Supplemental Material

Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair

McGinty, et al.

Contents:

Supplemental Methods

DNA extraction:

The DNA extraction protocol used is a slight modification of a classic ethanol precipitation. First, yeast cultures were grown overnight in 2 ml of complete media (YPD) and refreshed for 4 hours in an additional 8 ml YPD to achieve logarithmic growth. The cultures were spun down, and the pellets were resuspended in 290 µl of solution containing 0.9M Sorbitol and 0.1M EDTA at pH 7.5. 10 µl lyticase enzyme was added, and the mixture was incubated for 30 minutes at 37ºC in order to break down the yeast cell wall. The mixture was centrifuged at 8,000 rpm for two minutes, and the pellet was resuspended in 270 µl of solution containing 50mM Tris 20mM EDTA at pH 7.5, and 30 µl of 10% SDS. Following five minutes incubation at room temperature, 150 µl of chilled 5M potassium acetate solution was added. This mixture was incubated for 10 min at 4ºC and centrifuged for 10 min at 13,000 rpm. The supernatant was then combined with 900 µl of pure ethanol, which had been chilled on ice. This solution was stored overnight at -20ºC, and then centrifuged for 20 minutes at 4,000 rpm in a refrigerated centrifuge. The pellet was then washed twice with 70% ethanol, allowed to dry completely, and then resuspended in 50 µl TE (10mM Tris, 1mM EDTA, pH 8). The resuspended DNA was then treated with 20 µl RNaseA solution, incubated for 10 minutes at 37ºC and 30 minutes at room temperature. The above ethanol precipitation was then repeated in order to remove the digested RNA and enzymes. A gel was prepared with 0.4% agarose and 0.5X TBE. A portion of the sample, along with a high-range DNA ladder (GeneRuler High Range DNA Ladder – Thermo Scientific), was run at very low voltage and in a 4ºC cold room overnight. The resulting gel was stained with ethidium bromide and visualized using a BioRad GelDoc XR. This method of DNA preparation resulted in an average fragment size of 24-48 kb. (Fig. S1A) DNA quantity was measured via Qubit (Qubit dsDNA BR Assay kit – Thermo Scientific) and quality assessed via Nanodrop (Thermo Scientific).

Bioinformatics:

Raw current traces generated by ONT sequencing were basecalled via the Albacore basecalling software (ONT version 2.02), which produced barcode-separated FASTQ files. For the parent strain, reads were then aligned to the S288C reference genome (version R64-1-1, obtained from ensembl.org) using NGM-LR (Sedlazeck et al. 2017). This produced a sequence alignment map (SAM) file, which was then processed into a genome-coordinate-sorted BAM file using Samtools (Li et al. 2009). The BAM file was then analyzed by Sniffles (Sedlazeck et al. 2017). All bioinformatics steps were performed on a single Intel i7-based computer, with parameters set to maximize use of all processing cores. Compute times ranged from hours (Albacore) to minutes (NGM-LR, Samtools, Sniffles).

The Sniffles output and the BAM file were then imported to Ribbon (Nattestad et al.) as well as the bioinformatics software UGENE (Okonechnikov et al. 2012), for visualization. To avoid false positives, each structural variant call from Sniffles was examined in both Ribbon and UGENE. To avoid false negatives, regions with readily-apparent copy-number changes seen in UGENE were also closely examined in Ribbon, by typing in specific genome coordinates from which to select the reads. Sharp breakpoints that lined up in multiple reads, with either end mapping to a consistent region, were considered to be the hallmarks of true CGR events.

This analysis of the parent strain identified various deletions and insertions (Figs. 1B & S3), which were then incorporated into the reference genome. Deletion boundaries were identified via the UGENE alignment viewer and the reference sequence trimmed accordingly. Ty element insertions were initially identified via the UGENE alignment viewer. Sequences at the insertion boundaries were extracted and aligned against a database of Ty element sequences (Carr et al. 2012). A closely-related Ty element sequence, obtained from SGD (yeastgenome.org), was inserted into the reference FASTA at the determined boundaries to approximate the insertion. The inserted sequences were then refined by first re-aligning the Nanopore reads to the newly-created FASTA reference, followed by SNP calling in the region using the Samtools 'mpileup' command, and then incorporation of high-quality SNPs into the FASTA reference via the Bcftools ‘consensus’ command. This process was repeated several times, until no additional SNPs were detected.

After producing a FASTA reference representative of the parent strain, the above analysis pipeline from NGM-LR to Ribbon was repeated for each CGR-containing strain. Single base pair resolution of breakpoints within Ty elements was determined by analysis of SNPs within each Ty element of origin. Reference sequences of Ty elements involved in CGRs were obtained from SGD (yeastgenome.org) and aligned via the MUSCLE algorithm (Edgar 2004), revealing known SNPs. UGENE alignments were then examined for the presence of expected SNPs. SNPs were distinguished from random sequencing errors by the consistent alignment of SNPs in nearly all of the reads. For junctions involving a copy number change, SNPs were expected to appear in a particular proportion of the reads, and groups of SNPs were expected to consistently appear together in the same individual reads.

The length of $(GAA)_n$ repeats in Nanopore sequencing reads were determined as follows: First, reads were aligned to the reference genome, as described above. Reads aligning to the 5’ non-repetitive portion of the *URA3* cassette were then extracted in FASTA format via the UGENE alignment viewer. Note that these FASTA files contained the entire read, rather than just the sequences visible in the alignment viewer. The boundaries of the repeat were then determined by searching the FASTA file for the 5’ and 3’ surrounding non-repetitive sequences. The total number of base pairs between the 5’ and 3’ boundaries was then divided by three to approximate the number of $(GAA)_n$ triplets in each read.

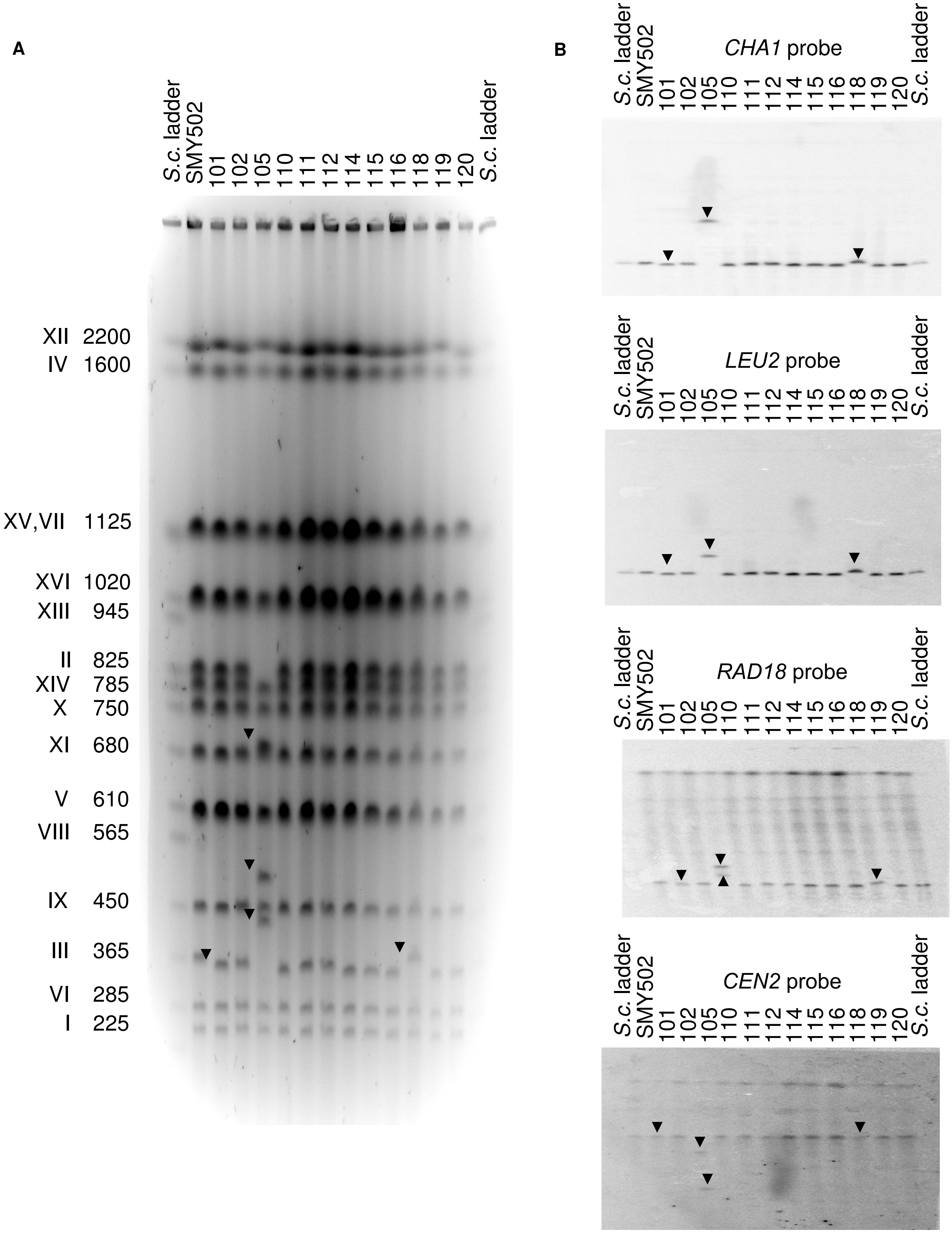
A
S.c. ladder
SMY502
101
102
105
110
111
112
114
115
116
118
119
120
S.c. ladder
XII 2200
IV 1600
XV,VII 1125
XVI 1020
XIII 945
II 825
XIV 785
X 750
XI 680
V 610
VIII 565
IX 450
III 365
VI 285
I 225
B
CHA1 probe
LEU2 probe
RAD18 probe
CEN2 probe

Fig. S1. CHEF gel and Southern analysis of multiple independent strains with $(GAA)_n$-induced chromosome rearrangements. A. EtBr-stained gel. The "*S. c.* ladder" lane contains commercially-available genomic DNA from the strain YY295; the approximate sizes of each chromosome in kb are indicated. SMY502 is the parental strain used in our study, and strains #101-120 are independent Ura^- derivatives. The strains relevant to this analysis are #101, 105, and 118, and the chromosomes relevant to our analysis are shown with arrows. Note that the remaining strains closely resemble strain #101, representing a commonly-observed class of rearrangements. **B.** These portions of the figures represent Southern analyses of the CHEF gels with various hybridization probes. The Southerns probed with *CHA1* and *LEU2* sequences were derived from the gel shown in A.; those probed with *RAD18* and *CEN2* sequences were derived from a different CHEF gel with the same DNA samples. *CHA1* is located on the left arm of Chromosome III, centromere-distal to the location of the $(GAA)_n$ tract. *LEU2* is located on the left arm of Chromosome III, centromere-proximal to the $(GAA)_n$ tract. *RAD18* is located on right arm of Chromosome III. *CEN2* is located at the centromere of Chromosome II.

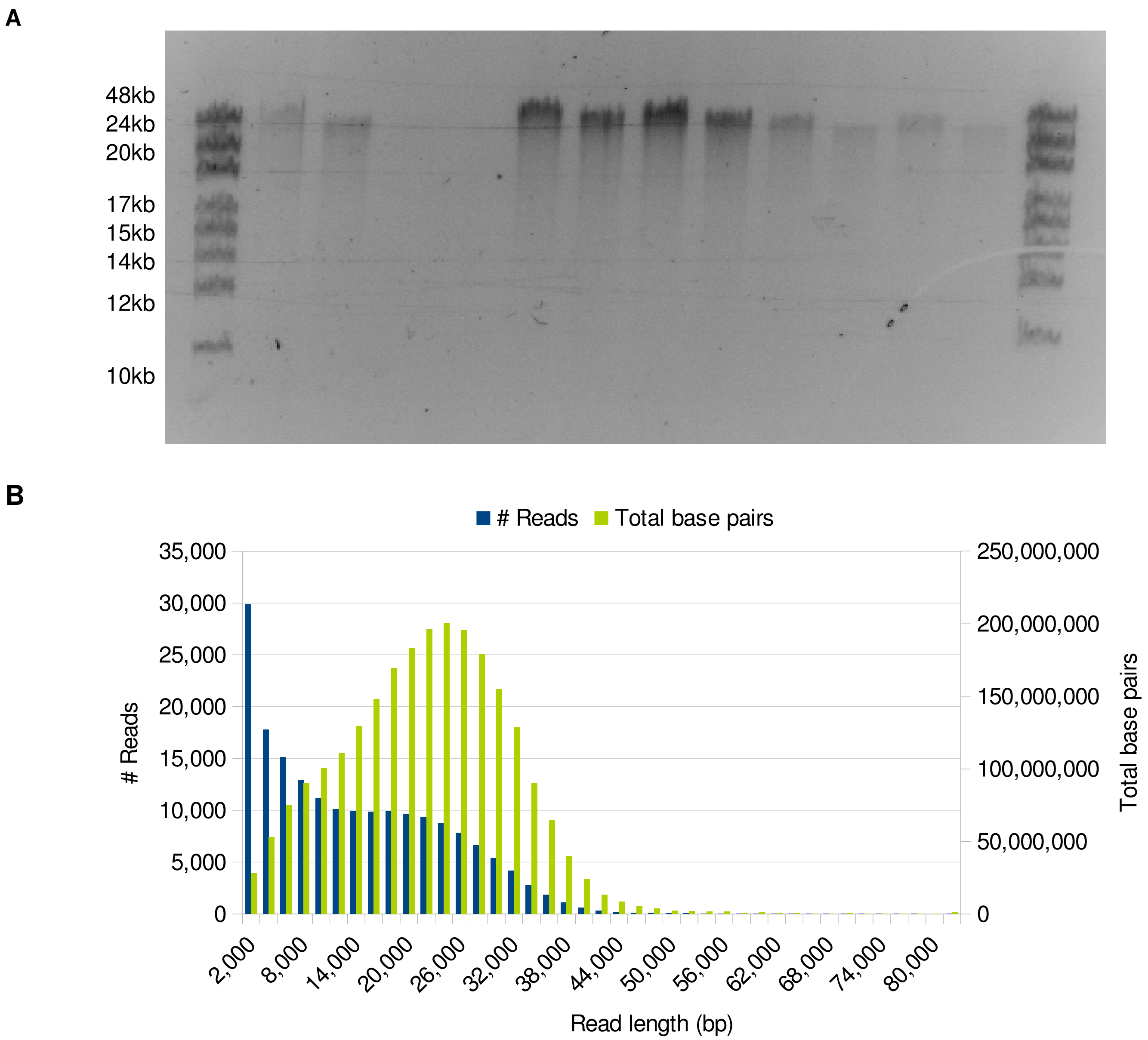
A
48kb
24kb
20kb
17kb
15kb
14kb
12kb
10kb
B
# Reads
Total base pairs
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
# Reads
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
Total base pairs
2,000
8,000
14,000
20,000
26,000
32,000
38,000
44,000
50,000
56,000
62,000
68,000
74,000
80,000
Read length (bp)

Fig. S2. Read lengths vs. DNA extraction. A. Gel electrophoresis following DNA isolation via ethanol precipitation. The gel contains 0.4% agarose and 0.5X TBE, and was run at low voltage overnight at 4ºC. First and last lanes are GeneRuler High Range DNA Ladder (Thermo Scientific). Middle lanes are various DNA samples (not necessarily those that were sequenced). **B.** Read lengths for the MinION sequence output. Blue bars correspond to the number of reads (left Y axis) for each length bin (X axis). Green bars correspond to the total base pair output (right Y axis) for each bin. Note that while shorter reads are common, longer reads contribute more to the overall data set.

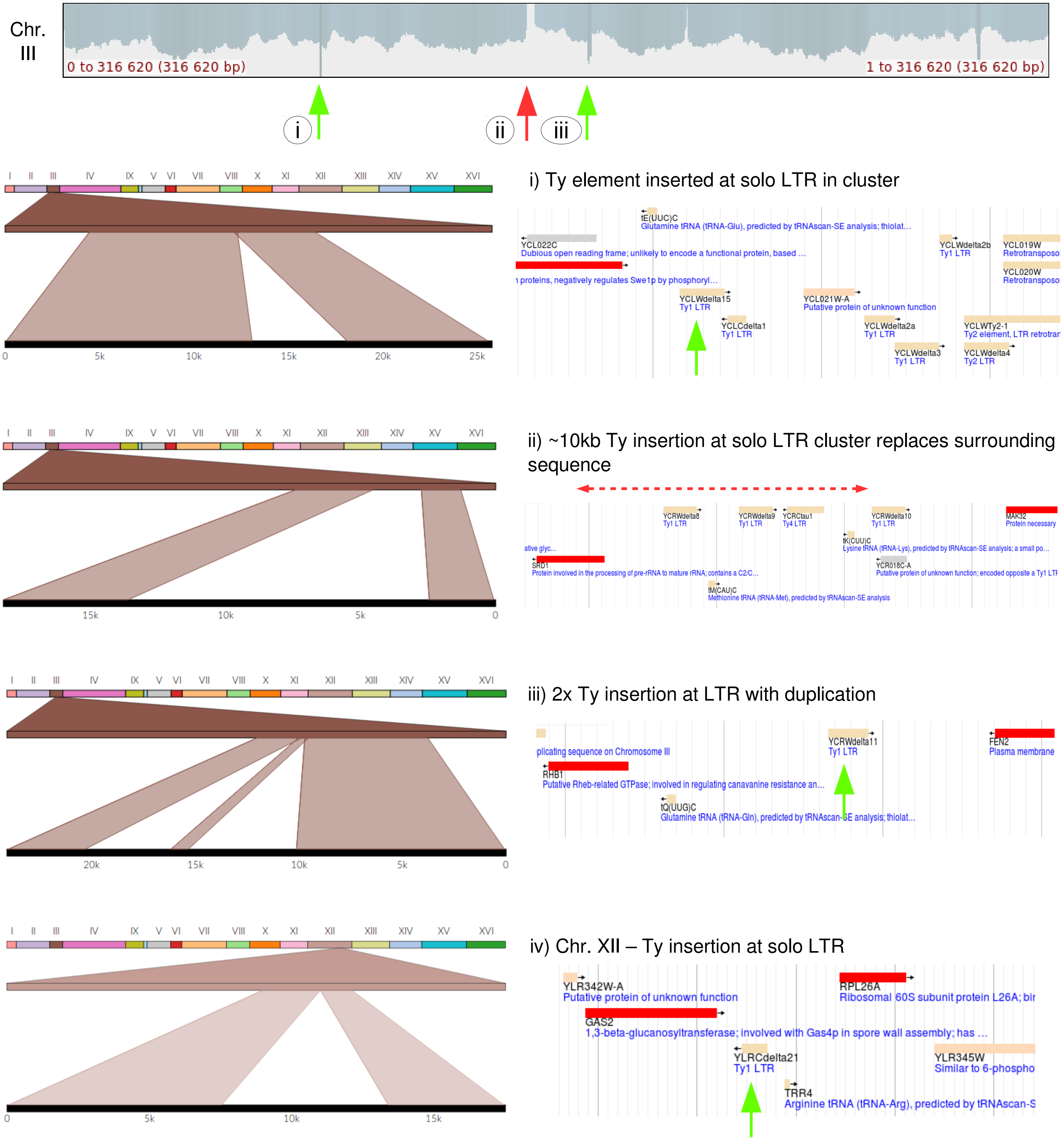

Fig. S3. Identifying novel Ty elements. Top: Nanopore sequencing coverage map of Chromosome III, generated via UGENE, for our starting strain, as aligned to the unaltered S288C reference genome. The chromosome position is represented on the X axis, and the read depth is indicated on the Y axis. Chromosome III contains three out of the four novel Ty elements identified in SMY502 that were not present in S288C. Arrows indicate the position of spikes or gaps in read-depth at the sites of the novel Ty insertions. **Left panels:** Ribbon single-read views highlighting split reads that correspond to each of the indicated regions (i-iii), as well as one additional region on Chromosome XII. Individual reads did not typically identify a particular Ty element as the donor for the insertion, and thus the portion of the read corresponding to the insertion is blank. The error rate of Nanopore sequencing is roughly on par with the divergence of the various Ty elements, some of which are still active and thus nearly identical. All Ty elements begin and end with a long terminal repeat (LTR) sequence of ~340bp, which are generally conserved within each Ty class 1 through 4. The entire Ty element is typically 5-6kb in length. Insertions i and iv are consistent with a single Ty element addition, while insertions ii and iii are consistent with tandem Ty insertions. **Right panels:** SGD Genome Browser views (yeastgenome.org), highlighting the location of each Ty insertion. Green arrows point to the specific LTRs used as insertion points, which were duplicated following the insertion. The red arrow in panel (ii) shows the location that was replaced by the insert, consisting of a cluster of LTR delta elements and some neighboring non-repetitive sequence.

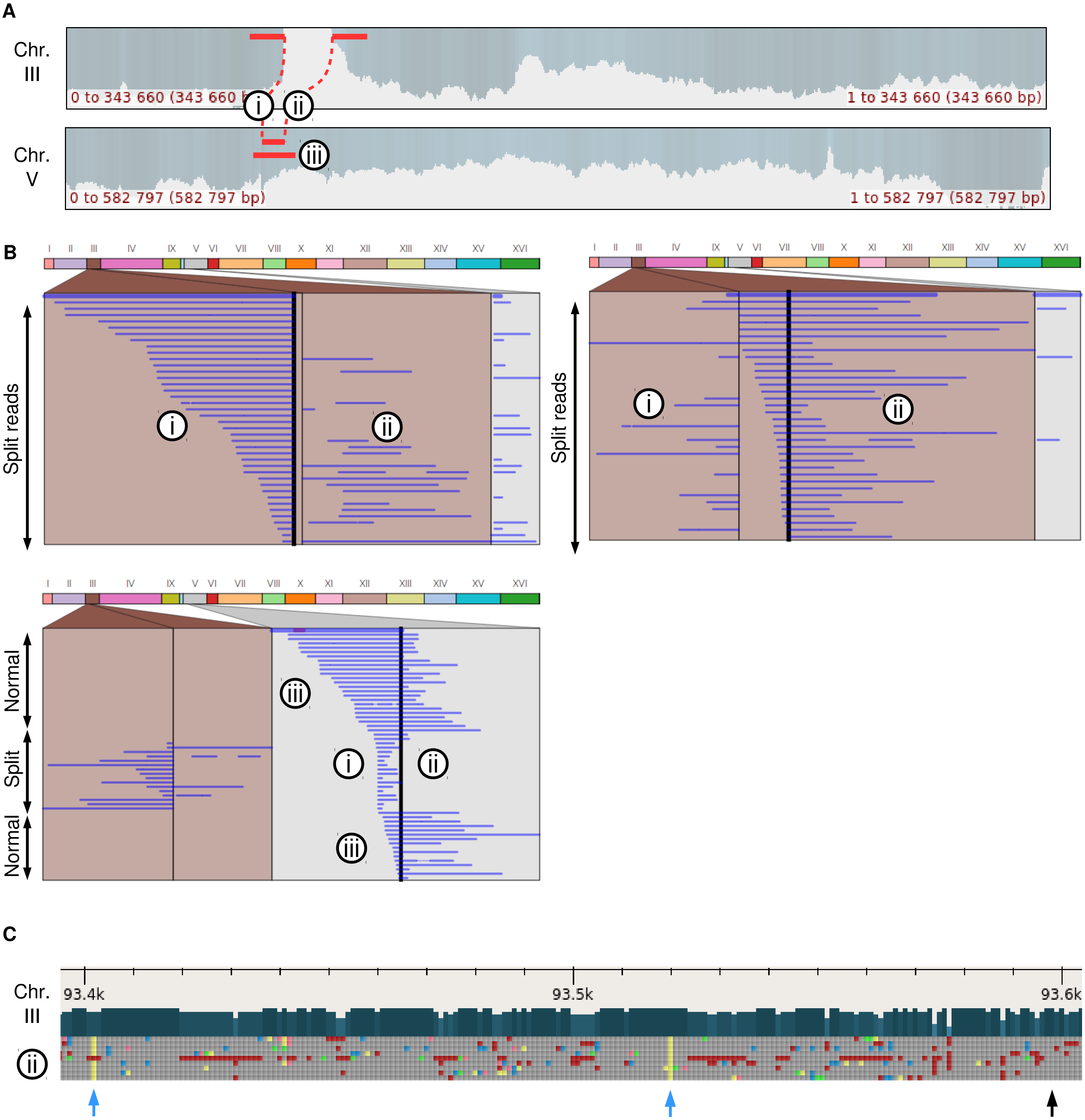
A
Chr. III
0 to 343 660 (343 660 b
1 to 343 660 (343 660 bp)
i
ii
Chr. V
iii
0 to 582 797 (582 797 bp)
1 to 582 797 (582 797 bp)
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Split reads
i
ii
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Split reads
i
ii
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal
Split
Normal
iii
i
ii
iii
C
Chr. III
ii
93.4k
93.5k
93.6k

Fig. S4. Strain #101: CGR identification. A. Nanopore sequencing coverage maps of Chromosomes III and V, generated via UGENE. Positions of observed split reads and normal reads at these same junctions are overlaid on the coverage map, and are labeled i-iii. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. In the multi-read view, the horizontal axis corresponds to genomic locations as indicated by the windows. It is important to note that, unlike the single-read view, the multi-read view does not show how each portion of the reference chromosome fits sequentially into each read. Rather, this view highlights regions found at any point within a read, and stacks multiple reads on the vertical axis. In the top two images, nearly all reads display the same 5' and 3' breakpoints on Chromosome III, indicating a consistent rearrangement. Portions of the reads mapping to Chromosome V at the *ura3-52* locus are seen on the right side. In the bottom panel, many reads are seen that span the *ura3-52* region but do not show a split-read pattern, indicating that the unmodified Chromosome V exists intact. **C.** View of the UGENE alignment zoomed in to the portion of Chromosome III at junction ii. The top portion contains a horizontal scale showing the chromosomal location, along with dark blue vertical bars indicating read depth. Below are individual reads running horizontally, which are stacked vertically. Gray boxes indicate bases that match the reference sequence. Colored boxes indicate bases that do not match the reference sequence (blue=G, green=C, yellow=A, light red=T, dark red=deletion). Blue arrows indicate sites of consistent SNPs which match to *ura3-52* on Chromosome V, rather than the novel Ty1 element adjacent to *YCLWTy2-1* on Chromosome III. The black arrow indicates the location of an expected SNP from *ura3-52* which does not appear on Chromosome III, thus establishing the junction boundaries with single-base pair resolution (Fig. 3D).

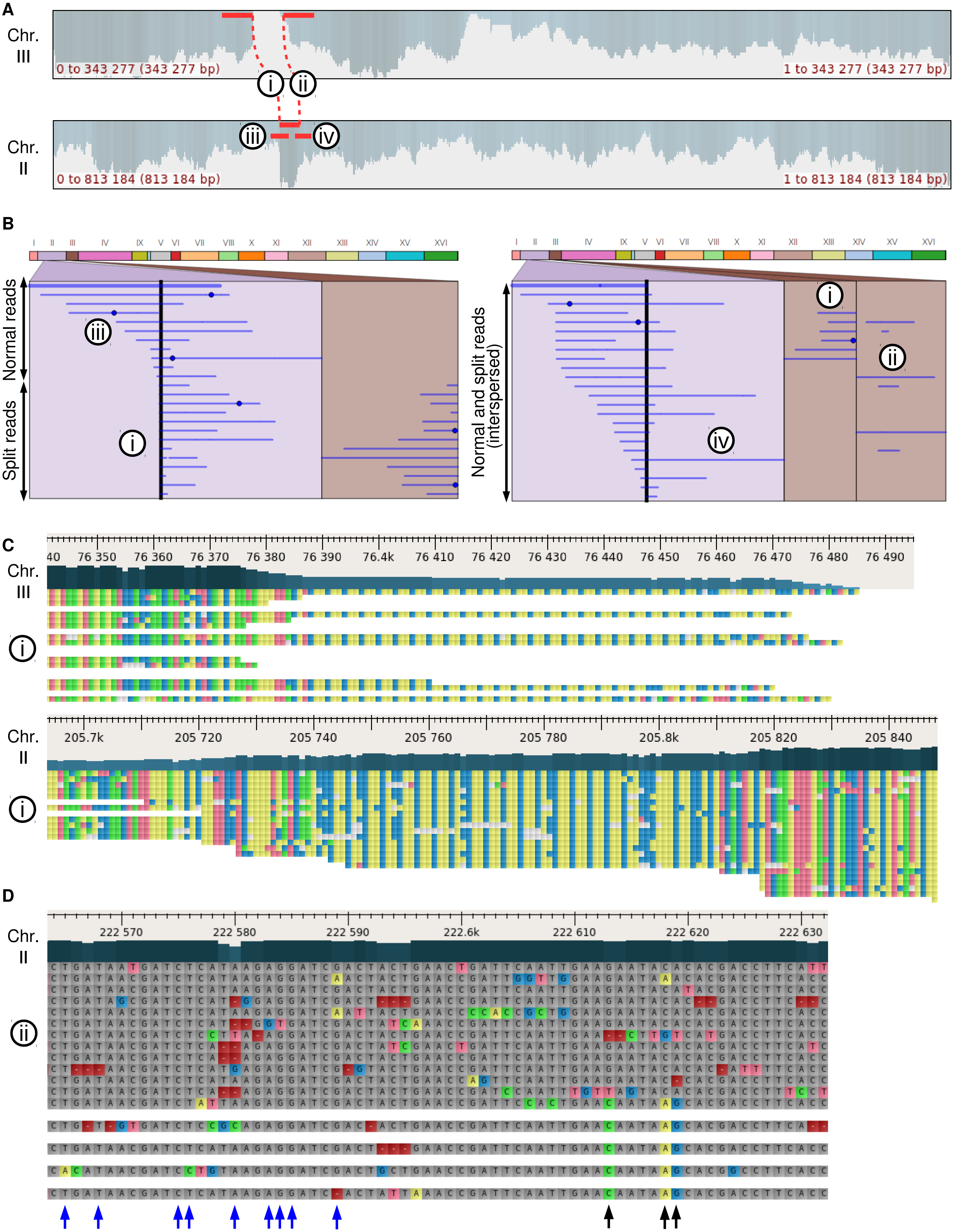
A
Chr. III
0 to 343 277 (343 277 bp)
1 to 343 277 (343 277 bp)
i
ii
Chr. II
iii
iv
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal reads
Split reads
iii
i
Normal and split reads (interspersed)
i
ii
iv
C
40 76 350 76 360 76 370 76 380 76 390 76.4k 76 410 76 420 76 430 76 440 76 450 76 460 76 470 76 480 76 490
Chr. III
i
205.7k 205 720 205 740 205 760 205 780 205.8k 205 820 205 840
Chr. II
i
D
222 570 222 580 222 590 222.6k 222 610 222 620 222 630
Chr. II
ii

Fig. S5. Strain #118: CGR identification. A. Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE. Positions of observed split reads and normal reads at these same junctions are overlaid on the coverage map, and are labeled i-iv. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. See Fig. S3B for explanation of Ribbon multi-read view. Both split reads and normal reads are observed at each of the junctions, indicating that the unmodified Chromosome II sequence exists intact. **C.** View of the UGENE alignment zoomed in to junction i. See Fig. S3C for explanation of the diagrams. In this particular view, all bases are displayed by color (blue=G, green=C, yellow=A, light red=T, dark red=deletion). Junction i is displayed for chromsomes III and II. The $(GAA)_n$ repeats are clearly visible by the blue-yellow-yellow pattern. Chromosome III shows that the reads stop aligning within the repeats, while Chromosome II shows that the read depth doubles within the repeats. **D.** View of the UGENE alignment zoomed in to the portion of Chromosome II at junction ii. See Fig. S3C for explanation of the diagrams. Black arrows indicate sites of consistent SNPs which match to the novel Ty1 element adjacent to *YCLWTy2-1* on Chromosome III, rather than *YBLWTy1-1* on Chromosome II. The blue arrows indicate the absence of SNPs on *YBLWTy1-1*, indicating that the junction was resolved within this 15bp window.

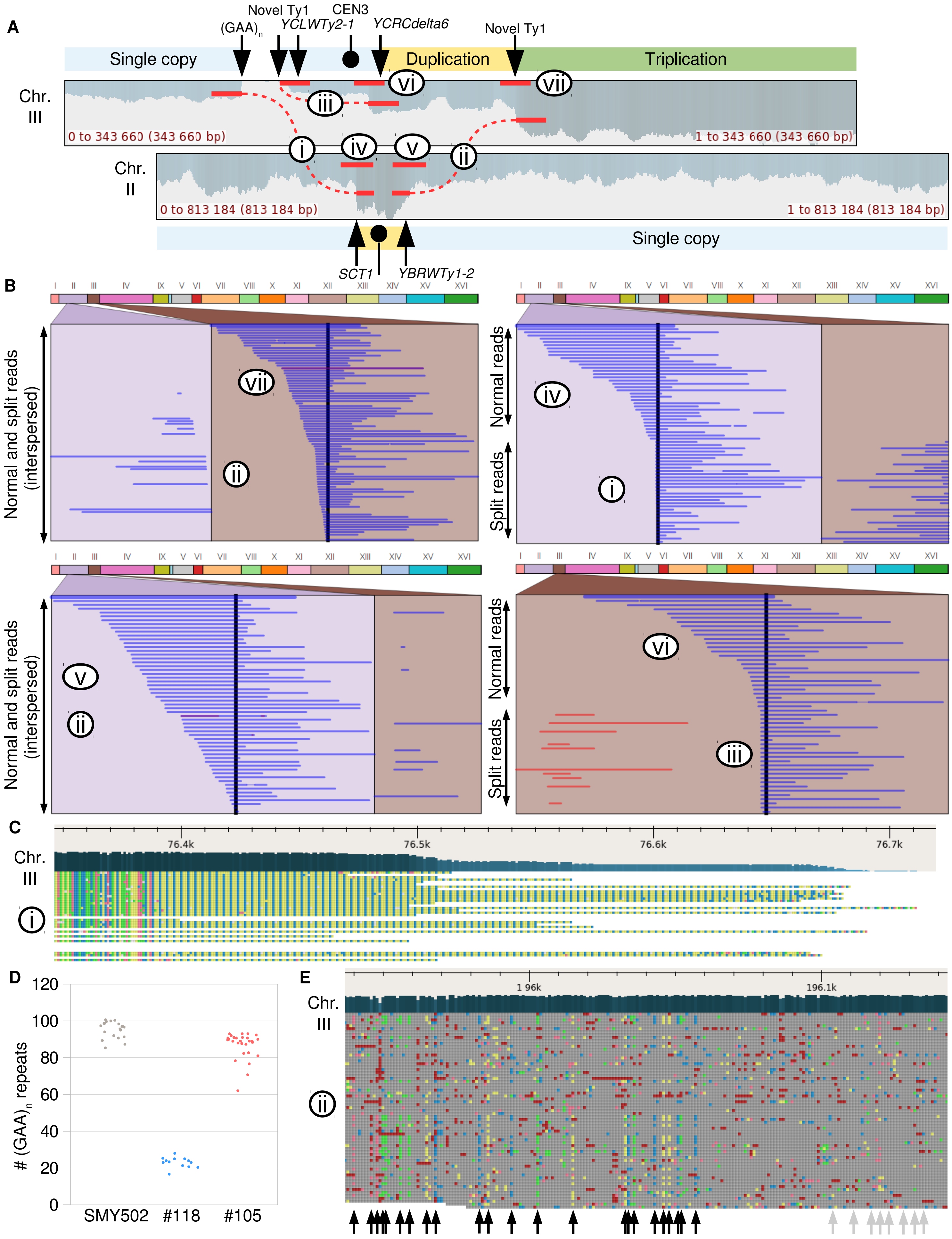
A
(GAA)$_n$
Novel Ty1
YCLWTy2-1
CEN3
YCRCdelta6
Novel Ty1
Single copy
Duplication
Triplication
Chr. III
vi
iii
vii
0 to 343 660 (343 660 bp)
1 to 343 660 (343 660 bp)
i
iv
v
ii
Chr. II
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
Single copy
SCT1
YBRWTy1-2
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal and split reads (interspersed)
vii
ii
Normal reads
Split reads
iv
i
Normal and split reads (interspersed)
v
ii
Normal reads
Split reads
vi
iii
C
76.4k
76.5k
76.6k
76.7k
Chr. III
i
D
120
100
80
60
40
20
0
# (GAA)$_n$ repeats
SMY502
#118
#105
E
1 96k
196.1k
Chr. III
ii

Fig. S6. Identifying genomic rearrangements in strain #105. A. Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE. Positions of relevant sequence features and large-scale copy number changes are indicated above/below the coverage maps. Positions of observed split reads and normal reads at these same junctions are overlayed on the coverage map, and are labeled i-vii. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. See Fig. S3B for explanation of Ribbon multi-read view. In the lower right image, red lines indicate reads that map in an inverted orientation. Both split reads and normal reads are observed at each of the junctions, indicating that the unmodified Chromosome II sequence exists intact in at least one of the altered chromosomes. However, because no reads span the entire ~50kb duplication on Chromosome II, it could not be determined whether a translocation existed by our sequence data alone. **C.** View of the UGENE alignment zoomed in junction i on Chromosome III. See Fig. S3C for explanation of the diagrams. In this particular view, all bases are displayed by color (blue=G, green=C, yellow=A, light red=T, dark red=deletion). The $(GAA)_n$ repeats are clearly visible by the blue-yellow-yellow pattern. Most reads show that the deletion begins at the very end of the repeat tract. **D.** $(GAA)_n$ repeat length analysis, comparing strains #118 and #105 to the reference strain, which contains 100 GAA repeats. Each dot represents the number of repeats found in an individual Nanopore read. **E.** View of the UGENE alignment zoomed in to the portion of Chromosome III at junction ii. See Fig. S3C for explanation of the diagrams. Black arrows indicate sites of consistent SNPs in approximately 1/3 of the reads (consistent with the triplication junction), which match to *YCLWTy2-1* on Chromosome III, rather than either *YBRWTy2-1* on Chromosome II or the novel Ty1 element replacing *YCRWdelta11* on Chromosome III (which is the reference sequence in this view). Gray arrows indicate the absence of SNPs from *YCLWTy2-1*, establishing the boundaries of the junction.

Supplemental References:

Carr M, Bensasson D, Bergman CM. 2012. Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. *PLoS One* **7**: e50978.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Nattestad M, Chin C, Schatz MC. Ribbon: Visualizing complex genome alignments and structural variation. *bioRxiv* doi:https://doi.org/10.1101/082123

Okonechnikov K, Golosova O, Fursov M, team U. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**: 1166-1167.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz M. 2017. Accurate detection of complex structural variations using single molecule sequencing. *BioRxiv*.