A comprehensive map of hotspots of de novo telomere addition in *Saccharomyces cerevisiae*

Katrina Ngo[1], Tristen H. Gittens[1], David I. Gonzalez[1], E. Anne Hatmaker[1,2], Simcha Plotkin[1], Mason Engle[1], Geofrey A. Friedman[1], Melissa Goldin [1], Remington E. Hoerr[1], Brandt F. Eichman[1,3], Antonis Rokas[1,2], Mary Lauren Benton[4] and Katherine L. Friedman[1*]

1. Department of Biological Sciences, Vanderbilt University
2. Evolutionary Studies Initiative, Vanderbilt University
3. Department of Biochemistry, Vanderbilt University
4. Department of Computer Science, Baylor University

Addresses:
1,2 and 3. 465 21st Avenue S, 1210 MRBIII, Nashville, TN 37232
4. One Bear Place #97141, Waco, TX, 76798

*Corresponding author: Katherine.friedman@vanderbilt.edu

Running head: Telomere addition hotspots in *S. cerevisiae*

Keywords: de novo telomere addition, DNA repair, DNA damage, genomic instability, yeast, *Saccharomyces cerevisiae*

Abstract

Telomere healing occurs when telomerase, normally restricted to chromosome ends, acts upon a double-strand break to create a new, functional telomere. De novo telomere addition on the centromere-proximal side of a break truncates the chromosome but, by blocking resection, may allow the cell to survive an otherwise lethal event. We previously identified several sequences in the baker's yeast, *Saccharomyces cerevisiae,* that act as hotspots of de novo telomere addition (termed Sites of Repair-associated Telomere Addition or SiRTAs), but the distribution and functional relevance of SiRTAs is unclear. Here, we describe a high-throughput sequencing method to measure the frequency and location of telomere addition within sequences of interest. Combining this methodology with a computational algorithm that identifies SiRTA sequence motifs, we generate the first comprehensive map of telomere-addition hotspots in yeast. Putative SiRTAs are strongly enriched in subtelomeric regions where they may facilitate formation of a new telomere following catastrophic telomere loss. In contrast, outside of subtelomeres, the distribution and orientation of SiRTAs appears random. Since truncating the chromosome at most SiRTAs would be lethal, this observation argues against selection for these sequences as sites of telomere addition per se. We find, however, that sequences predicted to function as SiRTAs are significantly more prevalent across the genome than expected by chance. Sequences identified by the algorithm bind the telomeric protein Cdc13, raising the possibility that association of Cdc13 with single-stranded regions generated during the response to DNA damage may facilitate DNA repair more generally.

45 **Introduction**

46 The maintenance of DNA integrity is essential for cell function. To maintain genomic integrity and
47 prevent sequence loss, most eukaryotic chromosomes terminate with nucleoprotein structures termed
48 telomeres that protect chromosomes from end-to-end fusion and block excessive nucleolytic resection.
49 Telomeres contain a characteristic, repetitive sequence that is rich in thymine and guanine (TG-rich) on
50 one strand (Blackburn 1991). While the majority of the telomere is double-stranded, the TG-rich strand
51 extends past the complementary cytosine and adenine (CA)-rich strand to create a 3' overhang.
52 Regeneration of this 3' overhang after each round of DNA replication results in progressive sequence
53 loss, but in cells that maintain telomere length over successive generations, this end-replication problem
54 is counterbalanced through extension of the 3' strand by telomerase (reviewed in Lingner et al. 1995;
55 Osterhage & Friedman, 2009; Bonnell et al. 2021). Telomerase uses an intrinsic RNA molecule as
56 template for synthesis of the TG-rich strand (Greider & Blackburn, 1989; Singer & Gottschling, 1994)
57 while the lagging strand polymerase machinery subsequently fills in the complementary, CA-rich strand
58 (reviewed in Gilson & Géli, 2007; Pfeiffer & Lingner, 2013).

59 Because telomeres are, by definition, the end of a DNA molecule, they resemble a DNA double-
60 strand break (DSB). Indeed, similar to telomeres, enzymatic resection at a DSB generates 3' overhangs
61 that can serve as substrates for homologous recombination. The specific sequence of the 3' overhang at
62 telomeres distinguishes it from 3' overhangs generated by resection at a double strand break, thereby
63 enforcing different outcomes at these otherwise similar structures (telomere elongation versus DNA
64 repair, respectively; reviewed in Casari et al. 2022; Doksani & de Lange, 2014). However, rarely, the 3'
65 overhang generated at a DSB is recognized by telomerase, resulting in addition of a new or de novo
66 telomere (reviewed in Hoerr et al. 2021; Pennaneach et al. 2006). De novo telomere addition (dnTA),
67 also termed telomere healing, causes loss of sequences distal to the site at which the telomere is added
68 but prevents additional resection that would ultimately be lethal.

69 Several human diseases (e.g. Phelan/Mcdermid syndrome and α-thalassemia) are associated
70 with terminal truncations generated by dnTA (Bonaglia et al. 2011; Guilherme et al. 2015; Lamb et al.
71 1993; Nevado et al. 2022). The observation of recurrent telomere addition events within a small
72 chromosome region suggests that sequences associated with these diseases are unusually prone to
73 telomerase action. This phenomenon is not limited to human cells, and has been observed in other
74 eukaryotic organisms including *S. cerevisiae* (Mangahas et al., 2001; Ouenzar et al., 2017; Stellwagen et
75 al., 2003). Although dnTA events are generally very rare, the *S. cerevisiae* genome contains hotspots
76 where dnTA occurs at frequencies estimated to be at least 200-fold above background (Obodo et al.
77 2016; Epum et al. 2020). These sequences present a unique opportunity to use yeast as a model to study
78 the consequences of such sequences for genome stability and evolution.
79 Telomeres in *S. cerevisiae* have a 3' terminating strand that consists of irregular repeats
80 containing a pattern of a single T followed by one, two, or three Gs. Despite this heterogeneity, the
81 telomere contains recognition sites for several sequence-specific DNA binding proteins that associate
82 with the double-stranded portion of the telomere and the single-stranded TG-rich overhang (Rap1 and
83 Cdc13, respectively; reviewed in Wellinger & Zakian, 2012). Rap1 participates in telomere length
84 homeostasis, telomere capping, and formation of telomeric chromatin (Hardy et al. 1992; Kyrion et al.
85 1993; Marcand et al. 1997; Negrini et al. 2007; Pardo & Marcand, 2005; Teixeira et al. 2004;
86 Vodenicharov et al. 2010), while Cdc13 interacts with the Est1 component of telomerase to recruit
87 telomerase to telomeres (Evans & Lundblad, 1999; Pennock et al. 2001, Chen et al. 2018). Cdc13
88 additionally interacts with Stn1 and Ten1 to limit nucleolytic resection and promote fill-in synthesis by
89 the lagging strand polymerase machinery (Pennock et al 2001; Lin et al. 2021).
90 Sequences serving as hotspots of dnTA in yeast were first observed as sites of telomere healing
91 in response to an induced DSB on chromosome VII (Mangahas et al. 2001). Subsequently, spontaneous
92 truncations of chromosome V occurring as a result of dnTA were shown to cluster in a small

93 chromosomal region (Myung et al. 2001; Stellwagen et al. 2003; Pennaneach et al. 2006). Following
94 structure/function analysis of the sequence on chromosome V and an additional hotspot on
95 chromosome IX, we named these sequences Sites of Repair-associated Telomere Addition or SiRTAs.
96 SiRTAs contain two TG-rich sequence tracts. One tract (the Core) serves as the direct substrate for
97 telomere addition by telomerase while the second tract (the Stim, located 5' to the Core on the TG-rich
98 strand) is required for high levels of dnTA at the Core sequence (Obodo et al. 2016). The Stim can be
99 functionally replaced with canonical Cdc13 binding sites or with a sequence designed to artificially
100 recruit Cdc13 (Obodo et al. 2016). Together, these observations support a model in which resection of
101 the 5'-terminating strand following a DSB exposes TG-rich sequences on the 3' overhanging strand that
102 are bound by Cdc13, with subsequent recruitment of telomerase. Telomere addition is favored at SiRTAs
103 even when the initiating break is artificially induced 2-3 kilobases distal to the eventual site of telomere
104 addition, suggesting that SiRTAs stimulate repair rather than serving as fragile sites *per se* (Obodo et al.
105 2016).
106 For the SiRTAs described above, the TG-rich sequence is on the same strand that terminates as a
107 TG-rich 3' overhang at the nearest telomere, a property we refer to as the "TG-orientation." On the left
108 arm of a chromosome, SiRTAs in this orientation are TG-rich on the bottom (3' to 5' or minus) strand
109 while on the right arm, the TG-rich sequence is on the top (5' to 3' or plus) strand. Telomere addition at
110 a SiRTA in the TG-orientation requires a DSB distal to the SiRTA and stabilizes the centromere-containing
111 side of the break. If the SiRTA is distal to all essential genes on that arm (as is true for the SiRTAs on
112 chromosomes V and IX), the resulting terminal deletion is compatible with viability, even in a haploid
113 strain. However, not all characterized SiRTAs are TG-oriented. We recently described a SiRTA in the
114 opposite or "CA-orientation" that promotes cell survival under sulfate-limiting conditions by facilitating
115 formation of an acentric fragment containing *SUL1*, encoding the primary sulfate transporter (Hoerr et

116 al. 2023). Despite identification of several SiRTAs in addition to those described above (Ngo et al. 2020),
117 understanding of the genome-wide frequency and distribution of these sequences is lacking.
118 Here, we validate the use of a computational algorithm, the Computational Algorithm for
119 Telomere Hotspot Identification (CATHI), to predict SiRTA function based on similarity with the TG_{1-3}
120 pattern of the yeast telomeric repeat. In parallel, we develop and validate a high-throughput sequencing
121 method that dramatically increases the number of putative sequences that can be characterized while
122 simultaneously yielding information about the site of telomerase action. Together, we use these
123 approaches to determine the overall locations and orientations of SiRTAs on a genome-wide scale. All
124 but one of the subtelomeric repetitive regions (defined as X and Y' elements; Louis et al. 1994; Louis &
125 Haber, 1992) contain at least one SiRTA in the TG-orientation. However, outside of the subtelomeric
126 regions, there is no apparent bias in the location or orientation of predicted SiRTAs, although these
127 sequences occur more frequently than expected by chance. SiRTA function correlates with the ability of
128 a sequence to bind Cdc13, but overall binding affinity is insufficient to explain all variation in the
129 frequency of dnTA. This work provides a foundation for developing a fuller understanding of how sites
130 with a propensity to stimulate dnTA impact genomic stability and evolution.
131
132 **Methods**
133 *Strain construction*
134 Strains were constructed in the S288C background as described (Ngo et al. 2020; Hoerr et al.
135 2023). The parental strain contains a *URA3* marker distal to an HO recognition site on chromosome VII
136 (YKF1975 *MAT**a**::ΔHOcs::hisG hmlαΔ::hisG HMRa::NAT ura3Δ851 trp1Δ63 leu2Δ::KAN^R ade3::GAL10::HO*
137 Chr VII, 15828-16027 (*adh4*)::HOcs::*HYG^R pau11*::*URA3*). 300 bp sequences to be tested for SiRTA
138 function were inserted using the CRISPR/Cas9 system as described in Anand et al. (2017) using a guide
139 sequence of 5'-TGCGGCAAGTTCATCTTCCA located ~2kb centromere-proximal to the HO recognition site.

140 PCR products for recombinational insertion were generated as follows. Forward primers were designed
141 by including 40 bases upstream of the gRNA recognition site
142 (5'TTTCTTTGGAAAACGTTGAAAATGAGGTTCTATGATCTAC) followed by the first ~20 bases on the 5' end
143 of the sequence of interest. Reverse primers were constructed by taking the reverse complement of the
144 40 bases downstream of the gRNA site (5'-AGAACATAGAATAAATTTGGTACTGGAACGTTGATTAACT)
145 followed by the last ~20 bases of the sequence of interest. Sequences tested are listed in Supplementary
146 File 1. The DNA fragment needed to insert SiRTA 6R210(+) onto chromosome VII using the CRISPR
147 system was synthesized and inserted into the pMX plasmid using Invitrogen GeneArt Gene Synthesis
148 services (Thermo Fisher Scientific). The 6R210(+) DNA fragment was amplified from the plasmid using
149 PCR designed as described above. One step gene replacement using template DNA from pFA6a-TRP1
150 (Longtine et al. 1998) was used to replace *RAD52*.
151 For testing on chromosome IX, the *URA3* marker and HO cleavage site were integrated on
152 chromosome IX to create strain YKF1752 (*MAT**a**::ΔHOcs::hisG hmlαΔ::hisG HMRa::NAT ura3Δ851*
153 *trp1Δ63 leu2Δ::KAN*^{R} *ade3::GAL10::HO* Chr9;35050-41450::HOcs::HPH^{R} *soa1*::*URA3*) as described in
154 Obodo et al. (2016). Yeast strains containing the BS Mut1 and BS Mut2 mutations are described in
155 Obodo et al. (2016) as YFK1610 and YFK1652, respectively.
156
157 *HO cleavage assay*
158 The HO cleavage assay was performed as described (Ngo et al. 2020; Hoerr et al. 2023). Briefly,
159 cells were grown in synthetic dropout media lacking uracil (SD-Ura) + 2% raffinose to an optical density
160 at 600nm (OD600) of 0.6-1.0. Cells were serially diluted and plated on yeast extract peptone medium
161 with either 2% dextrose (YEPD) or 2% galactose (YEPgal). After incubation at 30˚C for three days, colony
162 number was determined on at least two plates of each condition. The frequency of survival on YEPgal
163 was calculated as: (average number of colonies per plate on YEPgal x dilution factor)/(average number

164 of colonies per plate on YEPD x dilution factor). At least 100 colonies surviving on YEPgal were patched
165 to medium containing 1 mg/mL 5-fluoroorotic acid (5-FOA) to select for cells in which the *URA3* marker
166 was lost [gross chromosomal rearrangement (GCR) events]. The frequency of GCR events was
167 determined as: (frequency of survival on YEPgal*frequency of clones demonstrating 5-FOA resistance).
168 Thirty clones that displayed growth on medium containing 5-FOA were selected and inoculated in liquid
169 YEPD for genomic DNA extraction using the MasterPure™ Yeast DNA Purification Kit (Lucigen). Multiplex
170 PCR was used as described in Ngo et al. (2020) to map the approximate site of dnTA in relationship to
171 the sequence of interest. Primers for chromosome VII and IX are listed in Supplementary File 1. Colonies
172 where the DNA loss event mapped within the sequence of interest were tested for telomere addition
173 using one primer centromere proximal to the putative telomere addition site and a second primer
174 complementary to the telomeric repeat (Supplementary File 1).
175
176 *Pooled Tel-seq*
177 Thirty 5-FOA resistant clones isolated as described above were separately inoculated in 200 µL
178 of YEPD in a 96-well culture plate and incubated overnight at 30˚C to reach saturation. Equal volumes (at
179 least 30 µL) of each culture were pooled and DNA was extracted using the YeaStar™ genomic DNA kit
180 (ZYMO research).
181 Libraries were prepared using 50 ng of genomic DNA and a modified protocol using the Twist
182 Library Preparation kit (Twist Bioscience 106543). Denatured DNA templates in a 96-well plate were
183 randomly primed with 5' barcoded adapters. Samples were pooled, captured on streptavidin coated
184 magnetic beads, and washed to remove excess reactants. A second 5' adapter tailed primer with a
185 strand-displacing polymerase was utilized to convert the captured templates into dual adapter libraries.
186 Beads were washed to remove excess reactants. Four cycles of PCR were utilized to amplify the library
187 and incorporate the plate barcode in the index read position. Libraries were sequenced using the

188 NovaSeq 6000 with 150 bp paired end reads targeting 13 to 15 million reads per sample. Real Time
189 Analysis software (version 2.4.11; Illumina) was used for base calling and data quality control was
190 completed using MultiQC v1.7.

191 Sequencing data are available from the NIH Sequence Read Archive (SRA) under BioProject ID
192 PRJNA939836. Reads mapping to a 300 bp control sequence located in the essential gene *BRR6* (Chr VII:
193 36933 to 37233) or to the 300 bp sequence of interest [inserted on chromosome VII or at the
194 endogenous location of SiRTA 9L44(-)] were identified using Bowtie2 (Galaxy Version 2.5.0+galaxy0) with
195 the sensitive local setting (Langmead et al. 2009; Langmead and Salzberg 2012). Any remaining library
196 primer sequences were removed using the Trimmomatic tool (Galaxy Version 0.38.0; Bolger et al. 2014)
197 and reads mapping to the putative SiRTA that also contain telomere sequence (match to 5'-GGGTGTGG
198 or 5'-CCACACCC) were identified and tabulated. The number of individual reads with evidence of
199 telomere addition was normalized to the number of control reads at *BRR6* by expressing the number of
200 telomere reads as a percentage of control reads. Where applicable, sites of telomere addition were
201 mapped to the original SiRTA sequence to determine the location of the event.

202

203 *Purification of Cdc13-DBD*

204 Cdc13-DBD was expressed in *E. coli* using pET21a-Cdc13-DBD-His6, a gift from the Wuttke lab.
205 Purification was done as described (Anderson et al. 2002; Obodo et al. 2016).

206

207 *Fluorescence Polarization binding assays*

208 Binding assays were conducted using a fixed concentration of a 5'-6-carboxyfluorescein (FAM)
209 labeled tel-11 oligonucleotide (25 nM). Cdc13-DBD was added at final concentrations of 0, 6.25, 12.5,
210 25, 37.5, 50, 62.5 and 75 nM. Competition binding assays were conducted at fixed concentrations of
211 Cdc13-DBD (30 nM) and 5'-6-FAM labeled tel-11 oligonucleotide (25 nM). Unlabeled oligonucleotides of

212 75 bases each were used at final concentrations of 0, 6.25, 12.5, 25, 50, 150 and 200 nM.
213 Oligonucleotide sequences are listed in Supplementary File 1. Labeled and unlabeled oligonucleotides
214 and protein were mixed in binding buffer (50 µM Tris pH 8, 1 µM EDTA pH 8, 15% glycerol, 75 µM NaCl,
215 75 µM KCl) in a final volume of 80 µL and incubated at 4˚C for 30 minutes. Each reaction was measured
216 in triplicate (25 µl per measurement) in a Corning 384 well assay plate using the BioTek Synergy H1
217 hybrid reader. This procedure was repeated at least three times for each competitor. The relative
218 polarization (ΔP) was determined using the following equation: $\Delta P=P_0-P_x$, where P_0 is the polarization
219 value at a competitor concentration of 0 and P_x represents the polarization value at x competitor
220 concentration. For each experiment, technical replicates were averaged and the averaged data were fit
221 to the following equation: $\Delta P=(P_{max}[\text{competitor}])/(K_{i,app}[\text{competitor}])$, where P_{max} is the maximal
222 polarization value and $K_{i,app}$ is the apparent inhibition constant (Anderson et al. 2008; Vaasa et al. 2009).
223 An unlabeled 75mer containing the Tel11 sequence at the center of the oligonucleotide (Tel11-75) was
224 included in each experiment and normalized $K_{i,app}$ values are reported as the fold change relative to this
225 control ($K_{i,app}$ of Tel11-75/$K_{i,app}$ of experimental oligonucleotide)
226
227 *Implementation of the CATHI algorithm*
228 Initially, the program generates a series of sliding windows to be utilized in the score calculation.
229 The window and step size of the sliding windows can be customized using the --window and --step
230 options. For each window, the program searches for strings of at least 4 characters that begin with a G
231 and consist of only Gs and Ts. These become the set of candidate scoring regions. From these
232 candidates, regions that consist only of Gs are removed. The program then scans candidate regions for
233 any consecutive Ts, or four or more consecutive Gs. If either are encountered, that candidate region is
234 truncated after the first T or the third G, respectively. Once the set of candidates has been filtered, the
235 number of nucleotides remaining in the candidate set is counted and any applicable scoring penalties

236 are subtracted. There are no penalties applied by default, but users can choose to apply them. The --
237 penalty option imposes score deductions for any GGTGG sequences, and the --ttpenalty imposes score
238 deductions for Ts that flank the candidate regions.

239 CATHI is implemented in Python (version 3) using the BioPython (Cock et al. 2009), NumPy
240 (Harris et al. 2020), Pandas (McKinney 2010), and pybedtools (Dale et al. 2011) libraries. CATHI can
241 perform in two modes: (1) score mode; and (2) signal mode. The default score mode will return the
242 maximum CATHI score for each input sequence. For each sequence in the provided FASTA file, CATHI
243 will generate sliding windows and calculate the CATHI score for each window, returning only the
244 maximum score per sequence. In signal mode, CATHI will generate sliding windows and return the
245 genomic coordinates and CATHI score for each window. CATHI output is printed to the screen for easy
246 redirection and can be optionally printed in BED format. Code can be obtained from
247 https://github.com/bentonml/cathi. In this work, both strands of each chromosome in *S. cerevisiae*
248 were separately scanned in signal mode using a step size of 1 and window size of 75. Perfect telomeric
249 repeats representing *bona fide* terminal telomeres were trimmed prior to analysis (if present).
250 Coordinates used for each chromosome are in Supplementary File 2.

251 Overlapping and adjacent windows meeting or exceeding the threshold value can be merged
252 into a single region using the --cluster option, where the beginning is the start coordinate of the most
253 upstream window and the ending is end coordinate of the most downstream window. The score of the
254 merged region is the maximum CATHI score across all merged windows.

255

256 *Generation of a shuffled yeast genome*

257 A set of five randomized versions of the *S. cerevisiae* (sacCer3) genome was generated to
258 evaluate the number of SiRTAs expected when applying the CATHI algorithm to a null model. DNA
259 sequence was downloaded from the sacCer3 reference genome using the BedTools (version 2.30.0)

260 ‘getfasta’ command (Quinlan and Hall 2010) after adjusting the start and end coordinates of each
261 chromosome to exclude subtelomeric regions (Supplementary File 2). Nucleotides were randomly
262 shuffled within each adjusted chromosome using Python’s built-in randomization library. This procedure
263 maintains the nucleotide composition and length of each chromosome while randomizing the actual
264 DNA sequence.
265
266 *Enrichment for genomic annotations in putative SiRTAs*
267 Overlap between putative SiRTAs and other genomic annotations was determined using a
268 permutation-based enrichment test. Enrichment for SiRTAs with several different genomic annotations
269 was determined: (1) essential and non-essential genomic regions (Giaever et al. 2002); (2) Est2 binding
270 sites (Pandey et al. 2021); (3) Pif1 binding sites (Paeschke et al. 2011); (4) γH2AX binding sites (Capra et
271 al. 2010); (5) G-quadruplex regions (Capra et al. 2010) and (6) Rap1 binding sites (Rhee and Pugh 2011).
272 When the original dataset included strand information (as in the case of G4-sites) that information was
273 considered in the analysis.
274 Enrichment between the SiRTAs and the annotations was calculated as the fold change between
275 the observed and expected overlap. To ensure meaningful overlaps between the SiRTAs and the
276 genomic annotations, at least 50% of the binding site was required to overlap with the SiRTA or at least
277 50% of the SiRTA was required to overlap with the essential/non-essential region. To create the
278 distribution of expected overlap, 1000 permutations were performed by randomly shuffling regions
279 throughout the genome and calculating the amount of SiRTA overlap. Shuffled regions are non-
280 overlapping, length- and strand-matched (G4 sequences only) with the annotations. When specified,
281 telomeric and/or subtelomeric regions were excluded. Subtelomeric regions are defined in
282 Supplementary File 2. For G4 sites, overlap was only recorded if the G4-forming sequence and SiRTA are

283 on the same strand. An empirical p-value is calculated for the overlap using the expected distribution;
284 where relevant, p-values are corrected for multiple comparisons using the Bonferroni method.
285
286 *Determining overlap with genes*
287 The location of predicted SiRTAs was compared to the location of genes within the *S. cerevisiae*
288 genome to determine the number of predicted SiRTAs in both inter- and intragenic regions. Coordinates
289 for genes and subtelomeres (defined as X and/or Y' elements) were obtained from the *S. cerevisiae*
290 S288C annotation available from NCBI (accession numbers NC_001133.9, NC_001134.8, NC_001135.5,
291 NC_001136.10, NC_001137.3, NC_001138.5, NC_001139.9, NC_001140.6, NC_001141.2, NC_001142.9,
292 NC_001143.9, NC_001144.5, NC_001145.3, NC_001146.8, NC_001147.6, NC_001148.4); FASTA and
293 GFF3 files for the reference assembly of strain S288C (GCF_000146045.2) were downloaded from NCBI's
294 RefSeq database. The RefSeq genome annotation is identical to that in the *Saccharomyces* Genome
295 Database (SGD).
296 Coordinates for predicted SiRTAs were obtained from the CATHI program and manually
297 converted into GFF3 files, one for each chromosome. Overlap between predicted SiRTAs and genes was
298 calculated using the "intersect" function within bedtools v2.30.0 (Quinlan 2014) for each chromosome.
299 Predicted SiRTAs within annotated genes were manually assigned to the template or coding strand using
300 chromosome visualization in Geneious Prime v2020.1.2.
301
302 *Modeling SiRTA distribution*
303 Python programs to model the expected distribution of telomere addition events within a region
304 and to model the random distribution of SiRTAs between the forward and reverse strand are available at
305 https://github.com/geofreyfriedman/sirta. For the latter, random strand distributions were generated
306 for each chromosome based on the observed number of SiRTAs on each strand. The expected

307 distribution of SiRTAs between the forward and reverse strands was quantified by 1) determining the
308 number of consecutive SiRTAs on the same strand (run length) or 2) summing the number of times that
309 consecutive SiRTAs are found on different strands (number of strand switches). To avoid "edge effects"
310 generated at the ends of each chromosome, 10,000 iterations were generated for each chromosome
311 and run lengths or strand switches were summed across the 16 chromosomes (sum of iteration 1, sum
312 of iteration 2, etc). In each case, the observed value was compared to the random distribution
313 generated from 10,000 iterations.
314
315 **Results**
316 *High throughput sequencing of pooled samples accurately measures de novo telomere addition*
317 Measurement of the propensity for de novo telomere addition (dnTA) across the genome is
318 complicated by varied chromosome context (which can affect the frequency of competing repair
319 pathways) and our inability to capture dnTA addition events at sequences that are proximal to essential
320 genes and/or in the CA-orientation. To circumvent these limitations, we previously developed a "test
321 site" on the left arm of chromosome VII. CRISPR/Cas9 is used to insert sequences (typically 300 bp)
322 oriented such that the TG-rich sequence of interest is on the bottom (3' to 5') strand. A recognition site
323 for the homothallic switching (HO) endonuclease is located ~2kb distal to the CRISPR/Cas9 integration
324 site. A *URA3* marker located further distal to the HO cleavage site facilitates selection for cells carrying a
325 truncated chromosome VII-L (Figure 1a and b). Importantly, *RAD52* is deleted to prevent homology-
326 directed repair between the sequence inserted on chromosome VII and that same sequence at its
327 endogenous location.
328 Cells are plated on solid medium containing galactose to induce expression of the HO
329 endonuclease. To escape persistent cleavage and generate a colony, a cell must incur a mutation at the
330 HO site that blocks nuclease recognition or lose the HO site completely through formation of a gross

331 chromosomal rearrangement (GCR). To identify the latter, which include dnTA events, 100 clones arising
332 on the galactose plate are screened for loss of the *URA3* marker via growth on medium containing 5-
333 fluoroorotic acid (5-FOA). Thirty 5-FOA-resistant clones are then analyzed to determine the nature of the
334 resulting GCR event. In past work, we utilized a clone-by-clone mapping strategy that employed multiple
335 PCR reactions to identify the approximate location of each GCR event (Figure 1a). For colonies in which
336 the event maps to the sequence of interest, Southern blotting or PCR utilizing a telomeric primer is
337 utilized to determine if the event involved telomere addition (Figure 1b).

338 The efficiency of SiRTA function is expressed as the percent of 5-FOA-resistant clones (from a
339 total of 30) that contain a telomere-addition event within the sequence of interest at the insertion site
340 on chromosome VII. Typically, the experiment is done 2-3 times and the average values of the biological
341 replicates are reported. The 300 bp sequence analyzed represents ~1.4% of the 21,922 bp region within
342 which a GCR event can be recovered [between the HO cleavage site and the first essential gene on VII-L
343 (*BRR6*)]. To determine a threshold for SiRTA activity, we modeled the expectation for random repair
344 within this region (*Materials and Methods* and Supplementary Figure 1). Assuming random distribution
345 of 30 GCR events, two or more would be expected to occur within the 300 bp test sequence in 6.2% of
346 trials, while three or more would be expected in only 0.78% of trials. Therefore, we chose to define a
347 SiRTA as a sequence in which the average efficiency of dnTA is >6.6% (an average of greater than 2 out
348 of 30 clones containing dnTA within the 300 bp test sequence). Sequences tested are named using the
349 following scheme: chromosome number, chromosome arm (L for left and R for right), distance from the
350 left telomere rounded to nearest kilobase, and the strand on which the SiRTA is located [(+) for the
351 forward strand and (–) for the reverse strand].

352 To increase the throughput of this analysis pipeline and to map dnTA events with nucleotide
353 precision, we developed the Pool-Tel-seq (PT-seq) method (*Materials and Methods* and Figure 1c). As in
354 our original approach, a single inoculum is plated on a medium containing galactose to induce HO

355 cleavage and thirty 5-FOA resistant colonies are identified that have lost the chromosome VII terminus.
356 The 30 colonies are grown separately to saturation in liquid medium and equal volumes of each culture
357 are pooled to generate a single genomic DNA sample for library construction and high through-put
358 sequencing. The resulting sequence reads (>12 million) are analyzed for those that align at least partially
359 to the 300 bp putative SiRTA and show evidence of telomere addition (TG_{1-3} or $C_{1-3}A$ sequence). To
360 account for differences in read depth between experiments, the number of reads meeting these criteria
361 is normalized to the number of reads mapping to a 300 bp sequence within *BRR6,* an essential gene on
362 chromosome VII that lies centromere-proximal to the site at which the putative SiRTA is integrated
363 (Figure 1c). At least two biological replicates are generated for each strain. To benchmark SiRTA
364 efficiency based on our previous method, we applied PCR-based mapping and PT-seq to multiple 30-
365 colony samples derived from SiRTAs of a range of efficiencies. Using linear regression, we find a strong
366 correlation between the two methods (r^2=0.97), allowing us to use this standard curve to estimate the
367 number of colonies within a 30-colony sample that underwent dnTA at the putative SiRTA (Figure 1d,
368 closed triangles; Supplementary File 3). This method also yields information about the relative frequency
369 of telomere addition at each nucleotide position.
370 To verify that this method is applicable at other locations in the genome, we utilized PT-seq to
371 test SiRTA 9L44(-) at its endogenous location on chromosome IX. *Cis*- and *trans*-acting mutations with
372 effects on the efficiency of dnTA at SiRTA 9L44(-) were used to compare the PCR and PT-seq
373 methodologies over a range of SiRTA efficiencies. Again, results of the two methods are strongly
374 correlated (r^2=0.95; Figure 1d, open circles; Supplementary File 3). The slopes of the standard curves
375 generated at both chromosome locations are statistically indistinguishable ($p = 0.76$ by analysis of
376 covariance), suggesting that the percent of GCR events incurring dnTA (SiRTA efficiency) can be
377 accurately estimated from PT-seq results regardless of chromosome location.
378

379 *Putative SiRTAs are accurately identified using a computational method*

380 Visual inspection of sequences found to function as SiRTAs revealed similarity to yeast telomeric
381 sequences, consistent with prior work demonstrating that association of Cdc13 with the Stim sequence
382 is required for dnTA (Obodo et al. 2016). The Core sequence is also TG-rich, likely reflecting required
383 complementarity to the *TLC1* template sequence and (perhaps) the ability to associate with Cdc13. We
384 postulated that SiRTA function could be predicted by considering not only the TG-richness of a sequence
385 but also its similarity to the pattern of the yeast telomeric repeat (TG_{1-3}). SiRTA function does not require
386 a perfect match to the telomeric sequence, so we developed a strategy to score similarity to a telomeric
387 repeat while allowing divergence from that pattern. The Computational Algorithm for Telomere Hotspot
388 Identification (CATHI) identifies strings of consecutive Gs and Ts, awards one point for each base in that
389 string, and subtracts 1.5 points for each instance of GGTGG, a sequence that is not found in yeast
390 telomeres (Figure 2a). Calculations are done in a sliding window that can be varied in size (*Materials and*
391 *Methods*).

392 The algorithm was developed through an iterative process in which sequences were identified
393 and tested for SiRTA function. This dataset included several previously published SiRTAs, sequences
394 identified during the work described here, and negative control sequences that were not expected to
395 function as SiRTAs. To standardize measurements of SiRTA efficiency, all the tested sequences were
396 assayed on chromosome VII by inserting a 300 bp region encompassing the putative SiRTA. If boundaries
397 of the SiRTA sequence were previously established, the SiRTA was centered within the 300 bp region. All
398 sequences were tested at least in duplicate and the average percent of clones undergoing telomere
399 addition within the test sequence was determined in comparison to the chromosome VII standard curve
400 (Figure 1d). During initial testing, data were obtained for 37 sequences, seven of which had an average
401 SiRTA efficiency above the 6.6% cutoff. To optimize the algorithm, we calculated a score for each of the
402 37 sequences using varying window sizes (25 to 150 bp) and penalties (0 to 3) to identify a combination

403 generating the best fit by linear regression (Supplementary File 4). A window size of 75 and a penalty of
404 1.5 for GGTGG sequences yielded the highest correlation between CATHI score and SiRTA efficiency (r^2 =
405 0.63 for all sequences and 0.69 for the seven sequences exceeding the 6.6% cutoff for SiRTA function).
406 Testing of additional sequences after the algorithm parameters were established resulted in a
407 final dataset of 47 sequences (13 active as SiRTAs) that are graphed relative to CATHI score in Figure 2b
408 (Supplementary File 3). Using the threshold of 6.6% to define an active SiRTA, a CATHI score of 20
409 effectively separates active and inactive sequence with a false positive rate of ~2% (1/47) and a false
410 negative rate of ~4% (2/47). We conclude that the algorithm can be used to accurately identify
411 sequences with a propensity to stimulate dnTA. For those sequences with a CATHI score of 20 or
412 greater, the score is moderately predictive of SiRTA efficiency (r^2=0.43; p=0.015). The sequences with
413 the four highest CATHI scores tested are also the most efficient. However, scores between 20 and 30 are
414 less predictive of efficiency, suggesting that some aspects of SiRTA function are not captured by the
415 algorithm (see *Discussion*).
416
417 *Distribution of SiRTAs across the yeast genome*
418 Using the algorithm parameters established above, the 16 chromosomes (excluding any
419 terminal TG_{1-3} telomeric sequences; see Supplementary File 2 for coordinates) were scanned as a series
420 of 75 bp sliding windows with a step size of 1. Overlapping windows with scores of 20 or greater were
421 merged such that the starting and ending coordinates of a predicted SiRTA represent the maximum
422 distance between the first and last window meeting the threshold value. The final score assigned to a
423 set of overlapping windows is equivalent to the highest CATHI score in that set. The algorithm was
424 separately applied to the top and bottom strands and strand information was retained. Overall, we
425 identified 728 sequences in the *S. cerevisiae* genome with a CATHI score of 20 or greater
426 (Supplementary File 4).

427 We examined the overall distribution of these 728 sequences within the 16 yeast chromosomes
428 (Figure 3). SiRTAs on the top strand (5' to 3'; 342) are shown in blue and those on the bottom (3' to 5';
429 386) strand are shown in red (Figure 3b and c). The centromere of each chromosome is depicted as a
430 black circle. The overall distribution between the two strands is not different from random (p=0.25 by
431 chi-square test). However, there is a minor but significant tendency for the SiRTAs to cluster on the
432 same strand. This effect is quantified by measuring the number of times a SiRTA is found on the
433 opposite strand from its neighbor (number of "strand switches"). We observe 322 strand switches
434 across the genome, significantly fewer than the number expected by chance (351.8 ± 13.0; p=0.013;
435 Supplementary Figure 2a). This difference is driven almost entirely by a reduction in the number of
436 "singlet" SiRTAs compared to what would be expected by chance (148 observed versus 187.7 ± 15.0
437 expected; p=0.0043; Supplementary Figure 2b). In contrast, runs of longer length do not deviate
438 significantly from expectation. We conclude that there is a minor tendency for SiRTAs to cluster on the
439 same strand, but only with the nearest neighbor.
440 In a haploid cell, chromosome truncation proximal to the last essential gene on a chromosome
441 arm will be lethal. Therefore, we were interested in determining whether the distribution of putative
442 SiRTAs is different in essential versus nonessential regions. For our analysis, we defined the nonessential
443 region on each chromosome arm as comprising sequences distal to the last essential gene (Figure 3a;
444 Supplementary File 2). The last essential gene, in turn, is the most telomere-proximal gene for which
445 single gene deletion was reported to cause lethality during the systematic knockout of each open
446 reading frame in *S. cerevisiae* (Giaever et al. 2002). This definition does not take synthetic lethality into
447 account; some nonessential regions may be smaller than defined here if the combined loss of one or
448 more genes in that region results in cell death. In Figure 3b, nonessential regions are highlighted in gray;
449 those same regions are shown in expanded form in Figure 3c. The nonessential regions are divided into
450 sequences that are unique (in most cases) among the different chromosome arms and the highly

451 repetitive subtelomeric X and Y' elements found immediately adjacent to the telomeric repeats (Figure
452 3a). All chromosome arms contain at least a portion of the X element while some also contain one or
453 more Y' elements (~6 kb each; Louis and Haber 1990; Zhu and Gustafsson 2009). The transition to
454 subtelomeric sequence is shown on each chromosome arm as a black line (Figure 3c). Diagrams in Figure
455 3 are based on the published sequence of reference strain S288C. Recent long-read sequencing analyses
456 confirm that some subtelomeric regions contain additional terminal sequences (primarily Y' elements)
457 that were not included in the published reference genome (Yue et al. 2017), so our analysis likely
458 underestimates the number of subtelomeric SiRTAs.

459 To test the hypothesis that SiRTAs are preferentially located in nonessential terminal regions,
460 we utilized a permutation-based enrichment test to compare the distribution of SiRTAs between
461 essential and non-essential regions to that of randomly shuffled sequences matched in number and
462 length to the sequences identified by the CATHI algorithm. This analysis shows significant enrichment for
463 SiRTAs in the nonessential regions of the genome ($p<0.01$) but this enrichment disappears when the
464 subtelomeric regions (X and Y' elements) are excluded from the analysis (Figure 4a). We conclude that
465 SiRTAs are disproportionately enriched in nonessential regions, but that this effect is limited to the most
466 distal, highly repetitive subtelomeric sequences. Seventy-five of the 728 putative SiRTAs lie within the
467 subtelomeric sequences. Nine of those 75 sequences consist of perfect telomeric (TG_{1-3}) repeats located
468 between X and Y' elements and seven of these perfect repeats constitute the top scoring sites in the
469 genome (Figure 4b and Supplementary File 5). The remaining predicted SiRTAs in subtelomeric regions
470 are not comprised of perfect telomeric repeats and lie within the X or Y' elements. We were concerned
471 that the tracts of perfect telomeric repeats might affect our analysis. However, when the nine perfect
472 telomeric repeats are excluded, there remains significant enrichment for SiRTAs within the nonessential
473 regions ($p=0.001$; Supplementary Figure 3). Notably, all chromosome ends [with one exception: 6R)]
474 contain at least one region predicted to function as a SiRTA (Figure 3c). We conclude that SiRTAs are

475 disproportionately found within the subtelomeric regions but are otherwise not significantly enriched
476 within nonessential sequences.

477 As described in the *Introduction*, the strand on which a SiRTA is located has important
478 implications for the consequence of dnTA. SiRTAs in the TG-orientation (those oriented to stabilize the
479 centromere-containing fragment when a break occurs distal to the site) are on the bottom strand for the
480 left arm of a chromosome and on the top strand for the right arm. SiRTAs in the genome as a whole do
481 not show a bias for the TG- versus CA-orientation (372 versus 356; p=0.675 by chi-square test). To
482 examine the distribution more carefully, we examined SiRTA orientation within the nonessential regions,
483 where an excess of SiRTAs was already noted (Figure 4a). Interestingly, enrichment within nonessential
484 regions is observed only for SiRTAs in the TG-orientation ($p<0.01$), while SiRTAs in the CA-orientation are
485 not significantly enriched (Figure 4c). The same result is observed when the nine perfect TG_{1-3} repeats
486 are removed from the analysis (p=0.001; Supplementary Figure 3). As expected, enrichment is no longer
487 observed when subtelomeric sequences are excluded (Figure 4c). This differential distribution is
488 visualized in a plot showing the CATHI score of predicted SiRTAs within the nonessential chromosome
489 regions. In non-subtelomeric regions, distributions are indistinguishable for the two orientations (Figure
490 4b). In contrast, SiRTAs in the X and Y' elements are much more likely to be in the TG-orientation than in
491 the CA-orientation ($p<0.0001$ by chi-square test; Figure 4b). The enrichment of TG-oriented SiRTAs
492 within subtelomeres is also visually apparent in the clustering of red symbols at or near the subtelomere
493 junction on all left arms and of blue symbols on nearly all right arms (except 6R; Figure 3c).

494 We tested the ability of three sequences identified in subtelomeric regions to stimulate dnTA
495 using the HO cleavage assay. Two of these sequences overlap with an X element [14L07(-) and
496 15R1084(+)] and one overlaps with a Y' element [7R1089(-)]. All three sequences function as SiRTAs,
497 stimulating de novo telomere addition at frequencies of 10.0%, 13.3% and 36.6%, respectively
498 (Supplementary Figure 4). Although 7R1089(-) functions well as a SiRTA in our assay, it is worth noting

499 that it is in the CA orientation. Because it is part of a conserved Y' sequence, a very similar sequence
500 occurs at multiple chromosome ends (also in the CA orientation). The X element sequences are TG-
501 oriented and represent sequences found in two distinct regions of the X element sequence. Taken
502 together, these results support the interesting possibility that sequences capable of functioning as
503 SiRTAs have been retained near chromosome termini to facilitate chromosome rescue in the event of
504 telomere loss.
505
506 *TG-rich sequences identified by the algorithm are overrepresented in the yeast genome*
507 We were curious whether sequences predicted to stimulate dnTA are found in yeast at the
508 expected frequency given the nucleotide composition of the yeast genome. To address this question, we
509 generated five scrambled genomes identical in sequence composition to the yeast genome. To avoid
510 potential biases introduced by the subtelomeric regions, this analysis was done on sequences from
511 which the subtelomeric X and Y' elements were excluded (see *Materials and Methods*). The scrambled
512 genomes contain, on average, 283.2 ± 18.3 sequences that score 20 or higher compared to 653
513 sequences observed in the yeast genome, an excess of 2.3-fold. The differential becomes increasingly
514 apparent at higher scores with an excess of 1.8-fold at a score of 20 and an excess of 5.2-fold at a score
515 of 25. Among the scrambled genomes, an average of fewer than one sequence has a score of 30 or
516 higher (range 0 to 2), while 21 sequences scoring 30 or higher are observed in the *S. cerevisiae* genome
517 (Figure 5a and Supplementary File 6). In Figure 5b, putative SiRTAs with scores of 27 or higher are shown
518 to emphasize the strikingly different distributions in the simulated versus actual genomes. To address
519 whether the excess of higher scores might be functionally related to SiRTA function, we examined the
520 predicted and actual occurrence of CATHI scores less than 20, which are unlikely to stimulate increased
521 levels of dnTA (see Figure 2b). For sequences with CATHI scores of 15-19, we still observe an excess in

522 the actual genome, although the excess is less pronounced (1.3-fold, with 2703 ± 16.3 predicted
523 compared to 3485 observed; Figure 5c).
524
525 *TG-dinucleotide repeats stimulate dnTA and are among the strongest SiRTAs in the genome*
526 In analyzing predicted sites of dnTA, our attention was particularly drawn to SiRTA 6R210(+).
527 With a score of 61, this sequence represents the strongest predicted site outside the subtelomeric
528 regions (see outlier in Figure 5b). SiRTA 6R210(+) contains a nearly perfect 62 nucleotide TG-
529 dinucleotide repeat and is the longest TG-dinucleotide repeat in the *S. cerevisiae* genome (the next
530 longest is 41 nt; Supplementary File 5). Interestingly, the sequence is in the TG-orientation but lies
531 centromere-proximal to the last essential gene on the left arm of chromosome VI, implying that repair
532 by dnTA at this site in a haploid cell would be lethal.
533 To determine the efficiency of dnTA at SiRTA 6R210(+), we inserted the TG-dinucleotide repeat
534 (centered within a 300 bp region) at the test site on chromosome VII. Most strains that we monitor for
535 SiRTA function on chromosome VII generate GCR events at a frequency of ~0.001%, equivalent to a
536 negative control strain lacking a SiRTA. In contrast, a strain containing the 62 bp TG-dinucleotide repeat
537 generates 5-FOA resistant colonies at a 10-fold higher frequency of ~0.01% (Figure 6a). By PT-seq, 86%
538 of GCR events involve dnTA addition within the inserted sequence (Figure 6b). Therefore, although the
539 percent of events at the SiRTA that involve dnTA is similar in strains carrying the TG-dinucleotide repeat
540 versus the efficient SiRTA 14L35(-) (76%), the overall frequency with which dnTA occurs at the 62-nt
541 dinucleotide repeat is at least 10 times higher. We conclude that SiRTA efficiency alone (defined as the
542 percentage of GCR events in which telomere addition occurred within the sequence of interest)
543 underestimates the propensity of a sequence to stimulate telomere addition when SiRTA activity is very
544 high. For this reason, we did not include SiRTA 6R210(+) in the comparison of CATHI score and SiRTA
545 efficiency in Figure 2b.

546 Using the PT-seq results, we mapped the sites at which dnTA occurred at SiRTA 6R210(+) (Figure
547 6c). Each arrow corresponds to the last nucleotide that aligns between the chromosome and at least
548 one PT-seq read, representing the 3'-most nucleotide at which synthesis by telomerase may have
549 initiated. Given that ~86% of the 60 individual strains represented in the analysis contain evidence of
550 dnTA, these results correspond to the mapping of 50-52 independent telomere addition events. Sites
551 identified in a larger fraction of reads were likely targeted by telomerase in multiple independent
552 clones. Consistent with our prior observation that a 5' Stim sequence is required to stimulate telomere
553 addition in a 3' Core sequence (polarity relative to the TG-rich strand), the vast majority of telomere
554 addition events occur in the 3' half of the dinucleotide repeat or in sequences immediately downstream.
555 Two events occurred at least 50 bases downstream of the TG-repeat, consistent with prior reports that
556 TG-rich sequences can stimulate dnTA within neighboring sequences (Kramer and Haber 1993;
557 Mangahas et al. 2001).
558 Excluding the subtelomeric regions, TG-dinucleotide repeats comprise 11 of the 21 CATHI scores
559 of 30 or greater (Figure 6d). SiRTA 7L69(-) (CATHI score = 34) contains a 34-nt perfect TG repeat and was
560 identified by the Zakian laboratory as a site capable of stimulating dnTA in response to a DSB induced
561 more than 50kb distal to the eventual site of telomere addition (Mangahas et al. 2001). When
562 integrated at the test site on chromosome VII, this 34-nt repeat stimulates dnTA with an efficiency of
563 47.9% by PT-seq (Supplementary File 3). Together, these observations focus attention on TG-
564 dinucleotide repeats as potential mediators of genome instability.
565
566 *Sequences that function to stimulate de novo telomere addition bind Cdc13 in vitro*
567 Previous studies demonstrated that Cdc13 binding at the Stim sequence is required to promote
568 dnTA. We hypothesized that sequences with CATHI scores of 20 or greater will bind Cdc13 with greater
569 affinity than sequences with lower scores. Additionally, we predicted that the two sequences identified

570 as false negatives in our initial analysis (Figure 2b) should bind Cdc13 with higher affinity than the single
571 sequence identified as a false positive. To test these predictions, we utilized fluorescence polarization to
572 measure the ability of unlabeled, 75-base oligonucleotides to reduce association of the Cdc13 DNA
573 binding domain (Cdc13-DBD) with a 6-carboxyfluorescein (FAM) labeled 11-mer containing the canonical
574 Cdc13 binding site (5'-GTGTGGGTGTG; referred to here as Tel11). Cdc13-DBD binds with similar
575 sequence specificity and affinity to the Tel11 sequence as full-length Cdc13 (Lewis et al. 2014). The goal
576 of these analyses was not to identify individual Cdc13 binding sites, but rather to measure the relative,
577 cumulative ability of each sequence to bind Cdc13.

578 We first established that the FAM-labeled Tel11 oligonucleotide binds Cdc13-DBD
579 (Supplementary Figure 5a) and selected concentrations of 30 nM Cdc13-DBD and 25 nM labeled Tel11
580 for the competition analyses. The apparent inhibition constant ($K_{i,app}$) is defined as the concentration of
581 each competitor required to reduce binding to the FAM-labeled Tel11 by half. An unlabeled 75-mer
582 containing the Tel11 sequence at the center of the oligonucleotide (Tel11-75) was included in each
583 experiment and normalized $K_{i,app}$ values are reported as fold change relative to this control ($K_{i,app}$ of
584 Tel11-75/$K_{i,app}$ of experimental oligonucleotide). Sequences flanking the Cdc13 consensus binding site in
585 Tel11-75 lack any TG or GG motifs to minimize additional association of Cdc13-DBD. Oligonucleotides
586 used in these assays are found in Supplementary File 1 and representative competition curves are
587 shown in Supplementary Figure 5b. To validate the method, we determined the normalized $K_{i,app}$ of a
588 double-stranded version of Tel11-75 (0.5 +/- 0.3) and the inverse complement of the Tel11-75 sequence
589 (0.4 +/-0.2), both of which show the expected reduction in binding relative to Tel11-75 (Figure 7a). A 75-
590 mer sequence from chromosome VI previously shown to lack SiRTA activity (CATHI score=5) also
591 competes very weakly for Cdc13-DBD association (normalized $K_{i,app}$ = 0.5 +/- 0.2; Figure 7a). Finally, as
592 expected, a 2xTel11-75 sequence that contains two adjacent Tel11 sequences competes twice as well as
593 the Tel11-75 control oligonucleotide (normalized $K_{i,app}$ = 2.2 +/- 0.2; Supplementary Figure 6a).

594 We chose to test several sequences with CATHI scores over 20 that had been previously shown
595 to stimulate dnTA. Oligonucleotides were designed to correspond to the 75 bases with the highest
596 CATHI score within the 300bp sequence tested for SiRTA function. Both 14L35(-) and 14R131(+) compete
597 in a manner indistinguishable from the Tel11-75 control sequence (normalized $K_{i,app}$ of 1.0 +/- 0.3 and
598 0.9 +/-0.2, respectively) and bind more robustly than the negative control sequences (Figure 7a). The
599 two sequences identified as false negatives in Figure 2b [2R780(-) and 14R306(+)] both compete more
600 effectively than the Tel11-75 control sequence (normalized $K_{i,app}$ of 1.6 +/- 0.9 and 1.4 +/- 0.6,
601 respectively; Figure 7a). This observation is consistent with the ability of these sequences to stimulate
602 dnTA and suggests that the algorithm fails to predict Cdc13 binding in some cases. We also tested the
603 false positive sequence [12R330(+)] with a CATHI score of 22 and an average dnTA frequency of 2.3%
604 (below our cut-off for SiRTA function). This 75-base sequence has a normalized $K_{i,app}$ of 0.7 +/- 0.2,
605 intermediate to that of the Tel11-75 control sequence and the negative controls (Figure 7a).
606 Given the extremely high SiRTA activity of the 62-nt TG-dinucleotide repeat described above, we
607 tested the ability of a 75-mer containing this repeat to compete for Cdc13-DBD binding. The normalized
608 $K_{i,app}$ of 4.9 +/- 2.9 measured for this sequence is considerably higher than any other sequence tested
609 (Figure 7b). The first, second, and fourth base of the canonical Cdc13 binding site (5'-**G**T**GT**GGGTGTG)
610 contribute most strongly to Cdc13 affinity, comprising a GxGT motif that recurs in the TG-dinucleotide
611 motif. The 62-nt dinucleotide repeat is predicted to accommodate approximately five 11-mer binding
612 sites, remarkably close to the observed 4.9-fold increase in competition compared to the Tel11-75
613 control oligonucleotide with a single binding site.
614 Our prior analysis of SiRTA 2R780(-) presented an additional opportunity to test the correlation
615 between Cdc13 binding and SiRTA efficiency (Hoerr et al. 2023). SiRTA 2R780(-) contains a Stim
616 sequence of ~50 bases, deletion of which abrogates dnTA. We previously showed that mutation of
617 either one of two GxGT motifs located in this 50 nt Stim sequence greatly diminishes SiRTA function, an

618 effect that we attributed to reduced Cdc13 association (Hoerr et al. 2023). Consistent with this
619 hypothesis, we find that mutation of one or both motifs significantly reduces the ability of the
620 oligonucleotide to compete for Cdc13-DBD binding (Figure 7c).

621 The experiments described above provide evidence that sequences capable of stimulating dnTA
622 associate more robustly with Cdc13 than sequences that do not function as SiRTAs, although our ability
623 to distinguish borderline cases is limited. While there appears to be a threshold of binding required for
624 SiRTA function, the cumulative "affinity" of a sequence measured in this assay is not fully predictive of
625 SiRTA efficiency [e.g. 14L35(-) and 14R131(+) compete equivalently, but differ by a factor of two in SiRTA
626 efficiency; 80.5% versus 30.7%]. This discrepancy may result in part from our choice of 75-mer sequence
627 to test in each case, but also likely reflects the specific number, affinity, and distribution of Cdc13
628 binding sites within the sequence.

629 As another approach to benchmark the effect of high affinity Cdc13 binding on SiRTA function,
630 we tested the ability of either a single canonical Cdc13 binding site or two tandem sites to stimulate
631 dnTA at the test site on chromosome VII. One copy of the Tel11 site stimulated telomere addition in only
632 four or five of 60 GCR events analyzed by PT-seq (7.5%; Supplementary Figure 6b). Remarkably, adding a
633 second Tel11 (2xTel11) site increases the frequency GCR events undergoing dnTA to 83.3%
634 (Supplementary Figure 6b). Together these results demonstrate the importance of Cdc13 binding sites
635 for stimulating dnTA at SiRTAs.

636

637 *SiRTA distribution is not strongly associated with known protein binding sites or chromosome landmarks*

638 To gain insight into factors that may contribute to dnTA, we examined whether putative SiRTAs
639 preferentially overlap with the binding sites of proteins related to telomere addition such as Est2 (a
640 component of telomerase recently shown to associate with internal chromosome sites; Lendvay et al.
641 1996; Pandey et al. 2021), Rap1 (a transcription regulator that also binds telomeric repeats and affects

642 telomere length homeostasis; Conrad et al. 1990; Rhee & Pugh, 2011), and Pif1 (a helicase that
643 negatively regulates telomerase at telomeres and DNA double-strand breaks; Schulz and Zakian 1994;
644 Paeschke et al. 2011). We also examined the correlation between predicted SiRTAs and fragile sites,
645 identified as regions that associate with γH2AX even in the absence of exogenous damage (Downs et al.
646 2000; Capra et al. 2010), and between SiRTAs and sequences predicted to form G-quadruplex structures
647 (Capra et al. 2010). Using a permutation-based enrichment test under conditions that require overlap
648 with at least half of the predicted SiRTA sequence, we find statistically significant enrichment among
649 SiRTAs for Est2, Rap1, γH2AX binding sites, and G4-forming sequences (Supplementary Figure 7a and b).
650 However, overlap never exceeds 20% of the putative SiRTAs, arguing against a strong functional
651 relationship. Although the actual number of putative SiRTAs overlapping with a Pif1 binding site is the
652 highest of all characteristics tested (18.4%), this extent of overlap is not significant, likely because the
653 regions reported to bind Pif1 by chromatin immunoprecipitation are relatively broad. To determine
654 whether the predicted strength of the SiRTA affects these results, we divided putative SiRTAs into
655 tertiles based on CATHI score but found no strong relationship between CATHI score and the
656 significance of overlap (Supplementary Figure 7c). There is also no obvious enrichment among
657 sequences that are known to function as SiRTAs. Of the 14 SiRTAs known to be active, only one overlaps
658 with an Est2 binding site and one overlaps with a Rap1 binding site. Overall, these results fail to identify
659 any overlapping binding sites with evidence of strong functional significance and suggest that fragile
660 sites and G-quadruplex forming sequences are not strongly correlated with predicted hotspots of dnTA.
661

662 *SiRTAs are predominantly found within coding regions*

663 We were interested in determining whether SiRTAs are preferentially excluded from genic
664 regions due to higher levels of evolutionary constraint. Each of the 728 SiRTAs was categorized as genic
665 (any part of the SiRTA overlapped with a gene as defined by the start and stop codon of each annotated

666 gene) or intergenic (Supplementary File 9). Seventy-eight percent of all SiRTAs overlap with coding
667 regions and only 22% are exclusively found in intergenic regions (Supplementary Figure 8). Given that
668 approximately 30% of the yeast genome is intergenic (Hurowitz and Brown 2003; Lynch et al. 2008), we
669 conclude that SiRTAs are not excluded from expressed regions. There are two exceptions. When a SiRTA
670 contains long (>20 nt) TG-dinucleotide repeats, those repeats are virtually never found within a coding
671 region. This result is not surprising since expansion or contraction of a dinucleotide repeat is expected to
672 disrupt the open reading frame. Second, intergenic SiRTAs are disproportionately found in the
673 subtelomeric X and Y' elements. While only 8% of all SiRTAs are in the subtelomeric regions, 37% of
674 intergenic SiRTAs are subtelomeric, consistent with the presence of few transcribed regions in the
675 subtelomeres (Supplementary File 9). Interestingly, for those SiRTAs that overlap with open reading
676 frames, 58% are located on the template strand, which is different from the expectation of random
677 distribution (Supplementary Figure 8; $p<0.05$ by Chi square test) and suggests that the presence of these
678 sequences within genes may, in some cases, have consequences for cellular fitness.
679

680 **Discussion**

681 *Prediction of SiRTA function*

682 In this work, we predict the distribution of SiRTAs in the yeast genome, an important step in
683 understanding the role of these sequences in genome stability and function. The Zakian laboratory
684 initially proposed that hotspots of dnTA addition contain tracts of 15 or more nucleotides consisting
685 exclusively of T and G in a "telomere-like" pattern (Mangahas et al. 2001). Our subsequent analysis of
686 SiRTAs on chromosome V and IX revealed that these requirements are too strict. For example, SiRTA
687 5L35(-) (formerly called 5L-35) stimulates dnTA in response to both spontaneous and induced DSBs
688 (Stellwagen et al. 2003; Obodo et al. 2016; Ngo et al. 2020), but the longest uninterrupted string of TG

689 sequence is 14 nucleotides, including several instances of a TT motif that never occurs within telomeric
690 repeats.

691 Given this information, we set out to develop a method that could reliably predict whether a
692 particular sequence has the capacity to stimulate unusual levels of dnTA (Figure 2). The algorithm
693 described here prioritizes "telomere-like" sequences, but provides flexibility for some deviation from
694 that pattern. With a single exception (discussed below), sites previously identified to stimulate dnTA are
695 predicted by the algorithm to function as SiRTAs. For example, the Zakian lab identified three sites on
696 chromosome VII that stimulate dnTA following an induced DSB (Mangahas et al. 2001). Two of these
697 sites, now renamed SiRTA 7L67(-) and 7L69(-) (CATHI scores of 28 and 34, respectively), were originally
698 found to stimulate dnTA at a distance of more than 50 kb from the induced DSB. In the standardized
699 conditions of our chromosome VII test site where the break is induced ~2 kb from the sequence of
700 interest, these SiRTAs stimulate dnTA with efficiencies of 49.1% and 47.9% (Figure 2b and
701 Supplementary File 3). The third site identified by the Zakian lab lies within the *URA3* gene, integrated at
702 an ectopic location internal to the induced break. Although we did not test this sequence in our assay, it
703 has a CATHI score of 22 and is annotated as SiRTA 5R117(+) to reflect the native location of *URA3* on
704 chromosome V (Supplementary File 4).

705 Overall, the algorithm presented here correctly predicts SiRTA function (yes or no) with an
706 accuracy close to 95% (44/47). Because false positives and false negatives occur at similar frequency, our
707 estimate of ~650 SiRTAs (excluding subtelomeric X and Y' elements) is likely quite accurate, based on the
708 definition of a SiRTA proposed here.

709

710 *Sequences that stimulate dnTA associate with Cdc13*

711 Because prior work suggests that dnTA is stimulated by association of Cdc13 with single-
712 stranded DNA generated after a DSB, the CATHI algorithm likely identifies sequences with affinity for

713 Cdc13. To test this hypothesis, we developed a fluorescence polarization competition assay in which
714 sequences are tested for their relative ability to compete with a labeled oligonucleotide for binding to
715 the purified Cdc13 DNA binding domain. A 75-mer oligonucleotide containing two tandem copies of the
716 canonical Cdc13 binding site competes twice as well as an oligonucleotide containing a single site,
717 suggesting that the assay is sensitive to Cdc13 binding (Supplementary Figure 6a). Our goal is to
718 measure overall association of Cdc13, which arises as a combination of the number and affinity of
719 binding sites. We find that 75-mer oligonucleotides containing sequences that stimulate dnTA compete
720 as well or better than a 75-mer containing a single match to the telomeric consensus Cdc13 binding
721 sequence. In contrast, sequences that fail to support dnTA compete less well (Figure 7a). Importantly,
722 mutations previously demonstrated to reduce SiRTA function also reduce Cdc13 binding (Figure 7c).
723 Together with our previous demonstration that dnTA is stimulated through the artificial recruitment of
724 Cdc13 to the Stim sequence of a SiRTA (Epum et al. 2020; Hoerr et al. 2023; Obodo et al. 2016), these
725 results are consistent with the requirement for a threshold level of Cdc13 in stimulating dnTA.
726 Despite the observation at a single SiRTA that Cdc13 association correlates well with SiRTA
727 efficiency, the apparent overall affinity for Cdc13 measured by fluorescence polarization poorly predicts
728 SiRTA efficiency. For example, SiRTA 14L35(-) competes equivalently with the control sequence but
729 stimulates dnTA more strongly than most other SiRTAs, including those that compete more effectively
730 for Cdc13 binding. This apparent discrepancy may reflect an effect of the distribution or spacing of
731 Cdc13 binding sites on SiRTA function. In prior work, we observed that deletion of the ~30 nt spacer
732 region between the Stim and Core sequences of a SiRTA dramatically increases SiRTA efficiency (Obodo
733 et al. 2016). The highly efficient SiRTA 14L35(-) contains an unusually long region of TG-rich sequence
734 that likely acts as both a Stim and Core region with little or no spacer, a property that may account for
735 its ability to stimulate dnTA strongly despite an overall lower affinity for Cdc13.
736

737 *Limitations to the predictive capacity of the CATHI algorithm*

738 The well-characterized and functional SiRTA 2R780(-) has a CATHI score of only 13, despite
739 stimulating dnTA with an efficiency of 31.11% (Hoerr et al. 2023). By fluorescence polarization, this
740 sequence competes for Cdc13 binding more effectively than many sequences with higher CATHI scores,
741 suggesting that the failure of the algorithm to predict SiRTA function (at least in this case) is primarily a
742 failure to predict Cdc13 binding. One possible explanation is that the CATHI algorithm does not prioritize
743 matches to the GxGT motif identified as particularly impactful for Cdc13 binding (Anderson et al. 2002).
744 Indeed, we find that mutation of even one GxGT motif in SiRTA 2R780(-) strongly reduces Cdc13 binding
745 and nearly eliminates SiRTA function (Figure 7c) (Hoerr et al. 2023). The 226 bp minimal sequence of
746 SiRTA 2R780(-) contains seven GxGT sequences that may account for the ability of this sequence to
747 stimulate dnTA despite lacking sufficiently long/abundant telomere-like tracts to be identified by the
748 algorithm.

749 Although the presence of GxGT motifs is an attractive explanation for the activity of SiRTA
750 2R780(-), attempts to incorporate the motif into the algorithm did not improve the accuracy with which
751 SiRTA function could be predicted and instead increased the number of false positive results. For
752 example, the false negative SiRTA 14R306(+) (CATHI score = 15.5) contains five GxGT motifs (two of
753 which overlap), but the false positive SiRTA 12-330(+) (CATHI score = 22) also contains five distinct GxGT
754 motifs. There are at least three (non-exclusive) explanations for remaining discrepancies between the
755 predictive algorithm and measured rates of dnTA. First, it remains unclear how Cdc13 affinity is affected
756 by deviation from the consensus telomere binding site. Although extensive mutagenesis has been
757 conducted *in vitro*, these studies either altered single nucleotide sites (showing that positions 2 and 5-11
758 are tolerant of single changes) or simultaneously mutated the seven 3'-most nucleotides (showing that
759 the GxGT motif is insufficient; Lewis et al. 2014; Glustrom et al. 2018). Neither approach fully
760 recapitulates the sequences that Cdc13 will encounter at internal sites exposed by resection. Second, as

761 described above, the distance between Cdc13 binding sites likely contributes strongly to SiRTA function.
762 We have attempted to account for this property by using a window size of 75. However, some effects on
763 SiRTA efficiency likely arise from varied distributions of Cdc13 binding sites that we cannot fully capture
764 with the algorithm. Third, we suspect that dnTA can be stimulated either by a small number of high-
765 affinity Cdc13 binding sites or by a larger number of low-affinity sites. SiRTAs at either extreme of this
766 continuum may be difficult to identify using the current strategy.
767
768 *SiRTAs do not colocalize strongly with binding sites for other telomere/telomerase-associated proteins*
769 Our co-localization analysis failed to identify additional proteins that strongly impact SiRTA
770 function. Overall, we consider it unlikely that the observed enrichment represents a functional
771 relationship. For example, Rap1 binding sites are overrepresented among SiRTAs, but this result is not
772 surprising given that the consensus binding site for Rap1 is also TG-rich. In previous work, we showed
773 that Rap1 association *per se* is not required for SiRTA activity (Obodo et al. 2016). Our results suggest
774 that Est2 binding in undamaged conditions is not required for a sequence to function as a SiRTA. Only
775 8.1% of predicted SiRTAs overlap with experimentally determined sites of Est2 enrichment and only one
776 of the fourteen active SiRTAs is also an internal Est2 binding site.
777 Consistent with our observation that SiRTAs stimulate dnTA even when located several kilobases
778 from an induced DSB, we observe only a modest correlation between sequences that function as SiRTAs
779 and sites enriched for phosphorylated H2A (γH2AX), a marker of DNA damage. Since enrichment was
780 determined in undamaged cells, these sites represent regions of the genome that are prone to
781 spontaneous damage, likely due to difficulties encountered during DNA replication. It will be interesting
782 to determine whether SiRTAs that overlap with fragile sites are more likely to stimulate the spontaneous
783 formation of gross chromosomal rearrangements. For example, we have proposed that the generation
784 of acentric fragments through dnTA on chromosome II is facilitated by an unusually high density of

785 inverted repeats in this region combined with errors in the resolution of stalled replication forks (Hoerr
786 et al. 2023). In this light, it is interesting to note that sequences required to stimulate dnTA on
787 chromosome II [SiRTA 2R780(-), coordinates 779784 to 780009] overlap with a region of enhanced
788 γH2AX association (779987-780040; Capra et al. 2010).

789 Interestingly, we find a statistically significant tendency for predicted SiRTAs, if found within an
790 open reading frame, to be oriented with the TG-rich sequence on the template strand. We propose that
791 the presence of the TG-rich sequence within the exposed strand of the transcription bubble may be
792 deleterious. Interestingly, this bias is opposite to that observed for G-quadruplex forming sequences in
793 mammalian cells, which are more likely to be found on the coding strand (Agarwal et al. 2014; Rhodes
794 and Lipps 2015; Kim 2019). In yeast, Replication protein A (RPA)-bound single-stranded DNA at stalled
795 transcription complexes has been implicated as a major signal of DNA damage (Wang and Haber 2004;
796 Tapias et al. 2004; Fanning 2006). Conceivably, competition for binding to single-stranded DNA by Cdc13
797 could interfere with this process, leading to selection against Cdc13 target sequences on the exposed
798 coding strand. Despite the frequent presence of SiRTAs within genes, transcription does not appear to
799 be required for SiRTA function since the test site that we developed on chromosome VII is contained
800 within an intragenic region, more than 1.5 kilobases from the 3' end of the *ADH4* gene.

801

802 *Implications of SiRTA distribution in the yeast genome*

803 The compact and well-annotated yeast genome presents an opportunity to assess evidence of
804 selective pressures that might act upon sequences with a propensity to stimulate dnTA. Based on the
805 hypothesis that dnTA within the nonessential terminal region of a chromosome arm might provide a
806 selective advantage by allowing a cell to survive a persistent DSB, we examined the distribution of
807 predicted SiRTAs in essential and nonessential chromosome regions. As predicted, we found a significant
808 enrichment of putative SiRTAs in nonessential terminal regions. Furthermore, as expected for a role in

809 chromosome stabilization, only SiRTAs in the TG-orientation are overrepresented. However, both effects
810 disappear when the subtelomeric X and Y' elements are removed from the analysis (Figure 4a and c).
811 Nine of the TG-oriented SiRTAs in the subtelomeric regions correspond to stretches of TG_{1-3}
812 (perfectly "telomeric") sequence that are located predominantly between tandem Y' elements.
813 However, the vast majority, while TG-rich, deviate substantially from the TG_{1-3} pattern. Whether these
814 sequences are vestiges of ancient telomeric repeats is unclear. Because the Y' and X elements are similar
815 between chromosome ends, many of the subtelomeric SiRTAs have similar or identical CATHI scores and
816 represent a small number of unique sequences. Given the near ubiquity of TG-oriented SiRTAs within
817 subtelomeric regions (identified at 31 of 32 chromosome ends), we speculate that these sequences are
818 conserved, at least in part, due to an ability to stimulate dnTA in the event of catastrophic telomere loss,
819 most likely due to replication errors within the telomeric repeats. The subtelomeric region on the right
820 arm of chromosome VI contains a truncated X element followed immediately by TG_{1-3} telomeric repeats
821 and therefore lacks sequences predicted to function as a SiRTA (Figure 3c). This subtelomeric structure
822 may have resulted from a prior dnTA event within the X element. In the future, it would be interesting to
823 determine whether the right arm of chromosome VI is more sensitive to catastrophic loss of telomeric
824 repeats than other chromosome termini that contain intact X element repeats.
825 While the spatial distribution and orientation of putative SiRTAs outside the subtelomeres are
826 not strongly skewed, the number of sequences with potential to act as SiRTAs is significantly higher than
827 predicted by chance (Figure 5). This excess is observed at both low and high scores, but is increasingly
828 pronounced at higher CATHI scores. Scores of 30 or greater are approximately 20-fold overrepresented
829 in the yeast genome compared to the random expectation (Figure 5b). Our data do not provide evidence
830 that SiRTA function *per se* is driving this excess, particularly because we also observe an excess of scores
831 below 20, representing sequences that are not likely to stimulate dnTA at unusually high levels (Figure
832 5c). Since many SiRTAs are located within coding regions, we considered the possibility that codon bias

833 might explain this pattern. However, codons consisting of only G and T (or only C and A; TTT and AAA
834 excluded) collectively are not overrepresented among all codons (Supplementary File 10; Nakamura et
835 al. 2000). It is possible that particular amino acid repeats could result in this effect. For example, poly-
836 proline tracts consisting of CCA and CCC codons can generate a SiRTA signature. However, only a small
837 fraction of poly-proline tracts are also identified as potential sites of dnTA.

838 An intriguing possibility is that association of Cdc13 with single-stranded DNA revealed by
839 resection may be important to stimulate fill-in synthesis of the resected strand. At telomeres, Cdc13, in
840 association with its binding partners Stn1 and Ten1, recruits polymerase α-primase to facilitate
841 resynthesis of the 5' recessed strand (Grandin 2001; Rice and Skordalakes 2016; Ge et al. 2020). In
842 mammalian cells, the complex of Ctc1, Stn1, and Ten1 fulfills the same role at both telomeres and
843 double-strand breaks (Chastain et al. 2016; Giraud-Panis et al. 2010; Wang et al. 2007). In this model,
844 TG-rich sequences may be retained in the genome at a higher-than-expected frequency to facilitate DNA
845 repair, with elevated rates of dnTA resulting as a rare byproduct.

846 Persistence of sequences such as the long TG-dinucleotide repeat on chromosome VI that
847 support extremely high levels of dnTA is surprising since it seems likely that such sequences would
848 interfere with normal repair. Future work will address whether dnTA is inhibited at SiRTAs in some
849 contexts (for example, the SiRTA on chromosome VI may be less capable of stimulating dnTA in its
850 endogenous location than when that same sequence is integrated at the test site on chromosome VII).
851 Alternatively, the deleterious consequences incurred by dnTA at such a sequence may be insufficient to
852 result in purifying selection or the TG-dinucleotide repeat may contribute to cell fitness through some
853 other mechanism.

854 This work provides, for the first time, a genome-wide map of sites predicted to stimulate dnTA.
855 With the exception of sites clustered in subtelomeric regions, the largely random distribution and
856 orientation of SiRTAs throughout the genome stands in interesting contrast to the observation that

857 sequences with this capability are found much more often than expected by chance. The tools
858 presented here will facilitate studies to address this apparent contradiction and to determine the impact
859 of these sequences on genome stability and evolution.
860

861 **Data availability**

862 Strains and plasmids are available upon request. Code to run the CATHI algorithm can be found
863 at https://github.com/bentonml/cathi. Code to model random distribution can be found at
864 https://github.com/geofreyfriedman/sirta. Data summarized in Figures 1 and 2 are found in
865 Supplementary File 3. Data summarized in Figure 3 are found in Supplementary Files 2 and 4. Data
866 summarized in Figure 4 are found in Supplementary Files 2 and 5. Data summarized in Figure 5 are
867 found in Supplementary File 6. Data summarized in Figure 6 are found in Supplementary Files 3 and 5 .
868 Data summarized in Figure 7 are found in Supplementary File 7. Locations of data presented in
869 supplementary figures are referenced within the corresponding figure legends. Sequencing data are
870 available from the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA939836.
871

872 **Acknowledgements**

873 We thank James Haber and Deborah Wuttke for providing us with strains and plasmids. We also
874 thank Katherine Paulin for her guidance regarding the protein binding experiments. We are grateful to
875 the VANTAGE sequencing core for performing the Illumina sequencing in this work.
876

877 **Funding**

878 This work was supported by National Institutes of Health award R01 GM123292 and a Vanderbilt-Ingram
879 Cancer Center (VICC) Shared Resource Scholarship to K.L.F. E.A.H. is supported by the National Institutes
880 of Health National Eye Institute (F31 EY033235). Research in A.R.'s lab is supported by grants from the

881 National Science Foundation (DEB-2110404), the National Institutes of Health National Institute of
882 Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund. Research in B.F.E's
883 lab is supported by NIH grant R35 GM136401. R.E.H. and D.I.G. are supported by the National Institutes
884 of Health award T32 GM137793. MARC Scholars T.G. and S.P. are supported by a grant from the
885 National Institute of General Medical Sciences of the National Institutes of Health: T34 GM136451. The
886 content of this paper is solely the responsibility of the authors and does not necessarily represent the
887 official views of the National Institutes of Health.

888

889 **Conflicts of Interest**

890 A.R. is a scientific consultant for LifeMine Therapeutics, Inc.

891

892 **Figure legends**

893 Figure 1. Validation of Pooled Telomere sequencing (PT-seq) as a method to quantify de novo telomere
894 addition. a) Diagram depicting the structure of chromosome VII and the strategy for mapping GCR
895 events resulting from a DSB generated by the HO endonuclease. Locations of the essential gene *BRR6*
896 and the sequence to be tested (SiRTA) are shown. A *URA3* marker inserted distal to the HO cleavage
897 sites allows selection for cells that lose the chromosome terminus following HO cleavage (see text for
898 additional explanation). The approximate locations of PCR products utilized to map GCR events are
899 shown. Product 1 amplifies sequences within *BRR6* and is utilized as a positive control. Products 2 and 3
900 amplify regions internal to the putative SiRTA or across the SiRTA, respectively, and are used to identify
901 clones in which a GCR event occurred within the SiRTA. Product 4 is used to verify loss of the
902 chromosome terminus. b) Diagram of a GCR event in which telomere addition occurred in the putative
903 SiRTA. Addition of telomeric DNA at the SiRTA is verified by the presence of PCR product 5, generated
904 with a forward primer proximal to the SiRTA and a reverse primer complementary to the telomeric

905 repeat. c) Flow chart representing the PT-seq methodology. D) Correlation of the PCR-based and PT-seq
906 methodologies (Supplementary File 3). Results on chromosome VII (black triangles; r^2=0.97) and
907 chromosome IX (open circles; r^2=0.95) were analyzed by linear regression (lines are nearly
908 superimposed). Line equations (VII: y=3.961x-2.039 and IX: y=3.8022x-0.5505) are statistically
909 indistinguishable (p=0.94) by analysis of covariance.

910 Figure 2. Computational Algorithm for Telomere Hotspot Identification (CATHI) predicts SiRTA function.
911 a) Summary of methodology used to generate CATHI score. An example corresponding to SiRTA
912 14R131(+) is shown (300 bp sequence beginning at chromosome XIV nucleotide 131308). Although
913 multiple 75 bp windows within this sequence surpass the threshold score of 20, the calculation is shown
914 only for the highest-scoring window, starting at nucleotide 131471 (bold, bracketed text). Underlined
915 sequences correspond to strings of 4 or more guanine or thymine nucleotides conforming to the
916 patterns described in the flowchart and in more detail in *Materials and Methods*. Each underlined
917 nucleotide is awarded one point. Subsequently, each occurrence of a GGTGG pentanucleotide incurs a
918 1.5-point penalty to generate the final score. b) Correlation of CATHI score and the percentage of GCR
919 events that result from de novo telomere addition within the sequence of interest. Each value is the
920 average of at least two experiments, each with 30 GCR events. The standard curve for chromosome VII
921 (Figure 1d) is used to convert PT-seq values to the percentage of GCR events undergoing dnTA in the
922 SiRTA. Horizontal dashed line indicates a minimum telomere-addition efficiency of 6.6% used to define
923 an active SiRTA (see text for detail). Thirty-two of the sequences fall below this threshold and 15 are
924 above this threshold. Vertical dashed line illustrates a CATHI score of 20 that effectively separates active
925 and inactive sequences. Thirty-three sequences fall below this threshold and 14 are above this threshold
926 (Supplementary File 3). The open circles are false negatives or false positives. Linear regression analysis
927 on SiRTAs with a CATHI score of 20 or more yields a p-value of 0.01 (r^2 =0.43).

928 Figure 3. Summary of predicted SiRTAs across the *S. cerevisiae* genome. a) Diagram of chromosome
929 landmarks. Predicted SiRTAs are listed in Supplementary File 4. Non-essential regions (grey boxes) are
930 sequences located between the last essential gene (most distal gene on the chromosome arm that
931 causes lethality in a haploid strain when deleted) and the telomere on each chromosome arm.
932 Subtelomeres are located immediately adjacent to the telomeric repeats within the nonessential regions
933 and are composed of a single complete or partial X element (all telomeres) and one or more Y' elements
934 (a fraction of telomeres). Genomic coordinates are listed in Supplementary File 2. Location of the CA- or
935 TG-oriented sequences are indicated for the left and right chromosome arms (see text). b) Distribution
936 of SiRTAs on each of the 16 yeast chromosomes; distance in base pairs from the left telomere is
937 indicated at the bottom of the figure and corresponds to coordinates in the S288C reference genome.
938 Black circles mark the centromere position and nonessential regions are highlighted in grey. Blue lines in
939 the top half of each bar represent SiRTAs on the top (plus) strand and red lines in the bottom half of
940 each bar refer to SiRTAs on the bottom (minus) strand. c) Distribution of SiRTAs in the nonessential
941 regions of the left and right arms of each chromosome as defined in Supplementary File 2. The transition
942 between the subtelomeric X element and the remainder of the nonessential region is shown as a
943 horizontal black line. Red and blue lines are as described in (b). Diagrams in (a) and (b) were generated
944 using shinyChromosome (Yu et al. 2019).

945 Figure 4. SiRTAs are enriched in subtelomeric regions. a) Using a permutation strategy as described in
946 *Materials and Methods*, the enrichment of SiRTAs (Log2 fold change) was determined for the
947 nonessential and essential chromosome regions. Analysis utilized all genomic sequences (except the
948 terminal telomeric repeats) or genomic sequences from which the subtelomeric regions were excluded,
949 as indicated (coordinates in Supplementary File 2). *p-value <0.01 by chi-squared test with Bonferroni's
950 correction. b) Distributions of CATHI scores (20 and greater) for putative SiRTAs in the TG- or CA-
951 orientations. Analysis is presented separately for SiRTAs in nonessential regions (subtel excluded) versus

952 those in the subtelomeric repeats (subtel only). Nine SiRTAs containing perfect telomeric (TG_{1-3}) repeats
953 are indicated with open circles. ****p <0.0001 by chi-squared test. Coordinates of telomeric repeats are
954 found in Supplementary File 5. c) Enrichment analysis was conducted separately for SiRTAs in the TG or
955 CA orientation (see text and Figure 3a for definitions) as described in part (a). Results for the
956 nonessential regions are shown. *p-value<0.01 by chi-squared test with Bonferroni's correction.

957 Figure 5. CATHI scores are significantly elevated in the *S. cerevisiae* genome relative to expectation. a) As
958 described in *Materials and Methods*, the algorithm was applied to the *S. cerevisiae* genome (excluding
959 subtelomeric regions) and the number of sequences at each score (15 or higher; rounded down to the
960 nearest integer) was graphed (solid bars). Genomic sequences (excluding subtelomeric regions) were
961 scrambled five times and the identical procedure was applied (Supplementary File 6). Data are
962 presented as the average and standard deviation of the five trials (open bars). b) Distribution of CATHI
963 scores of 25 and above in the *S. cerevisiae* genome (closed circles) and shuffled genomes (open circles).
964 Subtelomeric sequences were excluded. c) As in (b), but data are shown for CATHI scores ranging from
965 15-19.

966 Figure 6. A 62 bp TG-dinucleotide repeat [SiRTA 6R210(+)] supports high levels of dnTA. a) Strains in
967 which a 300 bp sequence encompassing SiRTA 6R210(+) or 14L35(-) was integrated on chromosome VII
968 were subjected to the HO-cleavage assay as described in *Materials and Methods*. The percent of cells
969 that survived on galactose-containing medium and acquired 5-FOA resistance [cells containing a gross
970 chromosomal rearrangement (GCR)] is shown. A strain lacking any insertion (No SiRTA) was utilized as a
971 control. Error bars are standard deviation. b) The percent of GCR events that involve de novo telomere
972 addition in the sequence of interest was determined by PT-seq for each strain described in (a). Each data
973 point was generated by analysis of 30 GCR events. Average and standard deviation are shown. c) The
974 300 bp sequence encompassing 6R210(+) is shown. Sequence reads generated by PT-seq from a total of

975 60 GCR events [corresponding to the experiments shown in (b)] were filtered for those containing
976 evidence of de novo telomere addition within the 300 bp sequence. Sites at which de novo telomere
977 addition was observed are indicated (arrows). Arrow width indicates the percent of telomere-containing
978 reads that map to that particular site. d) CATHI scores are shown for all non-subtelomeric SiRTAs with
979 scores of 30 or higher, separated by CA- or TG-orientation. Open circles correspond to SiRTAs containing
980 TG-dinucleotide repeats (also listed in Supplementary File 5).

981 Figure 7. Sequences that function as SiRTAs bind Cdc13 *in vitro.* a) A competition fluorescence
982 polarization assay was utilized to measure the relative association of the Cdc13 DNA binding domain
983 with the indicated sequences (Supplementary File 1). Relative $K_{i,app}$ was determined as described in
984 *Materials and Methods*. Each point represents an independent measurement; error bars are standard
985 deviation (Supplementary File 7). The dotted line indicates normalization of values to the $K_{i,app}$ of a tel11-
986 75 oligonucleotide included in each experiment. b) Same as in (a). $K_{i,app}$ of the TG-dinucleotide repeat
987 analyzed in Figure 6 is shown [6R210(+)]. Data for 14L35(-) are repeated from (a) for comparison. c)
988 Same as in (a). 2R780(-) and its mutated variants are described in Hoerr et al. (2023).

989

990 **References Cited**

991 Agarwal T, Roy S, Kumar S, Chakraborty TK, Maiti S. 2014. In the sense of transcription regulation by G-
992 quadruplexes: Asymmetric effects in sense and antisense strands. Biochemistry. 53: 3711–3718.
993 https://doi.org/10.1021/bi401451q

994 Anand R, Memisoglu G, Haber J. 2017. Cas9-mediated gene editing in *Saccharomyces cerevisiae*. Protoc
995 Exch. https://doi.org/10.1038/protex.2017.021a

996 Anderson EM, Halsey WA, Wuttke DS. 2002. Delineation of the high-affinity single-stranded telomeric

997 DNA-binding domain of *Saccharomyces cerevisiae* Cdc13. Nucleic Acids Res. 30: 4305–4313.
998 https://doi.org/10.1093/nar/gkf554

999 Anderson BJ, Larkin C, Guja K, Schildbach JF. 2008. Chapter 12 Using fluorophore-labeled
1000 oligonucleotides to measure affinities of protein–DNA interactions. Methods in Enzymology, 450, 253–
1001 272. https://doi.org/10.1016/S0076-6879(08)03412-5

1002 Aylon Y, Kupiec M. 2004. DSB repair: The yeast paradigm. DNA Repair (Amst). 3: 797–815.
1003 https://doi.org/10.1016/j.dnarep.2004.04.013

1004 Blackburn EH. 1991 Structure and function of telomeres. Nature. 350: 569–573.
1005 https://doi.org/10.1038/350569a0

1006 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data.
1007 Bioinformatics. 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

1008 Bonaglia MC, Giorda R, Beri S, de Agostini C, Novara F, et al. 2011. Molecular mechanisms generating
1009 and stabilizing terminal 22q13 deletions in 44 subjects with Phelan/McDermid Syndrome. PLoS Genet. 7:
1010 e1002173. https://doi.org/10.1371/journal.pgen.1002173

1011 Bonnell E, Pasquier E, Wellinger RJ. 2021. Telomere replication: Solving multiple end replication
1012 problems. Front Cell Dev Biol. 9. https://doi.org/10.3389/fcell.2021.668171

1013 Capra JA, Paeschke K, Singh M, Zakian VA. 2010. G-quadruplex DNA sequences are evolutionarily
1014 conserved and associated with distinct genomic features in *Saccharomyces cerevisiae*. PLoS Comput Biol
1015 6: e1000861. https://doi.org/10.1371/journal.pcbi.1000861

1016 Casari E, Gnugnoli M, Rinaldi C, Pizzul P, Colombo CV, Bonetti D, Longhese MP. 2022. To fix or not to fix:
1017 Maintenance of chromosome ends versus repair of DNA double-strand breaks. Cells. 11(20), 3224.

1018 https://doi.org/10.3390/cells11203224

1019 Chastain M, Zhou Q, Shiva O, Fadri-Moskwik M, Whitmore L, Jia P, Dai X, Huang C, Ye P, Chai W. 2016.
1020 Human CST facilitates genome-wide RAD51 recruitment to GC-rich repetitive sequences in response to
1021 replication stress. Cell Reports. 16(5), 1300–1314. https://doi.org/10.1016/j.celrep.2016.06.077

1022 Chen H, Xue J, Churikov D, Hass EP, Shi S, et al. 2018. Structural insights into yeast telomerase
1023 recruitment to telomeres. Cell. 172: 331-343.e13. https://doi.org/10.1016/j.cell.2017.12.008

1024 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. 2009. Biopython: freely available Python tools
1025 for computational molecular biology and bioinformatics. Bioinformatics. 25: 1422–1423.
1026 https://doi.org/10.1093/bioinformatics/btp163

1027 Conrad MN, Wright JH, Wolf AJ, Zakian VA. 1990. RAP1 protein interacts with yeast telomeres in vivo:
1028 Overproduction alters telomere structure and decreases chromosome stability. Cell. 63: 739–750.
1029 https://doi.org/10.1016/0092-8674(90)90140-A

1030 Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: A flexible Python library for manipulating genomic
1031 datasets and annotations. Bioinformatics. 27: 3423–3424.
1032 https://doi.org/10.1093/bioinformatics/btr539

1033 Doksani, Y, de Lange T. 2014. The role of double-strand break repair pathways at functional and
1034 dysfunctional telomeres. Cold Spring Harbor Perspectives in Biology. 6(12), a016576–a016576.
1035 https://doi.org/10.1101/cshperspect.a016576

1036 Downs JA, Lowndes NF, Jackson SP. 2000. A role for *Saccharomyces cerevisiae* histone H2A in DNA
1037 repair. Nature. 408: 1001–1004. https://doi.org/10.1038/35050000

1038 Eldridge AM, Wuttke DS. 2008. Probing the mechanism of recognition of ssDNA by the Cdc13-DBD.

1039 Nucleic Acids Res. 36: 1624–1633. https://doi.org/10.1093/nar/gkn017

1040 Epum EA, Mohan MJ, Ruppe NP, Friedman KL. 2020. Interaction of yeast Rad51 and Rad52 relieves
1041 Rad52-mediated inhibition of de novo telomere addition. PLoS Genet. 16: e1008608.
1042 https://doi.org/10.1371/journal.pgen.1008608

1043 Evans SK, Lundblad V. 1999. Est1 and Cdc13 as comediators of telomerase access. Science. 286(5437):
1044 117–120. https://doi.org/10.1126/science.286.5437.117

1045 Fanning E. 2006. A dynamic model for replication protein A (RPA) function in DNA processing pathways.
1046 Nucleic Acids Res. 34: 4126–4137. https://doi.org/10.1093/nar/gkl550

1047 Ge Y, Wu Z, Chen H, Zhong Q, Shi S, et al. 2020. Structural insights into telomere protection and
1048 homeostasis regulation by yeast CST complex. Nat Struct Mol Biol. 27: 752–762.
1049 https://doi.org/10.1038/s41594-020-0459-8

1050 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B,
1051 Arkin AP, Astromoff A, El Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K,
1052 Johnston M. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature. 418(6896),
1053 387–391. https://doi.org/10.1038/nature00935

1054 Gilson E, Géli V. 2007. How telomeres are replicated. Nature Reviews Molecular Cell Biology. 8(10), 825–
1055 838. https://doi.org/10.1038/nrm2259

1056 Giraud-Panis MJ, Teixeira MT, Géli V, Gilson E. 2010. CST meets shelterin to keep telomeres in check.
1057 Mol Cell. 39: 665–676. https://doi.org/10.1016/j.molcel.2010.08.024

1058 Glustrom LW, Lyon KR, Paschini M, Reyes CM, Parsonnet NV, et al. 2018. Single-stranded telomere-
1059 binding protein employs a dual rheostat for binding affinity and specificity that drives function.
1060 Proceedings of the National Academy of Sciences. 115: 10315–10320.

1061 https://doi.org/10.1073/pnas.1722147115

1062 Grandin N. 2001. Ten1 functions in telomere end protection and length regulation in association with
1063 Stn1 and Cdc13. EMBO J 20: 1173–1183. https://doi.org/10.1093/emboj/20.5.1173

1064 Guilherme RS, Hermetz KE, Varela PT, Perez ABA, Meloni VA, Rudd MK, Kulikowski LD, Melaragno MI.
1065 2015. Terminal 18q deletions are stabilized by neotelomeres. Molecular Cytogenetics, 8(1), 32.
1066 https://doi.org/10.1186/s13039-015-0135-6

1067 Hardy CF, Sussel L, and Shore D. 1992. A RAP1-interacting protein involved in transcriptional silencing
1068 and telomere length regulation. Genes & Development. 6(5), 801–814.
1069 https://doi.org/10.1101/gad.6.5.801

1070 Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, et al. 2020. Array programming with
1071 NumPy. Nature. 585: 357–362. https://doi.org/10.1038/s41586-020-2649-2

1072 Hoerr RE, Eng A, Payen C, Di Rienzi SC, Raghuraman MK, Dunham MJ, Brewer BJ, Friedman KL. 2023.
1073 Hotspot of de novo telomere addition stabilizes linear amplicons in yeast grown in sulfate-limiting
1074 conditions. Genetics, iyad010. https://doi.org/10.1093/genetics/iyad010

1075 Hoerr RE, Ngo K, Friedman KL. 2021. When the ends justify the means: Regulation of telomere addition
1076 at double-strand breaks in yeast. Frontiers in Cell and Developmental Biology. 9.
1077 https://doi.org/10.3389/fcell.2021.655377

1078 Hurowitz EH, Brown PO. 2003. Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae.*
1079 Genome Biol. 5: R2. https://doi.org/10.1186/gb-2003-5-1-r2

1080 Hustedt N, Durocher D. 2017. The control of DNA repair by the cell cycle. Nat Cell Biol. 19: 1–9.
1081 https://doi.org/10.1038/ncb3452

1082 Kim N. 2019. The Interplay between G-quadruplex and Transcription. Curr Med Chem. 26: 2898–2917.

1083 https://doi.org/10.2174/0929867325666171229132619

1084 Kramer KM, Haber JE. 1993. New telomeres in yeast are initiated with a highly selected subset of TG1-3
1085 repeats. Genes Dev. 7: 2345–2356. https://doi.org/10.1101/gad.7.12a.2345

1086 Krenning L, van den Berg J, Medema RH. 2019. Life or Death after a Break: What Determines the Choice?
1087 Mol Cell. 76: 346–358. https://doi.org/10.1016/j.molcel.2019.08.023

1088 Kyrion G, Liu K, Liu C, Lustig AJ. 1993. RAP1 and telomere structure regulate telomere position effects in
1089 *Saccharomyces cerevisiae*. Genes Dev 7: 1146–1159. https://doi.org/10.1101/gad.7.7a.1146

1090 Lamb J, Harris PC, Wilkie AO, Wood WG, Dauwerse JG, Higgs DR. 1993. De novo truncation of
1091 chromosome 16p and healing with (TTAGGG)n in the alpha-thalassemia/mental retardation syndrome
1092 (ATR-16). American Journal of Human Genetics, 52(4), 668–676.

1093 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short
1094 DNA sequences to the human genome. Genome Biol. 10: R25. https://doi.org/10.1186/gb-2009-10-3-
1095 r25

1096 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9: 357–359.
1097 https://doi.org/10.1038/nmeth.1923

1098 Lendvay TS, Morris DK, Sah J, Balasubramanian B, Lundblad V. 1996. Senescence mutants of
1099 *Saccharomyces cerevisiae* with a defect in telomere replication identify three additional *EST* genes.
1100 Genetics. 144: 1399–1412. https://doi.org/10.1093/genetics/144.4.1399

1101 Lewis KA, Pfaff DA, Earley JN, Altschuler DE, Wuttke DS. 2014. The tenacious recognition of yeast
1102 telomere sequence by Cdc13 is fully exerted by a single OB-fold domain. Nucleic Acids Res. 42: 475–484.
1103 https://doi.org/10.1093/nar/gkt843

1104 Lin YY, Li MH, Chang YC, Fu PY, Ohniwa RL, et al. 2021. Dynamic DNA shortening by telomere-binding

1105 protein Cdc13. J Am Chem Soc. 143: 5815–5825. https://doi.org/10.1021/jacs.1c00820

1106 Lingner J, Cooper JP, Cech TR. 1995. Telomerase and DNA end replication: No longer a lagging strand
1107 problem? Science. 269: 1533–1534. https://doi.org/10.1126/science.7545310

1108 Lingner J, Cech TR, Hughes TR, Lundblad V. 1997. Three Ever Shorter Telomere (*EST*) genes are
1109 dispensable for in vitro yeast telomerase activity. Proceedings of the National Academy of Sciences. 94:
1110 11190–11195. https://doi.org/10.1073/pnas.94.21.11190

1111 Louis EJ, Haber JE. 1990. The subtelomeric Y' repeat family in *Saccharomyces cerevisiae*: an experimental
1112 system for repeated sequence evolution. Genetics 124: 533–545.
1113 https://doi.org/10.1093/genetics/124.3.533

1114 Louis EJ, Haber JE. 1992. The structure and evolution of subtelomeric Y' repeats in *Saccharomyces*
1115 *cerevisiae*. Genetics. 131(3), 559–574. https://doi.org/10.1093/genetics/131.3.559

1116 Louis EJ, Naumova ES, Lee A, Naumov G, Haber JE. 1994. The chromosome end in yeast: its mosaic
1117 nature and influence on recombinational dynamics. Genetics. 136(3), 789–802.
1118 https://doi.org/10.1093/genetics/136.3.789

1119 Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. 2008. A genome-wide view of the spectrum of
1120 spontaneous mutations in yeast. Proceedings of the National Academy of Sciences. 105: 9272–9277.
1121 https://doi.org/10.1073/pnas.0803466105

1122 Mangahas JL, Alexander MK, Sandell LL, Zakian VA. 2001. Repair of chromosome ends after telomere
1123 loss in *Saccharomyces*. Mol Biol Cell. 12: 4078–4089. https://doi.org/10.1091/mbc.12.12.4078

1124 Marcand S, Gilson E, Shore D. 1997. A protein-counting mechanism for telomere length regulation in

1125 yeast. Science. 275(5302), 986–990. https://doi.org/10.1126/science.275.5302.986

1126 McKinney W. 2010. “Data structures for statistical computing in python.” In Proceedings of the 9th
1127 Python in Science Conference. Edited by S. van der Walt and J. Millman. 56–61.
1128 https://doi.org/10.25080/Majora-92bf1922-00a

1129 Mersaoui SY, Wellinger RJ. 2019. Fine tuning the level of the Cdc13 telomere-capping protein for
1130 maximal chromosome stability performance. Curr Genet. 65: 109–118.
1131 https://doi.org/10.1007/s00294-018-0871-3

1132 Myung K, Chen C, Kolodner RD. 2001. Multiple pathways cooperate in the suppression of genome
1133 instability in *Saccharomyces cerevisiae*. Nature. 411: 1073–1076. https://doi.org/10.1038/35082608

1134 Nakamura Y, Ikemura T, Gojobori T. 2000. Codon usage tabulated from international DNA sequence
1135 databases: status for the year 2000. Nucleic Acids Res. 28: 292–292.
1136 https://doi.org/10.1093/nar/28.1.292

1137 Negrini S, Ribaud V, Bianchi A, Shore D. 2007. DNA breaks are masked by multiple Rap1 binding in yeast:
1138 implications for telomere capping and telomerase regulation. Genes Dev. 21: 292–302.
1139 https://doi.org/10.1101/gad.400907

1140 Nevado J, García-Miñaúr S, Palomares-Bralo M, Vallespín E, Guillén-Navarro E, Rosell J, Bel-Fenellós C,
1141 Mori MÁ, Milá M, Campo M, del Barrúz, P, Santos-Simarro F, Obregón G, Orellana C, Pachajoa H, Tenorio
1142 JA, Galán E, Cigudosa JC, Moresco A, Lapunzina P. 2022. Variability in Phelan-McDermid syndrome in a
1143 cohort of 210 individuals. Frontiers in Genetics. 13. https://doi.org/10.3389/fgene.2022.652454

1144 Ngo K, Epum EA, Friedman KL. 2020. Emerging non-canonical roles for the Rad51–Rad52 interaction in
1145 response to double-strand breaks in yeast. Curr Genet. 66: 917–926. https://doi.org/10.1007/s00294-

1146 020-01081-z

1147 Obodo UC, Epum EA, Platts MH, Seloff J, Dahlson NA, et al. 2016. Endogenous hot spots of de novo
1148 telomere addition in the yeast genome contain proximal enhancers that bind Cdc13. Mol Cell Biol. 36:
1149 1750–1763. https://doi.org/10.1128/MCB.00095-16.Address

1150 Osterhage JL, Friedman KL. 2009. Chromosome end maintenance by telomerase. Journal of Biological
1151 Chemistry. 284(24), 16061–16065. https://doi.org/10.1074/jbc.R900011200

1152 Ouenzar F, Lalonde M, Laprade H, Morin G, Gallardo F, Tremblay-Belzile S, Chartrand P. 2017. Cell cycle–
1153 dependent spatial segregation of telomerase from sites of DNA damage. Journal of Cell Biology, 216(8).
1154 2355–2371. https://doi.org/10.1083/jcb.201610071

1155 Paeschke K, Capra JA, Zakian VA. 2011. DNA replication through G-Quadruplex motifs is promoted by
1156 the *Saccharomyces cerevisiae* Pif1 DNA helicase. Cell. 145: 678–691.
1157 https://doi.org/10.1016/j.cell.2011.04.015

1158 Paeschke K, Bochman ML, Garcia PD, Cejka P, Friedman KL, et al. 2013. Pif1 family helicases suppress
1159 genome instability at G-quadruplex motifs. Nature. 497: 458–462. https://doi.org/10.1038/nature12149

1160 Pandey S, Hajikazemi M, Zacheja T, Schalbetter S, Neale MJ, et al. 2021. Telomerase subunit Est2 marks
1161 internal sites that are prone to accumulate DNA damage. BMC Biol 19(1): 247.
1162 https://doi.org/10.1186/s12915-021-01167-1

1163 Pardo B, Marcand S. 2005. Rap1 prevents telomere fusions by nonhomologous end joining. EMBO J. 24:
1164 3117–3127. https://doi.org/10.1038/sj.emboj.7600778

1165 Pennaneach V, Putnam CD, Kolodner RD. 2006. Chromosome healing by de novo telomere addition in
1166 *Saccharomyces cerevisiae*. Mol Microbiol. 59: 1357–1368. https://doi.org/10.1111/j.1365-

1167 2958.2006.05026.x

1168 Pennock E, Buckley K, Lundblad V. 2001. Cdc13 delivers separate complexes to the telomere for end
1169 protection and replication. Cell. 104: 387–396. https://doi.org/10.1016/S0092-8674(01)00226-4

1170 Pfeiffer V, Lingner J. 2013. Replication of telomeres and the regulation of telomerase. Cold Spring Harb
1171 Perspect Biol. 5: a010405–a010405. https://doi.org/10.1101/cshperspect.a010405

1172 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
1173 Bioinformatics. 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

1174 Quinlan AR. 2014. BEDTools: The Swiss-army tool for genome feature analysis. Curr Protoc Bioinform.
1175 47(1):11.12.11–11.12.34. doi:10. 1002/0471250953.bi1112s47.

1176 Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-
1177 nucleotide resolution. Cell. 147: 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013

1178 Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res. 43:
1179 8627–8637. https://doi.org/10.1093/nar/gkv862

1180 Rice C, Skordalakes E. 2016. Structure and function of the telomeric CST complex. Comput Struct
1181 Biotechnol J. 14: 161–167. https://doi.org/10.1016/j.csbj.2016.04.002

1182 https://doi.org/10.1038/s41598-019-55482-3

1183 Schulz VP, Zakian VA. 1994. The *Saccharomyces* PIF1 DNA helicase inhibits telomere elongation and de
1184 novo telomere formation. Cell. 76: 145–155. https://doi.org/10.1016/0092-8674(94)90179-1

1185 Singer MS, Gottschling DE. 1994. *TLC1* : Template RNA component of *Saccharomyces cerevisiae*
1186 telomerase. Science. 266: 404–409. https://doi.org/10.1126/science.7545955

1187 Stellwagen AE, Haimberger ZW, Veatch JR, Gottschling DE. 2003. Ku interacts with telomerase RNA to
1188 promote telomere addition at native and broken chromosome ends. Genes Dev. 17: 2384–2395.

1189 https://doi.org/10.1101/gad.1125903

1190 Tapias A, Auriol J, Forget D, Enzlin JH, Schärer OD, et al. 2004. Ordered conformational changes in
1191 damaged DNA induced by nucleotide excision repair factors. Journal of Biological Chemistry. 279:
1192 19074–19083. https://doi.org/10.1074/jbc.M312611200

1193 Teixeira MT, Arneric M, Sperisen P, Lingner J. 2004. Telomere length homeostasis is achieved via a
1194 switch between telomerase- extendible and -nonextendible states. Cell, 117(3), 323–335.
1195 https://doi.org/10.1016/S0092-8674(04)00334-4

1196 Vaasa A, Viil I, Enkvist E, Viht K, Raidaru G, et al. 2009. High-affinity bisubstrate probe for fluorescence
1197 polarization binding/displacement assays with protein kinases PKA and ROCK. Anal Biochem. 385: 85–
1198 93. https://doi.org/10.1016/j.ab.2008.10.030

1199 Vodenicharov MD, Laterreur N, Wellinger RJ. 2010. Telomere capping in non-dividing yeast cells requires
1200 Yku and Rap1. EMBO J. 29: 3007–3019. https://doi.org/10.1038/emboj.2010.155

1201 Wang X, Haber JE. 2004. Role of *Saccharomyces* single-stranded DNA-binding protein RPA in the strand
1202 invasion step of double-strand break repair. PLoS Biol 2: e21.
1203 https://doi.org/10.1371/journal.pbio.0020021

1204 Wang F, Podell ER, Zaug AJ, Yang Y, Baciu P, Cech TR, Lei M. 2007. The POT1–TPP1 telomere complex is a
1205 telomerase processivity factor. Nature. 445: 506–510. https://doi.org/10.1038/nature05454

1206 Wellinger RJ, Zakian VA. 2012. Everything you ever wanted to know about *Saccharomyces cerevisiae*
1207 telomeres: Beginning to end. Genetics. 191(4), 1073–1105.
1208 https://doi.org/10.1534/genetics.111.137851

1209 Yu Y, Yao W, Wang Y, Huang F. 2019. shinyChromosome: An R/Shiny application for interactive creation
1210 of non-circular plots of whole genomes. Genomics Proteomics Bioinformatics. 17: 535–539.

1211 https://doi.org/10.1016/j.gpb.2019.07.003

1212 Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino
1213 MC, Fischer G, Durbin R, Liti G. 2017. Contrasting evolutionary genome dynamics between domesticated
1214 and wild yeasts. Nature Genetics, 49(6), 913–924. https://doi.org/10.1038/ng.3847

1215 Zhu X, Gustafsson CM. 2009. Distinct differences in chromatin structure at subtelomeric X and Y'
1216 elements in budding yeast. PLoS One. 4: e6363. https://doi.org/10.1371/journal.pone.0006363

Figure 1

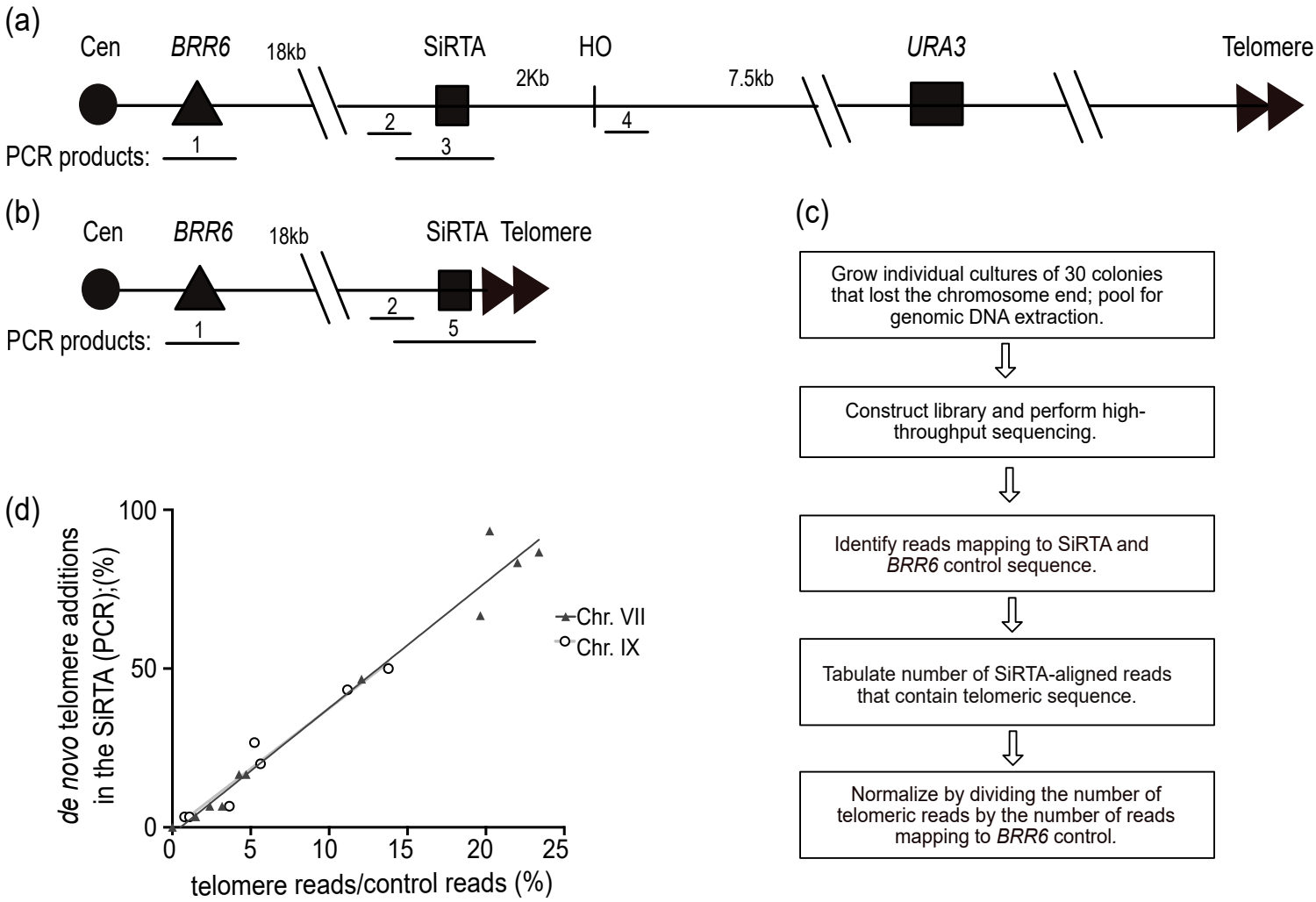

(a)
Cen
BRR6
18kb
SiRTA
HO
URA3
Telomere
2Kb
7.5kb
PCR products:
1
2
3
4
(b)
Cen
BRR6
18kb
SiRTA
Telomere
PCR products:
1
2
5
(c)
Grow individual cultures of 30 colonies that lost the chromosome end; pool for genomic DNA extraction.
Construct library and perform high-throughput sequencing.
Identify reads mapping to SiRTA and BRR6 control sequence.
Tabulate number of SiRTA-aligned reads that contain telomeric sequence.
Normalize by dividing the number of telomeric reads by the number of reads mapping to BRR6 control.
(d)
de novo telomere additions in the SiRTA (PCR);(%)
100
50
0
0
5
10
15
20
25
telomere reads/control reads (%)
Chr. VII
Chr. IX

Figure 2

(a)

Within 75bp windows, identify strings of 4 or more bases containing consecutive Gs (1-3) and a T (underline); the highest scoring window (bold) from overlapping windows is reported.

14R131(+)
TTTAATTTCTTGTGTGGACGAGGATACTTTTCAAATAATGAGGGCGATTGAGGAAGTAATGGTGATGTTACAGCTGATTTGAT
TTGCGATGCCGGCATCATAGGTGGAGCTGGCGGAATGACGCCATCAGTAGATGAAGATAATACAGATGGTAGCGGCGGCG
[GTGGTGGTGTTTCTCCCTTGGGTTTGCCGAACAATGCCATGGGTGGAGGAGGAGGTGGTGGTGGTGGAGGGGGAG]G
AGCAGGAGAATCCCCAGTAGTTTCTTGCTTGACAGTGTCGTCTTCCGACTTTTTCTCTTTCT

Starting at the first G award 1 point for each G and T

GTGGTGGTGTTTCTCCCTTGGGTTTGCCGAACAATGCCATGGGTGGAGGAGGAGGTGGTGGTGGTGGAGGGGGAG
10 4 6 14

Subtract 1.5 points for any instances of GGTGG

GTGGTGGTGTTTCTCCCTTGGGTTTGCCGAACAATGCCATGGGTGGAGGAGGAGGTGGTGGTGGTGGAGGGGGAG
-1.5 -1.5 -1.5 -1.5 -1.5 -1.5

CATHI=25

(b)

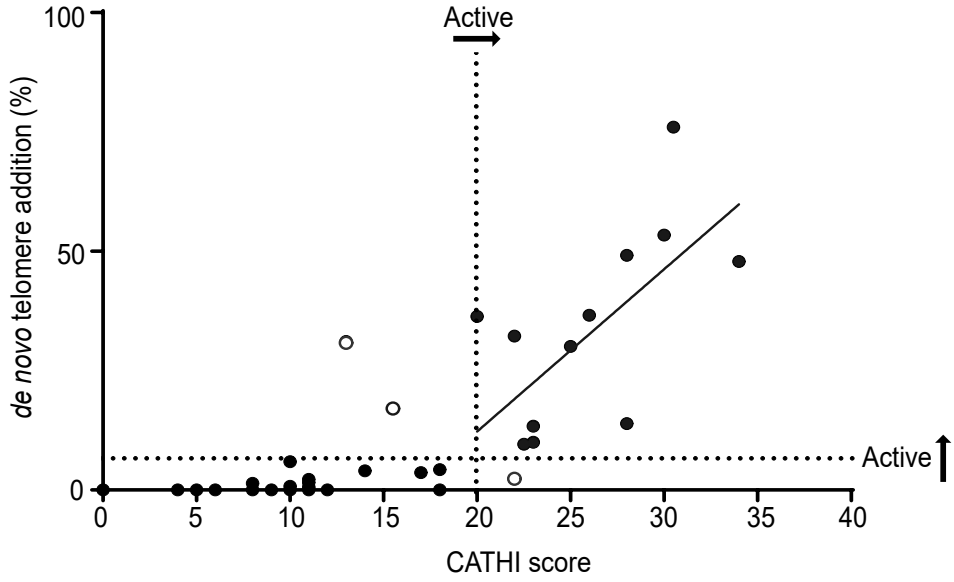


Figure 3

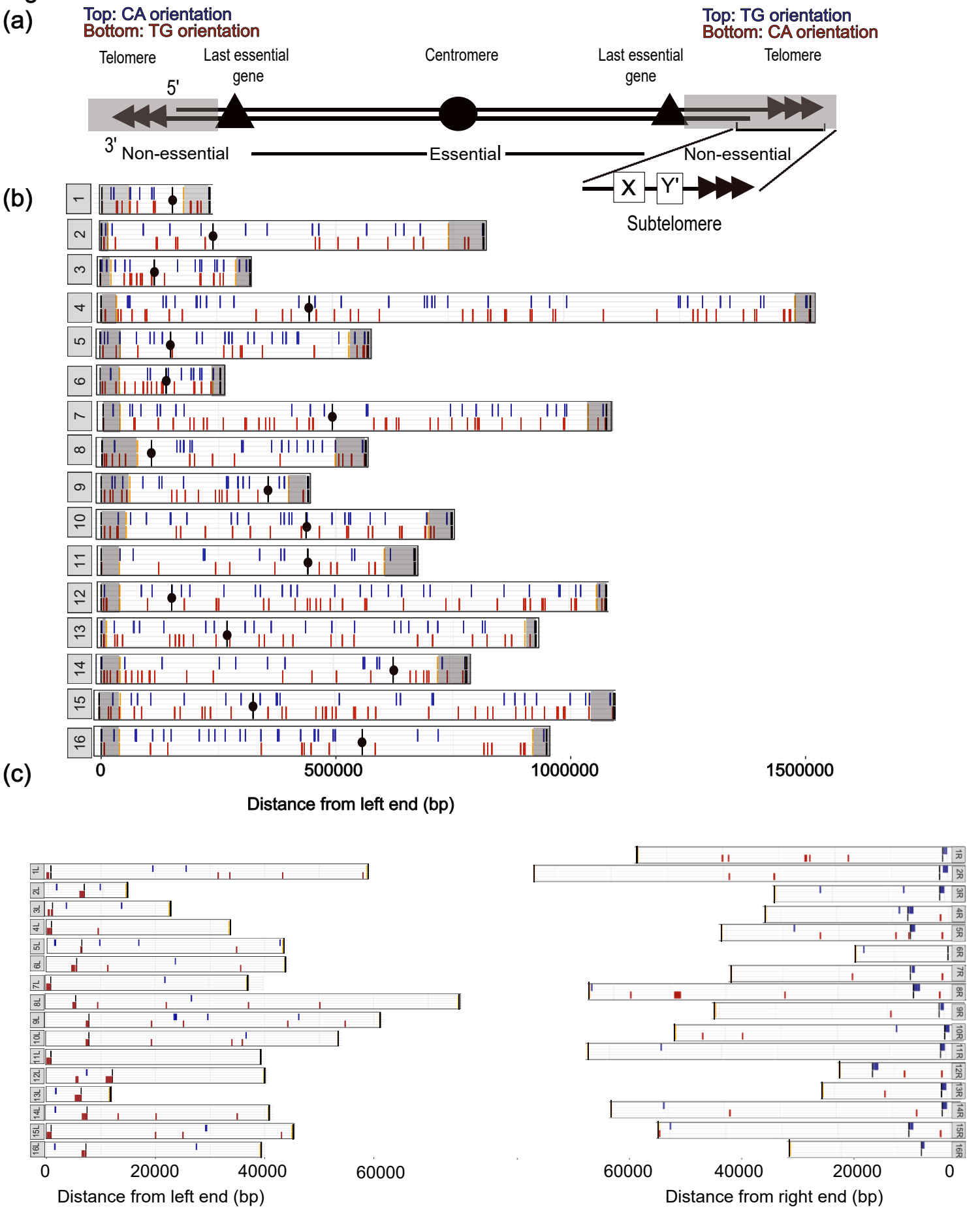
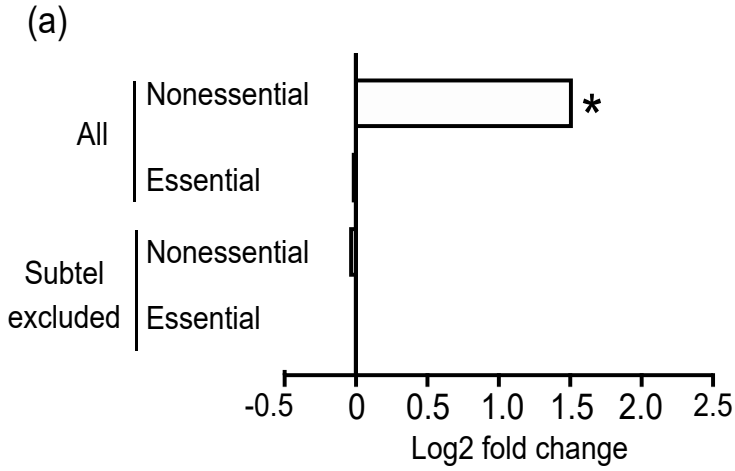
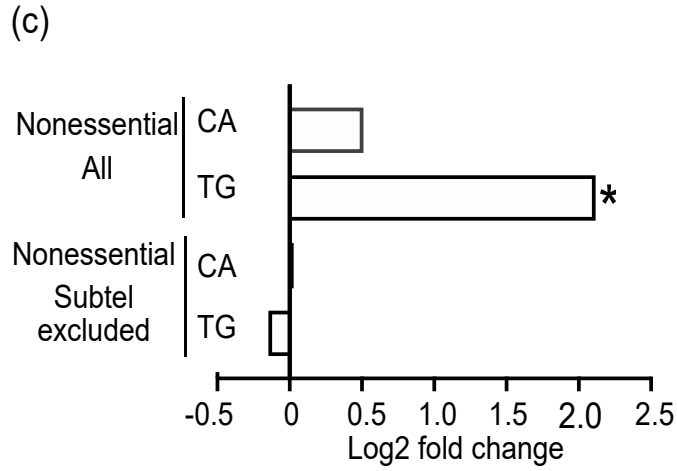

(a)
Top: CA orientation
Bottom: TG orientation
Top: TG orientation
Bottom: CA orientation
Telomere
Last essential gene
Centromere
Last essential gene
Telomere
5'
3'
Non-essential
Essential
Non-essential
X
Y'
Subtelomere
(b)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0
500000
1000000
1500000
Distance from left end (bp)
(c)
1L
2L
3L
4L
5L
6L
7L
8L
9L
10L
11L
12L
13L
14L
15L
16L
0
20000
40000
60000
Distance from left end (bp)
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
11R
12R
13R
14R
15R
16R
60000
40000
20000
0
Distance from right end (bp)

Figure 4


(a)
All
Nonessential
Essential
Subtel excluded
Nonessential
Essential
*
-0.5
0
0.5
1.0
1.5
2.0
2.5
Log2 fold change


(c)
Nonessential
All
CA
TG
Nonessential
Subtel excluded
CA
TG
*
-0.5
0
0.5
1.0
1.5
2.0
2.5
Log2 fold change

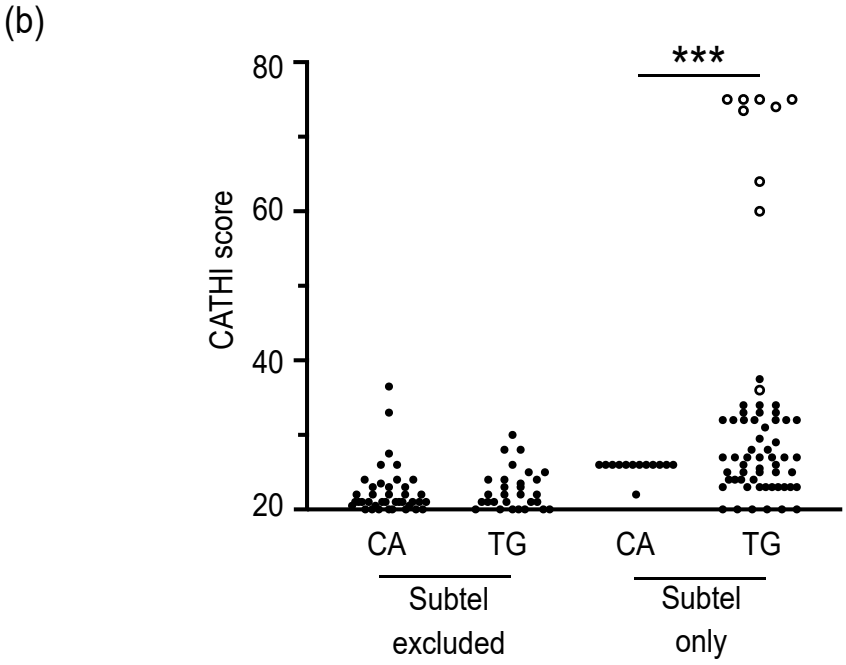

(b)
***
80
60
40
20
CATHI score
CA
TG
CA
TG
Subtel excluded
Subtel only

Figure 5

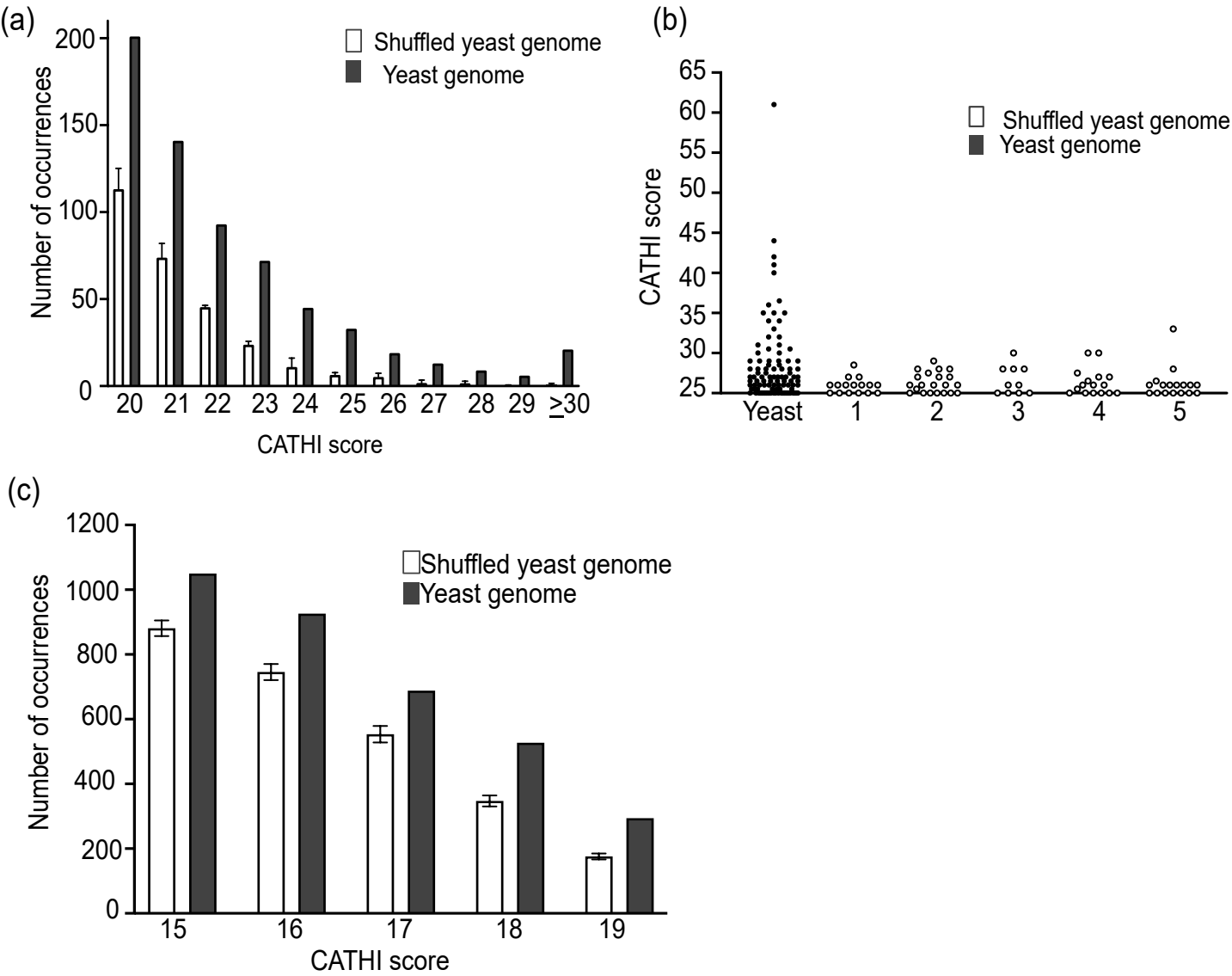
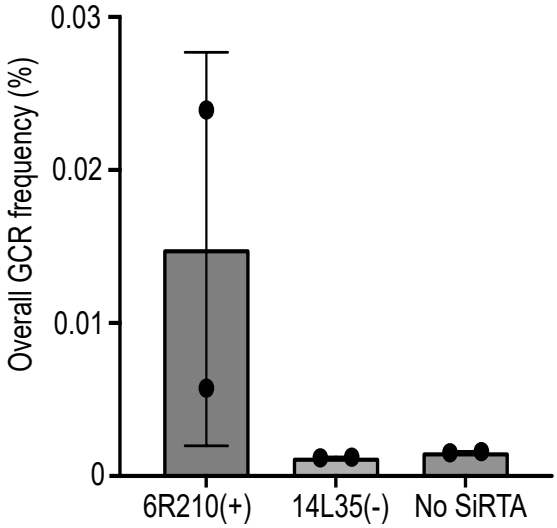
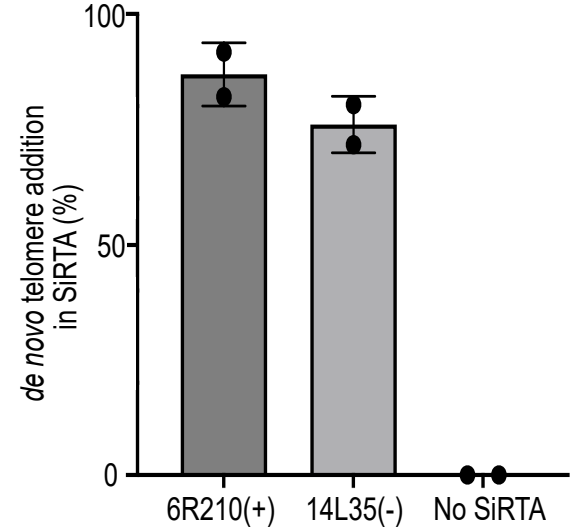

(a)
Shuffled yeast genome
Yeast genome
Number of occurrences
0
50
100
150
200
20
21
22
23
24
25
26
27
28
29
≥30
CATHI score
(b)
Shuffled yeast genome
Yeast genome
CATHI score
25
30
35
40
45
50
55
60
65
Yeast
1
2
3
4
5
(c)
Shuffled yeast genome
Yeast genome
Number of occurrences
0
200
400
600
800
1000
1200
15
16
17
18
19
CATHI score

Figure 6

(a)



(b)



(c)

(d)

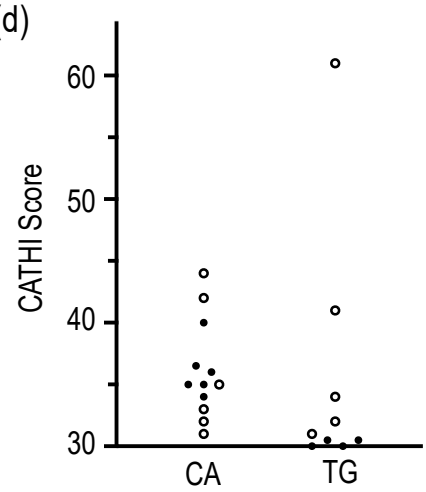
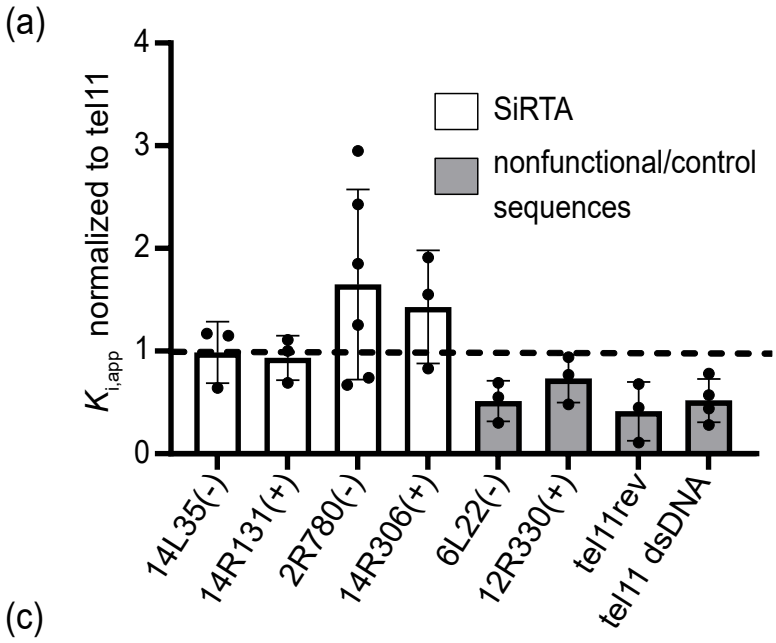
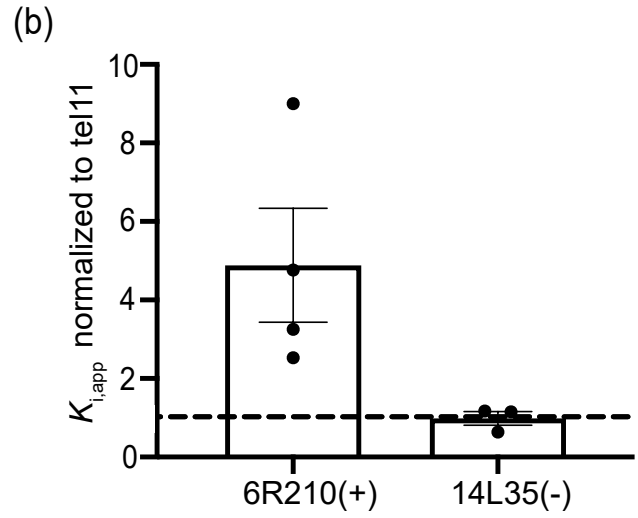
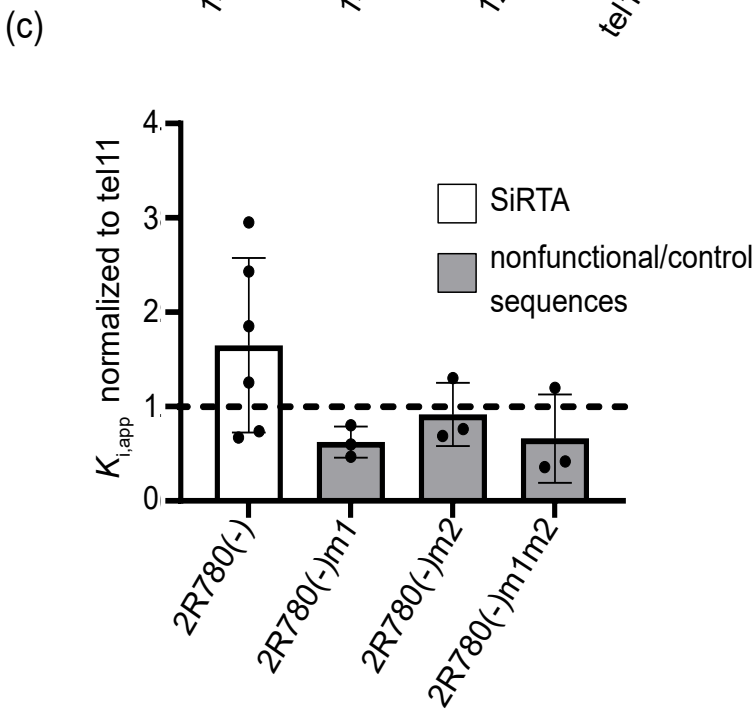


Figure 7


(a)
$K_{i,app}$ normalized to tel11
0
1
2
3
4
SiRTA
nonfunctional/control sequences
14L35(-)
14R131(+)
2R780(-)
14R306(+)
6L22(-)
12R330(+)
tel11rev
tel11 dsDNA


(b)
$K_{i,app}$ normalized to tel11
0
2
4
6
8
10
6R210(+)
14L35(-)


(c)
$K_{i,app}$ normalized to tel11
0
1
2
3
4
SiRTA
nonfunctional/control sequences
2R780(-)
2R780(-)m1
2R780(-)m2
2R780(-)m1m2