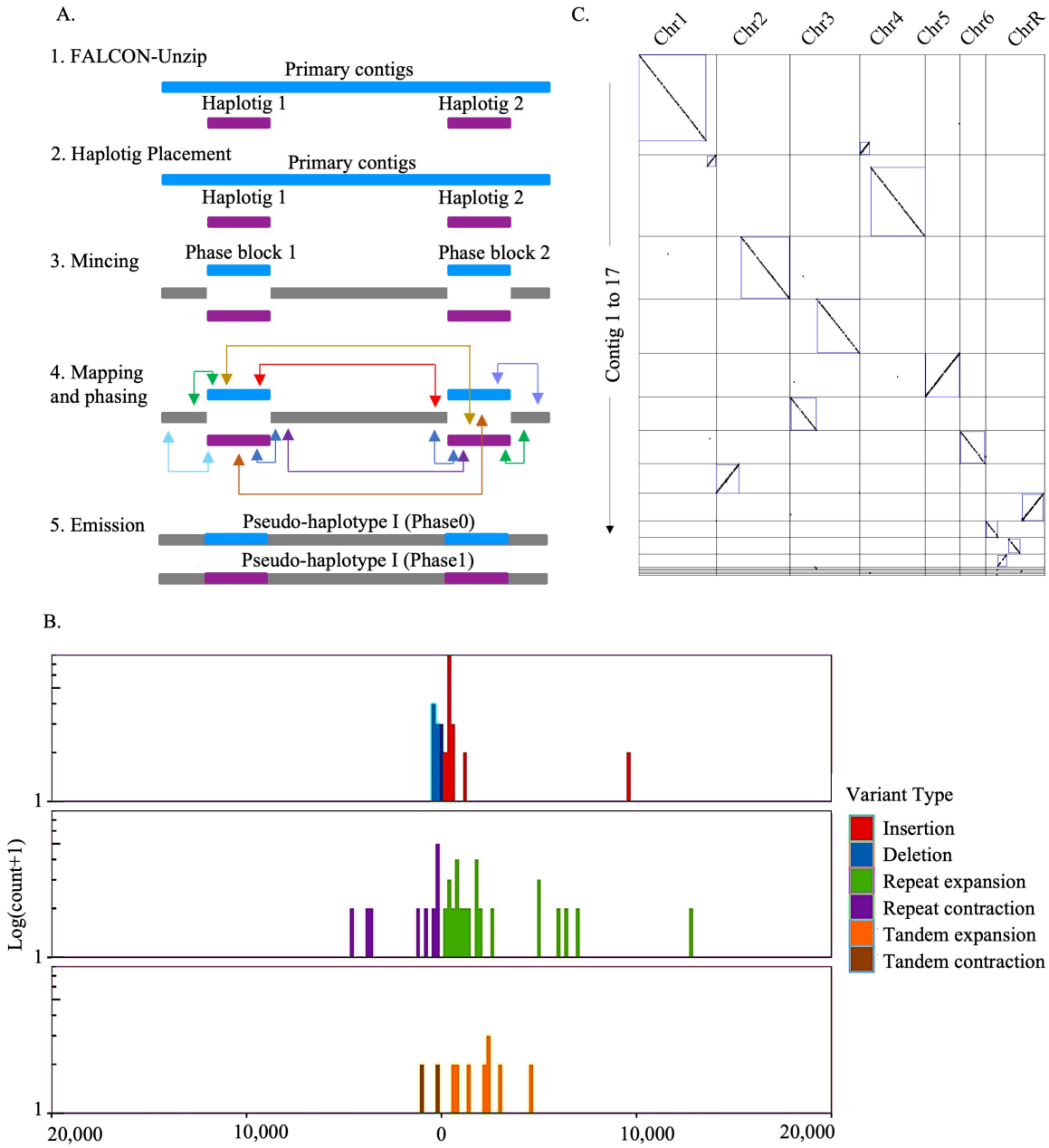


Figure 2.13 Phasing of diploid genome of *C. tropicalis* using FALCON.
A. The schematic of the steps followed by FALCON-Unzip during phasing of diploid variations. Modified from reference (325). B. The bar charts showing the number (*y*-axis) of haplotype-specific genomic variants of sizes ranging from 50 bp to 20000 bp (*x*-axis) (Materials and methods). This plot was generated by Assemblytics (343) after identification of haplotype specific differences by MUMmer (342). C. A synteny dot-plot between the chromosome level genome assembly and the *de novo* assembly obtained from FALCON pipeline, as generated by the Symap using the default parameters. The larger blocks were highlighted by a blue rectangle.

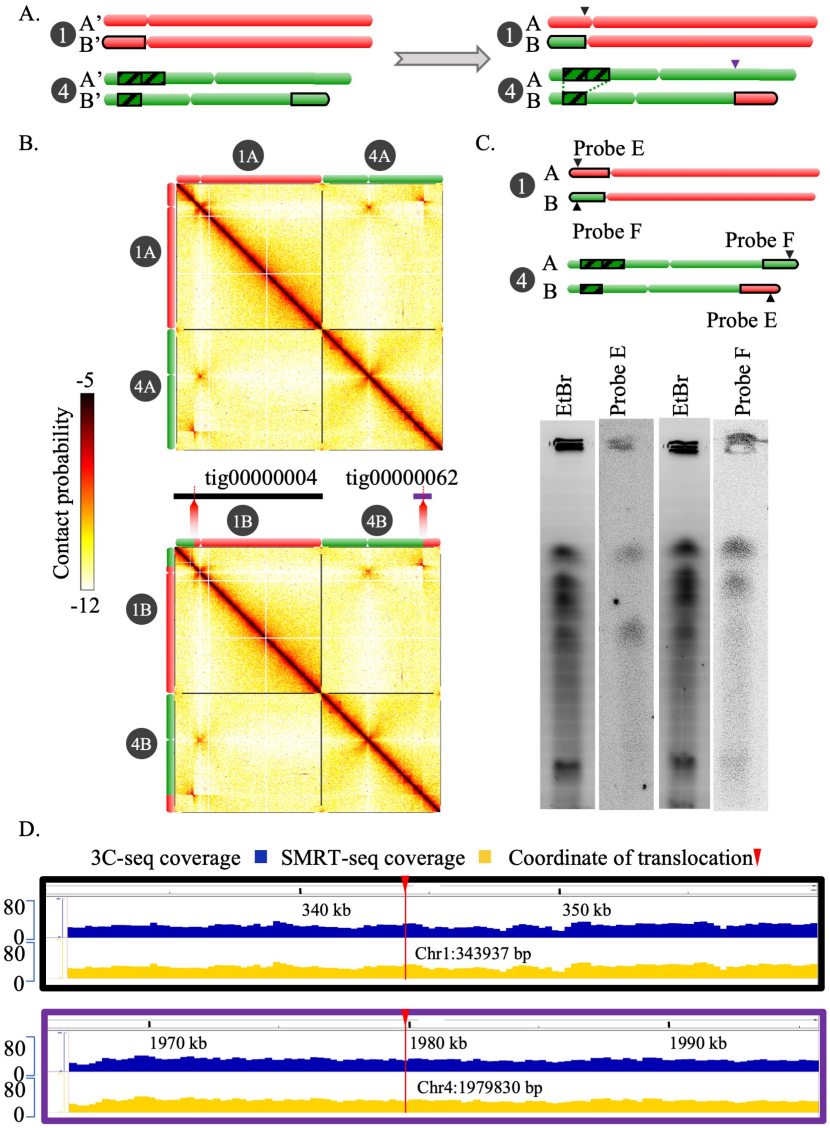


Figure 2.14 Analysis of chromoblots, 3C-seq contact probability data, *de novo* contigs and sequence coverage validates a balanced heterozygous translocation between Chr1 and Chr4.

A. Schematic of the balanced heterozygous translocation between Chr1B and Chr4B. *DUP4* locus is highlighted with the black striped box. The junction between Chr1 and Chr4 on Chr1B and Chr4B are marked with black and purple arrows, respectively. B. The contact probability heatmaps (bin size = 10 kb) of Chr1 and Chr4 of *C. tropicalis* showing a butterfly-like pattern (chromatin contacts split into two blocks) in the interchromosomal area. The 3C-seq reads were mapped to Assembly2020 (*top*) with Chr1A and Chr4A genomic sequences. We have also mapped the 3C-seq reads to an alternate assembly (*bottom*) with Chr1B and Chr4B sequences. Alternate assembly has been generated by exchanging the genomic sequences at the translocation breakpoint in Chr1 and Chr4. Coordinate of translocation was mapped using two *de novo* assembled contigs supporting the junctions. Chromosome labels and their corresponding ideograms are shown on the heatmap. Color-bar represents the contact probability in log2 scale. C. An ethidium bromide stained gel picture and phosphoriamges obtained from Southern hybridization using a probe from part of Chr1, which is exchanged with Chr4 (probe E) and another probe from part of Chr4, that is exchanged with Chr1 (Probe F). The black triangles point to the genomic coordinates of the probes used in this experiment (*top*). F. IGV tracks showing 3C-seq (blue) and SMRT-seq

coverage (yellow) across the translocation junctions on each of the unaltered homolog of Chr1 (black border) and Chr4 (purple border), respectively.

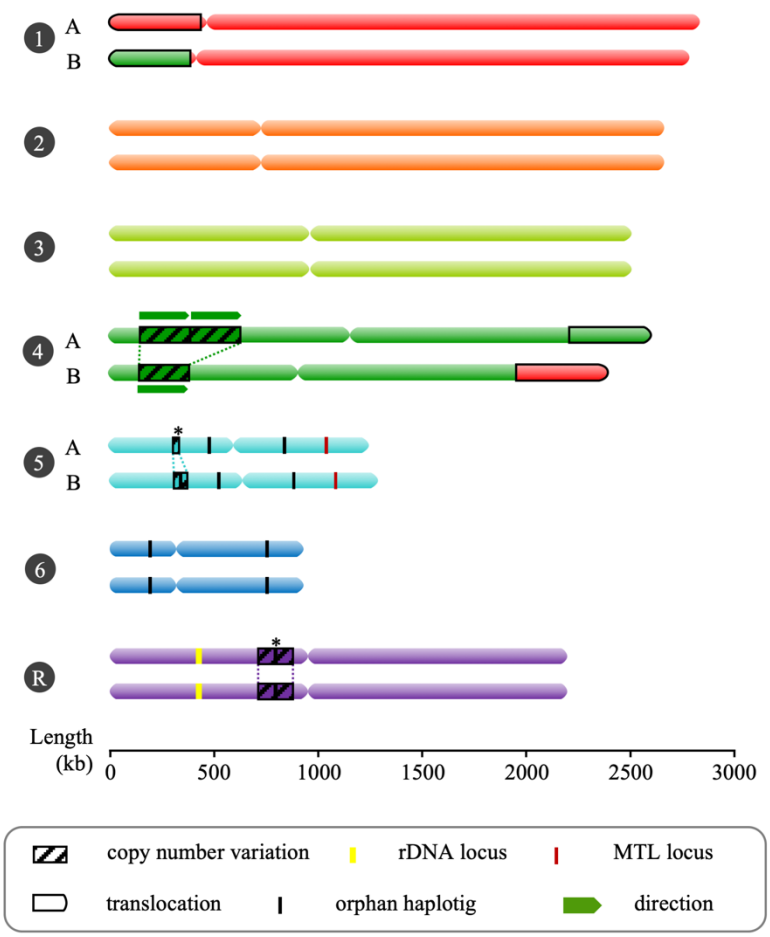


Figure 2.15 Chromosomal features of *C. tropicalis* strain MYA-3404 as revealed in Assembly2020.
An ideogram of seven chromosomes of *C. tropicalis* as deduced from Assembly2020 and drawn to scale. The genomic location of the three loci showing copy number variations (CNVs): *DUP4, DUP5* and *DUPR* located 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.

Next, the haplotigs generated by FALCON and the chromosomes of Assembly2020 were compared in a dot-plot analysis using Symap (326). This analysis confirmed the contiguity across six out of the seven centromeres (Figure 2.13C). Moreover, *de novo* contigs were found to be co-linear to the chromosomes except for one translocation between Chr1 and Chr4. To validate the possibility of heterozygous balanced translocation between Chr1

and Chr4 (Figure 2.14A), we analyzed the contact probability data obtained from the 3C-seq experiment (Figure 2.14B). This analysis, combined with the analysis of *de novo* contigs generated by diploid-aware assembler Canu (77), led to the precise identification of the translocation site. This translocation resulted in the exchange of ~392 kb region from the 3′ end of Chr4 with ~343 kb of the 5′ end of Chr1. In order to verify the translocation chromoblot experiments were performed. In a chromoblot strategy exploiting the size difference between the two homologs of Chr4, it was identified that the translocation occurred between the smaller homolog of Chr4 and one of the two homologs of Chr1 (Figure 2.14C). We also mapped both the 3C-seq and SMRT-seq data on Chr1 and Chr4 to show that one set of homologs were not involved in translocation (Figure 2.14C). Based on the results described so far, an ideogram was drawn to show various chromosomal features of *C. tropicalis* strain MYA-3404 (Figure 2.15).

Chapter 3

Results

Higher-order genome organization in *Candida tropicalis*

Conserved principle of genome organization in *C. tropicalis*

Although centromere DNA sequences change rapidly (191), kinetochore proteins remain relatively well conserved across species (344). Previous studies in *C. albicans* led to the identification of conserved kinetochore proteins such as CENP-A^{Cse4}, CENP-C^{Mif2}, Nuf2, and Dad1 (287). Localization of CENP-A^{Cse4} and CENP-C^{Mif2} and Nuf2 were studied in *C. albicans,* where all three proteins show a punctate localization throughout all stages of cell cycle (287, 345). Based on the sequence identity, the homologs of CENP-A^{Cse4}, CENP-C^{Mif2}, Nuf2, and Dad1 were identified and characterized as bona fide kinetochore proteins in *C. tropicalis* (193). We also studied the subcellular localization of inner and outer kinetochore proteins in *C. tropicalis.* Next, we analyzed the genome-wide 3C-seq data to study the physical contacts between the centromeres. In this analysis, we could find evidence that higher-order chromatin structures exist in *C. tropicalis* genome. Together, these experiments and analysis of 3C-seq data improve our understanding of the overall spatial organization of the nuclear genome in *C. tropicalis*.

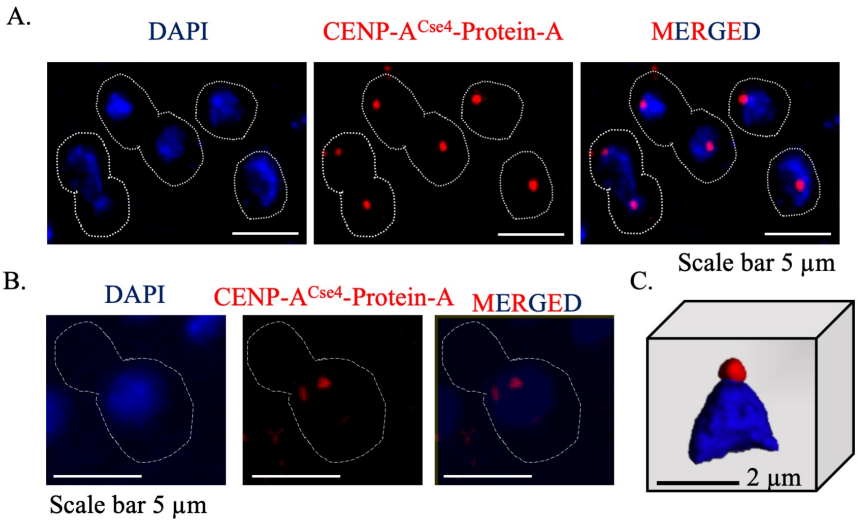


Figure 3.1 Nuclear localization of CENP-A^{Cse4} in *C. tropicalis.*
A. A representative field image of *C. tropicalis* (strain CtKS102) cells expressing Protein-A tagged CENP-A^{Cse4}. CENP-A^{Cse4} signals (red) were obtained using anti-Protein-A antibodies by indirect immuno-fluorescence microscopy. Nuclei of the corresponding cells were stained by DAPI (blue). The images were acquired using a DeltaVision imaging system (GE Healthcare Life Sciences) and processed using FIJI software (346). B. Representative image of a *C. tropicalis* cell (strain CtKS102) showing localization of CENP-A^{Cse4} as two closely spaced puncta within a single nucleus. C. A 3D reconstruction showing clustered kinetochores marked by CENP-A^{Cse4} (red) at the periphery of the DAPI-stained nucleus (blue) using Imaris software (Oxford Instruments) in *C. tropicalis*.

Co-localization of CENP-A^{Cse4} with DAPI stained genome in *C. tropicalis*

We studied the colocalization between the CENP-A^{Cse4} representing the clustered kinetochores and DAPI stained nuclear mass using fluorescence microscopy to study the spatial location of the centromere-kinetochore clusters in the nucleus. For this experiment, a strain CtKS102 (genotype) of *C. tropicalis* was constructed in which one of the two alleles of CENP-A^{CSE4} was tagged with a Protein-A epitope with *CaHIS1* marker and expressed from its native promoter at the native genomic locus. The colocalization of Protein-A tagged CENP-A^{Cse4} with the DAPI stained nuclear mass in CtKS102 was performed by indirect immune-fluorescence microscopy using a delta vision fluorescence microscope (GE Healthcare Life Sciences). The detailed protocol for this experiment is presented in the Materials & methods section. Examination of the colocalization pattern of CENP-A^{Cse4}-Protein-A and DAPI stained nucleus revealed two features of the centromere-kinetochore clustering in *C. tropicalis.* First, the presence of a single punctum per nucleus indicated clustering of all seven centromere-kinetochore complexes at a distance, which is below the resolution limit of the light microscope (Figure 3.1A). The observation of the single CENP-A^{Cse4}-Protein-A punctum during the different stages of the cell cycle proves that the clustered organization of the centromere-kinetochore complex remains intact during all stages of the cell. However, in rare instances, cells with two CENP-A^{Cse4} foci in a single nucleus were also observed, which probably represents a transient state during biorientation of the replicated centromere-kinetochore complexes (Figure 3.1B). This phenomenon was originally described as kinetochore breathing in *S. cerevisiae* (347). Second, the CENP-A^{Cse4}-Protein-A punctum was localized at the periphery of the nucleus during all stages of the cell cycle. This observation supports Rabl conformation of nuclear organization and indicates the maintenance of the Rabl configuration during different stages of the cell cycle in *C. tropicalis* (Figure 3.1C).

Localization of inner and outer kinetochore proteins in *C. tropicalis*

To extend our understanding of the spatial clustering of the centromere-kinetochore complex, we studied the localization patterns of inner kinetochore protein CENP-C^{Mif2} and Outer kinetochore protein Nuf2 in *C. tropicalis.* For this experiment, two separate strains were constructed. In each case, one of the two alleles of *NUF2* (strain CtKG500) or CENP-C^{MIF2} (strain CtKG501) was C-terminally tagged with a GFP epitope and expressed from

their native promoters (Figure 3.2A). These tagged strains CtKG106 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT NUF2/NUF2::NUF2-GFP* (*CaHIS1*)) and

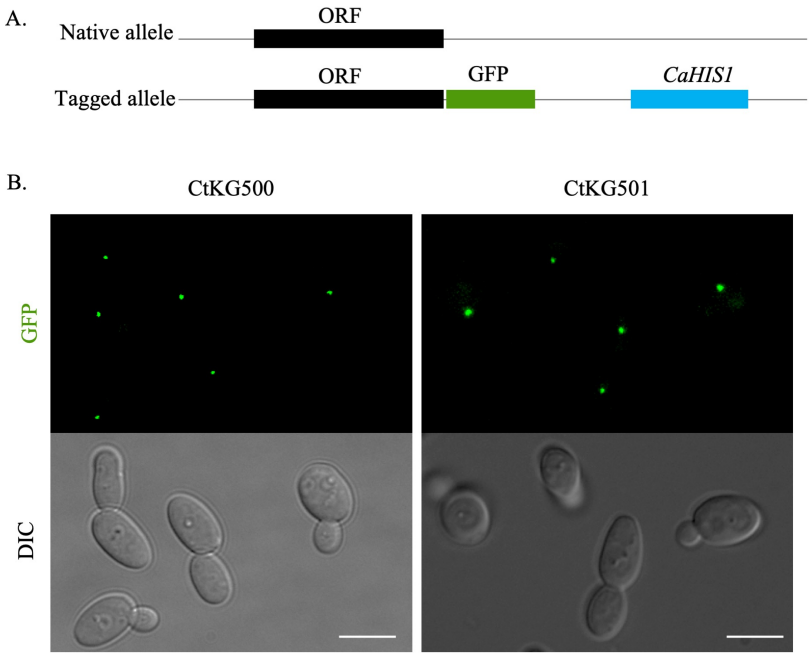


Figure 3.2 Localization of inner and outer kinetochore proteins in *C. tropicalis.*
A. Schematics showing strategy for GFP-tagging of kinetochore proteins, Nuf2 and CENP-C^{Mif2} in *C. tropicalis* strain CtKG500 and CtKG501, respectively. B. Representative field images showing localization of GFP tagged Nuf2 and CENP-C^{Mif2} in *C. tropicalis* strain CtKG500 and CtKG501, respectively. Images were acquired using a DeltaVision imaging system (GE Healthcare Life Sciences), and the images were processed using FIJI software (346). Scale, 5 µm. Genotype of each strain is mentioned in Appendix-I.

CtKG107 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT MIF2/MIF2::MIF2-GFP* (*CaHIS1*)) were used to perform fluorescent microscopy that revealed clustered kinetochores as a single punctum per nucleus. Similar to the localization of CENP-A^{Cse4}, both the kinetochore proteins remain clustered during various stages of the cell cycle (Figure 3.2B). These results indicate that the centromere-kinetochore complex in *C. tropicalis* follows a very similar pattern of clustering to what is observed in *C. albicans* (223).

Analysis of chromosome conformation capture sequencing (3C-seq) data reveals a conserved Rabl conformation of chromosomes in *C. tropicalis*

Two independent methods, namely Juicer (348) and Homer (349), were followed to analyze the spatial organization of the chromosomes. First, Juicer (348) was used for the analysis of 3C-seq data in a CPU based machine. The details of the parameters used, and the

script followed are presented in the materials and methods section. The Juicer output was visualized using a Java-based tool Juicebox (350). Analysis of the all versus all interactions

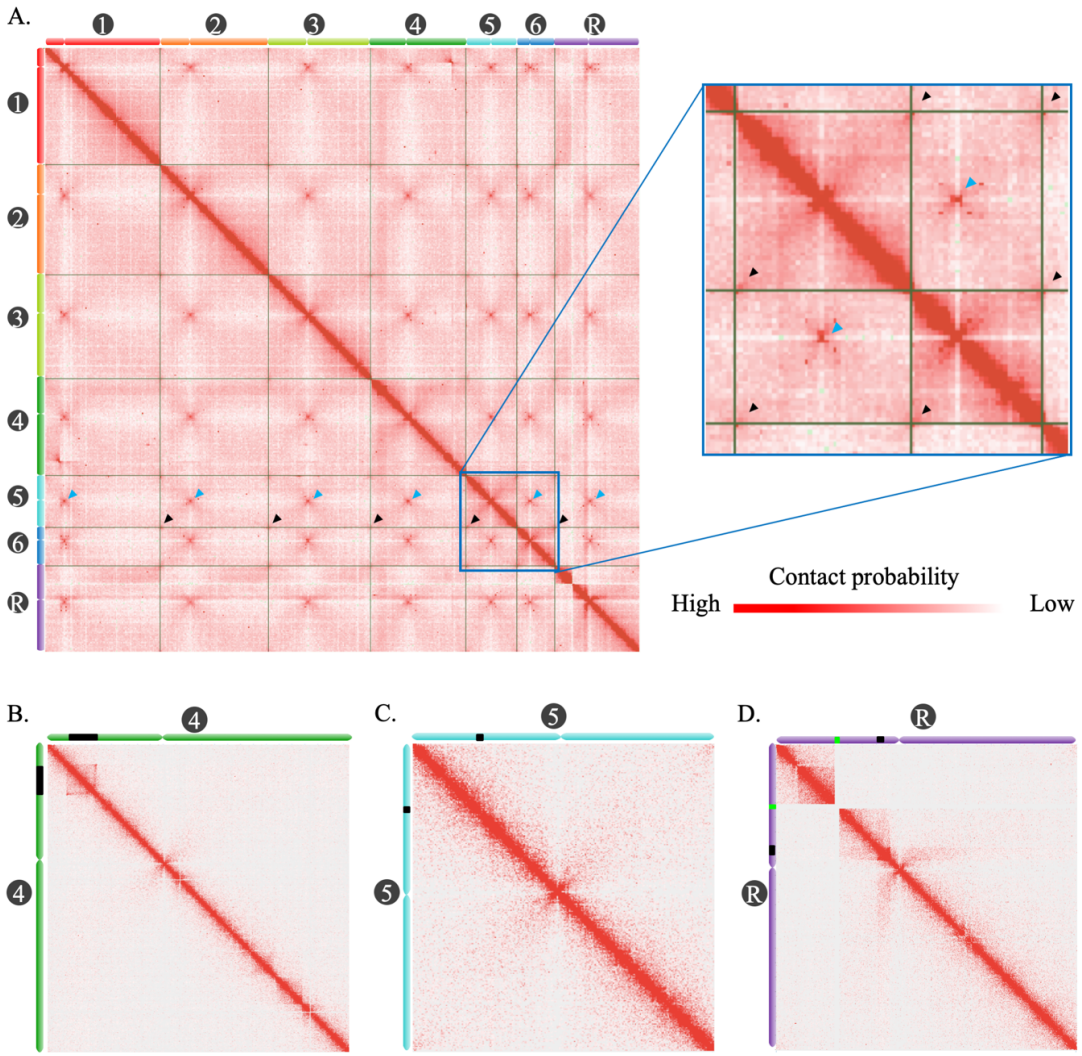


Figure 3.3 The contact probability matrix of the *C. tropicalis* genome reveals significant *CEN-CEN* and *TEL-TEL trans* interactions.
A. A genome-wide contact probability matrix (bin size = 10 kb) generated using 3C-seq data showing spatial contacts between seven chromosomes in the *C. tropicalis* genome. The *trans* interactions between *CEN5* and other centromeres are pointed with blue arrowheads. The *trans* contacts between the telomeres at 3′ end of Chr5 and the 5′ end of Chr6L with other telomeres are pointed with black arrowheads. Chromosome labels and their corresponding ideograms are shown on the heatmap. Color-bar represents the contact probability. A section of the heatmap is enlarged in the inset. B - D. Contact probability heatmap of Chr4, Chr5 and ChrR is presented after applying coverage-normalization using Juicebox. The location of CNV loci are marked (black) on the ideogram drawn. The rDNA locus on ChrR is marked in green.

---

across the seven chromosomes in *C. tropicalis* revealed a prominent diagonal across the contact probability matrix representing expected strong *cis* interactions between neighboring

regions of the genome (Figure 3.3A). However, a prominent cross-like pattern of contact-depleted bins was observed near the centromeres of each chromosome, indicating that the centromeres are excluded from interacting with the rest of the chromosomal arms (Figure 3.3A).

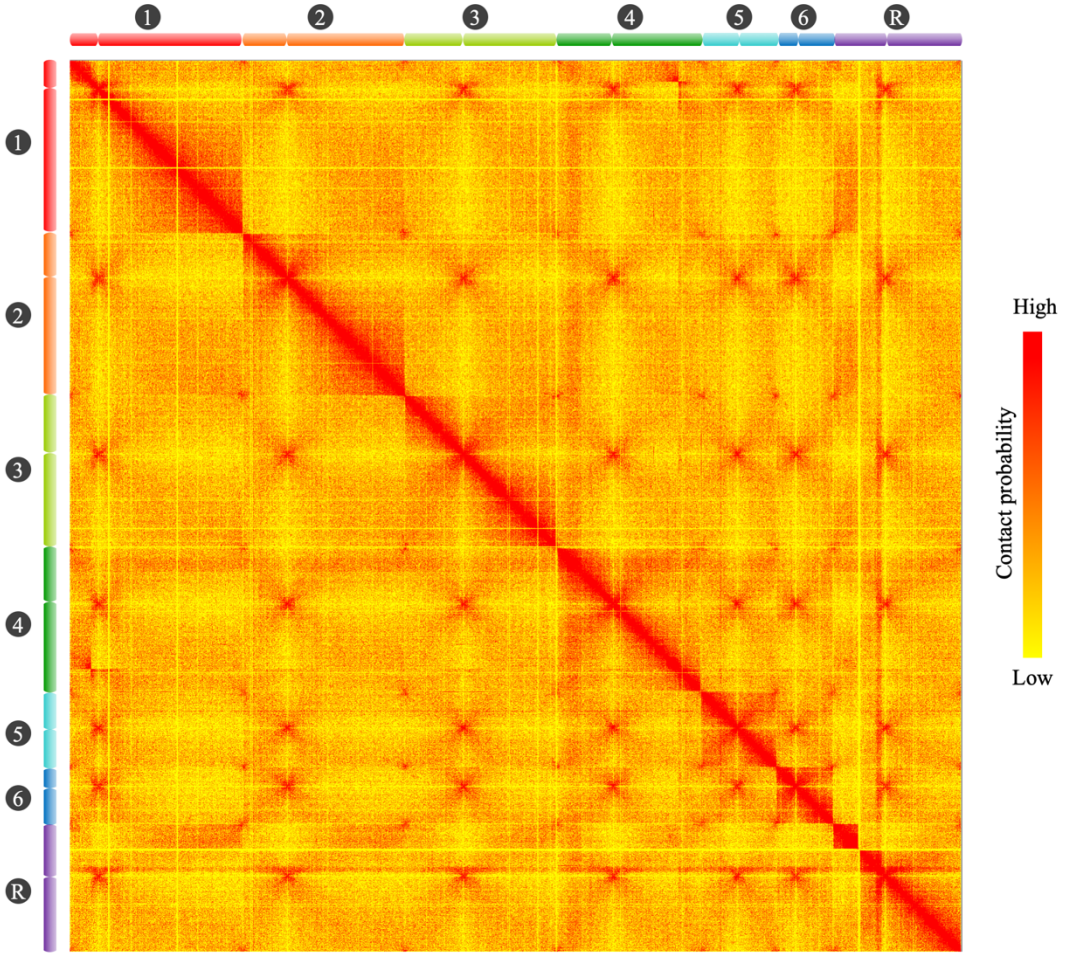


Figure 3.4 The contact probability matrix of the *C. tropicalis* genome obtained from analysis of 3C-seq data using Homer.
A genome-wide contact probability matrix (bin size = 5 kb) generated using 3C-seq data showing spatial contacts between seven chromosomes (cartoon ideogram shown on the top and left) in the *C. tropicalis* genome. The heatmap was generated using Java TreeView software (351). Color-bar represents the contact probability.

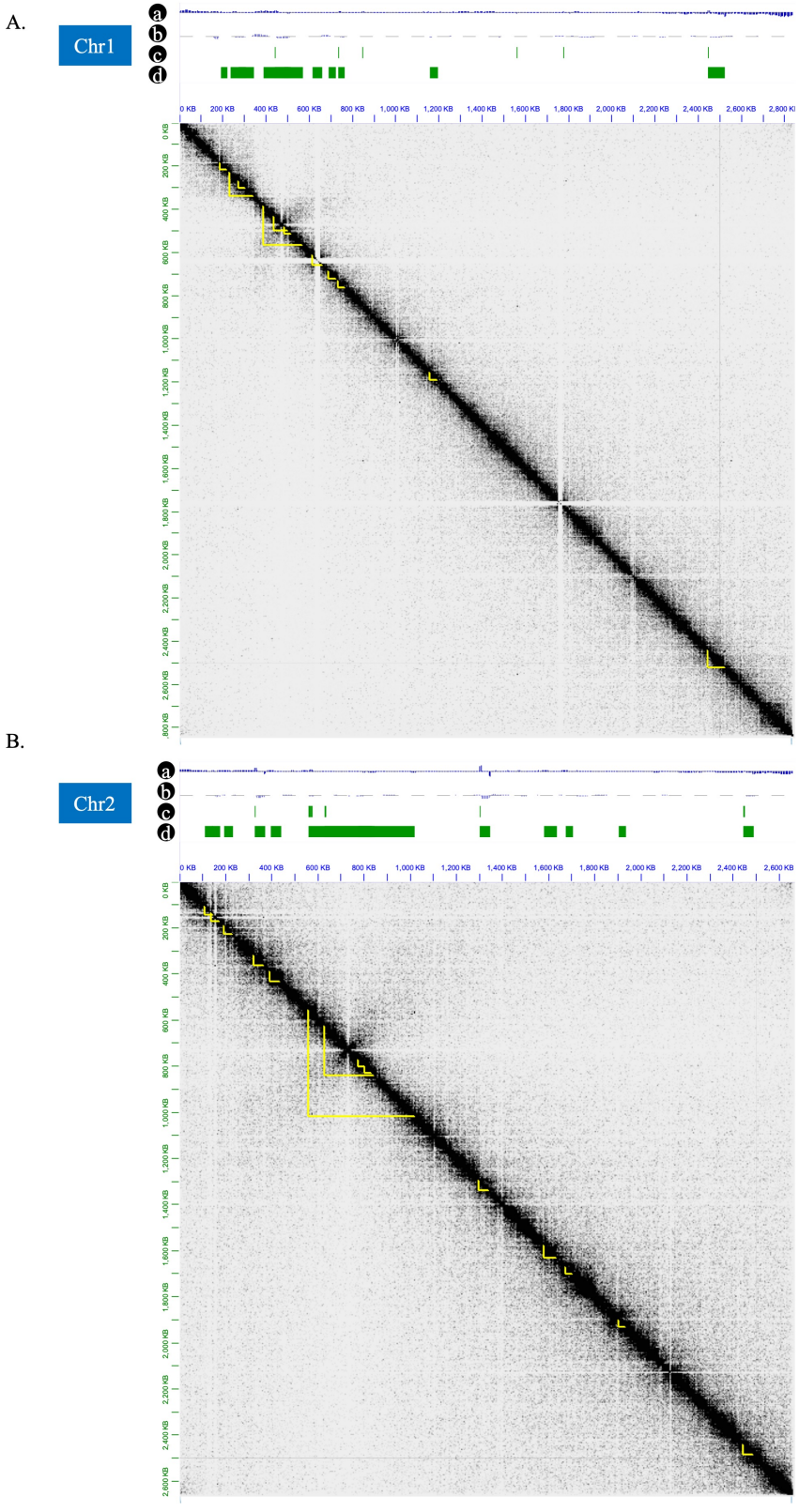
Apart from the *cis* interactions, an examination of the contact probability matrix revealed three distinct features of the genome-wide *trans* contacts across seven chromosomes in *C. tropicalis*. First, a prominent pattern in the contact probability matrix was observed at the interchromosomal areas corresponding to the centromere coordinates. This characteristic pattern of contact probability highlights the *trans* interactions among the centromeres (Figure 3.3A). Second, a cross-like pattern indicating a lack of *cis* contacts between a centromere and

chromosome arms (Figure 3.3B - D) was found to maintain a similar pattern for their contact with the arms of other chromosomes (Figure 3.3A). This observation suggests that each centromere is spatially positioned away from all chromosomal arms, including its own. Third, a significantly higher contact probability between the ends of the chromosomes was observed (Figure 3.3A). Thus, the telomeres of the chromosomes seem to interact with each other in *C. tropicalis*. The extent of interaction among the telomeres seems to be weaker than the *trans* interactions among the centromeres, as evident from the heatmap color density. However, a detailed analysis of the contact probability matrix revealed an interaction between the two telomeres of the same chromosome as well as interactions among telomeres of different chromosomes. These three distinct spatial contact patterns observed in the *C. tropicalis* genome indicates that the centromeres and telomeres are clustered away from each other (352). Independent analysis of the 3C-seq data using Homer (353) also revealed the clustering of centromeres as well as telomeres. (Figure 3.4).

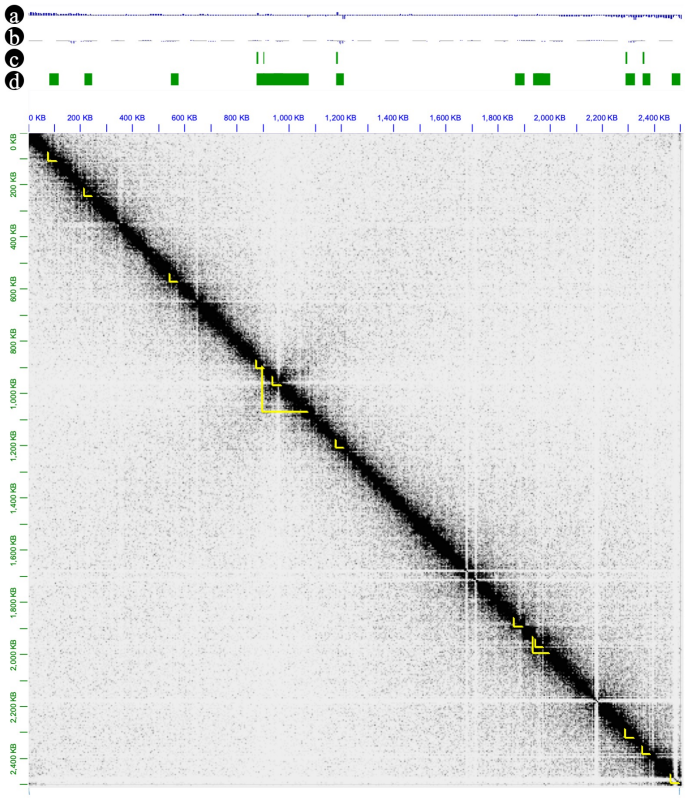
Evidence of topologically associated domains (TADs) in *C. tropicalis*

Analysis of the *cis* contacts across the chromosomes can identify higher-order local structures formed because of the interaction between two relatively distant but linked loci. Based on the nature of their organization, such structures can be classified as loops or topologically associated domains (TADs). The 3C-seq data was analyzed using Homer (353) to study the higher-order chromosomal organization in *C. tropicalis*. Based on the calculation of directionality index and insulation score parameters, analysis of our 3C-seq data at 10 kb resolution led to the identification of 44 TADs and 37 loops across seven chromosomes of *C. tropicalis* (Figure 3.5A - G; Materials and methods). Previously, the length of TADs in the *S. cerevisiae* genome was studied using Micro-C (128), which found the length of TADs is ~5 kb. However, in another study using Hi-C data (127), it was found that the TADs in G1 cells of *S. cerevisiae* can extend up to 400 kb in length. In our study, sorting of the TADs in different length categories reveals that the majority (23/44) of them are 20-40 kb in length (Figure 3.5H). Since the identification of smaller TADs requires contact probability data in the sub-kilobase resolution, we cannot rule out the possibility of additional TADs smaller than 10 kb being present in the *C. tropicalis* genome. However, this is the first evidence for the presence of TADs in any CUG-Ser1 clade species and further studies are required to understand structural and functional aspects of these higher-order chromatin structures. For example, high-resolution mapping of TADs in drug-resistant and sensitive clinical isolates of

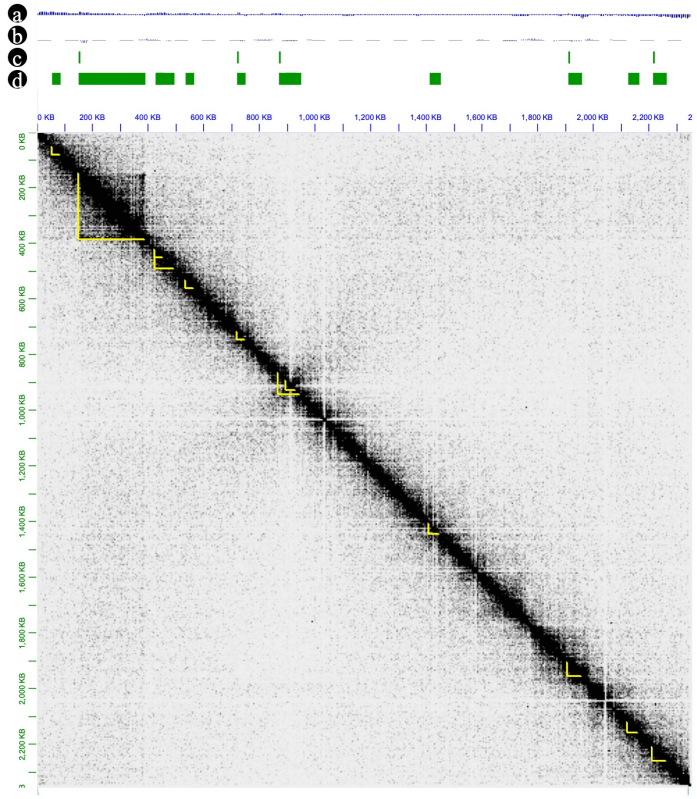
C. tropicalis might identify the contribution of the spatial genome organization on the development of drug resistance in this and related species.



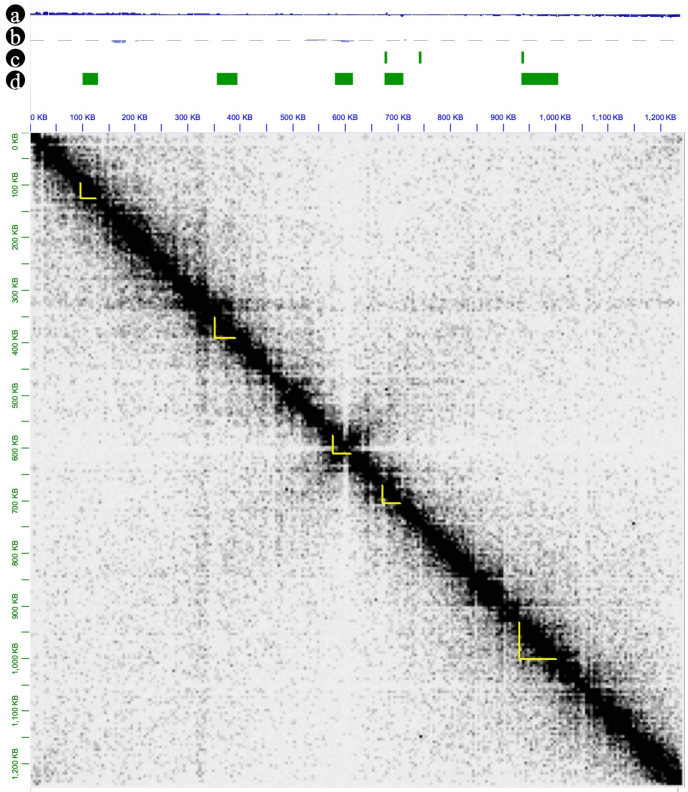
C.


Chr3
a
b
c
d
0 KB
200 KB
400 KB
600 KB
800 KB
1,000 KB
1,200 KB
1,400 KB
1,600 KB
1,800 KB
2,000 KB
2,200 KB
2,400 KB

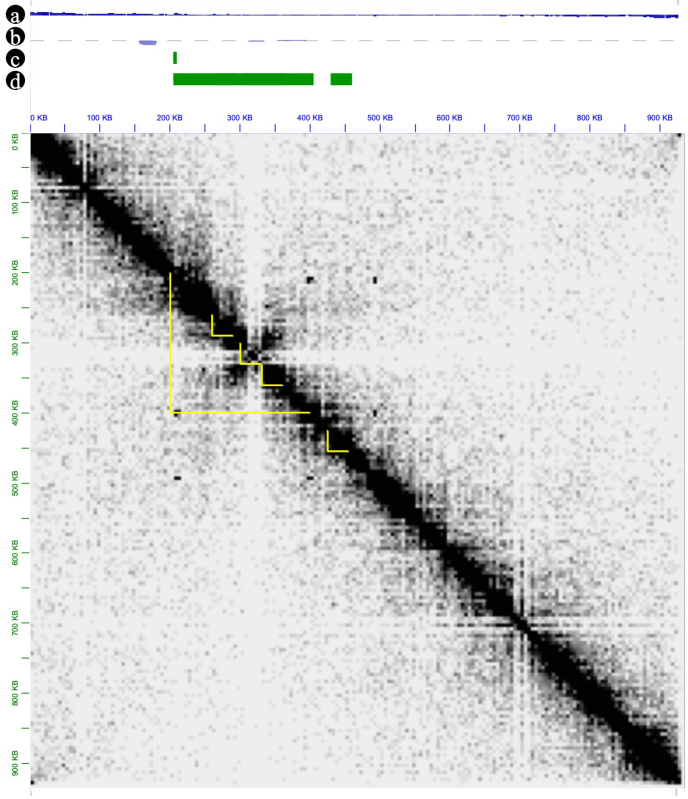
D.


Chr4
a
b
c
d
0 KB
200 KB
400 KB
600 KB
800 KB
1,000 KB
1,200 KB
1,400 KB
1,600 KB
1,800 KB
2,000 KB
2,200 KB

E.


Chr5
a
b
c
d
0 KB
100 KB
200 KB
300 KB
400 KB
500 KB
600 KB
700 KB
800 KB
900 KB
1,000 KB
1,100 KB
1,200 KB

F.


Chr6
a
b
c
d
0 KB
100 KB
200 KB
300 KB
400 KB
500 KB
600 KB
700 KB
800 KB
900 KB

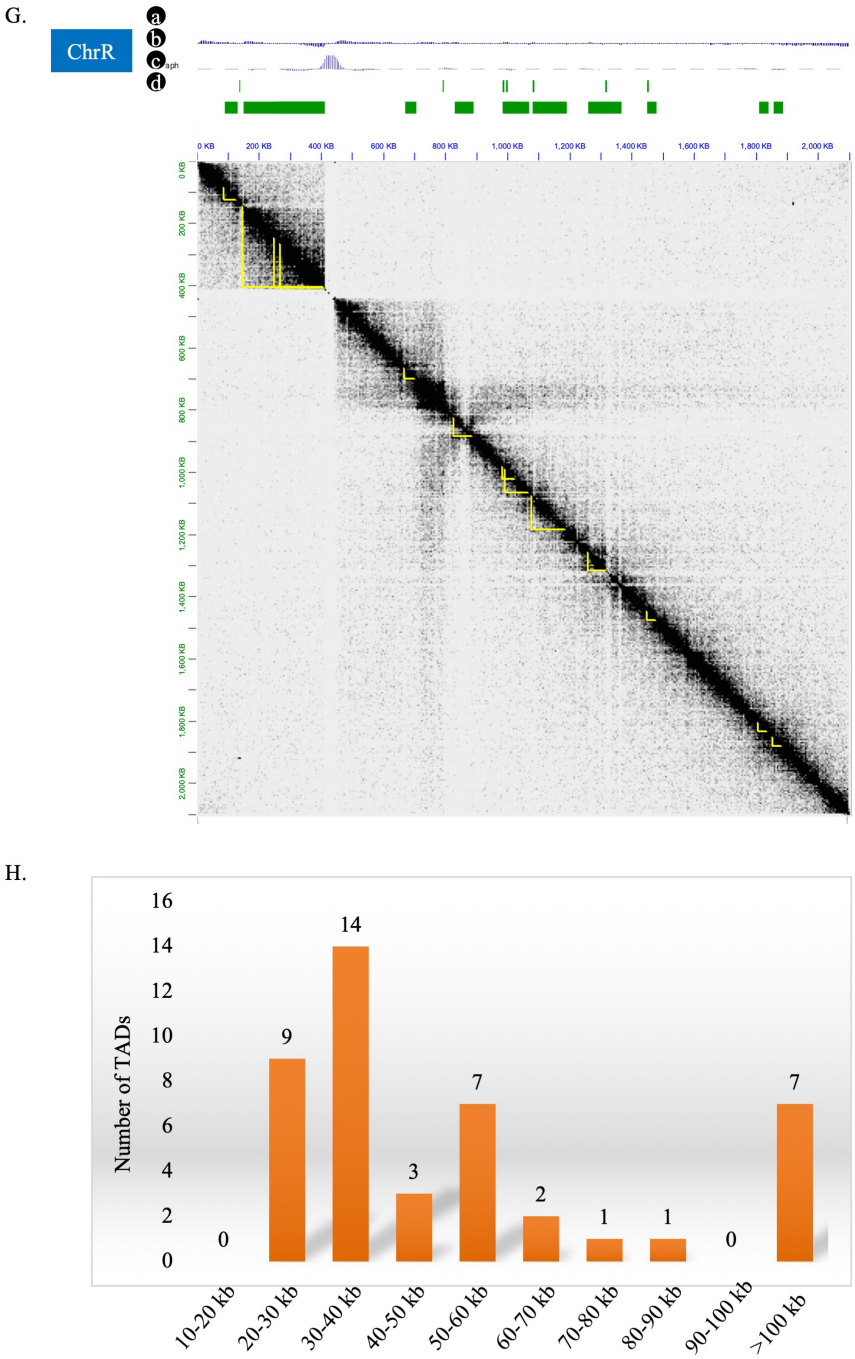


Figure 3.5 Identification of putative TADs in the *C. tropicalis* genome.
A. - G. The 3C-seq data was analyzed using Homer and contact probability matrix was generated at 5 kb resolution for each of the seven chromosomes of *C. tropicalis*. Chromatin compaction feature across the length of each chromosome was analyzed by Homer using directionality index (a) and insulation score (b) statistic and respective profiles are presented at the top of each matrix. Genomic location of each of the 37 loops (c) and 44 TADs (d) identified in this analysis are labelled on top of the contact probability matrix for each chromosome. Contact pattern at each TAD is highlighted using a yellow triangle. These plots were visualized in Juicebox by loading the matrix and individual annotation tracks (a, b, c, and d) generated by Homer. H. A bar chart showing a length-frequency distribution of the 44 TADs identified in *C. tropicalis* genome using Homer. The number of TADs belonging each of the length categories (*x*-axis) is mentioned on top of each bar.

Chapter 4

Results

Genomic rearrangements and centromere-type transition in the CUG-Ser1 clade species

Rapid evolution of centromere-types in closely related members of the CUG-Ser1 clade

Identification of HIR-associated small regional centromeres in *C. tropicalis* (193) in contrast to *C. albicans* (192) and *C. dubliniensis* (191), where the centromeres form on unique and different DNA sequences, is remarkable. Such a rapid transition in centromere structure has not been observed previously in any other closely related species complex. However, it remained unknown if the HIRs associated centromeres emerged in *C. tropicalis* or lost in *C. albicans* and *C. dubliniensis.* The identification of IR-associated centromeres in early diverging species *Komagataella pastoris* (196) suggests that IR-associated centromere-type can be present in the ancestral lineage. Therefore, an attractive hypothesis could be that the ancestral lineages retained DNA sequence dependent centromere, which could have been lost in the derived lineages with a transition into epigenetic centromere types. If this hypothesis is true, the HIR-associated centromere DNA sequences similar to what is identified in *C. tropicalis* should also be present at the centromeres of other closely related species. In addition, these HIR-associated centromere DNA sequences should be able to initiate *de novo* centromere function. However, answers to these questions critical to assess the validity of such a hypothesis remain unknown at this point. Even if the hypothesis is valid, the driving force behind the centromere type transition remains elusive at this point.

Therefore, the closely related members of the CUG-Ser1 clade present a unique opportunity to study the transition in centromere structure and function from an evolutionary standpoint. Our previous analysis suggested that certain centromeres of *C. tropicalis* are located near ICSBs (193). This observation indicates a possibility that interchromosomal recombination events occurred near the centromeres in the common ancestor of *C. tropicalis* and *C. albicans*. What are the factors that aided these translocations? The sub-cellular localization of the kinetochore proteins as a single punctum per cell indicated physical proximity between the centromeres in *C. tropicalis* (193). Therefore, the influence of the spatial proximity on the outcome of the translocations near the centromeres guiding the karyotype evolution in closely related human fungal pathogens of the CUG-Ser1 clade remains as a hypothesis to be tested. However, due to the nature of the then-available fragmented genome assembly, the genome-wide distribution of the ICSBs, as well as the spatial organization of the genome, remained unknown. Therefore, it could not be concluded if the ICSBs are specific to the centromeres, or they are also located elsewhere in the genome.

Mitotic stability assay of the centromeric constructs

The presence of 3-4 kb IRs with highly homogenized sequences on all seven centromeres of *C. tropicalis* suggests a conserved biological function of these sequence elements. We hypothesized that these IRs could initiate DNA sequence-dependent *de novo* activation of centromere function. Therefore, we used pCtCEN5 and pmid5, carrying entire *CtCEN5* DNA

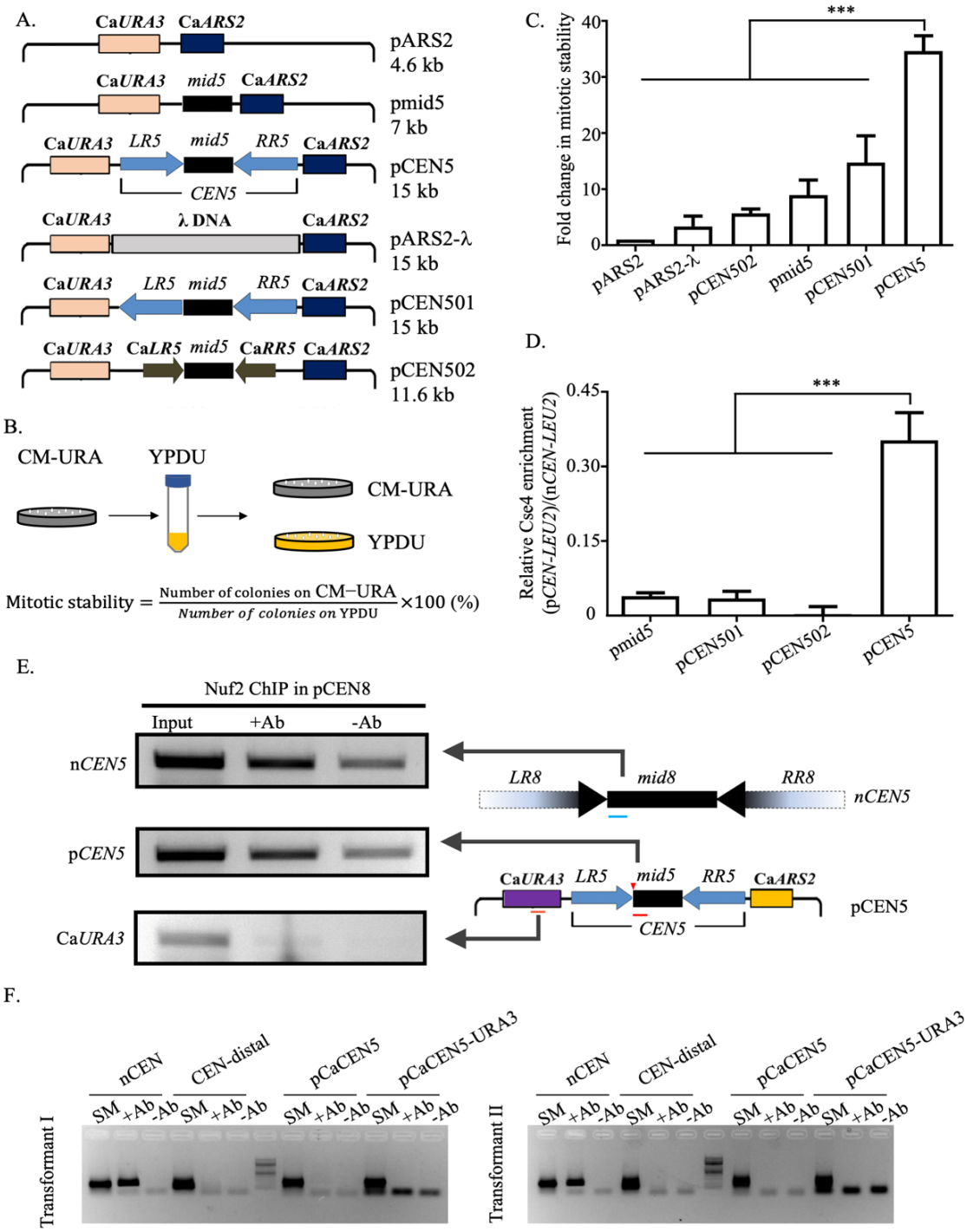


Figure 4.1 DNA sequence dependent regulation of centromere identity in *C. tropicalis*. A. Schematic representation of the plasmid constructs used for mitotic stability assay in *C. tropicalis*. B. Outline of the experimental procedure and the formula used for determination

of mitotic stability. C. A bar chart showing fold change in mitotic stability (*y*-axis) of the plasmid constructs (*x*-axis) compared to the pARS2 parental backbone. Significance of difference between pCEN5 and other constructs was tested using one-way ANOVA analysis. D. A bar chart showing relative CENP-A^{Cse4} enrichment on pmid5, pCEN501, pCEN502 and pCEN5 plasmids. Significance of difference between pCEN*5* and other constructs was tested using one-way ANOVA analysis. E. Ethidium bromide (EtBr) stained gel pictures showing PCR products obtained using input, IP (+Ab) and no antibody control (-Ab) fractions of recovered DNA in a chromatin immunoprecipitation (ChIP) experiment (Materials and methods). This ChIP experiment was performed using *C. tropicalis* strain CtKS300 (Materials and methods) to test relative enrichment of Nuf2-GFP on native centromere (n*CEN5*), plasmid borne centromere (pCEN5), and plasmid borne Ca*URA3* locus. Relative genomic location of the PCR amplified regions from each of n*CEN*5, pCEN5, and Ca*URA3* locus are shown using blue, red and orange horizontal lines, respectively. The location of the SalI restriction site on the pCEN5 is shown using red triangle. E. EtBr stained gel pictures showing PCR products obtained using input, IP (+Ab) and no antibody control (-Ab) fractions of recovered DNA in a ChIP experiment (Materials and methods). This ChIP experiment was performed using *C. albicans* strain CaKG001 (Materials and methods) to test relative enrichment of CENP-A^{Cse4} on native centromere, centromere distal locus, plasmid borne centromere (pCaCEN5), and plasmid borne Ca*URA3* locus. This experiment was performed using two independent transformants, Transformant I (*left*) and Transformant II (*right*) generated by transforming pCaCEN5 in CaKG001 strain. Genotype of each strain used in this experiment is mentioned in Appendix-I.

---

and only the central core (CC) DNA sequence of *CtCEN5,* respectively, in a yeast replicative plasmid (323) (Figure 4.1A) to assay for their mitotic stability. To negate the possibility of a size-dependent increase in mitotic stability and to test the contribution of both orientation and sequence of the IRs, we constructed three additional plasmids pARS2-λ, pCEN501 and pCEN502 (Figure 4.1A) (Materials and methods). We transformed these plasmids in the *C. tropicalis* strain CtKS06 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT*) and used the transformants to assay for the mitotic stability of each construct (Materials and methods) (Figure 4.1B). We found that pCtCEN5 showed a significant improvement in the mitotic stability compared to the pARS2. However, altering one of the IRs into a direct repeat orientation in pCEN501 or replacing both the IR sequences of *C. tropicalis* with those of *CEN5* of *C. albicans* in pCEN502 led to a significant reduction in the mitotic stability (Figure 4.1C).

Structure and sequence dependent *de novo* CENP-A^{Cse4} recruitment on the *C. tropicalis* *CEN-ARS* plasmid

Based on the results of the mitotic stability assay, we suspected that the full-length centromere DNA cloned in pCEN5 might activate *de novo* kinetochore assembly. We

designed a ChIP based experiment to test this possibility and check for *de novo* CENP-A^{Cse4} recruitment on pCEN5. For this experiment, a *C. tropicalis* strain CtKS102 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP* (*CaHIS1*)) was generated in which one of the two alleles of CENP-A^{CSE4} is C-terminally tagged with Protein-A epitope. Next, the centromeric plasmid constructs, pCEN5, pCEN501, pCEN502, and pmid5 were transformed into CtKS102, and three independent transformants carrying each of these constructs were taken forward to perform a ChIP assay to test CENP-A^{Cse4} recruitment on these plasmids. The presence of a unique SalI sequence in the constructs but not on the native centromeres allowed this ChIP assay to detect the presence of CENP-A^{Cse4}, specifically on the plasmid-borne centromere. The relative level of CENP-A^{Cse4} enrichment on the plasmid-borne centromere was calculated with respect to that of the native centromere. This experiment clearly demonstrated *de novo* recruitment of CENP-A^{Cse4} on pCEN5 but not on the other constructs with altered orientation or *CaCEN5IRs* (Figure 4.1D). Based on these results, we conclude that the IRs present at the centromeres of *C. tropicalis* facilitates *de novo* CENP-A^{Cse4} recruitment in both sequence- and orientation-dependent manner.

Next, to test whether the recruitment of CENP-A^{Cse4} on pCEN5 leads to the recruitment of other kinetochore proteins, we constructed a *C. tropicalis* strain CtKG500 (*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT NUF2/NUF2::NUF2-GFP* (*CaHIS1*)) which expresses the outer kinetochore protein Nuf2 with a C-terminal GFP epitope (Materials and methods). This strain was transformed with pCEN5, and the transformants were used to perform ChIP assay for the detection of Nuf2 on pCEN5. In this assay, we could detect the recruitment of Nuf2 on pCEN5 (Figure 4.1E). This experiment indicates *de novo* assembly of the kinetochore on the plasmid-borne centromere in *C. tropicalis*.

The *CaCEN5IR* fails to recruit CENP-A^{Cse4} in *C. tropicalis* cells. However, it remains unknown if the species-specific IR sequences can facilitate *de novo* CENP-A^{Cse4} recruitment through a separate mechanism in *C. albicans.* Therefore, to test the possibility of CENP-A^{Cse4} recruitment by the IR-associated *CEN5* present in *C. albicans,* a similar *CEN-ARS* construct pCaCEN5 was generated. Similar to pCEN5, while cloning of the CC element, we inserted a unique SalI restriction enzyme site, which is absent in the genome. Therefore, the SalI site specific primers could be used to distinguish between native and plasmid-borne centromere

sequence in the ChIP experiment. To test the *de novo* CENP-A^{Cse4} recruitment on pCaCEN5, a *C. albicans* strain CaKG001 (*arg4Δ/arg4Δ, leu2Δ/leu2Δ, his1Δ/his1Δ, ura3Δ::imm434/ura3Δ::imm434, iro1Δ::imm434/iro1Δ::imm434 CSE4/CSE4-TAP*(*CaSAT1*)) was constructed, in which, one of the two alleles of CENP-A^{CSE4} is C-terminally tagged with Protein-A epitope. Using this strain, a ChIP assay was performed, which could not detect CENP-A^{Cse4} recruitment over pCaCEN5 while the presence of CENP-A^{Cse4} on native Ca*CEN5* could be detected (Figure 4.1F).

Identification of homogenized inverted repeat (HIR)-associated centromeres in closely related CUG-Ser1 clade species

Recent advances in DNA sequencing technologies led to the development of high-quality genome assemblies of multiple closely related members of CUG-Ser1 clade and other ascomycetes. For example, chromosome-level genome assemblies are publicly available for *C. parapsilosis* (ASM18276v2) (354), which have diverged from the last common ancestor before the divergence between *C. albicans* and *C. tropicalis* (4). The presence of HIR-associated *CENs* in *C. parapsilosis* genome would prove that the last common ancestor of *C. albicans* and *C. tropicalis* had HIR-associated *CENs*. On the contrary, absence of HIR-associated *CENs* in *C. parapsilosis* would indicate that the HIRs were gained in *C. tropicalis* rather than being lost in *C. albicans* and therefore prove our hypothesis wrong. In such a case, presence or absence of HIR-associated centromere in other species, closely related to *C. tropicalis* should allow inference of the possible trajectory followed during the centromere type transition among the members of CUG-Ser1 clade. Two species, *C. sojae* and *C. viswanathii,* share a closer ancestry with *C. tropicalis* (4, 355). Genome assemblies for both *C. viswanathii* and *C. sojae* are available in the NCBI genome database. However, a highly fragmented assembly of *C. sojae* in 511 scaffolds is not suitable for the identification of the putative centromeres. Therefore, we generated an improved genome assembly of *C. sojae* and used the publicly available genome assembly of other closely related species to perform bioinformatic analyses for identification of the IR-associated putative centromeres among these CUG-Ser1 clade members.

To generate an improved genome assembly of *C. sojae* strain NCYC-2607 (equivalent strain designations are CBS 7871, JCM 1644, MUCL 46191), we performed Illumina and Oxford Nanopore sequencing and used both the datasets to develop an improved

genome assembly (Materials and methods). First, the Oxford Nanopore sequence reads were used to develop a *de novo* genome assembly using Canu (77). Canu produced 42 contigs, which included seven chromosome-length scaffolds. These contigs were then polished using paired-end Illumina sequence data using Pilon (324) to rectify base-pair level errors. This genome assembly of *C. sojae* in 42 contigs consists of a total of 15187968 bp with the N50 of 1778095 bp (Table 4.1). Details of the sequence data and the parameters used for Canu and Pilon run are presented in the materials and methods section.

Table 4.1 Statistics for the *C. sojae* genome assembly

Parameters	***C. sojae* assembly**
asm_contigs	42
asm_esize	1774739
asm_max	3035446
asm_mean	361618
asm_median	33579
asm_min	3595
asm_n50	1778095
asm_n90	310959
asm_n95	55971
asm_total_bp	15187968

With the new and improved genome assembly, we mapped the SNPs and indels present in the *C. sojae* genome. This analysis revealed that most parts of the *C. sojae* genome are homozygous. However, all seven centromere-proximal genomic loci retained heterozygosity (Figure 4.2). This genome-wide loss of heterozygosity in *C. sojae* compared to the mostly heterozygous genome of *C. tropicalis* and *C. albicans* is intriguing. Therefore, we performed an inter-species comparative analysis of the SNP/indel-poor loci. We found that in *C. albicans* and *C. sojae,* part of the genomic region syntenic to the LOH^R locus of *C. tropicalis* is also SNP/indel poor (Figure 4.3). Previous studies have revealed a link between LOH and transition from pathogenic to symbiotic lifestyle in *C. albicans* (356). Whether the genome-wide loss of heterogeneity in *C. sojae* is associated with a non-pathogenic lifestyle of this species remains to be explored.

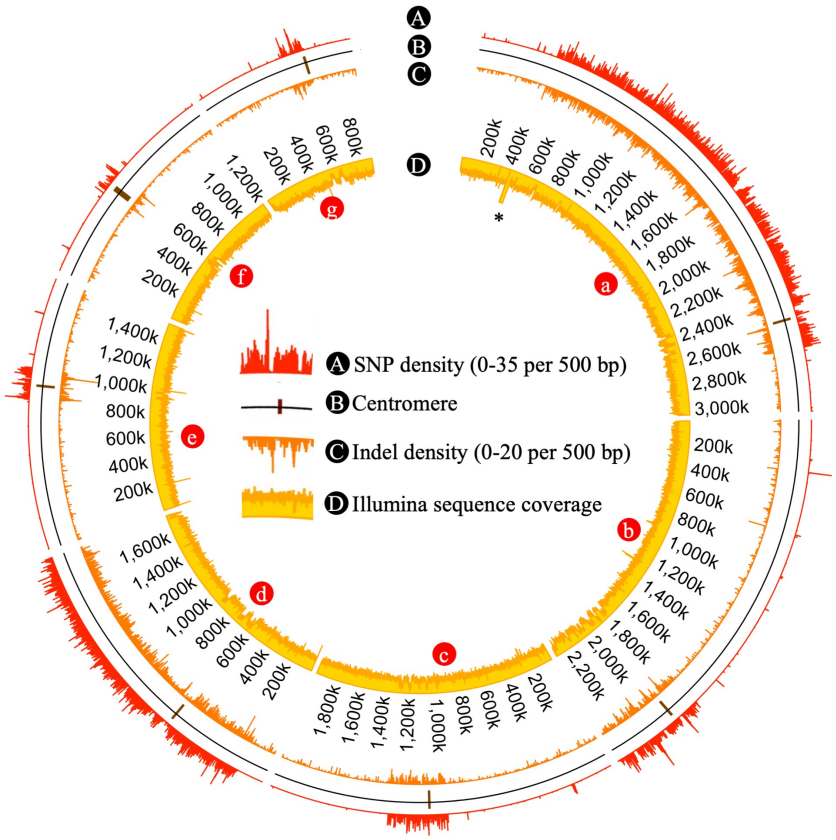


Figure 4.2 Identification of SNPs and indels in the *C. sojae* strain NCYC-2607.
The circos tracks represent the SNP density in red (A), positions of the centromeres (as brown bars) (B) Indel density in orange (C), Illumina sequence coverage in yellow (d). The sequence coverage at the rDNA loci is clipped for clearer representation and marked with an asterisk. The genomic contigs marked as a to g are tig00000002, tig00000008, tig00000017, tig00000001, tig00000038, tig00000050 and tig00016100, respectively.

The publicly available genome assembly of *C. viswanthii* (ASM332773v1) includes 30 contigs comprising of total 24170 kb in a partially diploid assembly that contains duplicated contigs, one of which carry identical DNA sequence mapping to either partially or entirely to another contig. However, the genome assemblies of other species used for comparative genomic analysis does not contain duplicated contigs. Therefore, to maintain uniformity of comparative analysis across all four species, we identified the duplicated contigs present in the genome assembly of *C. viswanthii* (ASM332773v1) (Materials and methods) and excluded them from the assembly to obtain ASM332773v1_modified assembly.

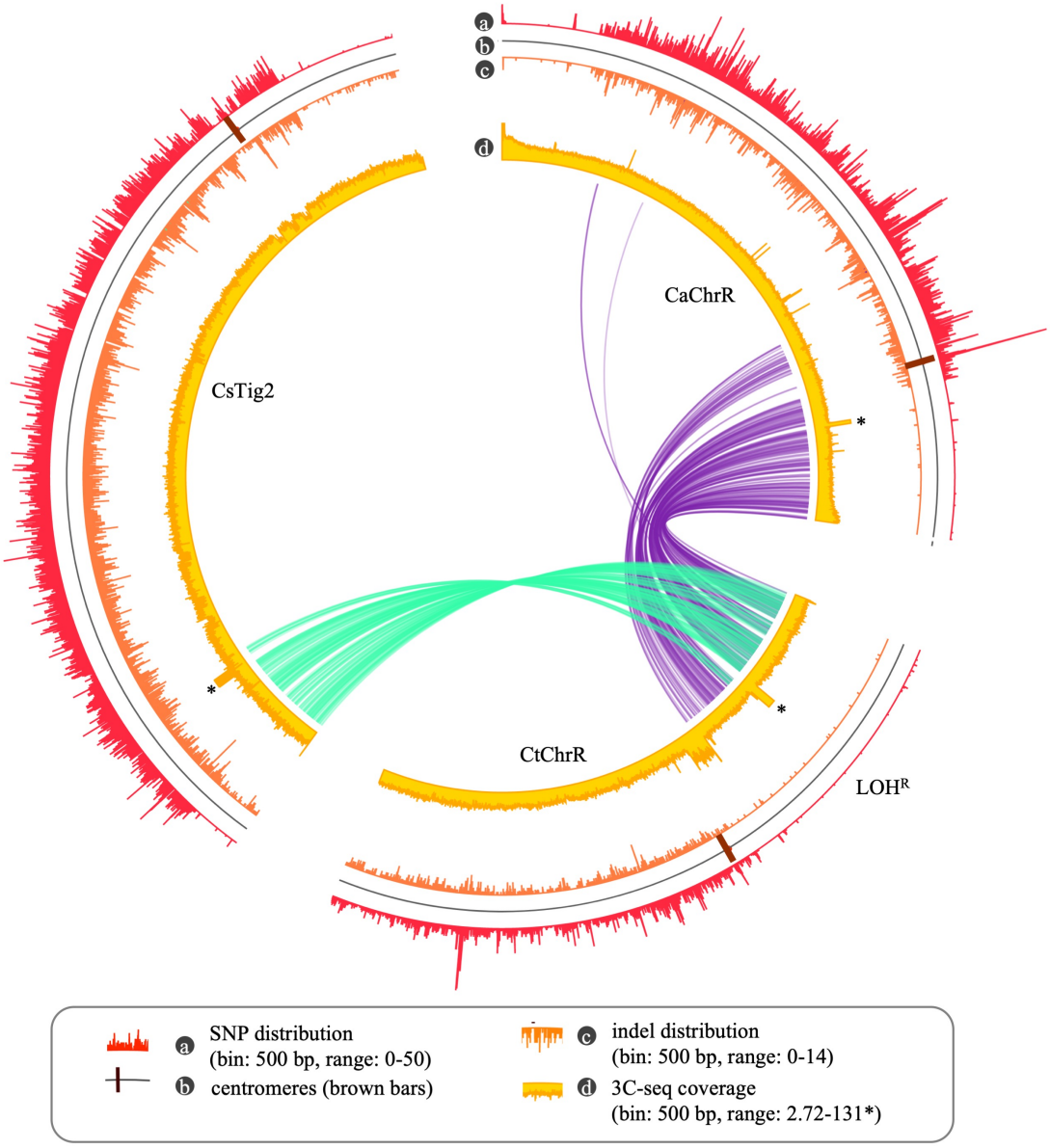


Figure 4.3 Partial conservation of a LOH block in each of the *C. albicans, C. tropicalis* and *C. sojae* genome.
The circos tracks represent the SNP density in red (a), positions of the centromeres (as brown bars) (b) Indel density in orange (c), Illumina sequence coverage (the sequence coverage at the rDNA loci is clipped for clearer representation and marked with an asterisk) (d). The ribbon plot is drawn by connecting the genomic coordinates of the conserved single copy orthologs between *C. tropicalis* and *C. sojae* (teal), and *C. tropicalis* and *C. albicans* (purple).

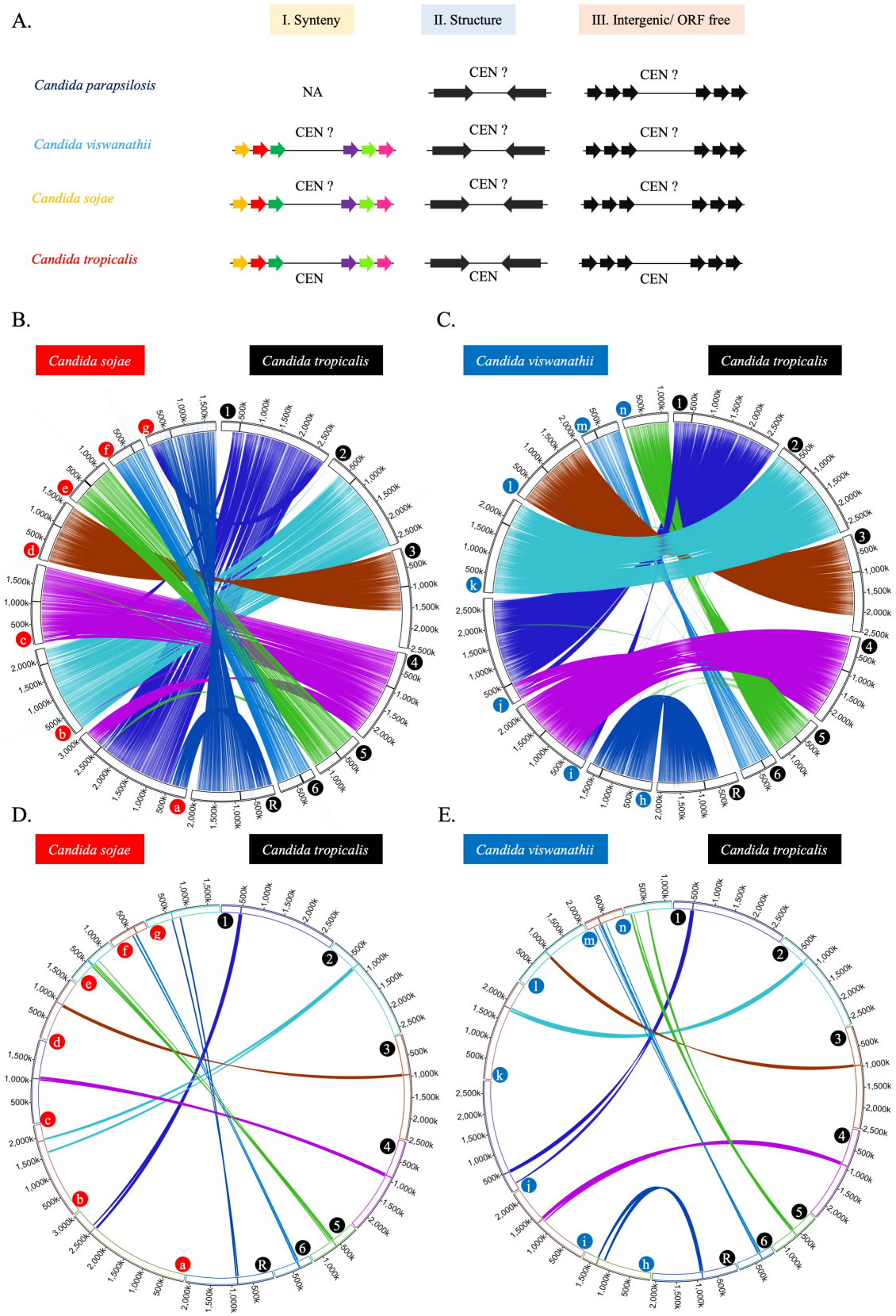


Figure 4.4 Identification of HIR-associated centromeres in the CUG-Ser1 clade.
A. Schematic of the method used for the identification of putative centromeres in *C. sojae*, *C. viswanathii,* and *C. parapsilosis.* Putative centromeric loci in these species were tested for gene synteny with *C. tropicalis* (for *C. sojae* and *C. viswanathii*), presence of IRs, and overlap with intergenic/ORF-free regions. B. Genome-wide synteny of conserved ortholog

pairs between *C. sojae* and *C. tropicalis.* C. Genome-wide synteny between conserved ortholog pairs between *C. viswanathii* and *C. tropicalis*. The location of the centromere on each chromosome is marked with a black bar. Chromosome numbers are marked at the beginning of each chromosome or contigs in colored filled circles. Here, a to g are tig00000002, tig00000008, tig00000017, tig00000038, tig00000050, tig00016100, tig00000001 and h to n are NW_020797881.1, NW_020797885.1, NW_020797858.1, NW_020797886.1, NW_020797884.1, NW_020797877.1 and NW_020797878.1, respectively. Contigs that are either <100 kb in length or do not carry putative centromeres (for *C. sojae*) or duplicated in the genome assembly (for *C. viswanathii*) were excluded from this analysis. The chromosomal coordinates of the ortholog pairs are connected using lines. D. and E. Circos plots similar to that of B and C, showing 10 ORFs on both sides of each centromere of *C. tropicalis* connected to the corresponding genomic loci carrying homologs in *C. sojae* and *C. viswanathii,* respectively.

Next, the newly developed genome assembly of *C. sojae,* modified assembly of *C. viswanathii* (ASM332773v1_modified) and the publicly available genome assemblies of *C. parapsilosis* (ASM18276v2) were used for identification of the putative centromeres in these species. We used three criteria to identify the putative centromeres in *C. sojae* and *C. viswanathii* (Figure 4.4A). First, the entire genome was scanned using YASS (357) for the presence of IR-associated structures. Second, the IRs identified in this search were manually inspected for overlap with the intergenic regions in the genome. Third, the synteny between the centromeres of *C. tropicalis* and theses IR-associated loci in *C. sojae* and *C. viswanathii* were analyzed. This analysis showed that these loci are orthologous to the centromeres of *C. tropicalis* (Figure 4.4B - E). Based on this analysis, seven putative centromeres were identified in *C. sojae* (Table 4.2), and six putative centromeres were identified in *C. viswanathii* (Table 4.2). The putative centromeres of *C. sojae* are comprised of ~2 kb central core (CC) region flanked by 2.6-12 kb long IRs (Table 4.3). The length of CC and IRs ranged from 5-10 kb and 2.6-3.7 kb, respectively, in *C. viswanathii* (Table 4.4). Similarly, eight unlinked IR-associated loci were also identified in *C. parapsilosis* genome (Table 4.1), which are present once in each of the eight chromosomes. Moreover, these putative centromeres overlap with ORF-free regions of the genome and they are poorly transcribed (Figure 4.5A), similar to the centromeres of *C. albicans* (Figure 4.5B). Taken together, identification of IR associated putative centromeres in *C. parapsilosis, C. sojae,* and *C. viswanathii* indicate that the centromeres of the last common ancestor of *C. tropicalis* and *C. albicans* were IR associated structures.

Table 4.2 Genomic coordinates of putative HIR associated centromeres in *C. sojae, C. viswanathii,* and *C. parapsilosis*

Species	*CENs*	Coordinate
C. sojae	1	Tig2:2493969-2501301
	2	Tig8:1905358-1914164
	3	Tig38:934048-941518
	4	Tig17:1062429-1069512
	5	Tig50:549697-575759
	6	Tig16100:638309-647019
	R	Tig1:623000-630276
C. viswanathii	1	NW_020797858.1:473076-485506
	2	NW_020797886.1:1829726-1842348
	3	NW_020797884.1:937333-949114
	4	NW_020797885.1:1322147-1333995
	6	NW_020797877.1:349856-362272
	R	NW_020797881.1:1345731-1360763
C. parapsilosis	1	HE605203.1:362,588-368,324
	2	HE605204.1:470,757-476,666
	3	HE605205.1:1,281,284-1,287,683
	4	HE605206.1:1,309,222-1,314,566
	5	HE605207.1:658,126-664,775
	6	HE605208.1:888,702-895,821
	7	HE605209.1:470,949-477,776
	R	HE605202.1:209,424-215647

Figure 4.5 Putative centromeres of *C. parapsilosis* are ORF-free and transcription poor loci similar to the centromeres of *C. albicans*.
IGV track images showing transcription status (number of transcripts mapped per kb of genomic region) at the genomic loci proximal to the putative centromeres (red) of *C. parapsilosis* (*left*) and *C. albicans* centromeres (*right*). Two replicates of mRNA-seq data are shown as top two tracks as lighter and darker shades of blue and teal for *C. parapsilosis* and *C. albicans,* respectively. The location and orientation of the ORFs (>300 bp) are shown in the lowermost track as black bars.

Table 4.3 Length of the centromere DNA elements in *C. sojae*

CEN*	**outer LR (bp)**	**left repeat (bp)**	**central core (bp)**	**right repeat (bp)**	**outer RR (bp)**
CENR	1389	2649	1967	2661	1386
CEN1	3957	2691	2138	2594	4009
CEN2	ND	3363	2111	3333	ND
CEN4	ND	2493	2068	2523	ND
CEN3	ND	2633	2207	2631	ND
CEN5[#]	12092		2174	11797	
CEN6	ND	3313	2086	3312	ND

*Syntenic to *CtCENs* ND: Not detected [#]Outer repeat is fused with the IR flanking the central core

Table 4.4 Length of the centromere DNA elements in *C. viswanathii*

CEN*	**left repeat (bp)**	**central core (bp)**	**right repeat (bp)**
CEN1	3238	5955	3238
CEN2	3792	5180	3651
CEN3	3228	5326	3228
CEN4	3115	5619	3115
CEN6	3165	6090	3162
CENR	2617	9798	2618

*Syntenic to *CtCENs*

All seven centromeres of *C. tropicalis* are highly homogenized (193). To test whether, the IR-associated putative centromeres identified in *C. sojae, C. viswanathii* and *C. parapsilosis* are also homogenized, pair-wise sequence alignments were performed between all pairs of the putative centromeres in each species. This analysis revealed a high level of sequence homology among the IR-associated centromeres present in each species. We noted

that the DNA sequences of the IR are more conserved than the DNA sequences of the CCs present within a species (Figure 4.6A).

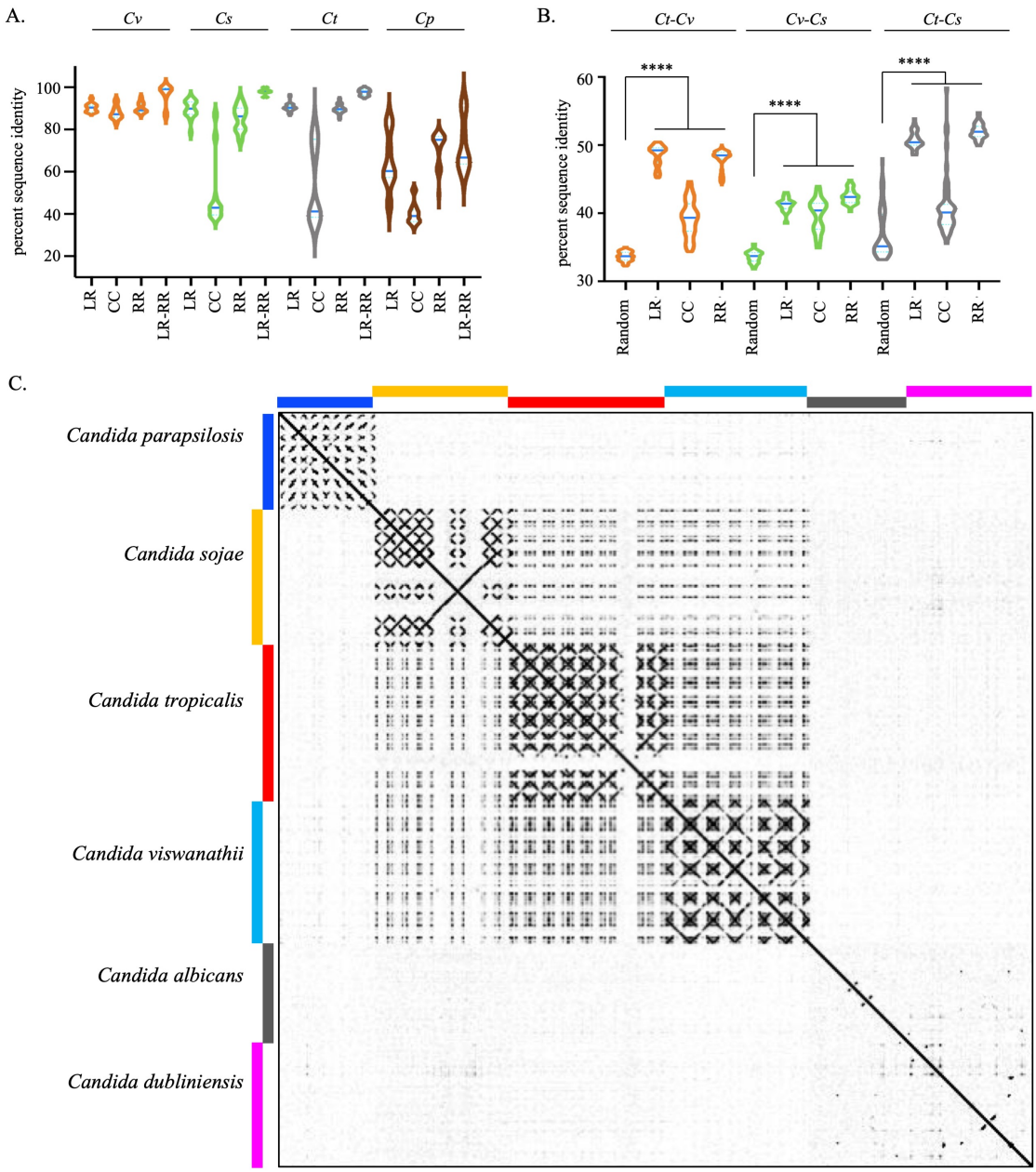


Figure 4.6 Intra- and inter-species conservation in homogenized inverted repeat-associated centromeres in the CUG-Ser1 clade.

A. Violin plots show the distribution of the percent sequence identity values obtained from pair-wise alignment of the left repeats (LR), the central cores (CC), and right repeats (RR) within *C. viswanathii* (*Cv*; orange), *C. sojae* (*Cs;* green), *C. tropicalis* (Ct; gray), and *C. parapsilosis* (*Cp;* brown). The sequence identity between all possible pairs of CC, LR, RR, and LR-RR pairs was calculated in blastn analysis using Clustal Omega. The median value for each dataset is depicted as a horizontal blue bar on each of the violin plots. B. Violin plots showing distribution and median (horizontal blue bar on each of the violin plots) of percent sequence identity values obtained from pairwise DNA sequence alignment of all possible combinations for each of seven random loci (Random), centromeric left repeats (LR), central cores (CC) and right repeats (RR) between species-pairs as indicated, using Clustal Omega. The significance of difference between percent sequence identity of centromere elements and

random loci for all three species pairs were tested using the Mann-Whitney U test ($P < 0.05$) and the P value summary for each comparison is represented with asterisks. C. A dot-plot matrix representing the sequence and structural homology of centromeres among species of the CUG-Ser1 clade was generated using Gepard (Materials and methods).

However, the sequence conservation across the CC elements of *C. viswanathii* centromeres are comparatively higher than other species (Figure 4.6A). Further, an examination of the cross-species conservation in the CC and IR DNA sequences revealed that the IRs remain well conserved across *C. tropicalis, C. viswanathii* and *C. sojae* but the CCs remain comparatively less conserved. However, both IRs and CCs across these three species show a significantly higher level of conservation compared to DNA sequences of random genomic loci (Figure 4.6B). Next, we performed a dot-plot analysis of the putative centromeres identified in *C. viswanathii, C. sojae,* and *C. parapsilosis* along with the known centromeres of *C. albicans, C. tropicalis,* and *C. dubliniensis* using Gepard (358) (Figure 4.6C). This analysis highlighted the inter-species conservation of structure as well as DNA sequences among HIR-associated centromeres. We also noted that each centromere DNA sequence is completely unique in *C. albicans* and *C. dubliniensis,* where HIRs are absent.

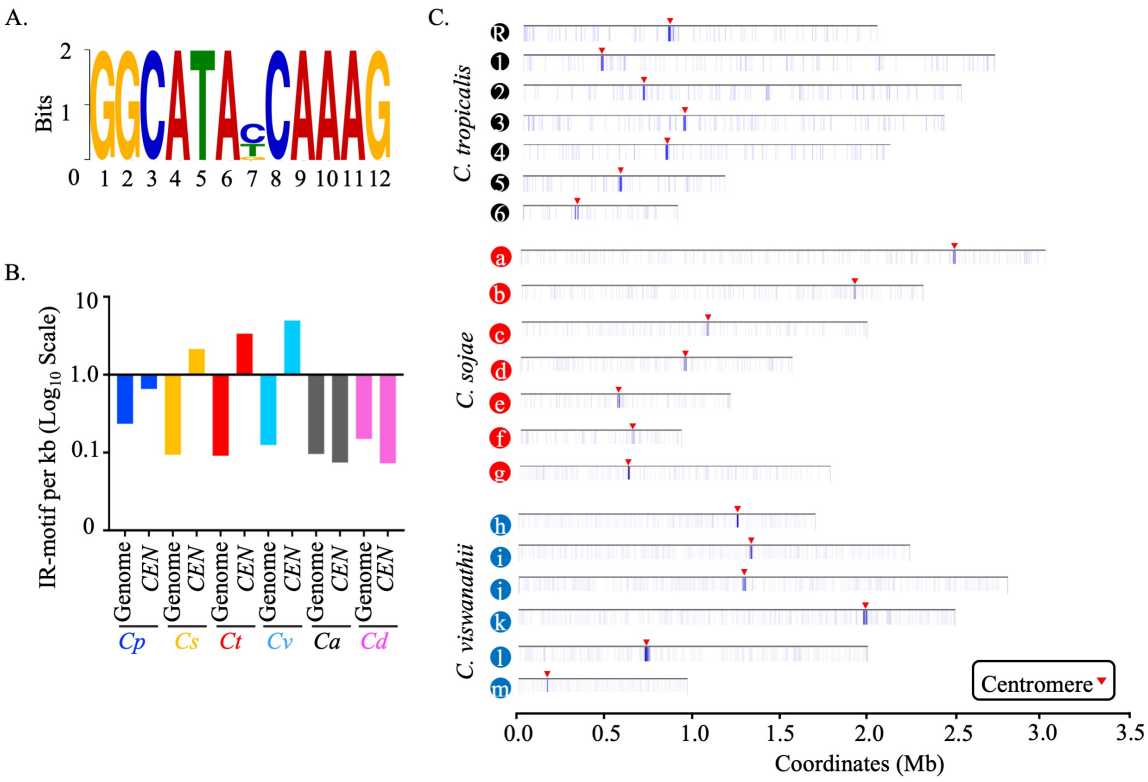


Figure 4.7 Identification of an inter-species conserved and centromere-enriched DNA sequence motif in CUG-Ser1 clade species.

A. A logo plot showing the 12 bp long inter-species conserved motif (IR-motif), identified using MEME-suit (Materials and methods). B. 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. Note that *C. albicans* and *C. dubliniensis* centromeres that form on unique and different DNA sequence do not contain the IR-motif. C. IGV track images showing the IR-motif density across seven chromosomes of *C. tropicalis,* and contigs containing the putative centromeres in *C. sojae,* and *C. viswanathii.* Location of the centromere is marked with a red arrowhead.

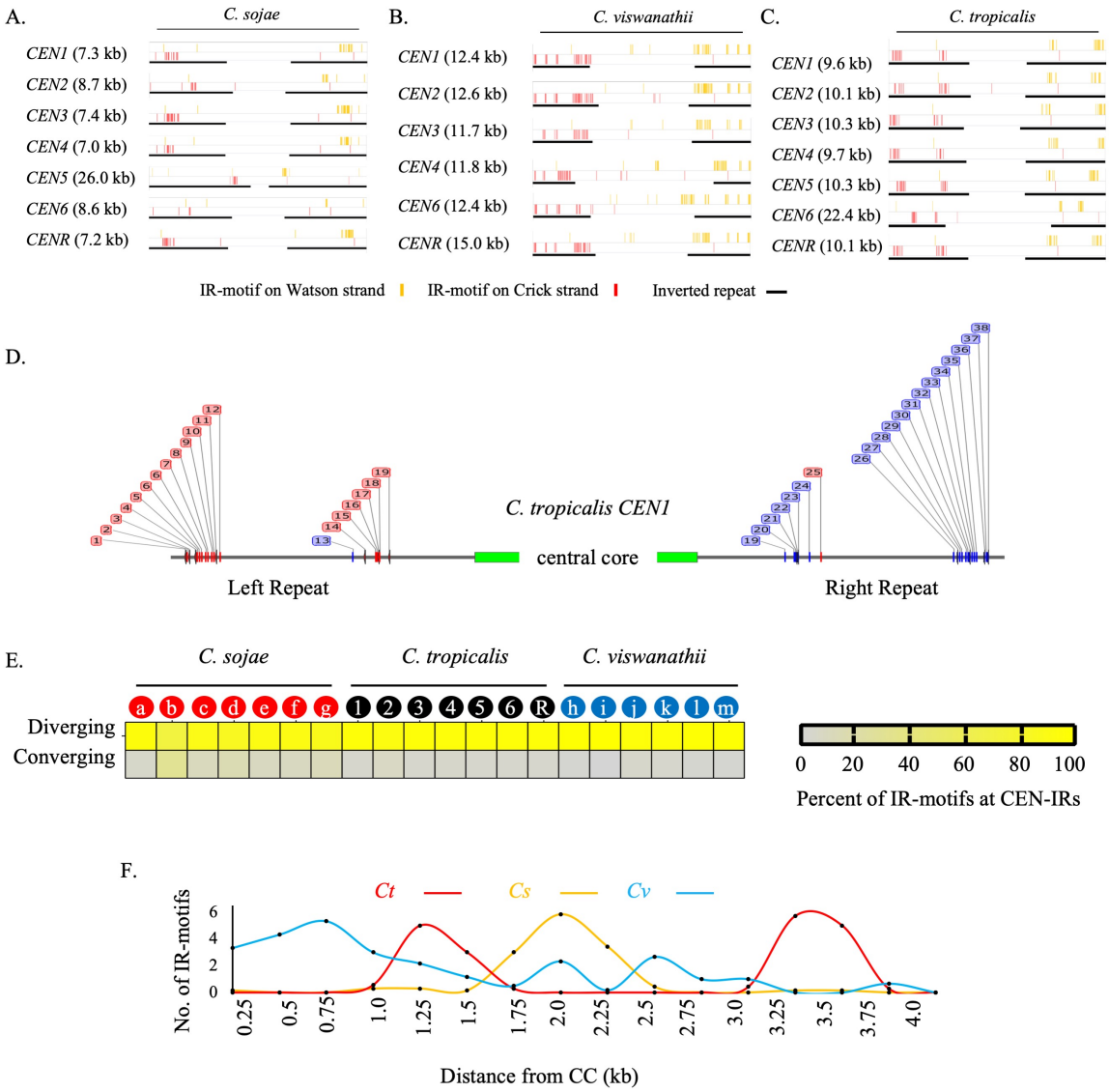


Figure 4.8 Organization of IR-motifs on homogenized inverted repeat-associated centromeres of *C. tropicalis, C. sojae* and *C. viswanathii.*

A., B., and C. shows IR-motif distribution across the HIR-associated centromeres of *C. tropicalis*, *C. sojae,* and *C. viswanthii,* respectively. D. A representative figure showing a zoomed view of the IR motif distribution on *C. tropicalis CEN1* DNA. The motifs on the Crick strand (red) and Watson strand (blue) are color-coded. E. A heatmap showing the percent of IR-motifs present in converging and diverging orientation with respect to the central core region for each of the HIR associated centromeres present in *C. sojae, C. tropicalis,* and *C. viswanathii.* F. Average number of IR-motifs per 250 bp on the IRs is

plotted in the *y*-axis as a function of the distance from the start of CC (*x*-axis) for *C. tropicalis* (red), *C. sojae* (yellow), and *C. viswanathii* (blue).

Inter-species conservation of the HIR-associated centromere DNA sequences indicates a conserved biological function, which may implement through conserved DNA motifs. Therefore, IR DNA sequences from all the centromeres of *C. tropicalis* and the putative centromeres of *C. sojae* and *C. viswanathii* were used as the input query sequences to identify conserved motifs using MEME suite (359). This analysis identified a conserved 12-bp motif, which we named as IR-motif (Figure 4.7A). We found that the IR motif is enriched at the centromeres of *C. tropicalis, C. viswanathii* and *C. sojae* but not in *C. albicans* or *C. dubliniensis* (Figure 4.7B). Therefore, our previous observation that the pCaCEN5 fails to facilitate *de novo* CENP-A^{Cse4} recruitment together with the absence of the IR-motifs in *C. albicans* possibly mean a functional significance of the inter and intra-species conserved IR-motif in CUG-Ser1 clade. Moreover, the *CEN*-enriched motif was found to be specifically concentrated on the IRs but not at the central core regions in HIR-associated centromeres present in *C. tropicalis* (Figure 4.8A) as well as at the putative centromeres in *C. soaje* (Figure 4.8B) and *C. viswanathii* (Figure 4.8C). Additionally, we detected that the direction of the IR-motif is diverging away from the central core of the centromeres in *C. tropicalis* (Figure 4.8D), and this pattern remains conserved in *C. soaje* and *C. viswanathii* as well (Figure 4.8E). However, we noted that the clusters of IR-motif are located at a variable distance from the central core in these three species (Figure 4.8F). The importance of this 12-bp conserved motif on the centromere function is yet to be determined.

Analysis of the interchromosomal synteny breaks (ICSBs) in *C. tropicalis* genome

Identification of the HIR-associated putative centromeres in the closely related members of the CUG-Ser1 clade revealed a rapid transition from the IR-associated ancestral centromere type to the unique centromere type. However, the driving force facilitating this transition remained unknown. Our previous analysis showed that none of the centromeres in *C. tropicalis* are orthologous to *C. albicans* (193). In addition, certain centromeres of *C. tropicalis* are located near the ICSBs when compared to the *C. albicans* genome (193). This observation indicates a possibility of centromere-proximal translocations associated with the

transition in centromere type. Now, with the identification of HIR-associated putative

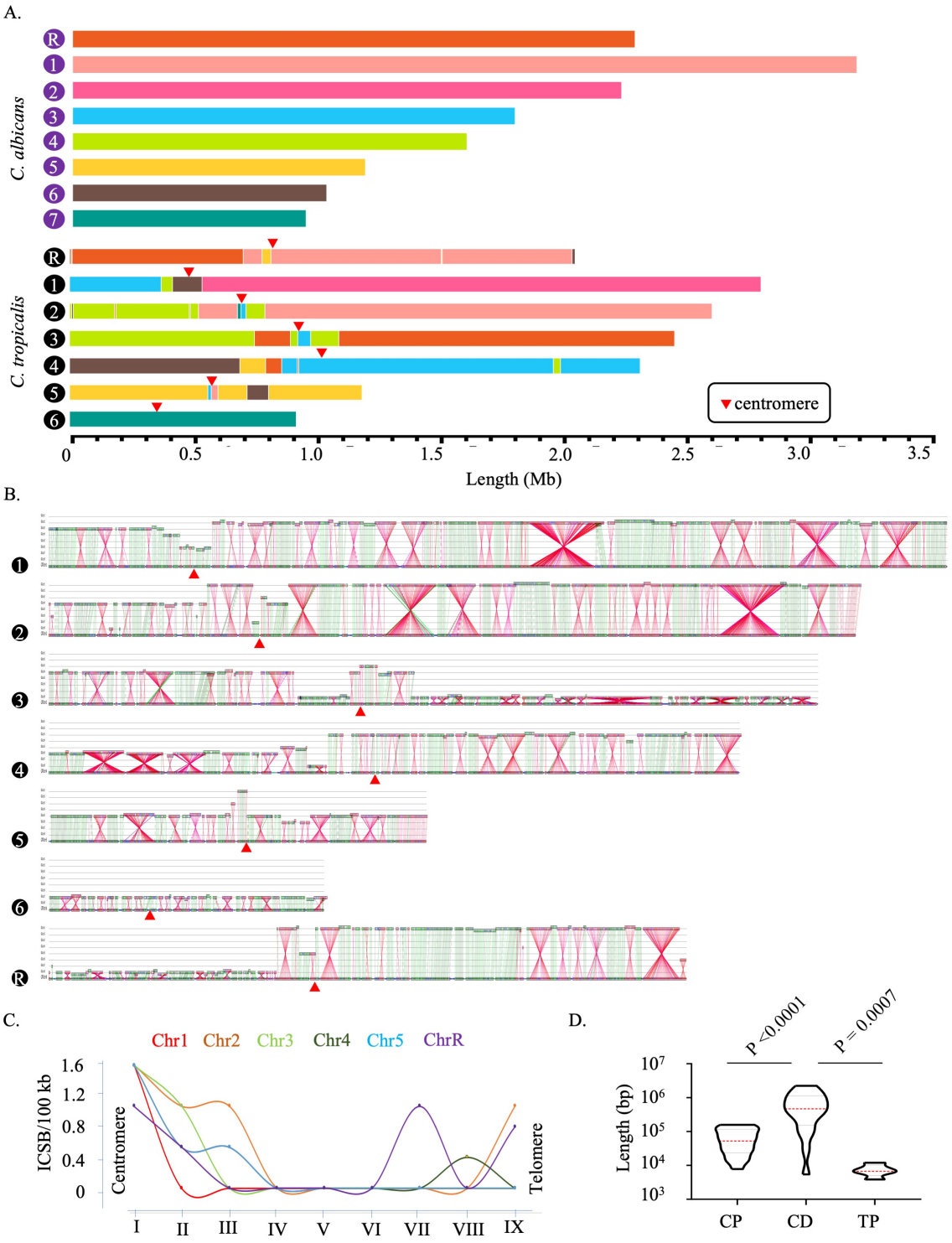


Figure 4.9 Genome-wide mapping of ICSBs on *C. tropicalis* genome reveals spatial regulation of centromere-proximal translocations in their common ancestor.

A. A scaled representation of the color-coded orthoblocks (relative to *C. albicans* chromosomes) and interchromosomal synteny breakpoints (ICSBs) (white lines) on *C. tropicalis* (Materials and 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. Synteny maps of *C. tropicalis* chromosomes (the lowermost line of each panel, marked by filled black circles numbered from 1 to R), with respect to *C. albicans* chromosomes (lines above the *C. tropicalis* chromosomes), in the order of Chr1 to ChrR (top to bottom) for all panels. Centromeres, red triangles. The ORFs (represented as beads) are color-coded: inverted, red and non-inverted, green. The more conserved the reciprocal best hits (RBH) are, the darker are the shades of red/green color. 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. D. 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 U test and the respective *P* values are presented.

---

centromeres in early-diverging *C. parapsilosis*, it is more logical to conclude that the common ancestor of *C. tropicalis* and *C. albicans* possessed HIR-associated centromeres. We speculate that the DNA sequence homology among the spatially clustered centromeres favored inter-centromeric translocations in the last common ancestor of *C. tropicalis* and *C. albicans*. Therefore, we attempted to map the genome-wide distribution of ICSBs to test if these relics of the ancient translocations are specifically enriched at the centromeres of *C. tropicalis*.

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 (326) and Satsuma synteny (360), and a custom approach (Materials and methods) to identify the ICSBs based on the synteny of the conserved single-copy orthologs (Figure 4.9A).

Next, we compared and validated the results obtained from our custom approach of analysis with another published tool Synchro (Figure 4.9B) (361). 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 4.9A).

Additionally, we found a convergence of orthoblocks from as many as four different chromosomes (for *CEN2*) within 100 kb of centromeres.

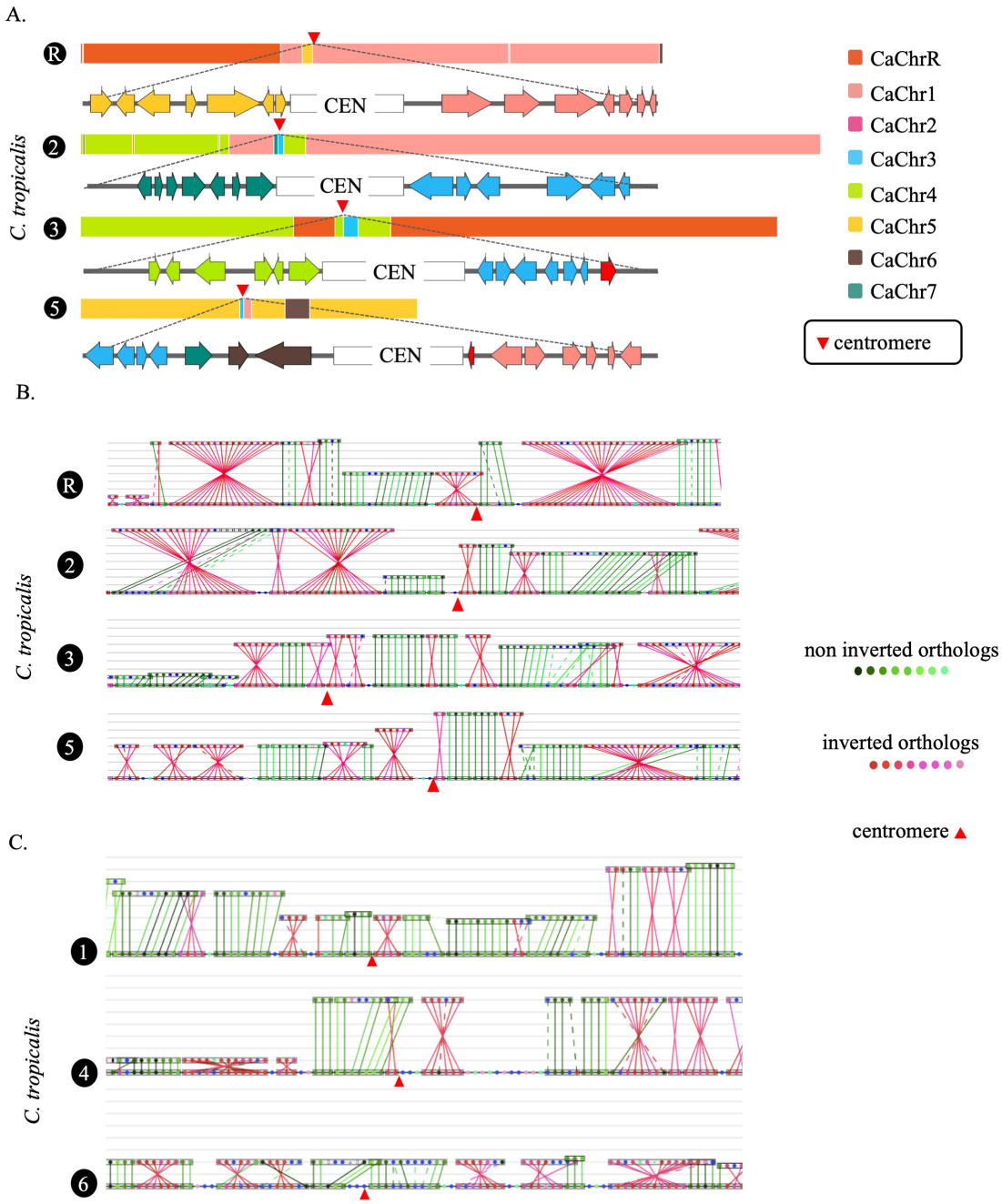


Figure 4.10 ORF-level synteny of the centromere proximal loci in *C. tropicalis* with respect to *C. albicans* genome.
A. Zoomed view of the centromere specific ICSBs on *CEN2*, *CEN3*, *CEN5* and *CENR* showing the color-coded (relative to *C. albicans* chromosomes) ORFs flanking each centromere. *C. tropicalis*-specific unique ORFs proximal to *CEN3* and *CEN5* are shown in red. B. The zoomed view of the reciprocal best hit (RBH) orthologs proximal to the centromeres of *C. tropicalis* for chromosome R, 2, 3, and 5 where each centromere is located at an ICSB. C. Zoomed view of centromere-proximal loci in Chr1, Chr4, and Chr6, in which centromere is located at an intra-chromosomal synteny breakpoint.

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) for different zones across the chromosome for all chromosomes except CtChr6 (Figure 4.9C). Our analysis reveals that the ICSB density is maximum at the centromere-proximal zones for all six chromosomes, but drops sharply at the chromosomal arms. However, the ICSB density near the telomere-proximal zone for Chr2, Chr4, and ChrR show an increase compared to the chromosomal arms, albeit at a lower magnitude than centromeres. We also compared the lengths of orthoblocks across three different genomic zones - the centromere-proximal (0 - 300 kb from the centromere on both sides), centromere-distal (>300 kb from the centromere to 200 kb away from the telomere ends), and telomere-proximal (0 - 200 kb from the telomere ends) zones. This analysis further reveals that the lengths of the orthoblocks located proximal to centromeres and telomeres are significantly smaller than orthoblocks located at the centromere-/telomere-distal zones (Figure 4.9D).

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 at least one ICSB is mapped within ~100 kb of *CEN1* and *CEN4* (Figure 4.10A - B). However, no ICSB could be identified on Chr6. Intriguingly, intra-chromosomal synteny breakpoints were found adjacent to *CEN1*, *CEN4* and *CEN6* (Figure 4.10C). This observation indicates that intra-chromosomal rearrangements led to the loss of HIR-associated DNA sequences.

Our observation indicates a possibility of 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 at least ten instances where such convergence is detected (Figure 4.11A). Intriguingly, four such loci are proximal to the centromeres (*CEN3, CEN4, CEN7,* and *CENR*) in *C. albicans* (Figure 4.11B - E). This observation supports the idea that HIR-associated centromeres in the common ancestor of *C. albicans* and *C. tropicalis* located at close proximity facilitated inter-centromeric translocation events. We also note that the other four centromeres (*CEN1, CEN2, CEN5,* and *CEN6*) are located proximal to ORFs, homologs of which are also proximal to certain centromeres in *C. tropicalis* (Figure 4.11A). 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 (192).

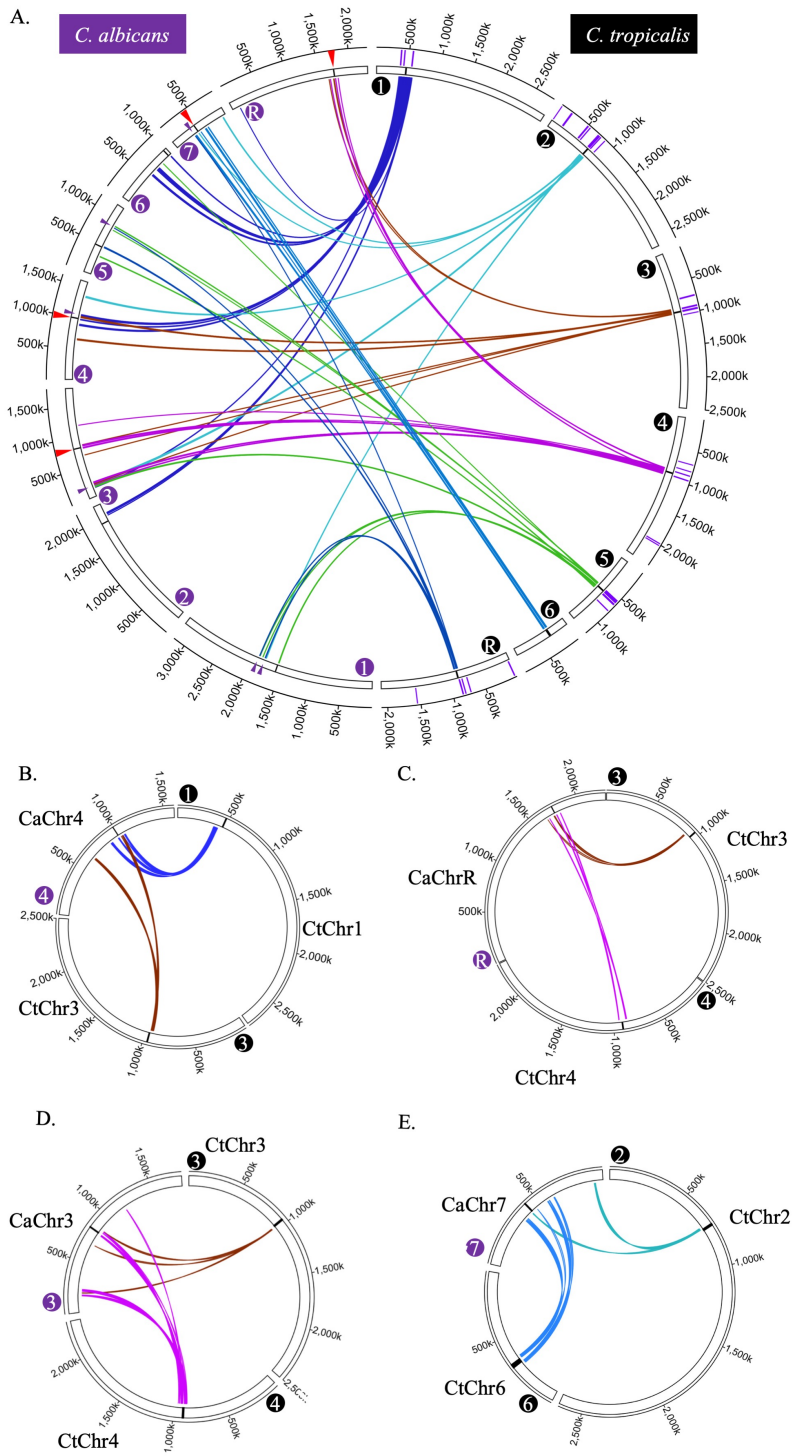


Figure 4.11 Genome-wide synteny analysis between *C. albicans* and *C. tropicalis* finds evidence of inter-centromere translocations in the last common ancestor.
A. Circos plot showing the ICSBs (purple lines on the outer-most circle) on *C. tropicalis* chromosomes (marked with black filled circles). The centromere-proximal ORFs present in *C. tropicalis* are connected to their homologs present on *C. albicans* chromosomes (marked by purple filled circles) by color-coded lines (based on their origin). The positions of

centromeres are marked with black lines of the inner-most circle in each chromosome. The genomic locations in *C. albicans* chromosomes showing the convergence of ORFs from at least two centromere-proximal loci of *C. tropicalis* are marked with red (proximal to the *C. albicans* centromere) and purple (a non-centromere locus) triangles. Note that all centromeres of *C. albicans* are proximal to ORFs, homologs of which are proximal to centromeres of *C. tropicalis.* B - E. Circos representation showing the convergence of centromere-proximal ORFs of *C. tropicalis* chromosomes near the centromeres on *C. albicans* Chr4, ChrR, Chr3, and Chr7, respectively. Chromosomes of *C. tropicalis* and *C. albicans* are marked with black and purple filled circles, respectively, at the beginning of each chromosome.

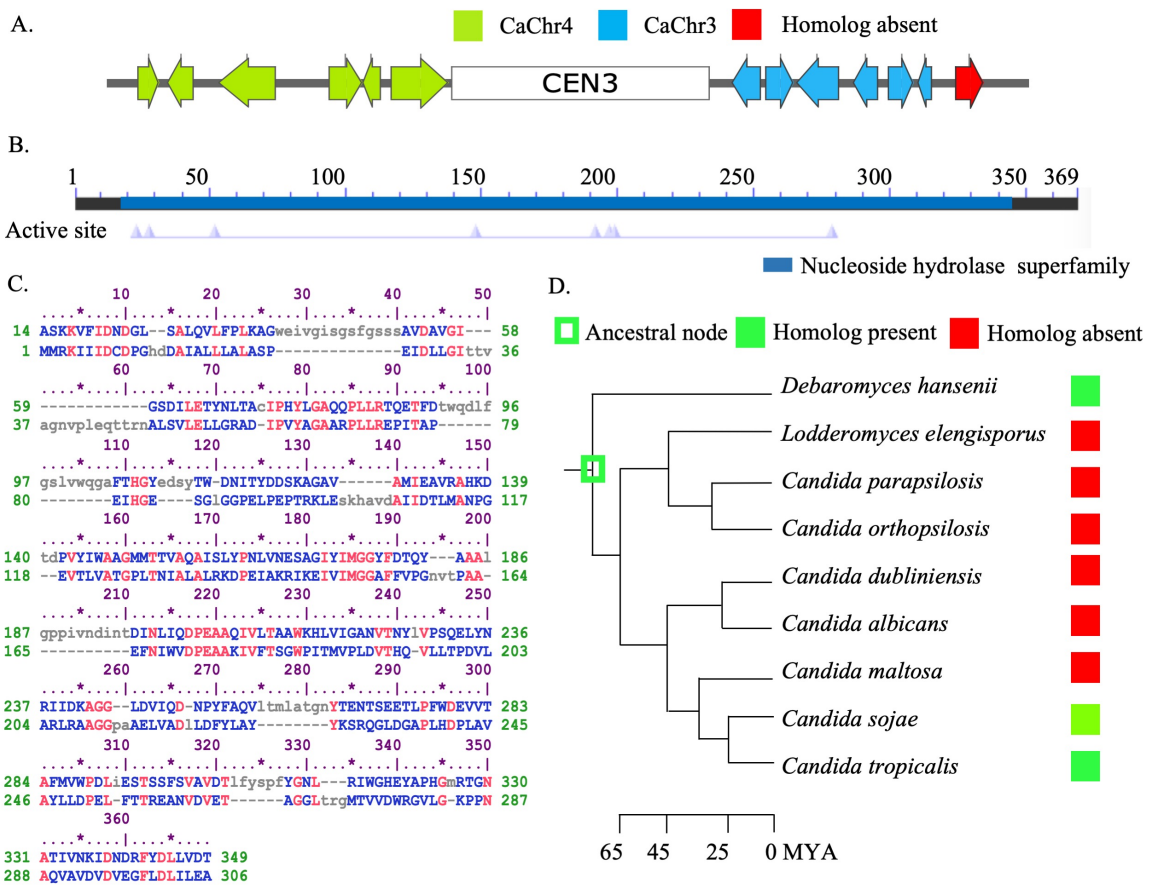


Figure 4.12 Centromere-proximal interchromosomal rearrangements leads to loss of inosine-uridine nucleoside N-ribohydrolase homolog.

A. Schematic of genes flanking centromere on CtChr3. The ORFs are color-coded based on the location of their orthologs on *C. albicans* chromosomes. The species-specific Inosine-uridine nucleoside N-ribohydrolase gene in *C. tropicalis* is shown in red. B. Line diagram of the polypeptide with the annotation of Nucleoside hydrolse superfamily domain obtained from conserved domain database. C. Pairwise sequence alignment of Inosine-uridine nucleoside N-ribohydrolase homolog (*top*) with cdd:COG1957 (PSSMID 224868) (*bottom*) as generated by conserved domain database. D. Distribution of this gene in closely related species complex. The phylogeny is adapted from reference (362) and the time scale was adapted from reference (4).

Do these extensive translocations lead to the loss of genes in the derived species? To answer this question, we scanned the centromere-proximal regions for the presence of ORFs unique to *C. tropicalis.* In this analysis, two unique ORFs proximal to the centromere, in CtChr3 and CtChr5 (Figure 4.10A), were detected. The unique ORF proximal to the *CEN3* of *C. tropicalis* (Figure 4.12A) is 1107 bp long and encodes for a 368 amino acid long protein containing a conserved nucleoside hydrolase domain (CDD ID: cl00226) (Figure 4.12B - C). BLAST search against *C. albicans* proteins failed to identify any ortholog. A more comprehensive search among other closely related species detected its homolog in *Debaromyces hansenii* and *C. sojae,* but not in the intermediate species (Figure 4.12D). This observation exemplifies the role of centromere mediated genomic rearrangements in loss or gain of species-specific genes creating variation among the majorly clonally propagated members of the CUG-Ser1 clade.

Chapter 5

Discussion

In this study, we improved the current genome assembly of the human fungal pathogen *C. tropicalis* by employing SMRT-seq, 3C-seq, and chromoblot experiments, and present Assembly2020, the first chromosome-level gapless genome assembly of this organism. We further identified three large-scale duplication events and a heterozygous balanced translocation in its genome, phased the diploid genome of *C. tropicalis*, and mapped SNPs and indels. We constructed a genome-wide chromatin contact map and identified significant centromere-centromere as well as telomere-telomere spatial interactions. Comparative genome analysis between *C. albicans* and *C. tropicalis* reveals that six out of seven centromeres of *C. tropicalis* are mapped precisely at or proximal to ICSBs. Strikingly, ORFs proximal to the centromeres of *C. tropicalis* are converged into specific regions on the *C. albicans* genome, suggesting that inter-centromeric translocations may have occurred in their common ancestor. Moreover, the presence of HIR-associated putative centromeres in *C. sojae* and *C. viswanathii,* like in *C. tropicalis*, suggests that such a centromere structure is plausibly the ancestral form in the CUG-Ser1 clade but lost both in *C. albicans* and *C. dubliniensis.* We propose that loss of such a centromere structure might have occurred during translocation events involving centromeres of homologous DNA sequences in the common ancestor, to give rise to ENCs on unique DNA sequences and facilitated speciation.

Unlike other centromeres, *CEN6* of *C. tropicalis* did not seem to undergo inter-centromeric translocations. A closer analysis revealed that three *CEN6*-associated ORFs of *C. tropicalis* are absent in the *C. albicans* genome while the other flanking ORFs remain conserved. This observation can be explained by a double-stranded DNA break at the centromere followed by the fusion of broken ends resulting in the loss of those ORFs.

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 this organism including the cause of antifungal drug resistance. Besides, this study sheds light on how genetic elements required for *de novo* centromere establishment in an ancestral species could be lost in the derived lineages to give rise to epigenetically-regulated centromeres.

C. tropicalis is a human pathogenic ascomycete, closely related to the well-studied model fungal pathogen *C. albicans* (363). These two species diverged from their common

ancestor ~39 million years ago (226) and evolved with distinct karyotypes (193), having different phenotypic traits (364), and ecological niches (365). 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 (43), Pakistan (44), and Brazil (366). Moreover, the occurrence of drug resistance, particularly multidrug resistance, in *C. tropicalis* is on the rise (43, 367, 368). Therefore, relatively less-studied *C. tropicalis* is emerging as a major threat for nosocomial candidemia with 29-72% broad spectrum mortality rate (369). Fluconazole resistance in *C. albicans* can be gained due to segmental aneuploidy of Chr5 containing long IRs at the centromere, by the formation of isochromosomes (25), which was also identified in Chr4 with IRs at its centromere (370). All seven centromeres in *C. tropicalis* are associated with long IRs with the potential to form isochromosomes.

Genetic variability is absolutely necessary for the successful survival of a pathogen. In the absence of true meiosis in majorly clonally propagated *Candida* species, the chance of appearance of new alleles is lower than other species with true meiosis. However, due to the highly plastic nature of its genome, mitotic recombination in *C. albicans* can lead to karyotypic changes (371) in addition to whole chromosome aneuploidy, or segmental aneuploidy, which confer specific selection advantage such as resistance against antifungal drugs (20). Such large scale chromosomal changes are often associated with a diseased state if not fatal in metazoans (372). Paradoxically, extensive genomic plasticity allows the human fungal pathogen *C. albicans* to thrive in challenging environments inside the host. The existence of such mechanisms in *C. tropicalis* is largely unknown. Identification of the long-duplicated regions, balanced heterozygous translocation, long track LOH, and recovery of Chr5 monosomic strains of *C. tropicalis* in this study provide evidence that *C. tropicalis* can tolerate considerable genomic changes. We also found that an increase in genomic copy number can lead to an increased expression of genes located on those loci. Moreover, we demonstrate that the increased copy number of *DUPR* locus confers extensive fluconazole resistance in MYA-3404 strain of *C. tropicalis.* What are the other genomic changes driving drug resistance in the clinical isolates of *C. tropicalis*? Our chromosome-level assembly of *C. tropicalis* can now be used to perform genome-wide association studies to understand the genomic alterations responsible for the emergence of drug resistance.

Since the mechanism of homology search during HR is positively influenced by spatial proximity and the extent of DNA sequence homology (321, 373), at least in the engineered model systems, it is expected that spatially clustered homologous DNA sequences undergo more translocation events 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, the bipolar to the tetrapolar transition of the mating type locus in the *Cryptococcus* species complex was associated with inter-centromeric recombination following pericentric inversion (206). Similar inter-centromeric recombination has been reported in the common ancestor of two fission yeast species, *S. cryophilus* and *S. octosporus* (205). These examples raise an intriguing notion that centromeres serve as sites of recombination, which may lead to centromere loss and/or the emergence of ENCs. This notion is supported by the fact that DSBs at centromeres following fusion of the acentric fragments to other chromosomes led to chromosome number reduction in *Ashbya* species (178) and *Malassezia* species (201). Genomic instability at the centromere can also lead to fluconazole resistance, as in the case of isochromosome formation on Chr5 of *C. albicans* (25). Additionally, breaks at the centromeres were reported to be associated with cancers in humans (374).

What would be the consequence of the spatial proximity of chromosomal regions with high DNA sequence homology in other domains of life? Interchromosomal contacts between chromosome pairs have been correlated with the number of translocation events in both naturally occurring populations and experimentally induced mammalian cells (375-384). It has been suggested that contacts between various chromosomal territories, as well as their relative positions in the nucleus, influence the sites and frequency of translocation events both in flies and mammals (147, 379, 385-388). While centromeres remained clustered either throughout the cell cycle or most parts of it in many fungal species, such is not the case in metazoan cells. Nevertheless, one of the well-studied translocation events, Robertsonian translocation (RT) involving fusion between arms of two different chromosomes near a centromere, is the most frequently detected chromosomal abnormality in humans (389). The occurrence of RT was first reported in grasshoppers (390), and subsequently, it has been implicated in the karyotype evolution in humans (389), mice (391, 392), and wheat (393). Moreover, RTs cause sterility in humans (394), often linked with the heterogeneity of carcinomas (300), and implicated in genetic disorders (395). Intriguingly, cytological and Hi-C based evidence (112) of spatial proximity (reviewed in (170)) among the repeat-associated

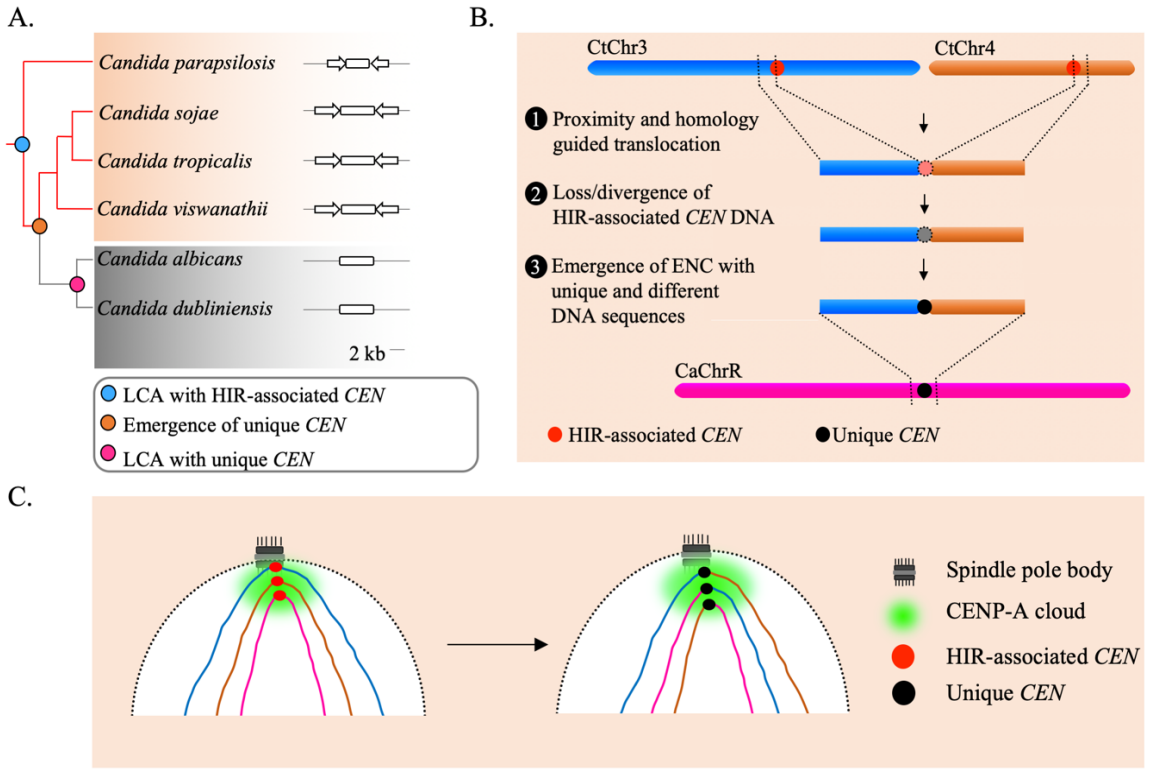


Figure 5.1 The spatial genome organization remained conserved in the CUG-Ser1 clade despite centromere type diversity.
A. A maximum likelihood-based phylogenetic tree of closely related CUG-Ser1 species analyzed in this study. Color-coded branches show presence (red) or absence (gray) of HIRs. The centromere structure of each species is shown and drawn to scale. B. A model showing possible events during the loss of HIR-associated centromeres and emergence of the unique centromere type through inter-centromeric translocations possibly occurred in the common ancestor of *C. tropicalis* and *C. albicans*. The model is drawn to show translocation events involving two *C. tropicalis* chromosomes (CtChr3 and CtChr4) as representatives, which can be mapped proximal to the centromere on *C. albicans* ChrR (CaChrR) as shown in Figure 4.11C. C. Rabl-like chromosomal conformation is maintained despite inter-centromeric translocations that facilitated centromere type transition.

---

centromere DNA sequences (396) in these species supports a possibility that RTs may have been guided by spatial proximity. Similarly, chromoplexy, involving a series of translocation events among multiple chromosomes without alterations in the copy number, was identified in prostate cancers (397, 398). Although fine mapping of translocation events at the repetitive regions in human cancer cells is challenging, the growing evidence that such events are associated with the formation of micronuclei (399) supports the idea that the spatial genome organization may influence chromoplexy as well (400).

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 (196) (Figure 5.1A**).** 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 but lack IR-motifs. Besides, *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-specific centromere function (193). Therefore, the loss of HIR-associated centromeres in *C. albicans* that are only epigenetically propagated (224) clearly shows how the ability of *de novo* establishment of kinetochore assembly in an ancestral lineage can be lost in a derived lineage. However, the mechanism through which IR-motifs may regulate centromere identity remains to be explored.

Loss of HIR-associated centromeres during inter-centromeric translocations must have been catastrophic for the cell, and the survivor was obligated 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 (241, 242), which then becomes part of the centromere cluster (281). This evidence supports the existence of a spatial determinant, known as the CENP-A cloud or CENP-A-rich zone (241, 401), influencing the preferential formation of neocentromere at loci proximal to the native centromere (241, 402). 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 centromere cluster. However, new centromeres of *C. albicans* are formed on loci with completely unique and different DNA sequences. Similar to centromeres of *C. albicans,* centromere repositioning events may lead to the formation of ENCs, which are often associated with speciation in mammals (315, 316). It was found that the location of one centromere in horse varies across individuals (317, 318). Although, there are cases where ENCs formed without genomic rearrangements, the driving force facilitating centromere relocation was proposed to be associated with chromosomal inversion and translocation in certain cases (319). Because of these reasons, it may be logical to consider the centromeres of *C. albicans* as ENCs (Figure 5.1B). Intriguingly, even after the catastrophic chromosomal rearrangements, the ENCs in *C. albicans* remain clustered similar

to *C. tropicalis* (Figure 5.1C). This observation identifies spatial clustering of centromeres as a matter of cardinal importance for the fungal genome organization.

Chapter 6

Materials and methods

Media, growth conditions and transformation

C. tropicalis and *C. sojae* strains (Genotypes are mentioned in Appendix-I) used in this study were grown in non-selective YPDU (2% dextrose, 2% peptone, 1% yeast extract, and 0.01% uracil), and incubated at 30°C at 180 rpm. For growing *C. albicans* strains, YPD media was supplemented with 0.1 mg/mL of uridine. The transformation of *C. tropicalis* was performed as described previously (193). The selection of transformants was based on prototrophy for the metabolic markers used. In the case of selection for the antibiotic marker (*CaSAT1*), conferring nourseothricin (NTC) resistance, growth media was supplemented with 100 µg/mL NTC (NTC; Werner Bioagents, CAS No. 96736-11-7). Recycling of the *CaSAT1* marker was done by growing the NTC resistant strains in YPMU (4% maltose, 2% peptone, 1% yeast extract, and 0.01% uracil) and segregants which are NTC sensitive were selected by patching them on YPDU and YPDU supplemented with NTC. For counter selection against *CaURA3*, the 5-Fluoroorotic Acid (5-FOA; Sigma-Aldrich, CAS No. 207291-81-4) was used at 1 mg/mL concentration. Transformation of *C. tropicalis* was performed using the lithium acetate mediated transformation technique, as described previously (193).

Construction and confirmation of strains and plasmids

a. Construction and confirmation of *C. tropicalis* strains (CtKS101 and CtKS102) expressing Protein-A tagged CENP-A^{Cse4}

To tag CENP-A^{Cse4} with TAP, an overlap PCR strategy was employed (Figure 6.1A). The CENP-A^{CSE4} ORF and its downstream sequence were PCR amplified. The Protein-A epitope (105) with Ca*URA3* fragment was PCR amplified from pPK335 (106). Using equimolar mixture of these fragments as template, an overlap PCR was setup to amplify a 3 kb CSE4-TAP-URA-DS cassette (Figure 6.1B). The cassette was transformed in CtKS06 strain and transformants were selected on CM-URA plates. The positive transformants were confirmed by both PCR and western blot analysis (Figure 6.1C - D). A cassette to TAP tag CENP-A^{Cse4} using *CaHIS1* marker was constructed as follows. First, *CaHIS1* was cloned into the EcoRI digested pBS to generate pBS-HIS. Then, CENP-A^{CSE4} ORF with tagged TAP epitope was amplified from CtKS101. The amplified fragment was digested by NotI and SpeI and cloned into respective sites of pBS-HIS to generate pCENP-A-TAP-HIS. The plasmid was then linearized by BstBI and transformed in CtKS06 to generate CtKG001. The positive integrant

was confirmed by PCR. Genotype of each strain is mentioned in Appendix-I and the primers used to construct and confirm these strains are mentioned in Appendix-II.

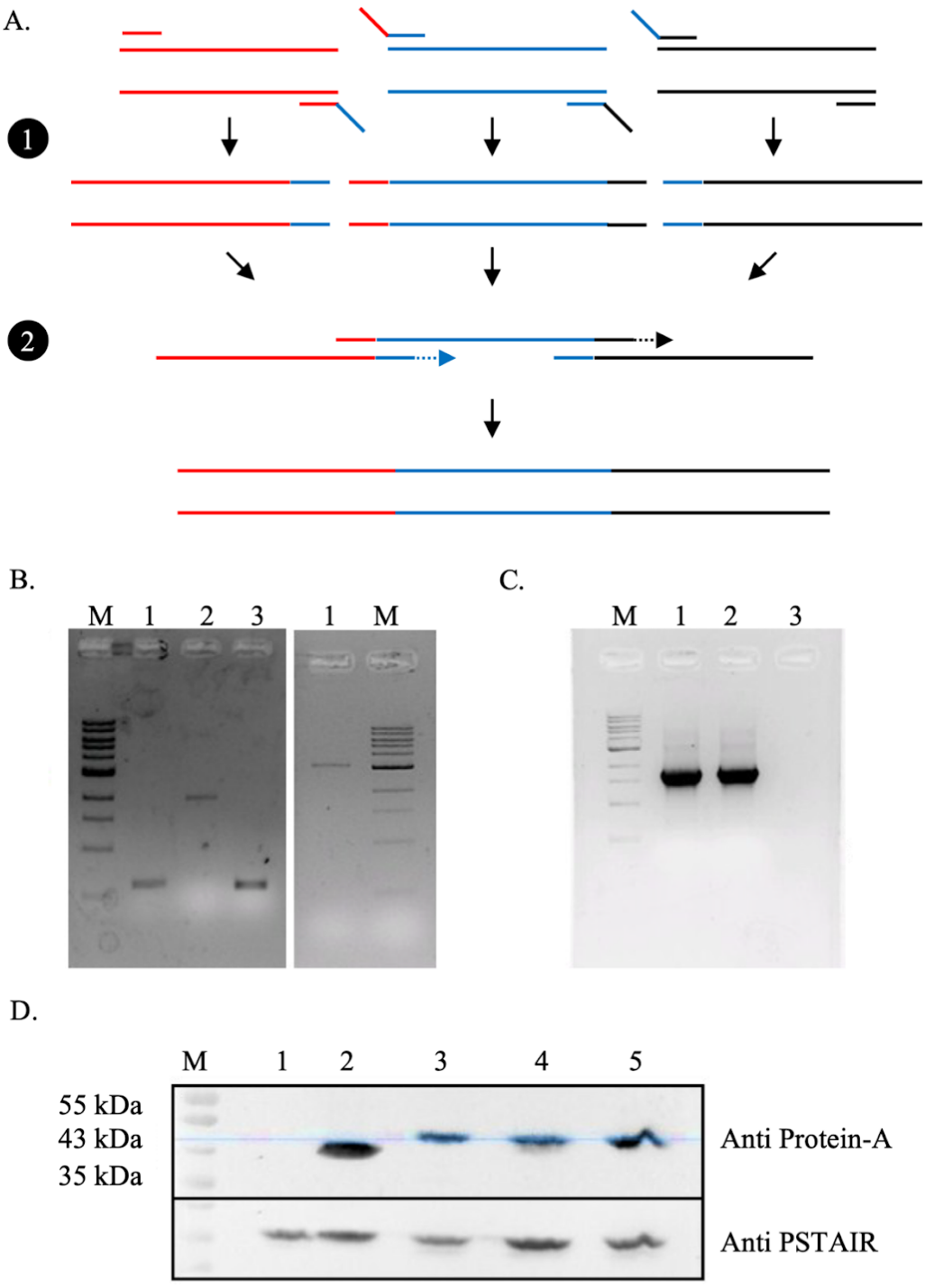


Figure 6.1 Construction and confirmation of CENP-A^{Cse4}-Protein-A tagged strain (CtKS101) of *C. tropicalis*.

A. Schematic drawing of overlap-PCR strategy for joining three fragments to construct a CENP-A^{CSE4}-Protein-A construct. This process involves two key steps: amplification of individual fragments with terminal homology (supplied by the primers) to the adjoining fragments (1) and an overlap reaction, in which the entire construct is synthesized from the

individual fragments (107). B. *Left*, ethidium bromide stained gel picture showing individual fragments amplified from CENP-A^{CSE4} C-terminus (lane 1), TAP-URA fragment (lane 2), and CENP-A^{CSE4} downstream fragment (lane 3). *Right*, ethidium bromide stained gel picture showing the overlap PCR product constructed from three individual fragments. C. Ethidium bromide stained gel picture showing PCR confirmation of the transformants obtained after transformation of the overlap PCR product in *C. tropicalis.* D. Western blots probed with anti-Protein-A (top) and anti-PSTAIR (bottom) antibody for confirmation of CENP-A^{Cse4}-Protein-A tagged transformants. Primers used in this experiment are mentioned in Appendix-II.

---

b. Construction of *C. tropicalis* strains expressing Nuf2-GFP (CtKG500) and CENP-C^{Mif2}-GFP (CtKG501)

NUF2-GFP (CtKG500): A 695 bp fragment of *NUF2* ORF was PCR amplified and cloned into SacII and SpeI sites of pGFP-HIS to obtain the plasmid pNUF2-ORF. A sequence downstream of *NUF2* (664 bp) was cloned into ApaI and KpnI sites of pNUF2-ORF to generate the plasmid pNUF2-GFP. The plasmid was digested with SacII and KpnI to release the cassette and then transformed in CtKS06 to obtain the strain CtKG500.

CENP-C^{Mif2}-GFP (CtKG501): A 486 bp fragment of the *MIF2* ORF was amplified and cloned into SacII and SpeI sites of pGFP-HIS to generate the plasmid pCENP-C-GFP. The plasmid was linearized by PmeI and transformed in CtKS06 to get the strain CtKG501. The primers used for generation of these strains are mentioned in Appendix-II. Genotype of each strain is mentioned in Appendix-I.

c. Construction of *dup4* and *dupR* mutant strains

The *dup4* (CtKG300S1, CtKG300S2, and CtKG300S3) and *dupR* (CtKG400S1, CtKG400S2, and CtKG400S3) mutant strains were constructed by double homologous recombination mediated replacement of the entire length of *DUP4* (~235 kb) and *DUPR* (~80 kb) by *CaSAT1* marker in the MYA-3404 strain. The constructs for replacing *DUP4* and *DUPR* loci were constructed using overlap PCR strategy, analogous to what was described in figure 6.1A. Approximately ~1 µg of these constructs were transformed in MYA-3404 strain and the transformants were selected on YPDU+NTC plate. These transformants were subjected to PCR analysis to conform if the target locus is replaced by *CaSAT1* marker (Figure 6.2). Genotype of each strain is mentioned in Appendix-I. The primers used to generate these mutants are listed in Appendix-II.

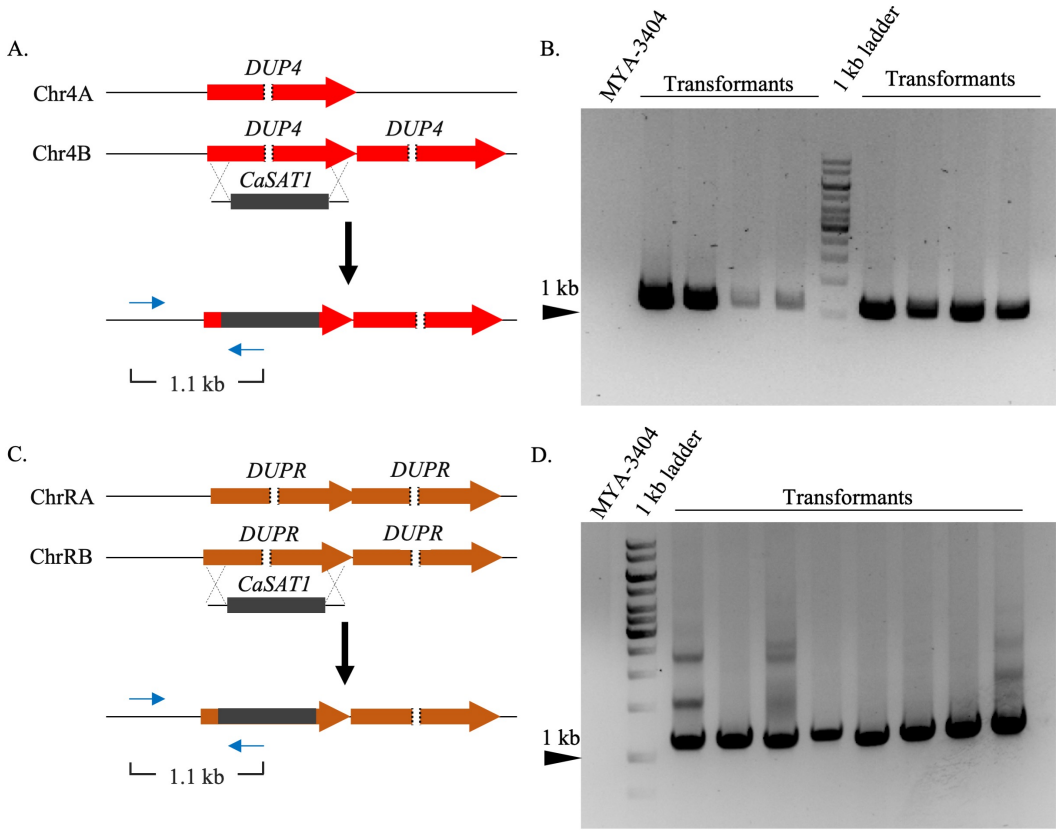


Figure 6.2 Double homologous recombination mediated deletion of CNV loci and PCR confirmation.
(A) and (B) represent schematics for the construction of *DUP4* (red) deletion mutants using *CaSAT1* (black), which confers resistance against nourseothricin. The ethidium bromide stained gel of PCR products confirmed the desired transformants. (C) and (D) show the same for the *DUPR* locus (orange). The relative location and orientation of the primer pairs used for PCR confirmation of the *DUP4* and *DUPR* mutants are shown using blue arrows.

---

d. Construction of pARS2-λ, pCEN501, pCEN502 and pCaCEN5 plasmids

pARS2-λ: Duplex DNA isolated from bacteriophage lambda (cI857ind 1 Sam 7) from New England Biolab (Cat# N3011S) was digested with BamHI and a ~12 kb fragment was cloned in the pARS2 backbone at BamHI restriction site.

pCEN501: To clone CtLR5 into direct orientation with respect to CtRR5, a 4032 bp CtLR5 was PCR amplified using *C. tropicalis* genomic DNA as the template. The PCR product was digested by SalI and PstI and cloned into respective sites of pmid+RR to generate pCEN501. The orientation of CtLR5 in pCEN501 was confirmed by NcoI digestion.

pCEN502: A 2254 bp left repeat of *CEN5* of *C. albicans* (CaLR5) was PCR amplified using *C. albicans* genomic DNA as the template and digested by PstI and SalI. The digested product was cloned between the PstI and SalI sites of pmid8 to generate pmid8-CaLR5. Subsequently, a 2317 bp right repeat of *CaCEN5* (CaRR5) was PCR amplified using *C. albicans* genomic DNA as the template, digested by BamHI and was cloned into the same site of pmid8-CaLR5 to generate pCEN502. The structure of each of these plasmids was verified by digesting with at least two different restriction enzymes.

pCaCEN5: To construct pCaCEN5, entire *CaCEN5* locus (Chr5A:466262-473964) was amplified in two fragments (3692 bp and 4011 bp) using the genomic DNA of MYA-3404 as template. These two fragments were cloned in pARS2 in PstI-SalI and SalI-BamHI restriction sites, respectively. During the process of cloning, a unique SalI site was introduced between these two fragments, which is not present in native *CaCEN5* DNA. This pCaCEN5 specific SalI site was used to design pCaCEN5 specific primers and distinguish it from the native *CaCEN5*. All primer pairs used to construct these plasmids are listed in Appendix-II. Plasmid constructs are listed in Appendix-III. We acknowledge the contribution of Dr. Gautam Chatterjee and Sundar Ram Shankaranarayanan for construction of pARS2-λ, pCEN501 and pCEN502 plasmids.

Cell lysate preparation and western blot

Whole cell protein lysates for western blot were prepared by the trichloroacetic acid (TCA) precipitation method (403). From overnight grown cultures, 3 OD^{600} equivalent cells were harvested, washed and resuspended in 400 µL of 12.5% ice cold TCA solution. The suspension was vortexed briefly and stored at -20°C for 12 h. The suspension was thawed on ice, pelleted at 14000 rpm for 10 min and washed twice with 500 µL of 80% acetone (ice cold). The washed pellets were air dried completely and resuspended in desired volume of lysis buffer (0.1 N NaOH, 1% SDS).

C. tropicalis cell lysates were electrophoresed on SDS-PAGE and blotted onto nitrocellulose membrane in a semi-dry apparatus (Bio-Rad). The blotted membranes were blocked by 5% skim milk containing PBS (pH 7.4) for 2 h at room temperature and then, were incubated with following dilutions of primary antibodies for western blot analysis: anti-Protein A antibodies (Sigma, Cat# P3775) 1:5000 and anti-PSTAIRE antibodies (Abcam, Cat No.

ab10345) 1:2000, for 3 h at room temperature. Next, the membranes were washed thrice with PBST (0.1% Tween-20 in PBS) solution. Anti-rabbit HRP conjugated antibodies (Bangalore Genei, Cat No. 105499) (for Protein A western blots) 1:5000 and anti-mouse HRP conjugated antibody (Bangalore Genei, Cat No. 105502) (for PSTAIRE western blot) 1:5000 dilution in 2.5% skim milk powder in 1X PBS were added and incubated for 2 h at room temperature followed by three to four washes with PBST solution. Signals were detected using the chemiluminiscence method (SuperSignal West Pico Chemiliminescent substrate, Thermo scientific, Cat# No. 34080) and imaged using Syngene G-Box gel doc system.

Indirect immunofluorescence microscopy

Subcellular localization of Protein-A tagged CENP-A^{Cse4} with DAPI (4, 6-Diamino-2-phenylindole) stained nuclear mass was performed in CtKG001 following the method described previously for *C. albicans* (38). Asynchronously grown CtKG001 cells were fixed with the 1/10th volume of formaldehyde (37%) for 1 h at room temperature. Antibodies used were diluted as follows: 1:1000 for rabbit anti-Protein A (Sigma, Cat No. P3775). The dilutions for secondary antibodies used were Alexa flour 568 goat anti-rabbit IgG (Invitrogen, Cat No. A11011) 1:1000. Cells were observed using a DeltaVision imaging system (GE Healthcare Life Sciences), and the images were processed using FIJI software (37).

Pulsed-field gel electrophoresis

C. tropicalis strain MYA-3404 and *C. albicans* strain SC5314 were grown until the exponential phase (~2×10^7 cells/mL). Cells were washed with 50 mM EDTA and counted with a hemocytometer. Approximately 6×10^8 cells were used for the preparation of 1 mL genomic DNA plugs. The plugs were made according to the instruction manual protocol (Bio-Rad, Cat No. 170–3593) with CleanCut Agarose (0.6%) and the lyticase enzyme provided by the kit. A 0.6% pulsed field certified agarose gel was prepared using 0.5x TBE buffer (0.1 M Tris, 0.09 M Boric acid, 0.01 M EDTA, pH 8.0) and PFGE was performed on contour-clamped homogeneous electric field (CHEF) system using CHEF-DR II (Bio-Rad) module. The running conditions used were as follows: block-I at 100-200 s for 24 h at 4.5 V/cm/120°, block-II at 200-400 s for 48 h at 2.5 V/cm/120°, block-III at 600-800 s for 120 h at 2.5 V/cm/120°. The gel was stained with ethidium bromide (EtBr) and analyzed by Quantity One software (Bio-Rad).

Preparation of high molecular weight genomic DNA

Briefly, 50 OD_{600} equivalent (1 OD_{600} = ~2×10^7 cells) cells were collected, washed with chilled 50 mM EDTA pH 8.0 and flash-frozen with liquid nitrogen. Next, the cell pellet was lyophilized. Then a volume equivalent to 5 mL of glass beads was added to the tube and vortexed till the pellet turns powdery. Then 20 mL Cetyltrimethyl ammonium bromide (CTAB) extraction buffer (100 mM Tris-HCl pH 7.5, 0.7 M NaCl, 10 mM EDTA, 1% CTAB powder, 1% 2-Mercaptoethanol) was added, and the tube was incubated at 65°C for ~30 min with occasional mixing by inverting the tube. Subsequently the tube was chilled on ice for 10 min, and the supernatant was transferred into another tube. An equal volume of chloroform was mixed with the supernatant gently inverting for 5 to 10 min. The mix was then centrifuged at 3200 rpm for 10 min, and the aqueous phase was carefully pipetted out using cut tips to a fresh tube. An equal volume of isopropanol was added into the supernatant and mixed gently until white thread-like structures appeared. The mix was incubated at −20°C for 1 h and centrifuged at 3200 rpm for 10 min to pellet the DNA. The pellet was washed twice with freshly prepared 70% ethanol and air-dried. The dried pellet was dissolved in 1 mL of 1x TE containing RNase A to a final concentration of 100 µg/mL and incubated at 37°C for 30 to 45 min. Sodium acetate solution was added into the mix to a final concentration of 0.5 M, and the solution was transferred to several 1.5 mL tubes in the aliquots of 0.4 mL each. An equal volume of isopropanol was added to each tube, mixed gently, and centrifuged at 13,000 rpm for 15 min. The supernatant was decanted, and the DNA pellet was washed with 70% ethanol. The pellet was air-dried and finally dissolved in 200 µL of 1x TE buffer. The quality of the isolated DNA was determined by performing PFGE analysis (switching time 1-25 s, at 5.8 V/cm/120° for 24 h, 1% agarose gel) on CHEF-DR II module (Bio-Rad).

SMRT sequencing of *C. tropicalis* strain MYA-3404 on PacBio sequel system

The genomic DNA fragments of ~20 kb length were size-selected and taken forward for library preparation using SMRTbell™ Template Prep Kit (Part No. 100-259-100). PacBio sequencing of the *C. tropicalis* MYA-3404 genome was performed by Sequel SMRT Cell 1M (Part No. 101-008-000) using Sequel™ Binding Kit 2.0 (Part no. 100-862-200) and SMRT Link version 5.0.1.9585. This run generated 996041 reads with an average read length of 5.8 kb.

Construction of the *de novo* SMRT assembly and contig stitching using SMIS

The *de novo* SMRT assembly using 996041 PacBio raw reads was generated using Canu 1.6 (77). The program was run in the trimming and correction mode with the '-pacbio-raw <input.fastq>' option that produced 135 contigs. For stitching the contigs from Assembly B using the PacBio raw reads, we used Single Molecular Integrative Scaffolding (https://github.com/fg6/smis) with the default options, to get a 12-contig assembly (Assembly C). Details of the assemblies produced by Canu and SMIS are presented in Table 2.3**.**

Filling N-gaps

The *de novo* SMRT contigs were used to fill the existing N-gaps in Assembly A. We used 500 bases upstream, and downstream regions of the N-gaps as queries against a custom BLAST (404) database generated using Geneious® software from the *de novo* assembled contigs and filled these N-gaps upon the mapping of upstream and downstream query sequences on the same contig with 100% coverage and more than 95% identity. Using this approach, we filled 78 out of 104 gaps leaving 26 gaps on seven chromosomes (Table 2.2, Figure 2.6A). We suspected that the remaining gaps were repetitive regions in the genome as immediate flanking regions identified multiple hits. To avoid this, we used a second strategy in which we used a 1-kb query sequence from either 10 kb upstream or downstream region of the N-gap, and performed a BLAST analysis against the *de novo* contigs generated using FALCON (78). All the remaining 26 gaps could be filled using this strategy (Table 2.2, Figure 2.6B). Further, to validate our claim, we confirmed the mapping of the Illumina and PacBio reads over the newly filled sequence.

Assembly of sub-telomeric regions

To assemble the sub-telomere regions, we performed a BLAST search using the terminal 5000 bp sequence of each chromosome as queries against the *de novo* SMRT contigs and identified the contigs containing the 23-bp telomeric repeats specific for *C. tropicalis* (5′-TGATCGTGACATCCTTACACCAA-3′) as reported previously (6). Schematic of the sub-telomere scaffolding has been shown (Figure 2.6C).

Mapping of the orphan haplotigs using the *de novo* SMRT assembly

Canu is a diploid-aware genome assembler (77), which generates two contigs from a heterozygous locus. Therefore, we used the Canu generated contigs (SMRT assembly) to map the orphan haplotigs as heterozygous regions of the genome (see Figure 2.4C). Heterozygosity of the orphan haplotigs was demonstrated by the Illumina read coverage (Figure 2.4B). For this analysis, the 3C-seq reads were mapped on the OHs and a control locus of Chr1 using Bowtie2 (332). The number of mapped reads were counted using the bamCoverage utility from deepTools2 (333) and plotted using boxplotR (405).

Pilon polishing of the genome assembly

The final telomere-to-telomere assembled chromosomes were polished through Pilon (324) using the Illumina reads obtained from the 3C-seq experiment. Pilon corrected base-pair level assembly errors and validated 99.5-99.8% bases of the seven chromosomes. The polishing step was repeated six times when the improvement stalled.

Construction of aneuploids for confirmation of heterozygosity of the OHs

We constructed *C. tropicalis* strains monosomic for Chr5 and used them to demonstrate that loss of one homolog of Chr5 leads to loss of one of the two alleles of the orphan contigs: contig14 and contig16, that are mapped on Chr5. Since the *sch9* mutants in *C. albicans* were viable but lost chromosomes at a significantly higher rate than the wild-type (327), we adopted the same strategy to delete both copies of *SCH9* homologs in *C. tropicalis.* Next, a reporter strain was created in this *sch9* mutant strain background of *C. tropicalis* to assay for loss of a Chr5 homolog. These strains (2n-1) that lacked one homolog of Chr5 were used to confirm the presence of heterozygosity of orphan haplotigs (OHs) of CtChr5.

a. Deletion of *SCH9* in *C. tropicalis*

The *SCH9* homolog in *C. tropicalis* was identified in a BLAST search using *CaSCH9* as the query sequence against the *C. tropicalis* proteome. A putative homolog of *SCH9* was located on Chr1:1994521-1996662 and encoded by the Crick strand. A deletion cassette (pKG1) for double homologous recombination-mediated deletion of *SCH9* ORF was constructed by cloning upstream and downstream homology regions in pSFS2a plasmid (328). This construct was transformed into CtKS102 for the deletion of both copies of *SCH9* ORF by recycling the *CaSAT1* marker after the deletion of the first copy of *SCH9* gene. Independent

colonies of the *sch9*/*sch9* null mutant strain (CtKG001) were confirmed using Southern hybridization (Figure 2.5B). Primers used in this study are mentioned in Appendix-II.

b. Construction of a reporter strain by integration of *URA3* on Chr5 for isolation of Chr5 monosomic isolates

Upstream and downstream homology regions of the target intergenic locus (Chr5_497_kb) in Chr5 were amplified, and cloned into pBSCaURA3 plasmid (193) to construct pKG2 (Appendix-III). This cassette was released by restriction digestion with BamHI and ApaI and transformed into the *sch9* mutant strain CtKG001 to construct the reporter strain CtKG002. Similarly, we integrated *CaURA3* into the target intergenic locus (Chr5_497_kb) of CtKS102 to create a control strain CtKG003. In both the strains (CtKG002 and CtKG003) the short arm (5′ end) of one of the two homologs is marked with *CaURA3* marker and the long arm (3′ end) carries the heterozygous *MTL* locus (*MTL***a** or *MTL*α) with two distinct alleles present on two homologs. Concomitant loss of one of the *MTL* alleles together with *CaURA3* marker would indicate loss of one homolog of Chr5.

c. Isolation and confirmation of the 2n-1 aneuploids for Chr5

Different cell numbers (10^5, 10^4, 10^3, and 10^2) of the reporter strain (CtKG002) and the wild-type control strain (CtKG003) were plated on complete media (CM) + 5-FOA and incubated for 48-72 h at 30°C. Multiple FOA^R colonies appeared for CtKG002 strain but no colonies appeared for the control strain CtKG003. The colonies were then patched on YPDU and CM-URA plates for confirmation of the loss of the *CaURA3* marker. Next, PCR was performed to confirm the loss of one of the *MTL* loci (*MTL***a** or *MTL*α) in these colonies using a multiplex PCR strategy described previously (Figure 2.5C) (13).

Library preparation and sequencing of the library DNA for chromosome conformation capture (3C-seq)

Wild-type *C. tropicalis* strain MYA-3404 was cultured in non-selective YPDU media and 500 OD_{600} equivalent cells were harvested for crosslinking. The cells were cross-linked with formaldehyde to a final concentration of 1.5% for 10 min and the cross-linking reaction was quenched by adding glycine to a final concentration of 400 mM. The crosslinked cells were centrifuged and the cell pellet was stored at -80°C till further use. For making the 3C library

of *C. tropicalis*, the cross-linked cell pellet was first resuspended in 5 mL of ice-cold 1x NEBuffer™ DpnII (50 mM Bis-Tris-HCl, 100 mM NaCl, 10 mM MgCl2, 1 mM DTT; pH 6 @ 25°C) and then lysed by liquid nitrogen grinding in a chilled mortar using a pestle to a fine powder. The powdered sample was scraped using a spatula into a pre-chilled tube and resuspended in 15 mL cold 1x NEBuffer™ DpnII. Cell lysate containing ~3×10^8 cells (4 mL) was processed for 3C library preparation. This lysate was centrifuged and the pellet was resuspended in 1.5 mL of cold 1x NEBuffer™ DpnII and then aliquoted equally into four 1.5 mL microcentrifuge tubes. Next, the chromatin was solubilized by adding SDS to a final concentration of 0.1% in each microcentrifuge tube and the sample was incubated at 65°C for exactly 10 min. The reaction was quenched by adding 45 µL of 10% Triton X-100 per tube with gentle mixing. Chromatin was then digested with 750 units of DpnII (NEB, Cat No. R0543M; 50,000 units/mL) per tube and incubated at 37°C overnight with gentle agitation (300 rpm) on a heating block. Next day, the restriction enzyme was heat-inactivated at 65°C for 20 min. The digested chromatin fraction in each tube was ligated with 50 U of T4 DNA ligase (Invitrogen Cat No.15224090; 1 U/µL) at 16°C for 6 h in a diluted condition (reaction volume 8 mL) to favor intra-molecular ligation of cross-linked restriction fragments. Reverse cross-linking was performed by adding 100 µL of 10 mg/mL Proteinase K (Invitrogen Cat No.25530031) per tube and incubating at 65°C overnight. Next, DNA, which constitutes the 3C library, was purified using conventional phenol-chloroform extraction and concentrated using Amicon Ultra-0.5 mL 30K centrifugal filters. About 1 µg of 3C library was used for size selection using Agencourt AMPure XP beads (Beckman Coulter) to select DNA fragments of 500-700 bp in length. The paired-end NGS library was prepared using NEBNext Ultra II kit, and sequencing was carried out using the Illumina HiSeq 2500 2×101 bp platform by a third party service provider.

3C-seq data analysis

a. Mapping of 3C-seq data, contact probability matrix generation, and visualization

The 3C-seq data was analysed using Juicer in a CPU based system as described before (348). The data was aligned on the Assembly2020 and contact probability matrix was generated at 10 kb resolution. The matrix was visualized using Juicebox (350). Analysis of 3C-seq data using Homer (349) was performed as described before. The contact probability matrix was generated at 5 kb resolution and visualized using Java TreeView software (351). Outline of the script used is presented in Appendix-IV.

b. Contig scaffolding

3C-seq reads were aligned to contig sequences and contact probability matrix was generated as described above. The 3C profile of a bin was plotted using values in a single row from the contact probability matrix. It is well-established that contact frequency generally shows a distance-dependent decay (97). Therefore, the connectivity between two contigs can be inferred by investigating the contact probabilities between the terminal bin of a contig and loci on the other contig. We acknowledge the contribution of Yao Chen from Dr. Amartya Sanyal's laboratory in School of Biological Sciences, Nanyang Technological University, for performing this analysis.

Identification of SNPs, indels and CNVs

a. Detection of SNPs and indels

The SNPs and indels were identified using GATK software (85) with the paired-end Illumina reads obtained from the 3C-seq experiment in a 12 cores Ubuntu 16.4 system with 96 GB memory. Briefly, the 3C-seq reads were mapped to Assembly2020 using Bwa-mem (406) paired-end alignment mode following sorting of the resulting SAM file with Picard (https://broadinstitute.github.io/picard/), SAM to BAM conversion using SAMtools (407), and duplicate marking using 'MarkDuplicates' utility of Picard. Next, we used GenomeAnalysisTK.jar (version 3.8.0) to call the variants with '-ploidy 2' option, SNPs were extracted, filtered with '--filterExpression 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0' --filterName "basic_snp_filter"' option following base quality score recalibration. Similarly indels were extracted, filtered with '--filterExpression 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 10.0' --filterName "basic_indel_filter"' option following base quality score recalibration. The data tracks were visualized using IGV (329) and presented using Circa software. The script used for this analysis is presented in Appendix-V.

b. Read coverage analysis for detection of large-scale CNVs

To generate a genome-wide read coverage plot, the 3C-seq reads were mapped to Assembly2020 using Bowtie2 (332) paired-end alignment mode with '--end-to-end' and '--very-sensitive' option. The resultant SAM file was converted to BAM format and sorted using SAMtools (407). Next, the mapped reads were counted using deepTools2 (333)

bamCoverage utility with the BPM normalization method, and the resulting BED file was used for downstream calculations or visualization in IGV (329).

Haplotype analysis

The FALCON, FALCON-Unzip (78), and FALCON-Phase (325) from the pb-assembly suite were run locally in a 12 core Ubuntu 16.4 system with 96 GB memory according to the instruction provided (https://github.com/PacificBiosciences/pb-assembly). The configuration files used for running FALCON, FALCON-Unzip and FALCON-Phase will be available upon request. Briefly, FALCON was run using modified fcrun.cfg with the input option 'pa_DBdust_option=true', and 'pa_fasta_filter_option=streamed-internal-median'. Next, data partitioning was performed with 'pa_DBsplit_option=-x500 -s100' and 'ovlp_DBsplit_option=-x500 -s100', repeat masking was performed using 'pa_HPCTANmask_option = -k18 -h480 -w8 -e.8 -s100', 'pa_HPCREPmask_option = -k18 -h480 -w8 -e.8 -s100', and 'pa_REPmask_code=0,300;0,300;0,300' options. Preassembly was generated using the following parameters: 'genome_size=15000000', 'seed_coverage=20', 'length_cutoff=100', 'pa_HPCdaligner_option=-v -B128 -M24', 'pa_daligner_option=-e.8 -l1000 -k18 -h480 -w8 -s100 -T10', 'falcon_sense_option=--output-multi --min-idt 0.70 --min-cov 2 --max-n-read 1800', 'falcon_sense_greedy=False'. Next, Pread overlapping was performed using 'ovlp_daligner_option=-e.96 -l1000 -k24 -h1024 -w6 -s100', and 'ovlp_HPCdaligner_option=-v -B128 -M24'. Next, the final assembly was generated using 'overlap_filtering_setting=--max-diff 100 --max-cov 100 --min-cov 2' and 'length_cutoff_pr=500'. Next, phasing of haplotypes was performed using FALCON-Unzip and FALCON-Phase as described (https://github.com/PacificBiosciences/pb-assembly). Script used in this analysis is presented in Appendix-VI.

Assessment of the genome assembly completeness using BUSCO

BUSCO (330) version 3.0.2 was run against ascomycota_odb9 database using the following script: 'python ./scripts/run_BUSCO.py -i genome.fasta -o BUSCO_output -l /Path_to_llineage_dir/ -m genome -c 1 -sp candida_tropicalis'.

Oxford Nanopore sequencing of *C. sojae* strain NCYC-2607

High molecular weight genomic DNA was isolated from yeast cells, and the average length of the DNA fragments of the genomic DNA was checked on a CHEF gel using a CHEF-DR

II system (Bio-Rad). Next, the DNA sample was quantified by NanoDrop (ND-1000 Spectrophotometer, NanoDrop Technologies) and Qubit 3 fluorometer (Thermo Fisher Scientific) using dsDNA HS assay kit (Thermo Fisher Scientific, Cat No. Q33230). An appropriate amount of DNA was taken forward for library preparation as per the manufacturer's instructions using reagents included in SQK-LSK109 and EXP-NBD103/EXP-NBD104 kits. DNA samples were then pooled together on a single R9 flow-cell, and sequenced by the MinION system (Oxford Nanopore Technologies). The fragmentation step was skipped to retain the longer fragments. The raw reads were taken forward for base calling using Guppy version 3.1.5. A total of 92320 reads containing 530421800 bp were generated.

Illumina sequencing of *C. sojae* strain NCYC-2607

DNA was quantified by Qubit 3 fluorometer (Thermo Fisher Scientific) using a dsDNA HS assay kit (Thermo Fisher Scientific, Cat No. Q33230). Approximately 100 ng of intact DNA was enzymatically fragmented by targeting 250-500 bp fragment size. The DNA fragments with overhangs resulting from fragmentation were end-filled. The 3′ to 5′ exonuclease activity of end-repair mix removed the 3′ overhangs, and polymerase activity filled in the 5′ overhangs. To the blunt-ended fragments, adenylation was performed by adding a single 'A' nucleotide to the 3′ ends. To the adenylated fragments, loop adapters were ligated and cleaved with uracil-specific excision reagent enzyme. The sample was further purified using AMPure XP beads (Beckman Coulter, Cat No. A63880), and DNA was then enriched by PCR with six cycles using NEBNext Ultra II Q5 master mix (NEB, Cat No. M0544S), Illumina universal primers, and sample-specific indexed Illumina primers. The amplified products were cleaned up by using AMPure XP beads, and the final DNA library was eluted in 15 µL of 0.1x TE buffer. One µL of the library was used to quantify the DNA concentration by Qubit 3 fluorometer using the dsDNA HS reagent. The fragment analysis was performed on Agilent 2100 Bioanalyzer (Agilent, Model G2939B), by loading 1 µL of the library into Agilent DNA 7500 chip. In this experiment, we generated 3501768 paired-end reads of 2×301 bp length.

***De novo* genome assembly of *C. sojae* strain NCYC-2607**

A total of 92320 reads containing 530421800 bp were used for the construction of a *de novo* assembly using Canu (77). Canu was run using default parameters in the trimming and the

correction mode with '-genomeSize <15m>', which produced the genome assembly of *C. sojae* in 42 contigs. Next, to rectify the base-pair level errors, we performed five rounds of polishing of the contigs using Illumina reads with Pilon (324).

Synteny analysis

Genome-wide synteny analysis was performed using Symap (326) with the parameters as (a) Min. dots 3 (minimum number of anchors required to define a synteny block), (b) top N 2 (retain the top N hits for every sequence region, as well as all hits with score at least 80% of the Nth), (c) BLAT args: '-minScore=30 -minIdentity=70 -tileSize=10 -qMask=lower -maxIntron=10000'. The Satsuma synteny and Synchro software were run using default parameters. For the custom approach to map the interchromosomal synteny breakpoints (ICSBs), first, the single-copy orthologs were identified using OrthoFinder (212), then the corresponding genomic coordinates of the ortholog pairs were sorted and the ICSBs were identified. For comparing the FALCON generated contigs with the Assembly2020 chromosomes, the dot-plot between the two assemblies was generated using the default options of Symap version 4.2.

Identification of the putative centromeres in the members of the CUG-Ser1 clade

The putative centromeres of *C. sojae* and *C. viswanathii* were identified as HIR-associated intergenic regions syntenic to centromeres of *C. tropicalis* centromeres. Briefly, the genomic loci in *C. sojae* and *C. viswanathii,* which are syntenic to the centromeres of *C. tropicalis* were scanned for the presence of inverted repeats falling in ORF-free regions using YASS (357) with the default parameters. Pair-wise alignments between seven random genomic loci of ~10 kb length, LR, CC, or RR DNA elements were performed using Clustal Omega (408). Synteny dot-plot analysis for centromere DNA sequences including the flanking ORF-free region in *C. albicans, C. dublininensis* and the HIR sequences of *C. tropicalis, C. sojae, C. viswanathii,* and *C. parapsilosis* was generated using Gepard (358) by running it in the simple mode with default parameters. The IR sequences from centromeres of *C. tropicalis* and the putative centromeres of *C. sojae* and *C. viswanathii* were analysed to identify the presence of conserved motifs using motif discovery tool MEME following the default parameters with 'ZOOPS: zero or one site per sequence' as the motif site distribution algorithm, and maximum motif width set to 12 bp. Next, we scanned for the presence of IR-motifs across the chromosomes including centromere DNA and flanking ORF-free regions in

C. albicans, C. dubliniensis, and putative centromeres of *C. parapsilosis* using FIMO with default parameters (359).

Construction of the phylogenetic tree

The publicly available genomes and the protein fasta files (when available) of *C. albicans* (ASM254v2), *C. dubliniensis* (ASM2694v1), *C. viswanathii* (ASM332773v1), and *C. parapsilosis* (ASM18276v2) were downloaded from NCBI database. The protein fasta files for *C. tropicalis* and *C. sojae* were generated using Augustus *ab initio* protein prediction software and the python script getAnnoFasta.py (340). Because of the partially diploid nature of *C. viswanathii* genome assembly, the duplicated contigs, that carried >100 kb of DNA sequence on another contig, were identified from dot-plot analysis (self) using Symap (326), and excluded from analysis. The protein fasta files were then used as input files for running OrthoFinder V2.3.1 (212). OrthoFinder was run using the default parameters except the -M msa option for the construction of maximum-likelihood trees using MAFFT (213) and FastTree (214). The tree topology was visualized using Evolview (3).

Total RNA isolation

Total RNA was isolated form the *C. tropicalis* yeast cells using Trizol reagent (Ambion, Cat# 15596018). Cells were grown in YPDU to an OD_{600} = 0.5. Approximately 4×10^7 cells were taken for spheroplasting. Cells were pelleted down at 4,000 rpm, were washed with 1 ml of Y1 buffer (2.5 M sorbitol, 0.5 M EDTA, pH 8.0) and were resuspended in 2 ml of Y1 buffer. Approximately 20 mg of lysing enzyme (Sigma, Cat# L1412) and 2 µl of β-Mercaptoethanol were added and spheroplasting was done at 30°C at 70 rpm. After 90% spheroplasting was achieved, spheroplasts were isolated by centrifugation at 1,800 rpm for 5 min. The spheroplasts were lysed in Trizol by vortexing and total RNA was extracted using chloroform (Fisher Scientific Cat# C607SK-1). The RNA was precipitated using equal volume of isopropanol (Merck Millipore, CAS# 67-63-0) and then the RNA pellet was washed in 70% ethanol twice and resuspended in nuclease free water (Thermo Scientific, Cat# AM9937). After Isolation of total RNA, the genomic DNA contamination in the samples was removed by treating with RNAse free DNAse (NEB, Cat# M0303S) as per the manufacturer's protocol. Removal of genomic DNA was confirmed by PCR, in which the 200 bp amplicon was detected in samples before DNAse treatment but not after the treatment even after 35 cycles of amplifications. A schematic of the steps followed is presented (Figure 6.3). Total

RNA samples were subjected to quality analysis using Bioanalyzer (Agilent, Model G2939B) and the only samples with >7 RNA integrity number (RIN) (409) were taken forward for library preparation.

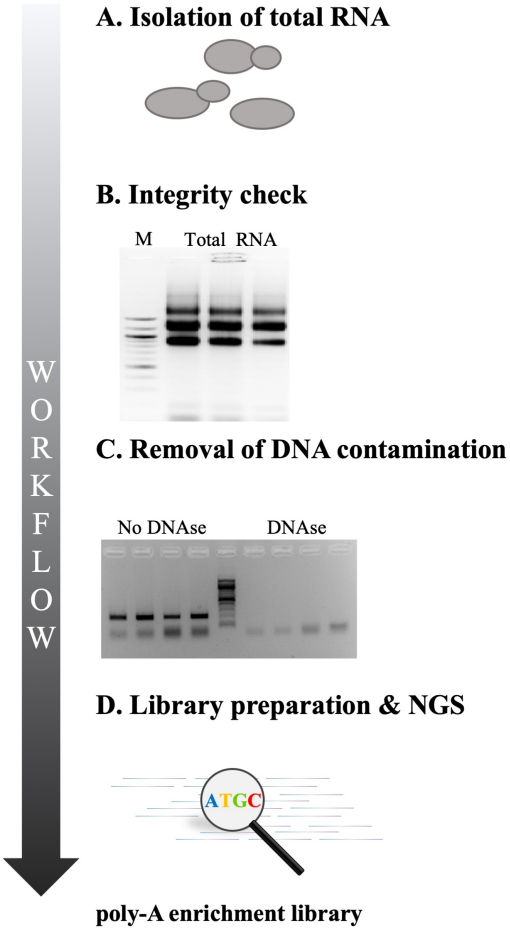


Figure 6.3 Schematic of the steps followed during isolation of DNA-free total RNA from *C. tropicalis* cells

---

Transcriptome analysis

The raw FASTQ files were processes with Trimgalore (102) to remove the sequencing adapters followed by removal of bases below quality score (q) 20 as well as reads shorter than 20 bp. The adapter and quality trimmed FASTQ files were used to map the paired end reads on the genome. For fast and efficient mapping of reads STAR (20) was chosen. The output sequence alignment map (SAM) files from STAR were then converted to binary alignment map (BAM) format and further processed to sort the reads according to the genomic coordinates using Samtools (103). Next, the sorted BAM files were processed with

Picard (104) to identify and flag the duplicates using MarkDuplicate utility. Genome-wide transcript mapping was quantified using deepTools2 (18) bamCoverage utility with bpm normalization option and the resulted browser extensible data (BED) files were used for downstream calculations or visualization in IGV (19). A generalized version of the script used for this analysis is presented in Appendix-VII.

Mitotic stability assay

The mitotic stability assay was performed to determine the loss rate of pARS2, pmid5, pCEN5, pARS2-λ, pCEN501 and pCEN502 in *C. tropicalis*. Briefly, the *C. tropicalis* strain, CtKG001 was transformed with above mentioned plasmids. The transformants were streaked on CM-Ura plates for single colony purification. Single colonies thus obtained were subsequently inoculated in a nonselective media (YPDU) and incubated at 30°C for overnight for 8-10 generations. Next day, equal numbers of cells were simultaneously plated on YPDU and CM-Ura and incubated at 30°C for 2 days. Colonies grown on both plates were counted and the mitotic stability was calculated in percentage as follows: Mitotic stability = (S/NS), where S and NS denote the number of colonies grown on selective and nonselective media respectively (Figure 4.1B).

Chromatin immunoprecipitation

The ChIP assays were done as described previously (108). Briefly, each strain was grown until exponential phase (~2×10^7 cells/mL) and cells were cross-linked with formaldehyde (final concentration 1%). Chromatin was isolated and sonicated to yield an average fragment size of 300–500 bp. Then the DNA was immunoprecipitated with anti-protein A antibodies (Sigma, Cat No. P3775) (final concentration is 24 µg/mL) or anti-GFP antibody (Santa Cruz Biotech, Cat. No. 9996) (final concentration is 4 µg/mL) and purified. The duration of cross-linking varies- 15 min for CENP-A and 1 h 45 min for Nuf2. The total, immunoprecipitated (IP) DNA, and beads only material were used to determine the binding of kinetochore proteins in all seven putative centromeres by semi-quantitative PCR. PCR conditions for primers (as listed in Appendix-II) were used as follows: 94°C for 2 min, T_m for 30 s (T_m varies with the primers), 72°C for 1 min, for 1 cycle; 94°C for 30 s, Tm for 30 s, 72°C for 1 min for 24 cycles in case of CENP-A and 27 cycles for Nuf2; 72°C for 10 min. Enrichment of a given genomic locus was calculated using qPCR analysis following percent input method

and the results were plotted using Graph-pad Prism software. One-way ANOVA was performed to analyze the statistical significance of difference between the samples.

Data access

All the genome sequencing data used in this work and the genome assemblies of *C. tropicalis* (Assembly2020) and *C. sojae* generated in this study have been submitted to NCBI under the BioProject accession number PRJNA596050.

References

1. Turner SA & Butler G (2014) The *Candida* pathogenic species complex. *Cold Spring Harb. Perspect. Med.* 4(9):a019778.
2. Pfaller MA, Diekema DJ, Turnidge JD, Castanheira M, & Jones RN (2019) Twenty years of the SENTRY antifungal surveillance program: results for *Candida* species from 1997–2016. *Open Forum Infect. Dis.*, (Oxford University Press US), pp S79-S94.
3. Subramanian B, Gao S, Lercher MJ, Hu S, & Chen WH (2019) Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47(W1):W270-W275.
4. Shen XX*, et al.* (2018) Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 175(6):1533-1545 e1520.
5. Chowdhary A, Sharma C, & Meis JF (2017) *Candida auris*: a rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally. *PLoS Pathog.* 13(5).
6. Butler G*, et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459(7247):657-662.
7. Ropars J*, et al.* (2018) Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.* 9(1):2253.
8. Wang JM, Bennett RJ, & Anderson MZ (2018) The genome of the human pathogen *Candida albicans* is shaped by mutation and cryptic sexual recombination. *mBio* 9(5):e01205-01218.
9. Reedy JL, Floyd AM, & Heitman J (2009) Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. *Curr. Biol.* 19(11):891-899.
10. Hull CM, Raisner RM, & Johnson AD (2000) Evidence for mating of the" asexual" yeast *Candida albicans* in a mammalian host. *Science* 289(5477):307-310.
11. Anderson MZ, Thomson GJ, Hirakawa MP, & Bennett RJ (2019) A 'parameiosis' drives depolyploidization and homologous recombination in *Candida albicans*. *Nat. Commun.* 10(1):1-10.
12. Du H, Zheng Q, Bing J, Bennett RJ, & Huang G (2018) A coupled process of same- and opposite-sex mating generates polyploidy and genetic diversity in *Candida tropicalis*. *PLoS Genet.* 14(5):e1007377.
13. Porman AM, Alby K, Hirakawa MP, & Bennett RJ (2011) Discovery of a phenotypic switch regulating sexual mating in the opportunistic fungal pathogen *Candida tropicalis*. *Proc. Natl. Acad. Sci. U.S.A.* 108(52):21158-21163.
14. Seervai RN, Jones SK, Hirakawa MP, Porman AM, & Bennett RJ (2013) Parasexuality and ploidy change in *Candida tropicalis*. *Eukaryot Cell* 12(12):1629-1640.
15. Zhang N*, et al.* (2015) Selective advantages of a parasexual cycle for the yeast *Candida albicans*. *Genetics* 200(4):1117-1132.
16. Popp C, Ramírez-Zavala B, Schwanfelder S, Krüger I, & Morschhäuser J (2019) Evolution of fluconazole-resistant *Candida albicans* strains by drug-induced mating competence and parasexual recombination. *mBio* 10(1):e02740-02718.
17. Anderson JM & Soll DR (1987) Unique phenotype of opaque cells in the white-opaque transition of *Candida albicans*. *J. Bacteriol.* 169(12):5579-5588.
18. Miller MG & Johnson AD (2002) White-opaque switching in *Candida albicans* is controlled by mating-type locus homeodomain proteins and allows efficient mating. *Cell* 110(3):293-302.
19. Bennett RJ (2010) Coming of age--sexual reproduction in *Candida* species. *PLoS Pathog.* 6(12):e1001155.

20. Wertheimer NB, Stone N, & Berman J (2016) Ploidy dynamics and evolvability in fungi. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371(1709).
21. Ford CB*, et al.* (2015) The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* 4:e00662.
22. Kwon-Chung KJ & Chang YC (2012) Aneuploidy and drug resistance in pathogenic fungi. *PLoS Pathog.* 8(11):e1003022.
23. Selmecki A, Gerami-Nejad M, Paulson C, Forche A, & Berman J (2008) An isochromosome confers drug resistance in vivo by amplification of two genes, ERG11 and TAC1. *Mol. Microbiol.* 68(3):624-641.
24. Selmecki AM, Dulmage K, Cowen LE, Anderson JB, & Berman J (2009) Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet.* 5(10).
25. Selmecki A, Forche A, & Berman J (2006) Aneuploidy and isochromosome formation in drug-resistant *Candida albicans*. *Science* 313(5785):367-370.
26. Muzzey D, Schwartz K, Weissman JS, & Sherlock G (2013) Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol.* 14(9):R97.
27. Jones T*, et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci. U.S.A.* 101(19):7329-7334.
28. Chibana H, Beckerman JL, & Magee PT (2000) Fine-resolution physical mapping of genomic diversity in *Candida albicans*. *Genome Res.* 10(12):1865-1877.
29. Legrand M*, et al.* (2008) Haplotype mapping of a diploid non-meiotic organism using existing and induced aneuploidies. *PLoS Genet.* 4(1):e1.
30. van het Hoog M*, et al.* (2007) Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biol.* 8(4):R52.
31. Lass-Flörl C (2009) The changing face of epidemiology of invasive fungal disease in Europe. *Mycoses* 52(3):197-205.
32. Krcmery V & Barnes A (2002) Non-albicans *Candida* spp. causing fungaemia: pathogenicity and antifungal resistance. *J. Hosp. Infect.* 50(4):243-260.
33. Tortorano A*, et al.* (2004) Epidemiology of candidaemia in Europe: results of 28-month European Confederation of Medical Mycology (ECMM) hospital-based surveillance study. *Eur. J. Clin. Microbiol.* 23(4):317-322.
34. Nucci M & Colombo AL (2007) Candidemia due to *Candida tropicalis*: clinical, epidemiologic, and microbiologic characteristics of 188 episodes occurring in tertiary care hospitals. *Diagn. Micr. Infec. Dis.* 58(1):77-82.
35. Álvarez-Lerma F*, et al.* (2003) Candiduria in critically ill patients admitted to intensive care medical units. *Intensive Care Med.* 29(7):1069-1076.
36. Colombo AL*, et al.* (2006) Epidemiology of candidemia in Brazil: a nationwide sentinel surveillance of candidemia in eleven medical centers. *J. Clin. Microbiol.* 44(8):2816-2823.
37. Eggimann P, Garbino J, & Pittet D (2003) Epidemiology of *Candida* species infections in critically ill non-immunosuppressed patients. *Lancet Infect. Dis.* 3(11):685-702.
38. Okawa Y MM, Kobayashi H (2008) Comparison of pathogenicity of various *Candida tropicalis* strains. *Bio. Pharma. Bulletin* 31(8):1507-1510.
39. Kontoyiannis DP*, et al.* (2001) Risk Factors for *Candida tropicalis* fungemia in patients with cancer. *Clin. Infect. Dis.* 33(10):1676-1681.
40. Nucci M CA (2007) Candidemia due to *Candida tropicalis*: clinical, epidemiologic, and microbiologic characteristics of 188 episodes occurring in tertiary care hospitals. *Diagn. Micr. Infec. Dis.* 58(1):77-82.

41. Chakrabarti A, *et al.* (2009) Recent experience with fungaemia: change in species distribution and azole resistance. *Scand. J. Infect. Dis.* 41(4):275-284.
42. Slavin MA & Chakrabarti A (2012) Opportunistic fungal infections in the Asia-Pacific region. *Med. Mycol.* 50(1):18-25.
43. Chakrabarti A, *et al.* (2015) Incidence, characteristics and outcome of ICU-acquired candidemia in India. *Intensive Care Med.* 41(2):285-295.
44. Farooqi JQ, *et al.* (2013) Invasive candidiasis in Pakistan: clinical characteristics, species distribution and antifungal susceptibility. *J. Med. Microbiol.* 62(Pt 2):259-268.
45. Marie C & White TC (2009) Genetic basis of antifungal drug resistance. *Curr. Fungal Infect. Rep.* 3(3):163-169.
46. Mancera E, Porman AM, Cuomo CA, Bennett RJ, & Johnson AD (2015) Finding a missing gene: *EFG1* regulates morphogenesis in *Candida tropicalis*. *G3 (Bethesda)* 5(5):849-856.
47. Watson JD & Crick FH (1953) The structure of DNA. *Cold Spring Harb. Symp Quant. Biol.*, (Cold Spring Harbor Laboratory Press), pp 123-131.
48. Meselson M & Stahl FW (1958) The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 44(7):671-682.
49. Sanger F, Nicklen S, & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12):5463-5467.
50. Maxam AM & Gilbert W (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74(2):560-564.
51. Bains W & Smith GC (1988) A novel method for nucleic acid sequence determination. *J. Theor. Biol.* 135(3):303-307.
52. Brenner S, *et al.* (2000) In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U.S.A.* 97(4):1665-1670.
53. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, & Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242(1):84-89.
54. Balasubramanian S, Klenerman D, & Bentley D (2004) U.S. Patent 6,787,308.
55. van Dijk EL, Jaszczyszyn Y, Naquin D, & Thermes C (2018) The Third Revolution in Sequencing Technology. *Trends Genet.* 34(9):666-681.
56. Jain M, Olsen HE, Paten B, & Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17(1):239.
57. Jain M, *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36(4):338-345.
58. Ip CLC, *et al.* (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res.* 4:1075.
59. Jain M, *et al.* (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12(4):351-356.
60. Myers EW, *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* 287(5461):2196-2204.
61. Smirnova JB & McFarlane RJ (2002) The unique centromeric chromatin structure of *Schizosaccharomyces pombe* is maintained during meiosis. *J. Biol. Chem.* 277(22):19817-19822.
62. Jaffe DB, *et al.* (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* 13(1):91-96.

63. Batzoglou S (2002) ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Res.* 12(1):177-189.
64. Huang X & Yang SP (2005) Generating a Genome Assembly with PCAP. *Curr. Protoc. Bioinformatics* 11(1):Unit11 13.
65. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
66. Hernandez D, Francois P, Farinelli L, Osteras M, & Schrenzel J (2008) De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* 18(5):802-809.
67. Miller JR, Koren S, & Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95(6):315-327.
68. Compeau PEC, Pevzner PA, & Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29(11):987-991.
69. Li R*, et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20(2):265-272.
70. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18(5):821-829.
71. Dohm JC, Lottaz C, Borodina T, & Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* 17(11):1697-1706.
72. Warren RL, Sutton GG, Jones SJ, & Holt RA (2006) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23(4):500-501.
73. Zhang W*, et al.* (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One* 6(3):e17915.
74. Phillippy AM, Schatz MC, & Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* 9(3):R55.
75. Koren S & Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* 23:110-120.
76. Koren S*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30(7):693-700.
77. Koren S*, et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27(5):722-736.
78. Chin CS*, et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13(12):1050-1054.
79. Weisenfeld NI, Kumar V, Shah P, Church DM, & Jaffe DB (2017) Direct determination of diploid genome sequences. *Genome Res.* 27(5):757-767.
80. Roach MJ, Schmidt SA, & Borneman AR (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19(1):460.
81. Kronenberg ZN, Hall, R. J., Hiendleder, S., Smith, T. P., Sullivan, S. T., Williams, J. L., & Kingan, S. B. (2018) FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *BioRxiv*.
82. Koren S*, et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36(12):1174-1182.
83. Abel HJ & Duncavage EJ (2013) Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet-Ny* 206(12):432-440.
84. Barriere A*, et al.* (2009) Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Res.* 19(3):470-480.

85. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297-1303.
86. DePristo MA, *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
87. Kitts PA, *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44(D1):D73-D80.
88. Hozier J, Renz M, & Nehls P (1977) The chromosome fiber: evidence for an ordered superstructure of nucleosomes. *Chromosoma* 62(4):301-317.
89. Rattner JB & Hamkalo BA (1979) Nucleosome packing in interphase chromatin. *J. Cell Biol.* 81(2):453-457.
90. Thoma F & Koller T (1981) Unravelled nucleosomes, nucleosome beads and higher order structures of chromatin: influence of non-histone components and histone H1. *J. Mol Biol.* 149(4):709-733.
91. Oudet P, Gross-Bellard M, & Chambon P (1975) Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* 4(4):281-300.
92. Cutter AR & Hayes JJ (2015) A brief review of nucleosome structure. *FEBS Lett.* 589(20):2914-2922.
93. Richmond T, Finch J, Rushton B, Rhodes D, & Klug A (1984) Structure of the nucleosome core particle at 7 Å resolution. *Nature* 311(5986):532-537.
94. Zhou K, Gaullier G, & Luger K (2019) Nucleosome structure and dynamics are coming of age. *Nat. Struct. Mol. Biol.* 26(1):3-13.
95. Tremethick DJ (2007) Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell* 128(4):651-654.
96. Luger K, Dechassa ML, & Tremethick DJ (2012) New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* 13(7):436-447.
97. Dekker J, Rippe K, Dekker M, & Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558):1306-1311.
98. Fudenberg G & Imakaev M (2017) FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat. Methods* 14(7):673-678.
99. Belton J-M & Dekker J (2015) Chromosome conformation capture (3C) in budding yeast. *Cold Spring Harb. Protoc.* 2015(6):pdb. prot085175.
100. Zhao Z, *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nat Genet* 38(11):1341-1347.
101. Dostie J, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16(10):1299-1309.
102. Van Berkum NL, *et al.* (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* (39):e1869.
103. Lieberman-Aiden E, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.
104. Sexton T, *et al.* (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148(3):458-472.
105. Rodley C, Bertels F, Jones B, & O'sullivan J (2009) Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genet. Biol.* 46(11):879-886.
106. Forcato M, *et al.* (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods* 14(7):679-685.

107. Barutcu AR, *et al.* (2016) C-ing the genome: a compendium of chromosome conformation capture methods to study higher-order chromatin organization. *J. Cell Physiol.* 231(1):31-35.
108. Schmitt AD, Hu M, & Ren B (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* 17(12):743.
109. Williamson I, *et al.* (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Gene Dev* 28(24):2778-2791.
110. Williamson I, *et al.* (2012) Anterior-posterior differences in HoxD chromatin topology in limb development. *Development* 139(17):3157-3167.
111. Yaffe E & Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 43(11):1059-1065.
112. Imakaev M, *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* 9(10):999-1003.
113. Rao SS, *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665-1680.
114. Servant N, Varoquaux N, Heard E, Barillot E, & Vert J-P (2018) Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics* 19(1):313.
115. Vidal E, *et al.* (2018) OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.* 46(8):e49-e49.
116. Maeshima K, Tamura S, Hansen JC, & Itoh Y (2020) Fluid-like chromatin: Toward understanding the real chromatin organization present in the cell. *Curr. Opin. Cell Biol.* 64:77-89.
117. Dixon JR, *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376-380.
118. Filippova D, Patro R, Duggal G, & Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* 9(1):14.
119. Weinreb C & Raphael BJ (2016) Identification of hierarchical chromatin domains. *Bioinformatics* 32(11):1601-1609.
120. Crane E, *et al.* (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* 523(7559):240-244.
121. Haddad N, Vaillant C, & Jost D (2017) IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res.* 45(10):e81-e81.
122. Nora EP, *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381-385.
123. Hou C, Li L, Qin ZS, & Corces VG (2012) Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. cell* 48(3):471-484.
124. Ikegami K, Egelhofer TA, Strome S, & Lieb JD (2010) *Caenorhabditis elegans* chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. *Genome Biol.* 11(12):R120.
125. Liu C, Cheng Y-J, Wang J-W, & Weigel D (2017) Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat. Plants* 3(9):742-748.
126. Mizuguchi T, *et al.* (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in *S. pombe*. *Nature* 516(7531):432-435.
127. Eser U, *et al.* (2017) Form and function of topologically associating genomic domains in budding yeast. *Proc. Natl. Acad. Sci. U.S.A.* 114(15):E3061-E3070.

128. Hsieh TH, *et al.* (2015) Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell* 162(1):108-119.
129. Le TB, Imakaev MV, Mirny LA, & Laub MT (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342(6159):731-734.
130. Minajigi A, *et al.* (2015) Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* 349(6245):aab2276.
131. Bintu B, *et al.* (2018) Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 362(6413):eaau1783.
132. Eagen KP, Aiden EL, & Kornberg RD (2017) Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc. Natl. Acad. Sci. U.S.A.* 114(33):8764-8769.
133. Szabo Q, Bantignies F, & Cavalli G (2019) Principles of genome folding into topologically associating domains. *Sci. Adv.* 5(4):eaaw1668.
134. Finn EH, *et al.* (2019) Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* 176(6):1502-1515. e1510.
135. Finn EH & Misteli T (2019) Molecular basis and biological function of variability in spatial genome organization. *Science* 365(6457):eaaw9498.
136. Zhu J, *et al.* (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 152(3):642-654.
137. Gizzi AMC, *et al.* (2019) Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. cell* 74(1):212-222. e215.
138. van Steensel B & Furlong EE (2019) The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.*:1.
139. Tang Z, *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163(7):1611-1627.
140. Stadhouders R, Filion GJ, & Graf T (2019) Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* 569(7756):345-354.
141. Flyamer IM, *et al.* (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 544(7648):110-114.
142. Hug CB, Grimaldi AG, Kruse K, & Vaquerizas JM (2017) Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* 169(2):216-228. e219.
143. Rajarajan P, *et al.* (2019) Spatial genome exploration in the context of cognitive and neurological disease. *Curr. Opin. Neurobiol.* 59:112-119.
144. Rajarajan P, Gil SE, Brennand KJ, & Akbarian S (2016) Spatial genome organization and cognition. *Nat. Rev. Neurosci.* 17(11):681.
145. Lupiáñez DG, Spielmann M, & Mundlos S (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genet.* 32(4):225-237.
146. Valton A-L & Dekker J (2016) TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* 36:34-40.
147. Meaburn KJ (2016) Spatial genome organization and its emerging role as a potential diagnosis tool. *Front. Genet.* 7:134.
148. Sanyal A, Lajoie BR, Jain G, & Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489(7414):109-113.
149. Pope BD, *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* 515(7527):402-405.
150. Sima J, *et al.* (2019) Identifying cis elements for spatiotemporal control of mammalian DNA replication. *Cell* 176(4):816-830 e818.

151. Ramírez F, *et al.* (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* 9(1):1-15.
152. Mourad R & Cuvier O (2018) TAD-free analysis of architectural proteins and insulators. *Nucleic Acids Res.* 46(5):e27.
153. Ong C-T & Corces VG (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* 15(4):234-246.
154. Nora EP, *et al.* (2017) Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 169(5):930-944. e922.
155. Ho JW, *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature* 512(7515):449.
156. Fudenberg G, *et al.* (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15(9):2038-2049.
157. Davidson IF, *et al.* (2019) DNA loop extrusion by human cohesin. *Science* 366(6471):1338-1345.
158. Ganji M, *et al.* (2018) Real-time imaging of DNA loop extrusion by condensin. *Science* 360(6384):102-105.
159. Tanizawa H, Kim K-D, Iwasaki O, & Noma K-i (2017) Architectural alterations of the fission yeast genome during the cell cycle. *Nat. Struct. Mol. Biol.* 24(11):965.
160. Lazar-Stefanita L, *et al.* (2017) Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *EMBO J.* 36(18):2684-2697.
161. Heun P, Laroche T, Shimada K, Furrer P, & Gasser SM (2001) Chromosome dynamics in the yeast interphase nucleus. *Science* 294(5549):2181-2186.
162. Stone EM, Heun P, Laroche T, Pillus L, & Gasser SM (2000) MAP kinase signaling induces nuclear reorganization in budding yeast. *Curr. Biol.* 10(7):373-382.
163. Gasser SM (2002) Visualizing chromatin dynamics in interphase nuclei. *Science* 296(5572):1412-1416.
164. Kitamura E, Blow JJ, & Tanaka TU (2006) Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. *Cell* 125(7):1297-1308.
165. Osborne CS, *et al.* (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 36(10):1065-1071.
166. Schober H, *et al.* (2008) Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. *Genome Res.* 18(2):261-271.
167. Berger AB, *et al.* (2008) High-resolution statistical mapping reveals gene territories in live yeast. *Nat. Methods* 5(12):1031.
168. Jin Q-W, Fuchs J, & Loidl J (2000) Centromere clustering is a major determinant of yeast interphase nuclear organization. *J. Cell Sci.* 113(11):1903-1912.
169. Duan Z, *et al.* (2010) A three-dimensional model of the yeast genome. *Nature* 465(7296):363-367.
170. Muller H, Gil J, Jr., & Drinnenberg IA (2019) The Impact of Centromeres on Spatial Genome Architecture. *Trends Genet.* 35(8):565-578.
171. Bunnik EM, *et al.* (2019) Comparative 3D genome organization in apicomplexan parasites. *Proc. Natl. Acad. Sci. U.S.A.* 116(8):3183-3192.
172. Flemming W (1882) *Zellsubstanz, kern und zelltheilung* (Vogel, Leipzig).
173. Clarke L & Carbon J (1980) Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 287(5782):504-509.
174. Bensasson D, Zarowiecki M, Burt A, & Koufopanou V (2008) Rapid evolution of yeast centromeres in the absence of drive. *Genetics* 178(4):2161-2167.
175. Varoquaux N, *et al.* (2015) Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Res.* 43(11):5331-5339.

176. Huberman JA, Pridmore RD, Jäger D, Zonneveld B, & Philippsen P (1986) Centromeric DNA from *Saccharomyces uvarum* is functional in *Saccharomyces cerevisiae*. *Chromosoma* 94(3):162-168.
177. Yamane S, Karashima H, Matsuzaki H, Hatano T, & Fukui S (1999) Isolation of centromeric DNA from *Saccharomyces bayanus*. *J. Gen. Appl. Microbiol.* 45(2):89-92.
178. Gordon JL, Byrne KP, & Wolfe KH (2011) Mechanisms of chromosome number evolution in yeast. *PLoS Genet.* 7(7):e1002190.
179. Pribylova L, De Montigny J, & Sychrova H (2007) Tools for the genetic manipulation of *Zygosaccharomyces rouxii*. *Fems Yeast Res* 7(8):1285-1294.
180. Souciet J-L*, et al.* (2009) Comparative genomics of protoploid Saccharomycetaceae. *Genome Res.* 19(10):1696-1709.
181. Kitada K, Yamaguchi E, & Arisawa M (1996) Isolation of a *Candida glabrata* centromere and its use in construction of plasmid vectors. *Gene* 175(1-2):105-108.
182. Dujon B*, et al.* (2004) Genome evolution in yeasts. *Nature* 430(6995):35.
183. Kobayashi N*, et al.* (2015) Discovery of an unconventional centromere in budding yeast redefines evolution of point centromeres. *Curr. Biol.* 25(15):2026-2033.
184. Vakirlis N*, et al.* (2016) Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* 26(7):918-932.
185. Dietrich FS*, et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304(5668):304-307.
186. Wendland J & Walther A (2011) Genome evolution in the eremothecium clade of the Saccharomyces complex revealed by comparative genomics. *G3 (Bethesda)* 1(7):539-548.
187. Wendland J & Walther A (2014) Chromosome number reduction in *Eremothecium coryli* by two telomere-to-telomere fusions. *Genome Biol Evol* 6(5):1186-1198.
188. Hens JJ, Zonneveld BJM, Yde Steensma H, & van den Berg JA (1993) The consensus sequence of *Kluyveromyces lactis* centromeres shows homology to functional centromeric DNA from *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* 236-236(2-3):355-362.
189. Iborra F & Ball MM (1994) *Kluyveromyces marxianus* small DNA fragments contain both autonomous replicative and centromeric elements that also function in *Kluyveromyces lactis*. *Yeast* 10(12):1621-1629.
190. Kapoor S, Zhu L, Froyd C, Liu T, & Rusche LN (2015) Regional centromeres in the yeast *Candida lusitaniae* lack pericentromeric heterochromatin. *Proc. Natl. Acad. Sci. U.S.A.* 112(39):12139-12144.
191. 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*. *Proc. Natl. Acad. Sci. U.S.A.* 105(50):19797-19802.
192. 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.
193. Chatterjee G*, et al.* (2016) Repeat-associated fission yeast-like regional centromeres in the ascomycetous budding yeast *Candida tropicalis*. *PLoS Genet.* 12(2):e1005839.
194. Marie-Nelly H*, et al.* (2014) Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics* 30(15):2105-2113.
195. Ravin NV*, et al.* (2013) Genome sequence and analysis of methylotrophic yeast *Hansenula polymorpha* DL1. *BMC Genomics* 14(1):837.

196. Coughlan AY, Hanson SJ, Byrne KP, & Wolfe KH (2016) Centromeres of the yeast *Komagataella phaffii* (*Pichia pastoris*) have a simple inverted-repeat structure. *Genome Biol Evol* 8(8):2482-2492.
197. Kunze G*, et al.* (2014) The complete genome of *Blastobotrys (Arxula) adeninivorans* LS3 - a yeast of biotechnological interest. *Biotechnol. Biofuels* 7(1):66.
198. Vernis L*, et al.* (2001) Only centromeres can supply the partition system required for ARS function in the yeast *Yarrowia lipolytica*. *J. Mol. Biol.* 305(2):203-217.
199. Schotanus K*, et al.* (2015) Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. *Epigenet Chromatin* 8(1):41.
200. Zhu Y*, et al.* (2017) Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of *Malassezia sympodialis*. *Nucleic Acids Res.* 45(5):2629-2643.
201. Sankaranarayanan SR*, et al.* (2020) Loss of centromere function drives karyotype evolution in closely related *Malassezia* species. *Elife* 9:e53944.
202. Winter DJ*, et al.* (2018) Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloe festucae. *PLoS Genet.* 14(10):e1007467.
203. Centola M & Carbon J (1994) Cloning and characterization of centromeric DNA from *Neurospora crassa*. *Mol. Cell Biol.* 14(2):1510-1519.
204. Clarke L, Baum M, Marschall L, Ngan V, & Steiner N (1993) Structure and function of *Schizosaccharomyces pombe* centromeres. *Cold Spring Harb. Symp. Quant. Biol.*, (Cold Spring Harbor Laboratory Press), pp 687-695.
205. Tong P*, et al.* (2019) Interspecies conservation of organisation and function between nonhomologous regional centromeres. *Nat. Commun.* 10(1):2343.
206. Sun S*, et al.* (2017) Fungal genome and mating system transitions facilitated by chromosomal translocations involving intercentromeric recombination. *PLoS Biol.* 15(8):e2002527.
207. Yadav V*, et al.* (2018) RNAi is a critical determinant of centromere evolution in closely related fungi. *Proceedings of the National Academy of Sciences of the United States of America* 115(12):3108-3113.
208. Yadav V*, et al.* (2019) Cellular dynamics and genomic identity of centromeres in cereal blast fungus. *mBio* 10:e01581-01519.
209. Jourdier E*, et al.* (2017) Proximity ligation scaffolding and comparison of two *Trichoderma reesei* strains genomes. *Biotechnol. Biofuels* 10(1):151.
210. Smith KM, Phatale PA, Sullivan CM, Pomraning KR, & Freitag M (2011) Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. *Mol. Cell Biol.* 31(12):2528-2542.
211. 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.
212. Emms DM & Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16(1):157.
213. Katoh K, Misawa K, Kuma K, & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14):3059-3066.
214. Poon AFY, Price MN, Dehal PS, & Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

215. Hieter P, *et al.* (1985) Functional selection and analysis of yeast centromeric DNA. *Cell* 42(3):913-921.
216. Fitzgerald-Hayes M, Clarke L, & Carbon J (1982) Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell* 29(1):235-244.
217. Gaudet A & Fitzgerald-Hayes M (1987) Alterations in the adenine-plus-thymine-rich region of CEN3 affect centromere function in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* 7(1):68-75.
218. Espelin CW, Simons KT, Harrison SC, & Sorger PK (2003) Binding of the essential *Saccharomyces cerevisiae* kinetochore protein Ndc10p to CDEII. *Mol. Biol. Cell* 14(11):4557-4568.
219. White CL, Suto RK, & Luger K (2001) Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *EMBO J.* 20(18):5207-5218.
220. Bechert T, Heck S, Fleig U, Diekmann S, & Hegemann JH (1999) All 16 centromere DNAs from *Saccharomyces cerevisiae* show DNA curvature. *Nucleic Acids Res.* 27(6):1444-1449.
221. Pietrasanta LI, *et al.* (1999) Probing the *Saccharomyces cerevisiae* centromeric DNA (CEN DNA)-binding factor 3 (CBF3) kinetochore complex by using atomic force microscopy. *Proc. Natl. Acad. Sci. U.S.A.* 96(7):3757-3762.
222. Meraldi P, McAinsh A, Rheinbay E, & Sorger P (2006) Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol.* 7(3):R23.
223. Sanyal K & Carbon J (2002) The CENP-A homolog CaCse4p in the pathogenic yeast *Candida albicans* is a centromere protein essential for chromosome transmission. *Proceedings of the National Academy of Sciences of the United States of America* 99(20):12969-12974.
224. Baum M SK, Mishra PK, Thaler N, Carbon J. (2006) Formation of functional centromeric chromatin is specified epigenetically in *Candida albicans*. *Proc. Natl. Acad. Sci. U.S.A.* 103(40)(Oct 3):14877-14882.
225. Fournier P, *et al.* (1993) Colocalization of centromeric and replicative functions on autonomously replicating sequences isolated from the yeast *Yarrowia lipolytica*. *Proc. Natl. Acad. Sci. U.S.A.* 90(11):4912-4916.
226. 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.
227. Nakaseko Y, Adachi Y, Funahashi Si, Niwa O, & Yanagida M (1986) Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. *EMBO J.* 5(5):1011-1021.
228. Fishel B, Amstutz H, Baum M, Carbon J, & Clarke L (1988) Structural organization and functional analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. *Mol. Cell Biol.* 8(2):754-763.
229. Chikashige Y, *et al.* (1989) Composite motifs and repeat symmetry in *S. pombe* centromeres: Direct analysis by integration of NotI restriction sites. *Cell* 57(5):739-751.
230. Baum M, Ngan VK, & Clarke L (1994) The centromeric K-type repeat and the central core are together sufficient to establish a functional *Schizosaccharomyces pombe* centromere. *Mol. Biol. Cell* 5(7):747-761.
231. Rhind N, *et al.* (2011) Comparative functional genomics of the fission yeasts. *Science* 332(6032):930-936.
232. Cambareri EB, Aisner R, & Carbon J (1998) Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Mol. Cell Biol.* 18(9):5465-5477.

233. Cambareri EB, Singer MJ, & Selker EU (1991) Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*. *Genetics* 127(4):699-710.
234. Selker EU (1990) Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* 24(1):579-613.
235. Yadav V*, et al.* (2019) Cellular dynamics and genomic identity of centromeres in cereal blast fungus. *mBio* 10(4).
236. Yadav V, Sreekumar L, Guin K, & Sanyal K (2018) Five pillars of centromeric chromatin in fungal pathogens. *PLoS Pathog.* 14(8):e1007150.
237. Akiyoshi B & Gull K (2014) Discovery of unconventional kinetochores in kinetoplastids. *Cell* 156(6):1247-1258.
238. Drinnenberg IA, deYoung D, Henikoff S, & Malik HS (2014) Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* 3:e03676.
239. Hahnenberger KM, Baum MP, Polizzi CM, Carbon J, & Clarke L (1989) Construction of functional artificial minichromosomes in the fission yeast *Schizosaccharomyces pombe*. *Proc. Natl. Acad. Sci. U.S.A.* 86(2):577-581.
240. Folco HD, Pidoux AL, Urano T, & Allshire RC (2008) Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. *Science* 319(5859):94-97.
241. 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.
242. Copenhaver GP*, et al.* (2009) Neocentromeres Form efficiently at multiple possible loci in *Candida albicans*. *PLoS Genet.* 5(3):e1000400.
243. Ishii K*, et al.* (2008) Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. *Science* 321(5892):1088-1091.
244. Hansen KR, Ibarra PT, & Thon G (2006) Evolutionary-conserved telomere-linked helicase genes of fission yeast are repressed by silencing factors, RNAi components and the telomere-binding protein Taz1. *Nucleic Acids Res.* 34(1):78-88.
245. Grewal SIS & Klar AJS (1996) Chromosomal inheritance of epigenetic states in fission yeast during mitosis and meiosis. *Cell* 86(1):95-101.
246. Yao J*, et al.* (2013) Plasticity and epigenetic inheritance of centromere-specific histone H3 (CENP-A)-containing nucleosome positioning in the fission yeast. *J. Biol. Chem.* 288(26):19184-19196.
247. Allshire RC, Javerzat J-P, Redhead NJ, & Cranston G (1994) Position effect variegation at fission yeast centromeres. *Cell* 76(1):157-169.
248. Allshire RC, Nimmo ER, Ekwall K, Javerzat JP, & Cranston G (1995) Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Gene Dev* 9(2):218-233.
249. Sreekumar L*, et al.* (2019) Cis- and trans-chromosomal interactions define pericentric boundaries in the absence of conventional heterochromatin. *Genetics* 212(4):1121-1132.
250. Scott KC, White CV, & Willard HF (2007) An RNA polymerase III-dependent heterochromatin barrier at fission yeast centromere 1. *PLos One* 2(10):e1099.
251. Jager D & Philippsen P (1989) Stabilization of dicentric chromosomes in *Saccharomyces cerevisiae* by telomere addition to broken ends or by centromere deletion. *EMBO J.* 8(1):247-254.
252. Sato H, Masuda F, Takayama Y, Takahashi K, & Saitoh S (2012) Epigenetic inactivation and subsequent heterochromatinization of a centromere stabilize dicentric chromosomes. *Curr. Biol.* 22(8):658-667.

253. Mythreye K & Bloom KS (2003) Differential kinetochore protein requirements for establishment versus propagation of centromere activity in *Saccharomyces cerevisiae*. *J. Cell Biol.* 160(6):833-843.
254. Steiner N (1994) A novel epigenetic effect can alter centromere function in fission yeast. *Cell* 79(5):865-874.
255. Black BE*, et al.* (2007) Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol. Cell* 25(2):309-322.
256. Kingston IJ, Yung JSY, & Singleton MR (2011) Biophysical characterization of the centromere-specific nucleosome from budding yeast. *J. Biol. Chem.* 286(5):4021-4026.
257. Bloom KS & Carbon J (1982) Yeast centromere DNA is in a unique and highly ordered structure in chromosomes and small circular minichromosomes. *Cell* 29(2):305-317.
258. Cho US & Harrison SC (2011) Ndc10 is a platform for inner kinetochore assembly in budding yeast. *Nat. Struct. Mol. Biol.* 19(1):48-55.
259. Furuyama T & Henikoff S (2009) Centromeric nucleosomes induce positive DNA supercoils. *Cell* 138(1):104-113.
260. Díaz-Ingelmo O, Martínez-García B, Segura J, Valdés A, & Roca J (2015) DNA topology and global architecture of point centromeres. *Cell Rep.* 13(4):667-677.
261. Camahort R*, et al.* (2009) Cse4 is part of an octameric nucleosome in budding yeast. *Mol. Cell* 35(6):794-805.
262. Mizuguchi G, Xiao H, Wisniewski J, Smith MM, & Wu C (2007) Nonhistone Scm3 and histones CenH3-H4 assemble the core of centromere-specific nucleosomes. *Cell* 129(6):1153-1164.
263. Pinto I (2000) Histone H2A is required for normal centromere function in *Saccharomyces cerevisiae*. *EMBO J.* 19(7):1598-1612.
264. Dalal Y, Wang H, Lindsay S, & Henikoff S (2007) Tetrameric structure of centromeric nucleosomes in interphase *Drosophila* cells. *PLoS Biol.* 5(8):e218.
265. Furuyama T, Codomo CA, & Henikoff S (2013) Reconstitution of hemisomes on budding yeast centromeric DNA. *Nucleic Acids Res.* 41(11):5769-5783.
266. Polizzi C & Clarke L (1991) The chromatin structure of centromeres from fission yeast: differentiation of the central core that correlates with function. *J. Cell Biol.* 112(2):191-201.
267. Moyle-Heyrman G*, et al.* (2013) Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. *Proc. Natl. Acad. Sci. U.S.A.* 110(50):20158-20163.
268. Thakur J, Talbert PB, & Henikoff S (2015) Inner kinetochore protein interactions with regional centromeres of fission yeast. *Genetics* 201(2):543-561.
269. Pidoux AL*, et al.* (2009) Fission yeast Scm3: a CENP-A receptor required for integrity of subkinetochore chromatin. *Mol. Cell* 33(3):299-311.
270. Kasinathan S & Henikoff S (2018) Non-B-form DNA is enriched at centromeres. *Mol. Biol. Evol.* 35(4):949-962.
271. Rountree MR & Selker EU (2010) DNA methylation and the formation of heterochromatin in *Neurospora crassa*. *Heredity* 105(1):38-44.
272. Freitag M*, et al.* (2004) DNA methylation is independent of RNA interference in *Neurospora*. *Science* 304(5679):1939-1939.
273. Mishra PK, Baum M, & Carbon J (2007) Centromere size and position in *Candida albicans* are evolutionarily conserved independent of DNA sequence heterogeneity. *Mol. Genet. Genomics* 278(4):455-465.

274. Stephens AD, *et al.* (2013) Pericentric chromatin loops function as a nonlinear spring in mitotic force balance. *J. Cell Biol.* 200(6):757-772.
275. Yeh E, *et al.* (2008) Pericentric chromatin is organized into an intramolecular loop in mitosis. *Curr. Biol.* 18(2):81-90.
276. Stephens AD, Haase J, Vicci L, Taylor RM, 2nd, & Bloom K (2011) Cohesin, condensin, and the intramolecular centromere loop together generate the mitotic chromatin spring. *J. Cell Biol.* 193(7):1167-1180.
277. Stephens AD, *et al.* (2013) Individual pericentromeres display coordinated motion and stretching in the yeast spindle. *J. Cell Biol.* 203(3):407-416.
278. Misteli T (2007) Beyond the sequence: cellular organization of genome function. *Cell* 128(4):787-800.
279. Jin QW, Fuchs J, & Loidl J (2000) Centromere clustering is a major determinant of yeast interphase nuclear organization. *J. Cell Sci.* 113:1903-1912.
280. Haase J, *et al.* (2013) A 3D map of the yeast kinetochore reveals the presence of core and accessory centromere-specific histone. *Curr. Biol.* 23(19):1939-1944.
281. 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.
282. Takahashi K, Chen ES, & Yanagida M (2000) Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. *Science* 288(5474):2215-2219.
283. Hou H, *et al.* (2012) Csi1 links centromeres to the nuclear envelope for centromere clustering. *J. Cell Biol.* 199(5):735-744.
284. Kozubowski L, *et al.* (2013) Ordered kinetochore assembly in the human-pathogenic basidiomycetous yeast *Cryptococcus neoformans*. *mBio* 4(5):e00614-00613.
285. Smith KM, Galazka JM, Phatale PA, Connolly LR, & Freitag M (2012) Centromeres of filamentous fungi. *Chromosome Res.* 20(5):635-656.
286. Yadav V & Sanyal K (2018) Sad1 spatiotemporally regulates kinetochore clustering to ensure high-fidelity chromosome segregation in the human fungal pathogen *Cryptococcus neoformans*. *mSphere* 3(4).
287. Thakur J & Sanyal K (2012) A coordinated interdependent protein circuitry stabilizes the kinetochore ensemble to protect CENP-A in the human pathogenic yeast *Candida albicans*. *PLoS Genet.* 8(4):e1002661.
288. Descorps-Declère S, *et al.* (2015) Genome-wide replication landscape of *Candida glabrata*. *BMC Biol.* 13(1):69.
289. Newlon CS, Pohl TJ, Brewer BJ, & Raghuraman MK (2012) Functional centromeres determine the activation time of pericentric origins of DNA replication in *Saccharomyces cerevisiae*. *PLoS Genet.* 8(5):e1002677.
290. Koren A, *et al.* (2010) Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. *PLoS Genet.* 6(8):e1001068.
291. Rhind N (2006) DNA replication timing: random thoughts about origin firing. *Nat. Cell Biol.* 8(12):1313-1316.
292. Kitamura E, Tanaka K, Kitamura Y, & Tanaka TU (2007) Kinetochore microtubule interaction during S phase in *Saccharomyces cerevisiae*. *Gene Dev* 21(24):3319-3330.
293. Feng W, Bachant J, Collingwood D, Raghuraman MK, & Brewer BJ (2009) Centromere replication timing determines different forms of genomic instability in *Saccharomyces cerevisiae* checkpoint mutants during replication stress. *Genetics* 183(4):1249-1260.
294. Aparicio OM (2013) Location, location, location: it's all in the timing for replication origins. *Gene Dev* 27(2):117-128.

295. Greenfeder SA & Newlon CS (1992) Replication forks pause at yeast centromeres. *Mol. Cell Biol.* 12(9):4056-4066.
296. Cook DM*, et al.* (2018) Fork pausing allows centromere DNA loop formation and kinetochore assembly. *Proc. Natl. Acad. Sci. U.S.A.* 115(46):11784-11789.
297. Bannister AJ*, et al.* (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410(6824):120-124.
298. Hayashi MT, Takahashi TS, Nakagawa T, Nakayama J, & Masukata H (2009) The heterochromatin protein Swi6/HP1 activates replication origins at the pericentromeric region and silent mating-type locus. *Nat. Cell Biol.* 11(3):357-362.
299. Frydman N*, et al.* (2001) Assisting reproduction of infertile men carrying a Robertsonian translocation. *Hum. Reprod.* 16(11):2274-2277.
300. Hermsen M*, et al.* (2005) Centromeric chromosomal translocations show tissue-specific differences between squamous cell carcinomas and adenocarcinomas. *Oncogene* 24(9):1571-1579.
301. Garagna S, Page J, Fernandez-Donoso R, Zuccotti M, & Searle JB (2014) The Robertsonian phenomenon in the house mouse: mutation, meiosis and speciation. *Chromosoma* 123(6):529-544.
302. Brown JD & O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annu. Rev. Genom. Hum. G.* 11(1):291-316.
303. Yadav V, Sun S, Coelho MA, & Heitman J (2020) Centromere scission drives chromosome shuffling and reproductive isolation. *Proc. Natl. Acad. Sci. U.S.A.* 117(14):7917-7928.
304. Luo J, Sun X, Cormack BP, & Boeke JD (2018) Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* 560(7718):392-396.
305. McClintock B (1941) The stability of broken ends of chromosomes in *Zea mays*. *Genetics* 26(2):234-282.
306. Gisselsson D*, et al.* (2000) Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proc. Natl. Acad. Sci. U.S.A.* 97(10):5357-5362.
307. Thomas R, Marks DH, Chin Y, & Benezra R (2018) Whole chromosome loss and associated breakage–fusion–bridge cycles transform mouse tetraploid cells. *EMBO J.* 37(2):201-218.
308. Levis RW (1989) Viable deletions of a telomere from a *Drosophila* chromosome. *Cell* 58(4):791-801.
309. Meier B*, et al.* (2014) *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* 24(10):1624-1636.
310. Croll D, Zala M, & McDonald BA (2013) Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet.* 9(6):e1003567.
311. Agudo M*, et al.* (2000) A dicentric chromosome of *Drosophila melanogaster* showing alternate centromere inactivation. *Chromosoma* 109(3):190-196.
312. Fisher AM*, et al.* (1997) Centromeric inactivation in a dicentric human Y; 21 translocation chromosome. *Chromosoma* 106(4):199-206.
313. Voullaire LE, Slater HR, Petrovic V, & Choo K (1993) A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *Am. J. Hum. Genet.* 52(6):1153.
314. Topp C*, et al.* (2009) Identification of a maize neocentromere in an oat-maize addition line. *Cytogenet. Genome Res.* 124(3-4):228-238.

315. Rocchi M, Archidiacono N, Schempp W, Capozzi O, & Stanyon R (2012) Centromere repositioning in mammals. *Heredity* 108(1):59-67.
316. Stanyon R, *et al.* (2008) Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res.* 16(1):17-39.
317. Purgato S, *et al.* (2015) Centromere sliding on a mammalian chromosome. *Chromosoma* 124(2):277-287.
318. Wade C, *et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326(5954):865-867.
319. Schubert I (2018) What is behind "centromere repositioning"? *Chromosoma* 127(2):229-234.
320. Lee CS, *et al.* (2016) Chromosome position determines the success of double-strand break repair. *Proc. Natl. Acad. Sci. U.S.A.* 113(2):E146-154.
321. 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.
322. Burgess SM & Kleckner N (1999) Collisions between yeast chromosomal loci in vivo are governed by three layers of organization. *Gene Dev* 13(14):1871-1883.
323. Chatterjee G (2014) Identification and characterization of the centromere in human pathogenic yeast *Candida tropicalis*. PhD Thesis (Jawaharlab Nehru centre for Advanced Scientific Research).
324. Walker BJ, *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.
325. Kronenberg ZN, *et al.* (2018) FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*:327064.
326. 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.
327. Varshney N, *et al.* (2015) A surprising role for the Sch9 protein kinase in chromosome segregation in *Candida albicans*. *Genetics* 199(3):671-674.
328. Reuss O, Vik A, Kolter R, & Morschhauser J (2004) The *SAT1* flipper, an optimized tool for gene disruption in *Candida albicans*. *Gene* 341:119-127.
329. Robinson JT, *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.* 29(1):24-26.
330. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, & Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210-3212.
331. Chaisson MJ & Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13(1):238.
332. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9(4):357-359.
333. Ramirez F, *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44(W1):W160-165.
334. Dobin A, *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.
335. Hoot SJ, Oliver BG, & White TC (2008) Candida albicans UPC2 is transcriptionally induced in response to antifungal drugs and anaerobicity through Upc2p-dependent and -independent mechanisms. *Microbiology* 154(Pt 9):2748-2756.
336. MacPherson S, *et al.* (2005) *Candida albicans* zinc cluster protein Upc2p confers resistance to antifungal drugs and is an activator of ergosterol biosynthetic genes. *Antimicrob Agents Ch* 49(5):1745-1752.

337. Hoot SJ, Brown RP, Oliver BG, & White TC (2010) The UPC2 Promoter in Candida albicans Contains Two cis-Acting Elements That Bind Directly to Upc2p, Resulting in Transcriptional Autoregulation. *Eukaryot Cell* 9(9):1354-1362.
338. Notman R, Noro M, O'Malley B, & Anwar J (2006) Molecular basis for dimethylsulfoxide (DMSO) action on lipid membranes. *J. Am. Chem. Soc.* 128(43):13982-13983.
339. Rodriguez RJ, Low C, Bottema CD, & Parks LW (1985) Multiple functions for sterols in Saccharomyces cerevisiae. *BBA-Mol. Cell Biol. L.* 837(3):336-343.
340. Stanke M MB (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33(suppl_2)(Jul 1):W465-467.
341. Cingolani P*, et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2):80-92.
342. Kurtz S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.* 5(2):R12.
343. Nattestad M & Schatz MC (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32(19):3021-3023.
344. 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.
345. Thakur J & Sanyal K (2011) The essentiality of the fungus-specific Dam1 complex is correlated with a one-kinetochore-one-microtubule interaction present throughout the cell cycle, independent of the nature of a centromere. *Eukaryot Cell* 10(10):1295-1305.
346. Schindelin J*, et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9(7):676-682.
347. Goshima G & Yanagida M (2000) Establishing biorientation occurs with precocious separation of the sister kinetochores, but not the arms, in the early spindle of budding yeast. *Cell* 100(6):619-633.
348. Durand NC*, et al.* (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* 3(1):95-98.
349. Heinz S*, et al.* (2018) Transcription elongation can affect genome 3D structure. *Cell* 174(6):1522-1536. e1522.
350. Durand NC*, et al.* (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 3(1):99-101.
351. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20(17):3246-3248.
352. Rabl C (1885) Uber zelltheilung. *Morphol. Jahrb.* 10:214-330.
353. Heinz S*, et al.* (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. cell* 38(4):576-589.
354. Guida A*, et al.* (2011) Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast *Candida parapsilosis*. *BMC Genomics* 12(1):628.
355. 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.

356. Tso GH R-CJ, Tan AS, Sem X, Le GT, Tan TG, Lai GC, Srinivasan KG, Yurieva M, Liao W, Poidinger M. (2018) Experimental evolution of a fungal pathogen into a gut symbiont. *Science* 362(6414)(Nov):589-595.
357. Noe L & Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* 33(Web Server issue):W540-543.
358. Krumsiek J, Arnold R, & Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23(8):1026-1028.
359. Bailey TL*, et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37(Web Server issue):W202-208.
360. 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.
361. 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.
362. Shen XX*, et al.* (2016) Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3 (Bethesda)* 6(12):3927-3939.
363. 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.
364. Cavalheiro M & Teixeira MC (2018) *Candida* Biofilms: threats, challenges, and promising strategies. *Front. Med.* 5:28.
365. Pappas PG, Lionakis MS, Arendrup MC, Ostrosky-Zeichner L, & Kullberg BJ (2018) Invasive candidiasis. *Nat Rev Dis Primers* 4:18026.
366. 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.
367. 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.
368. 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.
369. 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.
370. Todd RT, Wikoff TD, Forche A, & Selmecki A (2019) Genome plasticity in *Candida albicans* is driven by long repeat sequences. *Elife* 8:e45954.
371. Gusa A & Jinks-Robertson S (2019) Mitotic recombination and adaptive genomic changes in human pathogenic fungi. *Genes* 10(11):901.
372. Torres EM, Williams BR, & Amon A (2008) Aneuploidy: cells losing their balance. *Genetics* 179(2):737-746.
373. Seeber A, Hauer MH, & Gasser SM (2018) Chromosome dynamics in response to DNA damage. *Annu. Rev. Genet.* 52(1):295-319.
374. 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.
375. Arsuaga J*, et al.* (2004) Chromosome spatial clustering inferred from radiogenic aberrations. *Int. J. Radiat. Biol.* 80(7):507-515.

376. Bickmore WA & Teague P (2002) Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosome Res.* 10(8):707-715.
377. Branco MR & Pombo A (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* 4(5).
378. Canela A*, et al.* (2017) Genome organization drives chromosome fragility. *Cell* 170(3):507-521. e518.
379. Engreitz JM, Agarwala V, & Mirny LA (2012) Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One* 7(9).
380. Hlatky L, Sachs RK, Vazquez M, & Cornforth MN (2002) Radiation-induced chromosome aberrations: Insights gained from biophysical modeling. *Bioessays* 24(8):714-723.
381. Holley W, Mian I, Park S, Rydberg B, & Chatterjee A (2002) A model for interphase chromosomes and evaluation of radiation-induced aberrations. *Radiat. Res.* 158(5):568-580.
382. Klein IA*, et al.* (2011) Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* 147(1):95-106.
383. Roukos V, Burman B, & Misteli T (2013) The cellular etiology of chromosome translocations. *Curr. Opin. Cell Biol.* 25(3):357-364.
384. Zhang Y*, et al.* (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148(5):908-921.
385. Aten JA*, et al.* (2004) Dynamics of DNA double-strand breaks revealed by clustering of damaged chromosome domains. *Science* 303(5654):92-95.
386. Foster HA*, et al.* (2013) Relative proximity of chromosome territories influences chromosome exchange partners in radiation-induced chromosome rearrangements in primary human bronchial epithelial cells. *Mutat. Res.* 756(1-2):66-77.
387. Savage JR (1998) A brief survey of aberration origin theories. *Mutat. Res.* 404(1-2):139-147.
388. Savage JR (2000) Proximity matters. *Science* 290(5489):62-63.
389. Therman E, Susman B, & Denniston C (1989) The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Ann. Hum. Genet.* 53(1):49-65.
390. Robertson WRB (1916) Chromosome studies. I. Taxonomic relationships shown in the chromosomes of *Tettigidae* and *Acrididae*: V-shaped chromosomes and their significance in *Acrididae*, *Locustidae*, and *Gryllidae*: chromosomes and variation. *J. Morphol.* 27(2):179-331.
391. Castiglia R & Capanna E (2002) Chiasma repatterning across a chromosomal hybrid zone between chromosomal races of *Mus musculus domesticus*. *Genetica* 114(1):35-40.
392. Dumas D & Britton-Davidian J (2002) Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and Robertsonian populations of the house mouse. *Genetics* 162(3):1355-1366.
393. Friebe B, Zhang P, Linc G, & Gill B (2005) Robertsonian translocations in wheat arise by centric misdivision of univalents at anaphase I and rejoining of broken centromeres during interkinesis of meiosis II. *Cytogenet. Genome Res.* 109(1-3):293-297.
394. Guichaoua M*, et al.* (1990) Infertility in human males with autosomal translocations: meiotic study of a 14; 22 Robertsonian translocation. *Hum. Genet.* 86(2):162-166.

395. Mattei M, Souiah N, & Mattei J (1984) Chromosome 15 anomalies and the Prader-Willi syndrome: cytogenetic analysis. *Hum. Genet.* 66(4):313-334.
396. Kalitsis P, Griffiths B, & Choo KA (2006) Mouse telocentric sequences reveal a high rate of homogenization and possible role in Robertsonian translocation. *Proc. Natl. Acad. Sci. U.S.A.* 103(23):8786-8791.
397. Zhang CZ, Leibowitz ML, & Pellman D (2013) Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Gene Dev* 27(23):2513-2530.
398. Baca SC*, et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell* 153(3):666-677.
399. Crasta K*, et al.* (2012) DNA breaks and chromosome pulverization from errors in mitosis. *Nature* 482(7383):53-58.
400. Meaburn KJ, Misteli T, & Soutoglou E (2007) Spatial genome organization in the formation of chromosomal translocations. *Seminars in cancer biology*, (Elsevier), pp 80-90.
401. Fukagawa T & Earnshaw WC (2014) Neocentromeres. *Curr. Biol.* 24(19):R946-947.
402. Scott KC & Sullivan BA (2014) Neocentromeres: a place for everything and everything in its place. *Trends Genet.* 30(2):66-74.
403. Link AJ & LaBaer J (2011) Trichloroacetic acid (TCA) precipitation of proteins. *Cold Spring Harb. Protoc.* 2011:993-994.
404. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215(3):403-410.
405. Spitzer M, Wildenhain J, Rappsilber J, & Tyers M (2014) BoxPlotR: a web tool for generation of box plots. *Nat. Methods* 11(2):121-122.
406. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997v2 [q-bio.GN]*.
407. Li H*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
408. Sievers F & Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple sequence alignment methods*, (Springer), pp 105-116.
409. Schroeder A*, et al.* (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol.* 7(1):3.

Appendix-I: List of strains used in this study

Strain name	Genotype
SC5314	*C. albicans* reference strain (Wild-type)
CaKG001	*arg4Δ/arg4Δ, leu2Δ/leu2Δ, his1Δ/his1Δ, ura3Δ::imm434/ura3Δ::imm434, iro1Δ::imm434/iro1Δ::imm434 CSE4/CSE4-TAP(CaSAT1)*
MYA3404	*C. tropicalis* Clinical isolate (Wild-type)
NCYC-2606	Environmental isolate of *C. sojae* (wild-type)
CTKS06	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT*
CtKS101	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaURA3)*
CtKS102	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1)*
CtKG001	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT*
CtKG002	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT Chr5-497kb/ Chr5-497kb::CaURA3*
CtKG003	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) Chr5-497kb/ Chr5-497kb::CaURA3*
CtKG101	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy Transformant 1
CtKG102	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy Transformant 2
CtKG103	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy Transformant 3
CtKG104	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy Transformant 4
CtKG105	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT CSE4/CSE4::CSE4-TAP (CaHIS1) sch9::FRT/sch9::FRT* Chr5 monosomy Transformant 5
CtKG300S1	*DUP4/DUP4/DUP4::CaSAT1* Transformant S1
CtKG300S2	*DUP4/DUP4/DUP4::CaSAT1* Transformant S2
CtKG300S3	*DUP4/DUP4/DUP4::CaSAT1* Transformant S3
CtKG400S1	*DUPR/DUPR/DUPR/DUPR::CaSAT1* Transformants S1
CtKG400S2	*DUPR/DUPR/DUPR/DUPR::CaSAT1* Transformants S2
CtKG400S3	*DUPR/DUPR/DUPR/DUPR::CaSAT1* Transformants S3
CtKG500	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT NUF2/NUF2::NUF2-GFP (CaHIS1)*
CtKG501	*ura3::FRT/ura3::FRT his1::FRT/his1::FRT arg4::FRT/arg4::FRT MIF2/MIF2::MIF2-GFP (CaHIS1)*

Appendix-II List of primers used in this study

Primer name	Sequence	Purpose of use
Construction and confirmation of CtKG001		
KG1	ATGCGGTACCGTTGGTACCTTCTACAGATGC	Amplification and cloning of upstream homology region of SCH9 ORF
KG2	AGTCCTCGAGAATGGGTGAGCAGATGATGG	
KG3	ATGCCCGCGGGATGAAGAAATGCAACCAGCAG	Amplification and cloning of downstream homology region of SCH9 ORF
KG4	ATGAGAGCTCCAAAATTGGAATCGTTAGAAACGG	
KG5	CAACAATTTAACTTAACATGTGGCAC	PCR confirmation of Sch9::CaSAT1
KG6	TTCTGAAACTTGAAGGATTAGATAC	
KG7	CAGTGGCTACAACTCAGAGCACGC	
KG8	TTAGAGACACAAACGAACAATGTACC	
Construction and confirmation of CtKG002 and CtKG003, *MTL* PCR		
KG9	GATCGGGCCCGGGGAACACCAACTTCAAAA	Amplification and cloning of upstream homology region of Chr5-497kb locus
KG10	AGTCCTCGAGGAGAGTCATGACACACCACTTGTTG	
KG11	ATCGCTGCAGGACTGGAACCTTATGTGAGGAGACAG	Amplification and cloning of downstream homology region of Chr5-497_kb locus
KG12	AGCTGGATCCGACATCATGGATGAGCCTTGGTAG	
KG13	GAGAAAAAGAAAGAGAAGGATTCTAAGG	PCR confirmation of Chr5-497_kb::URA3 transformants
KG14	ATCCTTCTTCTTGGCCACCC	
KG15	GGACTGGGAGGGTGCATTGG	Probe for Southern hybridization confirmation of the sch9/sch9 mutants
KG16	CTATGTGGGCGTGTGATTGCGC	
KG17	GATTTGGTATGAAAAGAGGAACTCTAAC	Amplification of *MTL***a** locus
KG18	CTACTAATTTTGAAACCATTTGGAGTCT	
KG19	TAAAACATTAAGCATAGAGGACAAAGAA	Amplification of the *MTL*α locus
KG20	AACTTCAAATGCAAAATGTAAAACATAC	

Amplification of probes used in Southern hybridization experiments		
KG21	GTTATTGAAGAACCTAGAGGG	Contig14_Probe
KG22	AGCTTGTAATTCAGGTGACA	
KG23	CTACTCCACAGAAACAATCTCC	Contig16_Probe
KG24	CAGGGATACACTTCTTATGACC	
KG25	TGCAGTTGAAATCTCTTGGACC	Probe A
KG26	GGATGTTGCGATCACTTTGG	
KG27	GGAAAGATTCAGGATAAACCAATTG	Probe B
KG28	CCATTTGACTTGCCCACTC	
KG29	ATCATGAGAATACAAAGAGAAAGTT	Probe C
KG30	AACTTTTGTAGTCTATCCAACTCTG	
KG31	GTGCATACGTCACAGTTTGG	Probe D
KG32	GATGCTATCATCTCAAACCAAGG	
KG33	GGTGGTTACGGTACCAGATTG	Probe E
KG34	CCGGCATTGATTCTGTTACC	
KG35	GTTGTTATCCAGTGTACCAGTTG	Probe F
KG36	GGGATTTCTGGTGGTTCAAC	
Construction of Mif2-GFP strain (CtKG501)		
KG113	TCCCCGCGGAGACCACAAACACCAACTAGCG	Amplification and cloning of C-terminus region of the *MIF2* ORF
KG119	CTAAACCTGAAGAAGTTGAAGATACTTGGATGAGTAAGGGAGAAGAACTTTTCACTGGAG	
Construction of Nuf2-GFP strain (CtKG500)		
KG127	ATCACCGCGGGATTATAAACAGGAGAAAACCAATTTAGC	Amplification and cloning of C-terminus region of the *NUF2* ORF
KG128	ATCAACTAGTTTTGATATTTTTGTTTAATTCTGTCATATATC	
KG129	ATCGGGGCCCCACTAACCTGTATAGACCAAATAATTTG	Amplification and cloning of downstream region of the *NUF2* ORF
KG1	GGGGTACCATTTTATCACCTTTGGGAACAGG	

Construction of pCEN501		
KG235	ATGCGTCGACCAATATTTCATCGTGTTTCACCCG	Amplification and cloning of LR from *CtCEN5*
KG236	CAGTCTGCAGAATACTTTGAATCAAGGTTAGCAATG	
Construction of pCEN502		
KG229	AGTCCTGCAGGCATTCGAAGGACATTAATTAACG	Amplification and cloning of *CaLR5*
KG230	ATGCGTCGACCAGTACGTTGTGTTTTGAAGTCCTC	
KG231	AGCGGATCCCTTTTTATTCCAGTATTCTGATTGATCTATTTATC	Amplification and cloning of *CaRR5*
KG232	ATGCGGATCCGATGTTGTTGTGGTAGCCATAGTGTG	
Construction of pCaCEN5		
KG229	AGTCCTGCAGGCATTCGAAGGACATTAATTAACG	Amplification and cloning of *CaCEN5*
KG234	ATGCGTCGACTGGTGTGTTGCTGCTGCCCTTAG	
KG231	AGCGGATCCCTTTTTATTCCAGTATTCTGATTGATCTATTTATC	
KG233	ATGCGTCGACCATGTTCCAACTCTCTCATGCGATC	
Construction of CENP-A^{Cse4}-Protein-A strain		
CSE4 1F	GAACAGCTACTAGAGAGAGATAG	CENP-A ORF fragment
CSE4 2R	CTTTTTCCATCTTCTCTTTTCTAGAATCCAGGACTG	
CSE4-3UTR F	CAATTCGCCCTATAGTGAGTCGTAGTGTACCATATAGAATGTAAGAG	CENP-A downstream sequences
CSE4 6R	GTACCAATAGAGAATTCTAGG	
CSE4 3F	CCAGTCCTGGATTCTAGAAAAGAGAAGATGGAAAAAG	TAP-Ca*URA3* amplification
CSE4-TAP R	CTCTTACATTCTATATGGTACACTACGACTCACTATAGGGCGAATTG	
KG121	ATG CGC GGC CGC GTG GGC ATC TAT CGA AAT CAG	CENP-A^{CSE4} along with *TAP* amplification
KG78	GAC TAG TGG CCA ATT ATA AAT GTG AAG GGG G	

Deletion of *DUP4* locus		
KG518US-F	CTCGTGACAAGTATCAGAATGTC	Construction of overlap PCR construct to delete *DUP4* locus with *CaSAT1* marker
KG519US-R	GTATTCTGGGCCTCCATGTCGATTCGAGGGC GAACTGACAC	
KG520NATF	GTGTCAGTTCGCCCTCGAATCGACATGGAGG CCCAGAATAC	
KG521NAT-R	CATAGAGGCAGTCAAGGCTGCAGTATAGCG ACCAGCATTCAC	
KG522DS-F	GTGAATGCTGGTCGCTATACTGCAGCCTTGA CTGCCTCTATG	
KG523DS-R	CTCTCTGAGAATCCTTTCTAAGC	
KG531_Chr4-conf	ATCGTGGTTGTTTGGGCC	PCR confirmation of deletion
KG378	GCCAGAGAAAGAGGTGCTGG	
Deletion of *DUPR* locus		
KG524_ChrRDDFP	CCTTGTTAGTTTCATAAATTGCCGC	Construction of overlap PCR construct to delete *DUPR* locus with *CaSAT1* marker
KG525_ChrR-USR	GGTATTCTGGGCCTCCATGTCGAACACAGTT TCAGCGGCCTG	
KG526_chrRNATF	CAGGCCGCTGAAACTGTGTTCGACATGGAG GCCCAGAATACC	
KG527_Chr-R-NATR	GTATGGGGAAAAGAGATTCACGTCCAGTAT AGCGACCAGCATTCAC	
KG528_ChR-DSF	GTGAATGCTGGTCGCTATACTGGACGTGAAT CTCTTTTCCCCATAC	
KG529_ChrR-DSR	CATCAATTGACTGCTACTAGCTTTG	
KG530_ChrR-conf	GGCAACTTTTGGGCAACC	PCR confirmation of deletion
KG378	GCCAGAGAAAGAGGTGCTGG	
ChIP-PCR and qPCR assay		
SalI CC only FP	ACGCGTCGACGTAGTTATCTAGATGCAATTT GTTTG	For qPCR on native *CEN5*
RT-pCtCEN RP	TATTACCTACAAATAACTTCATCAAGTC	
pCEN8 ChIP FP	AACCTTGATTCAAAGTATTGTGTCGAC	For qPCR on pCEN5
RT-pCtCEN RP	TATTACCTACAAATAACTTCATCAAGTC	

RT- CC+IR and CC FP	CTTGCATGCCTGCAGGTCGAC	For qPCR on pmid5
RT-pCtCEN RP	TATTACCTACAAATAACTTCATCAAGTC	
Leu2-3FP (F)	TAAAAATCATTTAATTGGTGGTG	For qPCR on *CtLEU2*
Leu2-3RP (R)	ACAGCATCTGATGATTTAGCACAT	
Ca21	CTGGTGCAAGACCCTCATAGAAGC	For PCR on *CaCEN7*
Ca22	CCTGACACTGTCGTTTCCCATAGC	
Ca7a	ACTCGCCTTCCCCTCCTTTAAAT	For PCR on *CEN7*-distal locus
Ca7b	CCACTACTACGACTGTGGATTCA	
KG506	AGAGTTGGAACATGGTCGAC	For PCR on pCaCEN5
KG507	CCACCTTAAAATACGGTCCC	
KG200	ATGCTCATGGTGTCACTGGG	For PCR on *CaURA3*
KG201	ATCCTTCTTCTTGGCCACCC	

Appendix-III List of plasmids used in this study

Name	Vector/ backbone	Modification	Length (bp)	References
pKG1	pSFS2a	Upstream and downstream homology region of *SCH9* ORF is cloned	8542	This study
pKG2	pBSCaURA3	Upstream and downstream homology region of intergenic locus (Chr5_497_kb) is cloned	5039	This study
pCENP-A-TAP-HIS	pBS-HIS	CtCENP-A^{Cse4}-TAP fragment is cloned	5975	This study
pCENP-C-GFP	pGFP-HIS	C-terminus region of the *MIF2* ORF is cloned	5803	This study
pNUF2-ORF	pGFP-HIS	C-terminus region of the *NUF2* ORF is cloned	6012	This study
pNUF2-GFP	pNUF2-ORF	downstream homology region of *NUF2* ORF is cloned	6676	This study
pARS2-λ	pARS2	~11 kb λ -DNA is cloned	15000	This study
pCEN501	pmid+RR	*CtCEN5* LR is cloned is a direct orientation to the RR	15149	This study
pCEN502	pmid8	IRs from *CaCEN5* are cloned	11818	This study
pCaCEN5	pARS2	*CaCEN5* is cloned	12448	This study

Appendix-IV Script used for analysis of 3C-seq data using Homer

```
#trimming
homerTools trim -3 GATC -mis 0 -matchStart 20 -min 20 1.fq
homerTools trim -3 GATC -mis 0 -matchStart 20 -min 20 2.fq

#bowtie2 index
bowtie2-build Ctrop.fasta Ctrop.fasta

#bowtie2 mapping
bowtie2 -p 40 -x Ctrop.fasta -U 1.fq.trimmed > 1.sam
bowtie2 -p 40 -x Ctrop.fasta -U 2.fq.trimmed > 2.sam

#make tag_directory
makeTagDirectory tag_directory_ct 1.sam,2.sam -tbp 1 -restrictionSite GATC -unique -genome Ctrop.fasta -
removeSpikes 10000 8

#analyzeHiC
analyzeHiC tag_directory_ct -res 5000 -cpu 40 > matrix_tag_dir_ct_5k.txt

#Find_tads_and_loops
findTADsAndLoops.pl find tag_directory_ct/ -cpu 40 -res 5000 -window 10000 -genome Ctrop.fasta -
minTADsize 5000 -keepOverlappingTADs -minLoopDist 5000
```

Appendix-V Script used in SNP/indel analysis using GATK software

Appendix-V Script used in SNP/indel analysis using GATK softwate

```
#indexing the reference
bwa index Ctrop.fasta

#Alignment – Map to Reference
bwa mem -t 42 -M -R '@RG\tID:sample_MYA3404\tLB:3C-seq\tPL:ILLUMINA\tPM:MISEQ\tSM:sample_Number' Ctrop.fasta
/media/mml2/candida2/all-gap-filled_Ct_chr/Homer_on_Ct/1.fq.gz /media/mml2/candida2/all-gap-
filled_Ct_chr/Homer_on_Ct/1.fq.gz > Aligned.sam

#Sort SAM file by coordinate, convert to BAM
java -jar ~/software1/picard/build/libs/picard.jar SortSam INPUT=Aligned.sam OUTPUT=Aligned_sorted.bam
SORT_ORDER=coordinate

#Collect Alignment & Insert Size Metrics
java -jar ~/software1/picard/build/libs/picard.jar CollectAlignmentSummaryMetrics R=Ctrop.fasta I=Aligned_sorted.bam
O=read_alignment_metrics.txt

#generate histogram
java -jar ~/software1/picard/build/libs/picard.jar CollectInsertSizeMetrics I=Aligned_sorted.bam O=insert_metrics.txt
HISTOGRAM_FILE=histogram.pdf

#Depth_check of bam file
samtools depth -a Aligned_sorted.bam > depth_out.txt

#Mark Duplicates
java -jar ~/software1/picard/build/libs/picard.jar MarkDuplicates INPUT=Aligned_sorted.bam
OUTPUT=dedup_Aligned_sorted.bam METRICS_FILE=metrics.txt

#Build BAM Index
java -jar ~/software1/picard/build/libs/picard.jar BuildBamIndex INPUT=dedup_Aligned_sorted.bam

#make sure the .jar files are in working directory
#R=reference
#o=out directory
#index reference
#step 01
java -jar -Xmx64G CreateSequenceDictionary.jar R=Ctrop.fasta O=Ctrop.dict

#samtools index fasta
samtools faidx Ctrop.fasta

#Create Realignment Targets
java -jar -Xmx64G GenomeAnalysisTK.jar -T RealignerTargetCreator -R Ctrop.fasta -I dedup_Aligned_sorted.bam -o
readalignment_target.list

#Realign Indels
java -jar GenomeAnalysisTK.jar -T IndelRealigner -R Ctrop.fasta -I dedup_Aligned_sorted.bam -targetIntervals
readalignment_target.list -o realigned_reads.bam

#Call Variants
java -jar -Xmx64G GenomeAnalysisTK.jar -T HaplotypeCaller -R Ctrop.fasta -ploidy 2 -I dedup_Aligned_sorted.bam -o
realigned_raw.vcf

#Extract SNPs
java -jar -Xmx64G GenomeAnalysisTK.jar -T SelectVariants -R Ctrop.fasta -V realigned_raw.vcf -selectType SNP -o SNPs.vcf

#Extract Indels
java -jar -Xmx64G GenomeAnalysisTK.jar -T SelectVariants -R Ctrop.fasta -V realigned_raw.vcf -selectType INDEL -o
indels.vcf

#Filter SNPs
java -jar -Xmx64G GenomeAnalysisTK.jar -T VariantFiltration -R Ctrop.fasta -V SNPs.vcf --filterExpression 'QD < 2.0 || FS >
60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0' --filterName "basic_snp_filter" -o
filtered_SNPs.vcf
```

```
#-------------------------------------------------------------------------------------------------------------------
#QD:variant confidence
#FShred-scaled p-value using Fisher's Exact Test to detect strand bias
#RMSMappingQuality (MQ) 40.0:Root Mean Square of the mapping quality of the reads across all samples.
#MappingQualityRankSumTest (MQRankSum) -12.5
#ReadPosRankSumTest (ReadPosRankSum) -8.0
#StrandOddsRatio (SOR) 3.0:The StrandOddsRatio annotation is one of several methods that aims to evaluate whether there is strand bias in the data
#-------------------------------------------------------------------------------------------------------------------

#Filter Indels
java -jar GenomeAnalysisTK.jar -T VariantFiltration -R Ctrop.fasta -V indels.vcf --filterExpression 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 10.0' --filterName "basic_indel_filter" -o filtered_indels.vcf

#Base Quality Score Recalibration (BQSR) #1
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R Ctrop.fasta -I realigned_reads.bam -knownSites filtered_SNPs.vcf -knownSites filtered_indels.vcf  -o recal_data.table

#Base Quality Score Recalibration (BQSR) #2
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R Ctrop.fasta -I realigned_reads.bam -knownSites filtered_SNPs.vcf -knownSites filtered_indels.vcf  -BQSR recal_data.table -o post_recal_data.table

#Analyze Covariates
#not working
java -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R Ctrop.fasta -before recal_data.table -after post_recal_data.table -plots recalibration_plots.pdf

#Apply BQSR
java -jar GenomeAnalysisTK.jar -T PrintReads -R Ctrop.fasta -I realigned_reads.bam -BQSR recal_data.table -o recal_reads.bam

#Call Variants
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R Ctrop.fasta -ploidy 2 -I recal_reads.bam -o raw_variants_recal.vcf


#Extract SNPs
java -jar GenomeAnalysisTK.jar -T SelectVariants -R Ctrop.fasta -V raw_variants_recal.vcf -selectType SNP -o raw_SNPs_recal.vcf


#Extract Indels
java -jar GenomeAnalysisTK.jar -T SelectVariants -R Ctrop.fasta -V raw_variants_recal.vcf -selectType INDEL -o raw_indels_recal.vcf

#Filter SNPs
java -jar GenomeAnalysisTK.jar -T VariantFiltration -R Ctrop.fasta -V raw_SNPs_recal.vcf --filterExpression 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0' --filterName "basic_snp_filter" -o filtered_SNPs_recal_final.vcf

#Filter Indels
java -jar GenomeAnalysisTK.jar -T VariantFiltration -R Ctrop.fasta -V raw_indels_recal.vcf --filterExpression 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 10.0' --filterName "basic_indel_filter" -o filtered
```

Appendix-VI Script used in haplotype analysis using FALCON, FALCON-Unzip and FALCON-Phase software

A. fc_run.cfg

```
#### Input
[General]
input_fofn=input.fofn
input_type=raw
pa_DBdust_option=true
pa_fasta_filter_option=streamed-internal-median
target=assembly
skip_checks=False
LA4Falcon_preload=false

#### Data Partitioning
pa_DBsplit_option=-x500 -s100
ovlp_DBsplit_option=-x500 -s100

#### Repeat Masking
pa_HPCTANmask_option = -k18 -h480 -w8 -e.8 -s100
pa_HPCREPmask_option = -k18 -h480 -w8 -e.8 -s100
pa_REPmask_code=0,300;0,300;0,300

####Pre-assembly
genome_size=15000000
seed_coverage=20
length_cutoff=100
pa_HPCdaligner_option=-v -B128 -M24
pa_daligner_option=-e.8 -l1000 -k18 -h480  -w8 -s100 -T10
falcon_sense_option=--output-multi --min-idt 0.70 --min-cov 2 --max-n-read 1800
falcon_sense_greedy=False

####Pread overlapping
ovlp_daligner_option=-e.96 -l1000 -k24 -h1024 -w6 -s100
ovlp_HPCdaligner_option=-v -B128 -M24

####Final Assembly
overlap_filtering_setting=--max-diff 100 --max-cov 100 --min-cov 2
fc_ovlp_to_graph_option=
length_cutoff_pr=500

[job.defaults]
job_type= local
pwatcher_type=blocking
JOB_QUEUE=default
MB=32768
NPROC=10
njobs=1
submit = bash -C ${CMD} >| ${STDOUT_FILE} 2>| ${STDERR_FILE}
```

```
[job.step.da]
NPROC=10
MB=32768
njobs=1
[job.step.la]
NPROC=10
MB=32768
njobs=1
[job.step.cns]
NPROC=10
MB=65536
njobs=1
[job.step.pda]
NPROC=10
MB=32768
njobs=1
[job.step.pla]
NPROC=10
MB=32768
njobs=1
[job.step.asm]
NPROC=10
MB=32768
njobs=1
```

B. fc_unzip.cfg

```
[General]
max_n_open_files = 1000

[Unzip]
input_fofn=input.fofn
input_bam_fofn=input_bam.fofn

[job.defaults]
job_type=local
pwatcher_type=blocking
JOB_QUEUE=default
MB=8192
NPROC=10
njobs=8
submit = bash -C ${CMD} >| ${STDOUT_FILE} 2>| ${STDERR_FILE}

[job.step.unzip.track_reads]
njobs=2
NPROC=10
MB=65536
```

```
# uses minimap2 now
[job.step.unzip.blasr_aln]
njobs=16
NPROC=10
MB=16384
[job.step.unzip.phasing]
njobs=16
NPROC=10
MB=16384
[job.step.unzip.hasm]
njobs=4
NPROC=10
MB=65536
# uses arrow now
[job.step.unzip.quiver]
njobs=4
NPROC=10
MB=65536
```

C. fc_phase.cfg

```
[General]
#target=minced

[job.defaults]
NPROC=10
njobs=100
MB=64000
pwatcher_type=blocking
job_type=local
JOB_QUEUE=default
submit = bash -C ${CMD} >| ${STDOUT_FILE} 2>| ${STDERR_FILE}

[Phase]
cns_p_ctg_fasta = ./4-polish/cns-output/cns_p_ctg.fasta
cns_h_ctg_fasta = ./4-polish/cns-output/cns_h_ctg.fasta
reads_1=/media/mml/6f60ef75-45fb-4532-9f2a-
1a5d642a3093/Candida_tropicalis/Candida_tropicalis_Chromosomes/M_Y_WT_1_val_1
.fq.gz
reads_2=/media/mml/6f60ef75-45fb-4532-9f2a-
1a5d642a3093/Candida_tropicalis/Candida_tropicalis_Chromosomes/M_Y_WT_2_val_2
.fq.gz
min_aln_len=3000
iterations=10000000
enzyme="GATC"
output_format=pseudohap
```

Appendix-VII Script used for RNA-seq data analysis

```
#make geneome1 dir
mkdir genome1

#genomegenerate (STAR)
STAR --runMode genomeGenerate --runThreadN 46 --genomeDir genome1 --genomeFastaFiles Ctrop.fasta
--sjdbGTFfile Ctrop.gtf --sjdbGTFfeatureExon CDS --sjdbOverhang 100

#readmapping (STAR)
STAR --runThreadN 16 --genomeDir genome1 --readFilesIn
/media/mml2/candida2/mml_system_data_4.3.19/Candida_tropicalis/Ct_RNAseq_19.2.19/READS_TRIM
MED/MYAD2_R1_val_1.fq
/media/mml2/candida2/mml_system_data_4.3.19/Candida_tropicalis/Ct_RNAseq_19.2.19/READS_TRIM
MED/MYAD2_R2_val_2.fq --outFileNamePrefix MYAD2_26.04.19 --outSAMtype BAM
SortedByCoordinate --quantMode GeneCounts --limitBAMsortRAM 6534549647 --twopassMode Basic
ulimit -n 10000

#readmapping2
STAR --runThreadN 16 --genomeDir genome1 --readFilesIn
/media/mml2/candida2/mml_system_data_4.3.19/Candida_tropicalis/Ct_RNAseq_19.2.19/READS_TRIM
MED/MYAD3_R1_val_1.fq
/media/mml2/candida2/mml_system_data_4.3.19/Candida_tropicalis/Ct_RNAseq_19.2.19/READS_TRIM
MED/MYAD3_R2_val_2.fq --outFileNamePrefix MYAD3_26.04.19 --outSAMtype BAM
SortedByCoordinate --quantMode GeneCounts --limitBAMsortRAM 6534549647 --twopassMode Basic
ulimit -n 10000

#samtools sort
samtools sort -m 1G -@ 46 MYAD2_26.04.19Aligned.sortedByCoord.out.bam -o
MYAD2_26.04.19Aligned.sortedByCoord.out_sorted.bam
samtools sort -m 1G -@ 46 MYAD3_26.04.19Aligned.sortedByCoord.out.bam -o
MYAD3_26.04.19Aligned.sortedByCoord.out_sorted.bam

#samtools index
samtools index MYAD2_26.04.19Aligned.sortedByCoord.out.bam
samtools index MYAD3_26.04.19Aligned.sortedByCoord.out.bam

#coverage
bamCoverage --bam MYAD2_26.04.19Aligned.sortedByCoord.out.bam -o MYAD2@250bp.bdg --binSize
250 --normalizeUsing BPM --outFileFormat bedgraph -p 46
bamCoverage --bam MYAD3_26.04.19Aligned.sortedByCoord.out.bam -o MYAD3@250bp.bdg --binSize
250 --normalizeUsing BPM --outFileFormat bedgraph -p 46
```

List of publications

Research articles and reviews

1. Chatterjee, G., S. R. Sankaranarayanan, **Guin**, **K**., Y. Thattikota, S. Padmanabhan, R. Siddharthan and K. Sanyal (2016). Repeat-associated fission yeast-like regional centromeres in the ascomycetous budding yeast *Candida tropicalis*. *PLoS Genet.* 12(2): e1005839. DOI:10.1371/journal.pgen.1005839

2. Yadav, V., Sreekumar, L., **Guin, K**., & Sanyal, K. (2018). Five pillars of centromeric chromatin in fungal pathogens (PEARLS Review). *PLoS Pathog.* 14(8): e1007150 DOI:10.1371/journal.ppat.1007150

3. Sreekumar, L., K. Kumari, Bakshi, A., Varshney, N., Thimmappa, B. C., **Guin, K**., Narlikar, L., Padinhateeri, R., Siddharthan R., and Sanyal, K. (2019). Orc4 spatiotemporally stabilizes centromeric chromatin. bioRxiv. 465880. DOI**:** https://doi.org/10.1101/465880

4. **Guin, K**., Chen, Y., Mishra, R., Muzaki, S.R.B., Thimmappa, B.C., O'Brien, C.E., Butler, G., Sanyal, A. and Sanyal, K., (2020). Spatial inter-centromeric interactions facilitated the emergence of evolutionary new centromeres. bioRxiv. 2020.02.07.938175 DOI:10.1101/2020.02.07.938175

5. **Guin, K**., Sreekumar L., & Sanyal, K. (*in press*) Implications of the evolutionary trajectory of centromeres in the fungal kingdom. *Ann. Rev. Microbiol.*

Book Chapter

1. Sridhar, S., A. Dumbrepatil, L. Sreekumar, S. R. Sankaranarayanan, Guin, K., and Sanyal, K. (2017) Centromere and kinetochore: Essential components for chromosome segregation in *Gene Regulation, Epigenetics and Hormone Signaling*, edited by S. S. Mandal. DOI:10.1002/9783527697274.ch9

Spatial inter-centromeric interactions facilitated the emergence of evolutionary new centromeres

Krishnendu Guin[1], Yao Chen[2], Radha Mishra[1], Siti Rawaidah B. M. Muzaki[2], Bhagya C. Thimmappa[1], Caoimhe E. O'Brien[3], Geraldine Butler[3], Amartya Sanyal[2*], Kaustuv Sanyal[1, 4*]

[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]School of Biomolecular & Biomed Science, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. [4]Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka 565 0871, Japan

*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.
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, QC H3T 1J4, Canada.
Radha Mishra, Department of Cellular & Molecular Medicine, University of Ottawa, ON K1H 8M5, Canada.

Classification:
Biological Science, Genetics

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

Implications of the Evolutionary Trajectory of Centromeres in the Fungal Kingdom

Krishnendu Guin[1#], Lakshmi Sreekumar[1#], Kaustuv Sanyal[1*]

[1] Molecular Mycology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India

[#] These authors contributed equally to this article.

Email ID and ORCID Number:

Krishnendu Guin krishnendu@jncasr.ac.in (0000-0001-6957-465X)
Lakshmi Sreekumar lsree@jncasr.ac.in (0000-0002-1849-4374)
Kaustuv Sanyal sanyal@jncasr.ac.in (0000-002-6611-4073)

*corresponding author

Kaustuv Sanyal
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
Homepage: http://www.jncasr.ac.in/sanyal

Keywords: CENP-A, centromere clustering, chromosome conformation capture, karyotype evolution, heterochromatin, RNAi

PLOS | GENETICS

RESEARCH ARTICLE

Repeat-Associated Fission Yeast-Like Regional Centromeres in the Ascomycetous Budding Yeast *Candida tropicalis*

Gautam Chatterjee[1¤a], Sundar Ram Sankaranarayanan[1], Krishnendu Guin[1], Yogitha Thattikota[1¤b], Sreedevi Padmanabhan[1¤c], Rahul Siddharthan[2], Kaustuv Sanyal[1]*

1 Molecular Mycology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, India, **2** The Institute of Mathematical Sciences, C.I.T. Campus, Taramani, Chennai, India

¤a Current address: Department of Agricultural Biotechnology, Faculty Centre of Integrated Rural Development and Management, Ramakrishna Mission Vivekananda University, Narendrapur, Kolkata, India
¤b Current address: Institute for Research in Immunology and Cancer, Université de Montréal, Station Centre-Ville, Montreal, Canada
¤c Current address: Molecular Biology Laboratory, Department of Biotechnology, Veer Bahadur Singh Purvanchal University, Jaunpur, India
* sanyal@jncasr.ac.in

OPEN ACCESS

Citation: Chatterjee G, Sankaranarayanan SR, Guin K, Thattikota Y, Padmanabhan S, Siddharthan R, et al. (2016) Repeat-Associated Fission Yeast-Like Regional Centromeres in the Ascomycetous Budding Yeast *Candida tropicalis*. PLoS Genet 12(2): e1005839. doi:10.1371/journal.pgen.1005839

Editor: Beth A. Sullivan, Duke University, UNITED STATES

Received: August 4, 2015

Accepted: January 11, 2016

Published: February 4, 2016

Data Availability Statement: The sequencing data generated by the ChIP-seq experiments in this study have been submitted to the NCBI under the accession number SUB432163. The sequencing data generated by Sanger sequencing of Scnt 1, Scnt 7 and Scnt 9 have been deposited to the NCBI with following accession numbers KJ398406, KJ425116 and KJ398405 respectively.

Funding: This work is supported by a grant from DBT (Grant number: BT/PR14840/BRB/10/880/2010), Govt. of India and intramural funding of JNCASR to KS. RS acknowledges the PRISM project

Abstract

The centromere, on which kinetochore proteins assemble, ensures precise chromosome segregation. Centromeres are largely specified by the histone H3 variant CENP-A (also known as Cse4 in yeasts). Structurally, centromere DNA sequences are highly diverse in nature. However, the evolutionary consequence of these structural diversities on *de novo* CENP-A chromatin formation remains elusive. Here, we report the identification of centromeres, as the binding sites of four evolutionarily conserved kinetochore proteins, in the human pathogenic budding yeast *Candida tropicalis*. Each of the seven centromeres comprises a 2 to 5 kb non-repetitive *mid* core flanked by 2 to 5 kb inverted repeats. The repeat-associated centromeres of *C*. *tropicalis* all share a high degree of sequence conservation with each other and are strikingly diverged from the unique and mostly non-repetitive centromeres of related *Candida* species—*Candida albicans*, *Candida dubliniensis*, and *Candida lusitaniae*. Using a plasmid-based assay, we further demonstrate that pericentric inverted repeats and the underlying DNA sequence provide a structural determinant in CENP-A recruitment in *C*. *tropicalis*, as opposed to epigenetically regulated CENP-A loading at centromeres in *C*. *albicans*. Thus, the centromere structure and its influence on *de novo* CENP-A recruitment has been significantly rewired in closely related *Candida* species. Strikingly, the centromere structural properties along with role of pericentric repeats in *de novo* CENP-A loading in *C*. *tropicalis* are more reminiscent to those of the distantly related fission yeast *Schizosaccharomyces pombe*. Taken together, we demonstrate, for the first time, fission yeast-like repeat-associated centromeres in an ascomycetous budding yeast.

9
Centromere and Kinetochore: Essential Components for Chromosome Segregation

Shreyas Sridhar, Arti Dumbrepatil, Lakshmi Sreekumar, Sundar Ram Sankaranarayanan, Krishnendu Guin, and Kaustuv Sanyal

Jawaharlal Nehru Centre for Advanced Scientific Research, Molecular Biology and Genetics Unit, Molecular Mycology Laboratory, Jakkur, Bangalore 560 064 India

9.1
Introduction

Perpetuation of life occurs by the fundamental property of cells to divide. A somatic cell undergoes a cell cycle that is comprised of essentially two periods: interphase and mitosis. Interphase can be further divided into G1, S, and G2. G1 and G2 constitute gap phases, involving cell growth that prepare cells for genome duplication in synthesis (S) phase and subsequent segregation in mitotic (M) phase, respectively. The mitotic cell cycle ensures equal division of the duplicated genetic content of the mother nucleus with the help of the kinetochore and centromere. The kinetochore is a proteinaceous structure that assembles on centromere (*CEN*) DNA. The centromere/kinetochore generally appears as a constricted region of a metaphase chromosome (Figure 9.1a). The kinetochore complex interacts with microtubules on one side and centromeric chromatin on the other (Figure 9.1a). In most metazoans, multiple microtubules bind to each kinetochore, with an exception of certain budding yeasts where only a single microtubule appears to be associated with each kinetochore [1–4].

Apart from these general features of mitosis, organism-specific variations also exist. Mitosis is broadly classified in two types: closed mitosis and open mitosis (Figure 9.1b). This distinction primarily refers to the permeability of the nuclear envelope (NE), a bilayered membrane which along with the nuclear pore complexes (NPCs) regulate the entry and exit of molecules to and from the nucleus. Closed mitosis is considered to be the more primitive form of eukaryotic cell division, whereas open mitosis seems to have appeared several times during evolution. Plants and animals share open mitosis predominantly, while most fungi employ closed mitosis and variations of it. During closed mitosis, the NE

Gene Regulation, Epigenetics, and Hormone Signaling, First Edition. Edited by Subhrangsu S. Mandal.

PLOS | PATHOGENS

PEARLS

Five pillars of centromeric chromatin in fungal pathogens

Vikas Yadav, Lakshmi Sreekumar, Krishnendu Guin, Kaustuv Sanyal*

Molecular Mycology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, India

* sanyal@jncasr.ac.in

Citation: Yadav V, Sreekumar L, Guin K, Sanyal K (2018) Five pillars of centromeric chromatin in fungal pathogens. PLoS Pathog 14(8): e1007150. https://doi.org/10.1371/journal.ppat.1007150

Editor: Donald C Sheppard, McGill University, CANADA

Published: August 23, 2018

Funding: The Senior Research Fellowship from the Council of Scientific and Industrial Research (CSIR), Government of India (grant number 09/733 (0179)/2012/EMR-I) was received by VY. The Senior Research Fellowship from the Council of Scientific and Industrial Research (CSIR), Government of India (grant number 09/733(0178)/ 2012-EMR-I) was received by LSK. The Shyama Prasad Mukherjee Fellowship from the Council of Scientific and Industrial Research (CSIR), Government of India (grant number 07/733(0181)/ 2013-EMR-I) was received by KG. The Tata innovation fellowship (grant number BT/HRT/35/ 01/03/2017) was received by KS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

"*The greater the diversity, the greater the perfection.*"

—*Thomas Berry*

A centromere is classically defined as the primary constriction on a metaphase chromosome [1] that holds the sister chromatids together, binds to spindle microtubules, and brings about their separation during anaphase. Despite having a conserved and essential function, centromeres are among the fastest evolving DNA sequence loci in eukaryotic genomes [2]. With the advent of molecular biology techniques, centromeres could be mapped and sequenced in a large number of fungal species. The length of centromere DNA in fungi is found to be highly variable, classifying them as point (<400 bp), short regional (>400 bp, –20 kb), and large regional (>20 kb) [3]. Such diversity is achieved by different regulatory factors that have overlapping functions required for loading of the centromere-specific histone H3 variant centromere protein A/chromosome segregation 4(CENP-A/Cse4) to DNA to define centromere identity. Although genetic and epigenetic mechanisms of centromere formation across eukaryotes are largely conserved, there are examples of molecular innovation and genetic improvisation that help fungal species to maintain their ploidy across generations. In this review, we highlight five such genetic and epigenetic factors that define centromere identity in pathogenic fungi.

DNA sequence and organization of DNA sequence elements

DNA sequence features provide the necessary template to act as a binding platform for kinetochore proteins. The genus *Candida*, which harbors several pathogenic species, presents a diverse array of centromere types. *Candida glabrata* carries point genetic centromeres, much like the 125-bp DNA sequence that serves as a fully functional point centromere in the budding yeast *Saccharomyces cerevisiae* [3–5]. Typically, genetic centromeres have specific and conserved DNA sequence motifs and confer mitotic stability to otherwise unstable plasmids carrying an autonomous replicating sequence (ARS) during cell division. Despite high-structural homology in DNA sequence elements, the point centromeres of *C. glabrata* are not fully functional in *S. cerevisiae*, suggesting that centromere function is species-specific [5, 6]. Short regional genetic centromeres of *Candida tropicalis* comprise a central core flanked by inverted repeats, similar to those of the fission yeast *Schizosaccharomyces pombe* [3, 7]. The sequence and orientation of these repeats are important for centromere function. Due to the presence of inverted repeats, the centromeres in *C. tropicalis* can acquire a hairpin loop-like secondary structure that might be crucial for kinetochore assembly. *Candida albicans* and *Candida dubliniensis*, on the other hand, possess unique and different centromere DNA sequences on each of their chromosomes [8, 9]. While *C. tropicalis* centromeres can stabilize an ARS plasmid,

Orc4 spatiotemporally stabilizes centromeric chromatin

Lakshmi Sreekumar[1], Kiran Kumari[2,3,4], Asif Bakshi[1], Neha Varshney[1], Bhagya C. Thimmappa[1], Krishnendu Guin[1], Leelavati Narlikar[5], Ranjith Padinhateeri[2], Rahul Siddharthan[6], Kaustuv Sanyal[1]

[1]Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India; [2]Department of Biosciences and Bioengineering, Indian Institute of Technology, Bombay, Mumbai, India; [3]IITB-Monash Research Academy, Mumbai, India; [4]Department of Chemical Engineering, Monash University, Melbourne, Australia [5]Department of Chemical Engineering, CSIR-National Chemical Laboratory, Pune, India; [6]The Institute of Mathematical Sciences/HBNI, Taramani, Chennai, India

*corresponding author

Kaustuv Sanyal
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
Homepage: http://www.jncasr.ac.in/sanyal

Present address: Asif Bakshi, Laboratory of Drosophila Neural Development, Centre for DNA Fingerprinting and Diagnostics, Inner Ring Road, Uppal, Hyderabad 500039, India

Bhagya C. Thimmappa, Department of Biochemistry, Robert-Cedergren Centre for Bioinformatics and Genomics, University of Montreal, 2900 Edouard-Montpetit, Montreal, H3T1J4, QC, Canada