Candida albicans isolates contain frequent heterozygous structural variants and transposable elements within genes and centromeres

Authors

Ursula Oggenfuss[1], Robert T. Todd[1,2], Natthapon Soisangwan[1], Bailey Kemp[1], Alison Guyer[1], Annette Beach[1], Anna Selmecki[1,#]

1 Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA

2 Department of Biology, Bard College, Annandale-on-Hudson, New York, NY, USA

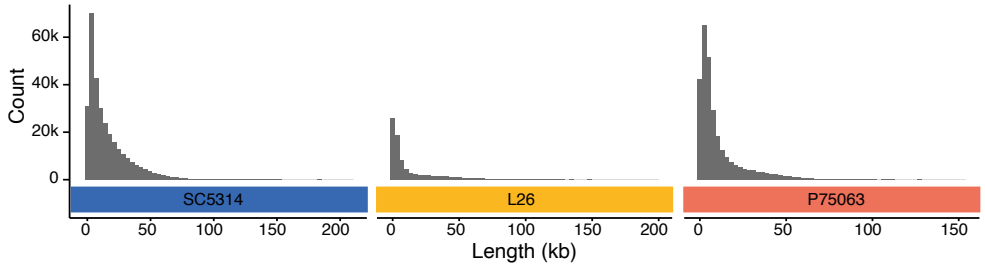
Supplementary Figures and Tables

A

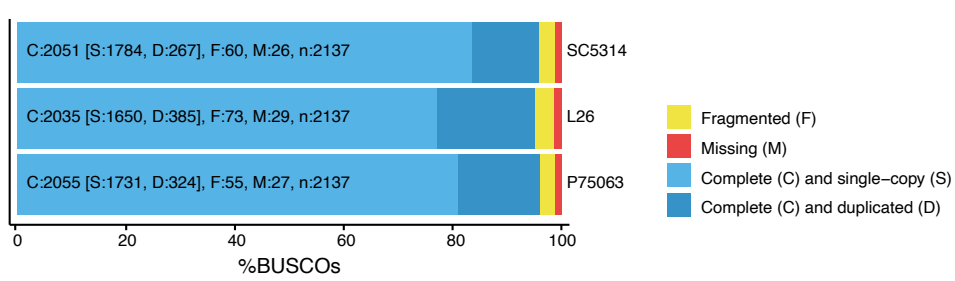
SC5314	contigs
# contigs (>= 0 bp)	37
# contigs (>= 1000 bp)	37
# contigs (>= 5000 bp)	36
# contigs (>= 10000 bp)	36
# contigs (>= 25000 bp)	36
# contigs (>= 50000 bp)	36
Total length (>= 0 bp)	17822073
Total length (>= 0 bp)	17822073
Total length (>= 0 bp)	17817691
Total length (>= 0 bp)	17817691
Total length (>= 0 bp)	17817691
Total length (>= 0 bp)	17817691
# contigs	37
Largest contig	3230252
Total length	17822073
GC (%)	33.92
N50	1162837
N90	141193
auN	1451250.2
L50	5
L90	21
# N's per 100 kbp	0.00

L26	contigs
# contigs (>= 0 bp)	38
# contigs (>= 1000 bp)	38
# contigs (>= 5000 bp)	37
# contigs (>= 10000 bp)	37
# contigs (>= 25000 bp)	37
# contigs (>= 50000 bp)	37
Total length (>= 0 bp)	17971502
Total length (>= 0 bp)	17971502
Total length (>= 0 bp)	17967896
Total length (>= 0 bp)	17967896
Total length (>= 0 bp)	17967896
Total length (>= 0 bp)	17967896
# contigs	38
Largest contig	3194456
Total length	17971502
GC (%)	33.63
N50	1825850
N90	120856
auN	1761115.9
L50	4
L90	19
# N's per 100 kbp	0.00

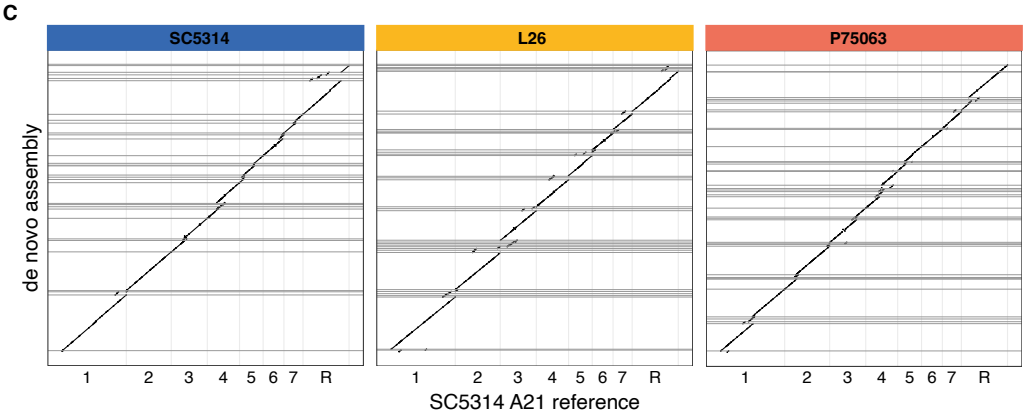
P75063	contigs
# contigs (>= 0 bp)	48
# contigs (>= 1000 bp)	48
# contigs (>= 5000 bp)	48
# contigs (>= 10000 bp)	48
# contigs (>= 25000 bp)	47
# contigs (>= 50000 bp)	44
Total length (>= 0 bp)	18442963
Total length (>= 0 bp)	18442963
Total length (>= 0 bp)	18442963
Total length (>= 0 bp)	18442963
Total length (>= 0 bp)	18432275
Total length (>= 0 bp)	18289107
# contigs	48
Largest contig	1698430
Total length	18442963
GC (%)	33.91
N50	1029629
N90	123931
auN	1010537.6
L50	7
L90	27
# N's per 100 kbp	0.00



B

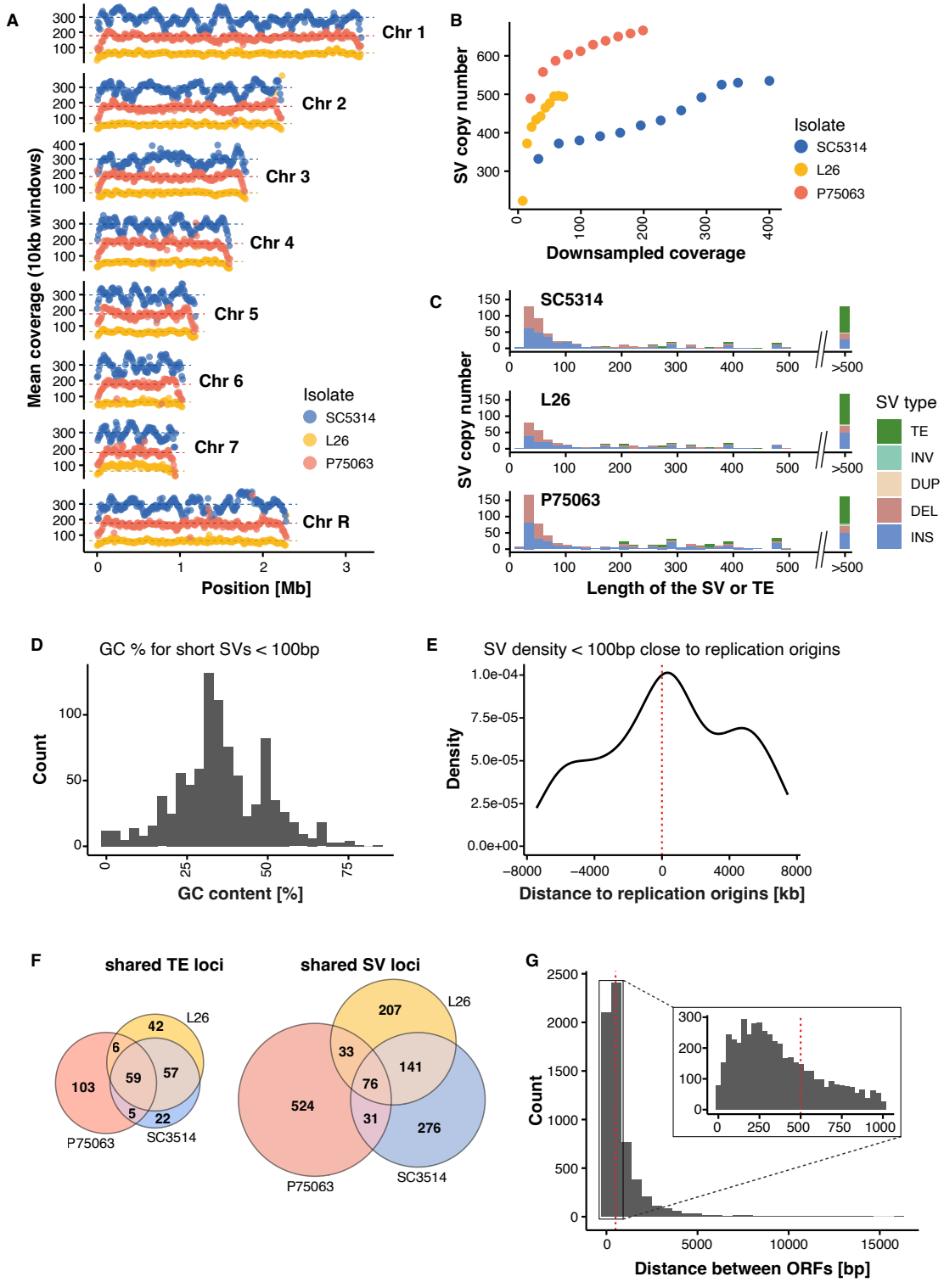


C

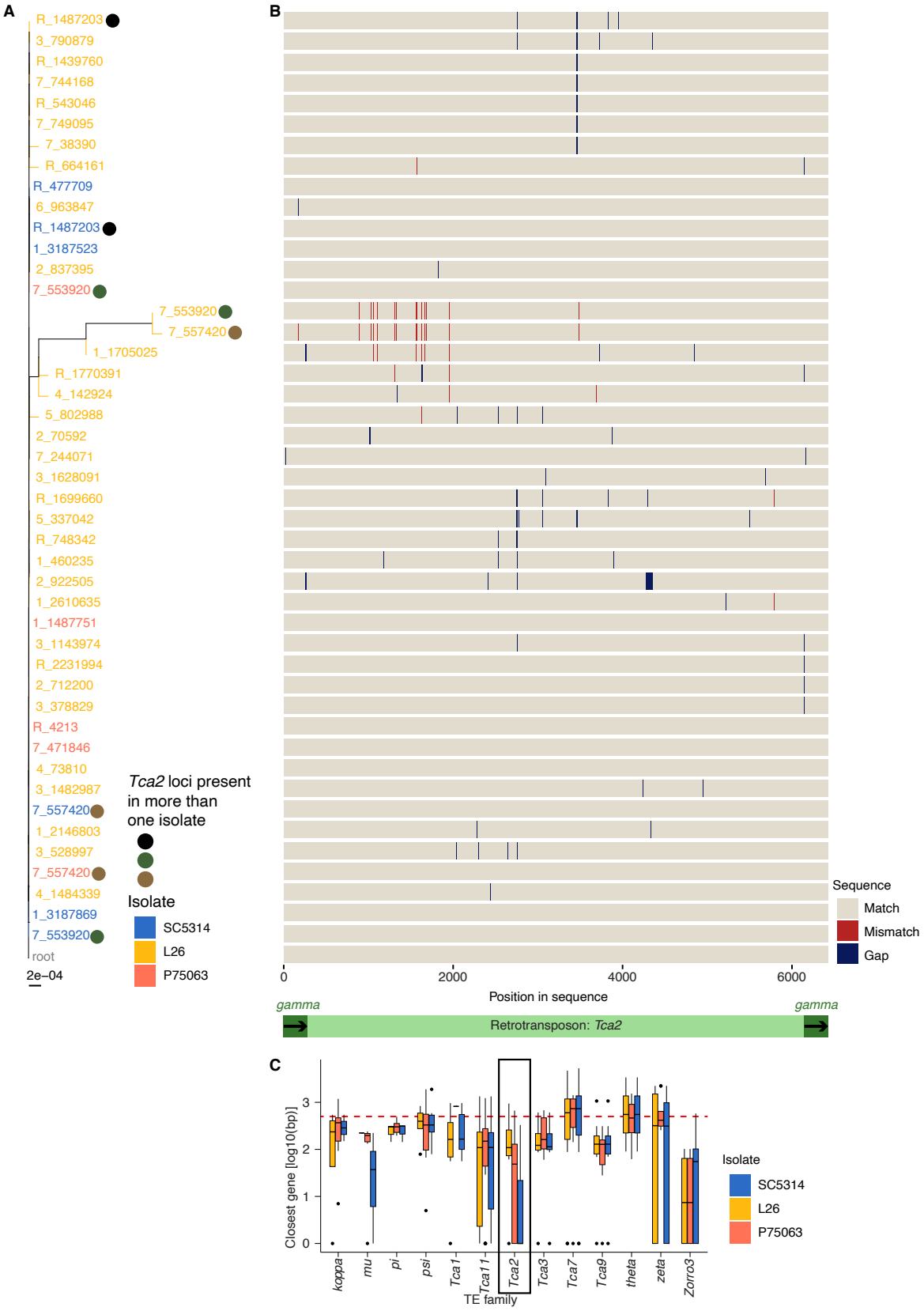


Supplementary Figure S1. ***De novo*** **assembly**. **A)** Assembly quality assessment with QUAST. The Histogram shows read length distribution per isolate. The longest reads have a length of 208.83 kb

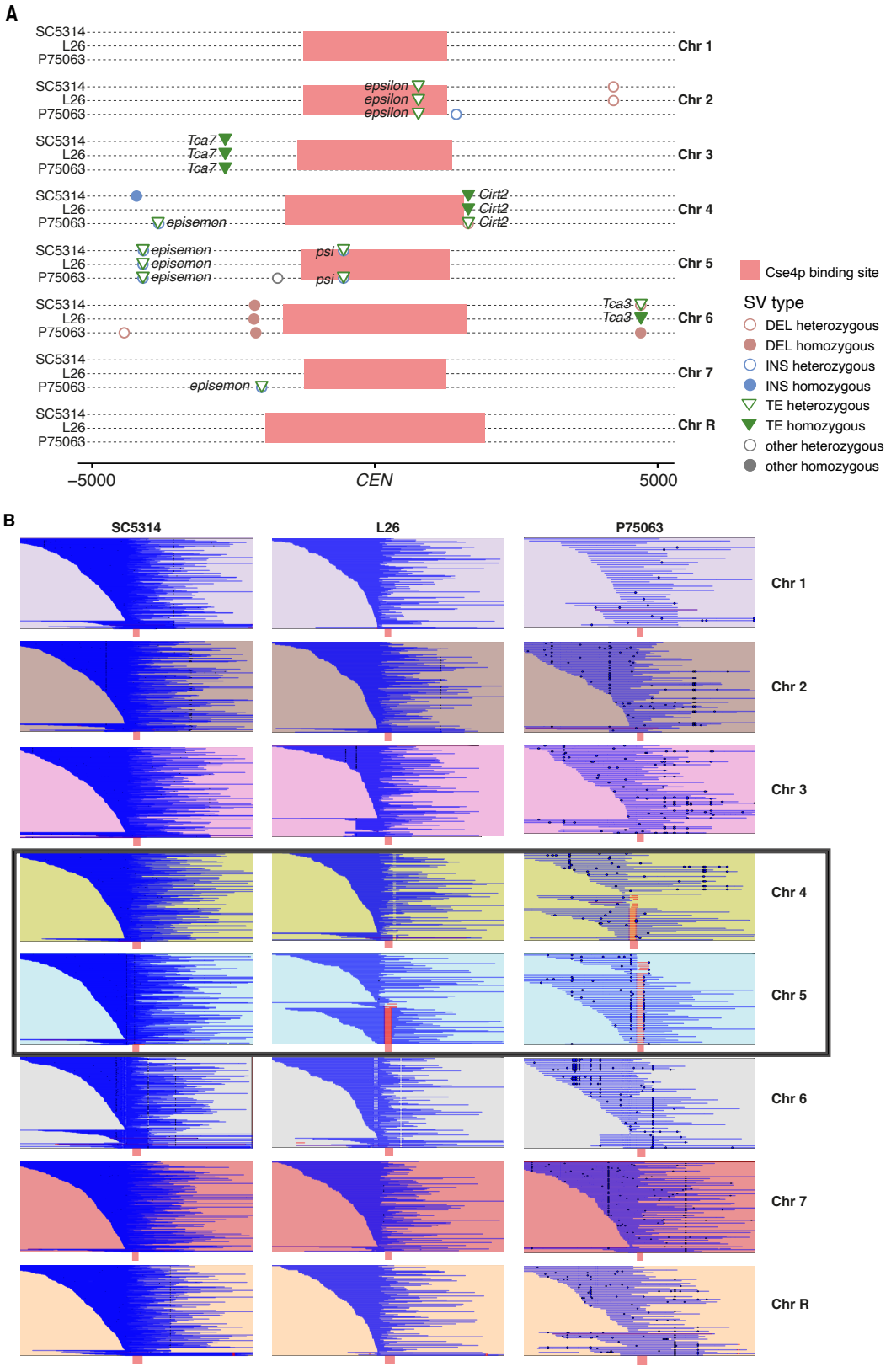
(SC5314), 198.49 kb (L26) and 153.01 kb (P75063). **B)** Genome completeness with BUSCO. BUSCO was tested on 2137 genes c conserved in Saccharomycotina. **C)** Contigs of the de novo assembly aligned to the reference genome SC5314 A21 (x-axis).



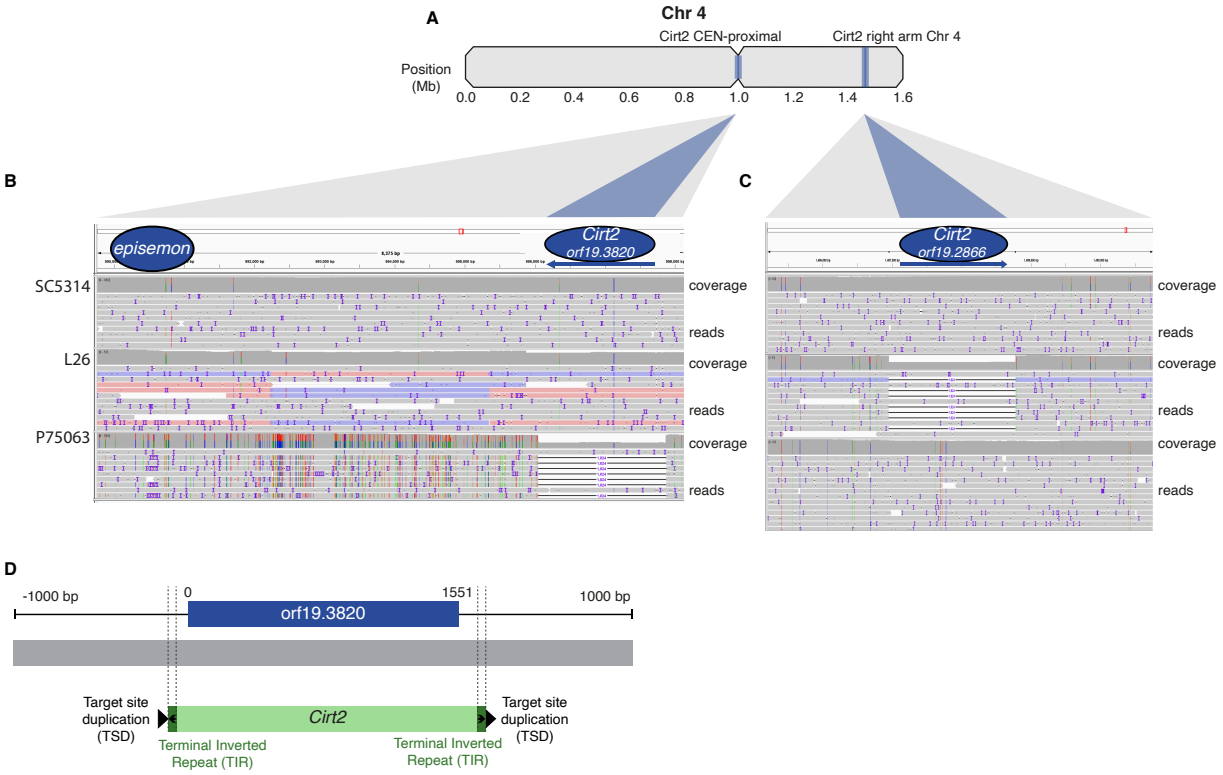
Supplementary Figure S2. Genomic landscape, TEs and SVs in three clinical isolates. A) Mean read depth coverage for 10,000 bp windows for SC5314, L26 and P75063. Dotted lines indicate the genome-wide mean coverage for each genome (73.7x in L26, 331.5x in SC5314 and 203.9x in P75063). **B)** Read depth downsampling analysis. Each dot represents the number of detected SVs with DELLY on a reduced set of reads. **C)** Length distribution of SVs. For SVs, the length is taken from the output of DELLY, for TEs the length of the consensus sequence. Colors by SV type (green: TE = transposable element; turquoise: INV = inversion; apricot: DUP = duplication; red: DEL = deletion; blue: INS = insertion). **D)** GC content distribution for short SVs (< 100 bp) **E)** Density of SVs < 100 bp detected around replication origins in a window of -7.5kb to 7.5kb. Replication origin positions were extracted from CHIPseq data (Tsai et al (2014). **F)** TEs and SV identical in location between the isolates SC5315, L26 and P75063. **G)** Distribution of distance between ORFs. The red line indicates a distance of 500bp. The box shows a more fine-scale zoom in to a distance of 100 bp.


A
R_1487203
3_790879
R_1439760
7_744168
R_543046
7_749095
7_38390
R_664161
R_477709
6_963847
R_1487203
1_3187523
2_837395
7_553920
7_553920
7_557420
1_1705025
R_1770391
4_142924
5_802988
2_70592
7_244071
3_1628091
R_1699660
5_337042
R_748342
1_460235
2_922505
1_2610635
1_1487751
3_1143974
R_2231994
2_712200
3_378829
R_4213
7_471846
4_73810
3_1482987
7_557420
1_2146803
3_528997
7_557420
4_1484339
1_3187869
7_553920
root
2e−04
Tca2 loci present in more than one isolate
Isolate
SC5314
L26
P75063
B
0
2000
4000
6000
Position in sequence
gamma
gamma
Retrotransposon: Tca2
Sequence
Match
Mismatch
Gap
C
Closest gene [log10(bp)]
0
1
2
3
koppa
mu
pi
psi
Tca1
Tca11
Tca2
Tca3
Tca7
Tca9
theta
zeta
Zorro3
TE family
Isolate
SC5314
L26
P75063

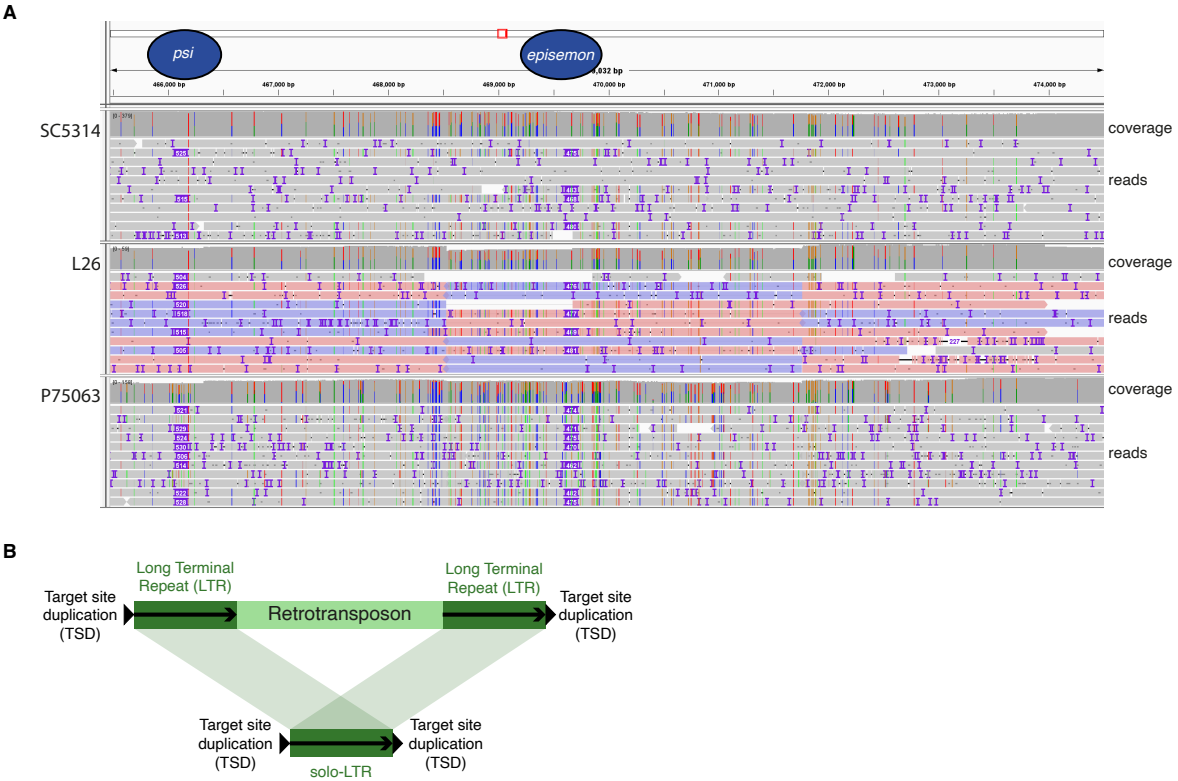
Supplementary Figure S3. Positions and schematic of the *Tca2* TE. A) Phylogenetic tree based on the multiple sequence alignment. Circles indicate copies that are present in more than one isolate (Fixed = present in all isolates; 2 copies = present in two isolates). Colors indicate the isolate. The tree is rooted with the consensus sequence. **B)** Multiple sequence alignment for the whole sequence of *Tca2*. Colors indicate sequence similarity (grey = match; red = mismatch; blue = gap). The schematic at the bottom represents *Tca2* and its *gamma* LTRs **C)** Distance to the closest annotated ORF for each TE copy. Only TE families with ten or more copies were used. Colors indicate the isolate.


A
SC5314
L26
P75063
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr R
epsilon
Tca7
Cirt2
episemon
psi
Tca3
–5000
CEN
5000
Cse4p binding site
SV type
DEL heterozygous
DEL homozygous
INS heterozygous
INS homozygous
TE heterozygous
TE homozygous
other heterozygous
other homozygous
B
SC5314
L26
P75063
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr R

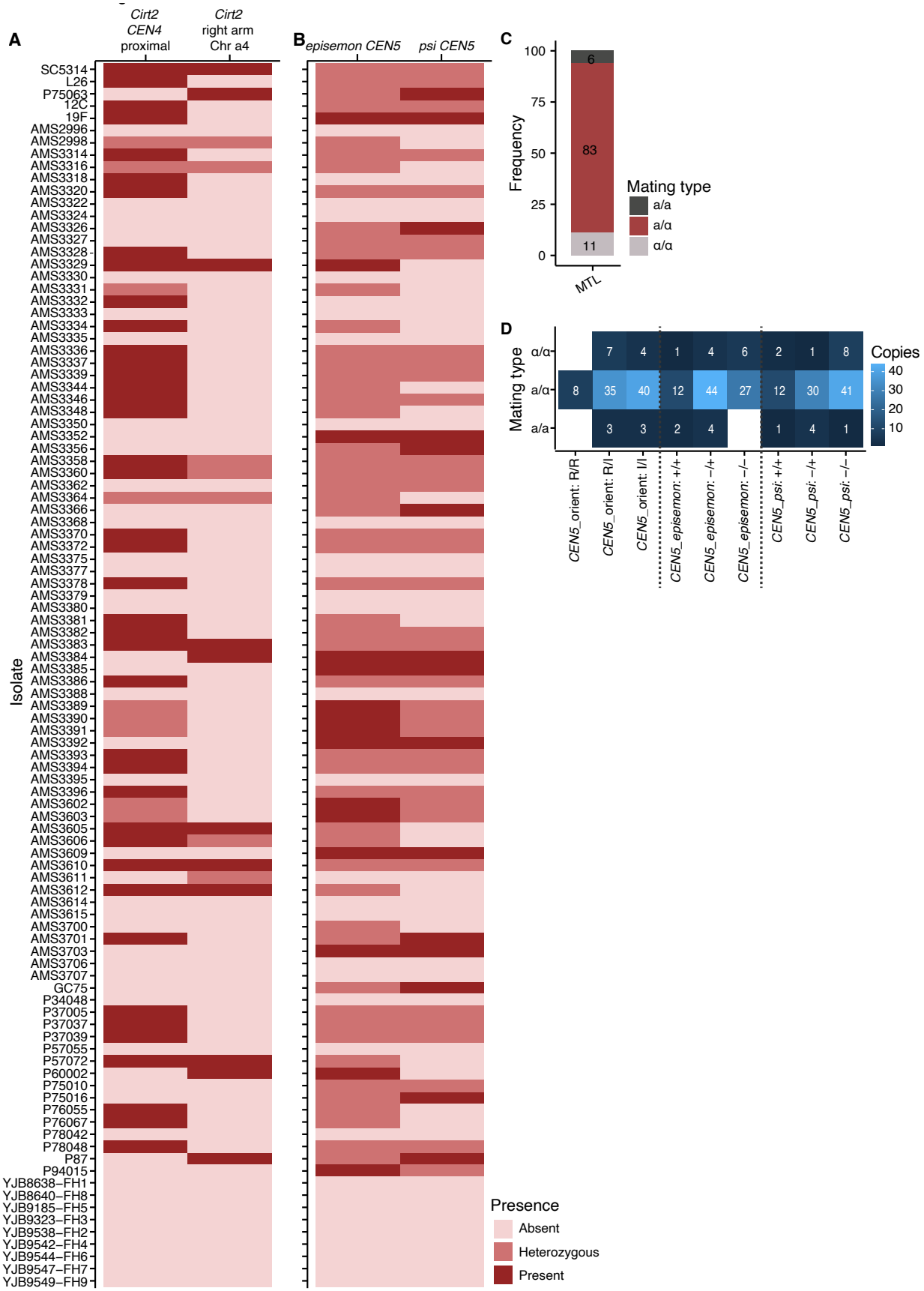
Supplementary Figure S4. SVs at *CEN* positions. **A)** Schematic of SVs at each Cse4p binding site (pink region) and each 5000 bp upstream and downstream. Figure adapted from main Figure 1. Colors indicate the SV type in comparison with the reference genome SC5314 A21 (green: TE = transposable element; red: DEL = deletion; blue: INS = insertion, grey = other types of SV (inversions, duplications, translocations)). The fill represents zygosity (filled symbol = homozygosity, unfilled symbol = heterozygosity). The names of TE families are added. **B)** Ribbon plots for each *CEN* position for each isolate. The pink box represents the Cse4p binding site for each chromosome. Cse4p/CENP-A binding delineates the central core sequence of the centromere (Sanyal et al. 2004; Ketel et al. 2009).



Supplementary Figure S5. Positions and schematic of the *Cirt2* TE. A) Schematic of chr4 and *Cirt2* positions (blue lines). **B)** Mapped reads (IGV screenshot) of *CEN4* and nearby features for a subset of SC5314, L26, and P75063 reads. The *CEN4* solo-LTR *episemon* and *Cirt2* are indicated above the reads. **C)** Mapped reads in the region around *orf19.2866*, right arm copy of *Cirt2* in SC5314, L26, and P75063. The chr4:1,486,953-1,488,776 deletion ("*orf19.2866* deletion") is indicated above the reads, which is 1,824 bp and spans the annotated *orf19.2866* plus some bases upstream and downstream. **D)** Schematic of the *Cirt2* transposason, that contains a transposase as a coding region (blue), terminal inverted repeats at both sides of the element (green), and is flanked by a short target site duplication (typically TA; black).



Supplementary Figure S6. *CEN5* structural variants depicted by whole genome sequencing (WGS) visualization. A) Mapped reads (IGV screenshot) of *CEN5* and nearby features for SC5314, L26, and P75063. The *CEN5* solo-LTRs are indicated in the track above the reads. The locations of the ~470 bp *psi* and ~523 bp *episemon* solo-LTRs are indicated above the reads with a triangle and in the reads with a purple bar. **B)** Schematic of retrotransposon non-homologous recombination that leads to the presence of solo-LTRs. Full length LTR-retrotransposons contain two long terminal repeats (LTR; dark green), and superfamily specific coding regions (bright green). Upon insertion, a short target site duplication (TSD) is created. After the non-homologous recombination, only the TSD and a single copy of the LTR remains. Non-homologous recombination can also happen between two full length copies or between two LTRs of independent copies, leading to a deletion beyond the TE.


A
Cirt2
CEN4
proximal
Cirt2
right arm
Chr a4
B
episemon CEN5
psi CEN5
C
Frequency
100
75
50
25
0
6
83
11
MTL
Mating type
a/a
a/α
α/α
D
Mating type
α/α
a/α
a/a
7 4 1 4 6 2 1 8
8 35 40 12 44 27 12 30 41
3 3 2 4 1 4 1
CEN5_orient: R/R
CEN5_orient: R/I
CEN5_orient: I/I
CEN5_episemon: +/+
CEN5_episemon: −/+
CEN5_episemon: −/−
CEN5_psi: +/+
CEN5_psi: −/+
CEN5_psi: −/−
Copies
40
30
20
10
Isolate
SC5314
L26
P75063
12C
19F
AMS2996
AMS2998
AMS3314
AMS3316
AMS3318
AMS3320
AMS3322
AMS3324
AMS3326
AMS3327
AMS3328
AMS3329
AMS3330
AMS3331
AMS3332
AMS3333
AMS3334
AMS3335
AMS3336
AMS3337
AMS3339
AMS3344
AMS3346
AMS3348
AMS3350
AMS3352
AMS3356
AMS3358
AMS3360
AMS3362
AMS3364
AMS3366
AMS3368
AMS3370
AMS3372
AMS3375
AMS3377
AMS3378
AMS3379
AMS3380
AMS3381
AMS3382
AMS3383
AMS3384
AMS3385
AMS3386
AMS3388
AMS3389
AMS3390
AMS3391
AMS3392
AMS3393
AMS3394
AMS3395
AMS3396
AMS3602
AMS3603
AMS3605
AMS3606
AMS3609
AMS3610
AMS3611
AMS3612
AMS3614
AMS3615
AMS3700
AMS3701
AMS3703
AMS3706
AMS3707
GC75
P34048
P37005
P37037
P37039
P57055
P57072
P60002
P75010
P75016
P76055
P76067
P78042
P78048
P87
P94015
YJB8638−FH1
YJB8640−FH8
YJB9185−FH5
YJB9323−FH3
YJB9538−FH2
YJB9542−FH4
YJB9544−FH6
YJB9547−FH7
YJB9549−FH9
Presence
Absent
Heterozygous
Present

Supplementary Figure S7. Clinical isolates. A) Presence/absence and zygosity of *Cirt2* in the pericentromeric region and on the right arm of chr4. **B)** Mating type distribution of the 100 clinical isolates. **C)** Presence/absence and zygosity of the solo-LTRs *episemon* and *psi* in the pericentromeric region of chr5. **D)** Distribution of mating types over the different pericentromeric TEs.