Predicting and characterizing the distribution of sites of repair associated telomere addition in
Saccharomyces cerevisiae

By

Katrina Ngo

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biological Sciences

August 11, 2023

Nashville, Tennessee

Approved:

Antonis Rokas, Ph.D.

Houra Merrikh, Ph.D.

Jared Nordman, Ph.D

James Dewar, Ph.D.

Katherine L. Friedman, Ph.D.

.

DEDICATION

To me, who worked so hard all without drinking water. You are not like the other girls.

ACKNOWLEDGEMENTS

I am especially thankful for my advisor and mentor Dr. Katherine Friedman who has worked tirelessly to support my work and teach me how to be an effective scientist as well as a decent human and mentor to others. I am also thankful for my committee, Dr. Antonis Rokas, Dr. Jared Nordman, Dr. James Dewar and Dr. Houra Merrikh for their support and advice throughout this process. Thank you also to Dr. Mary Lauren Benton who was the most patient with me as I struggled through running things in python. My fellow lab members throughout this process have been especially helpful as well for their support both intellectual and emotional particularly Remington Hoerr, David Gonzalez and all of the undergraduate students I was given the privilege of mentoring. Speaking of undergraduates, a special thank you to the SyBBURE program and team SyBBURE for giving me the opportunity to mentor people and do cool stuff. I would not have made it through graduate school without the support of my friends and family. Thank you to my parents who have been endlessly supportive and without whom I would not exist I will never be able to thank you enough. I also appreciate the support from my extended family who have sent care packages and nice messages throughout my time in school. Thank you also to Sadie, Hayley and Amanda my childhood best friends who have always believed in me and also thank you to Annie, Anne and Jennie the friends I have made here at Vanderbilt and who I hope to be friends with even after our time at Vanderbilt comes to an end. Finally thank you to my husband Alexander Bohlin for his support, his dental insurance and for feeding me snacks as I got hangry through the writing process. I love you.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1[1,2]

Introduction

Telomeres are nucleoprotein complexes that mark the ends of most eukaryotic chromosomes. Telomeres function to protect DNA ends, offer a solution to the end replication problem and distinguish naturally occurring 3'overhangs at the ends of chromosomes from 3' overhangs that occur as a result of a double-strand break (DSB). Rarely, telomerase acts upon a DSB to add a *de novo* telomere. Telomere addition prevents normal repair processes from occurring and contributes to genomic instability, but this process also prevents further damage to the DNA by blocking resection. Sequences that facilitate *de novo* telomere addition (*dn*TA) in response to a break at a higher rate than other genome sequences are termed sites of repair-associated telomere addition (SiRTAs). *De novo* telomere addition in response to a break results in loss of chromosomal sequences distal to the newly added telomere. However, *dn*TA also protects the region centromere proximal to the newly added telomere from further exonucleolytic degradation. In the case of an unrepaired break, SiRTAs may provide a mechanism of survival wherein the protection provided by a telomere mitigates the cost of some sequence loss. If this is the case, then SiRTAs would be under positive selection, and we would predict enrichment of SiRTAs in the terminal regions of chromosomes where they could function to minimize loss of

---

[1] This work is adapted from the review article: Hoerr RE, Ngo K, Friedman KL. When the Ends Justify the Means: Regulation of Telomere Addition at Double-Strand Breaks in Yeast. Front Cell Dev Biol. 2021 Mar 18;9:655377. doi: 10.3389/fcell.2021.655377 I wrote the text and Remington Hoerr made the figure shown here as Figure 1. (Hoerr *et al.* 2021)

[2] This work is adapted from the introduction of the research article: Ngo K, Gittens TH, Gonzalez DI, Hatmaker EA, Plotkin S, Engle M, Friedman GA, Goldin M, Hoerr RE, Eichman BF, Rokas A, Benton, ML, and Friedman KL. 2023. A comprehensive map of hotspots of *de novo* telomere addition in Saccharomyces cerevisiae. *Genetics*, iyad076. https://doi.org/10.1093/genetics/iyad076 (Ngo et al. 2023)

genes essential for cell survival. We would also expect to see SiRTAs oriented such that capping

by *dn*TA would occur on the portion of the chromosome containing the centromere. These

hypotheses can only be addressed through the development of strategies to map the location and

orientation of SiRTAs genome wide. This dissertation discusses novel strategies to elucidate the

distribution of SiRTAs within the *Saccharomyces cerevisiae* genome.

1.1 The history of telomere research in general and in *S. cerevisiae*

The maintenance of DNA integrity is essential for cell function. To maintain genomic

integrity and prevent sequence loss, most eukaryotic chromosomes terminate with nucleoprotein

structures termed telomeres that protect chromosomes from end-to-end fusion and block

excessive nucleolytic resection. Telomeres contain a characteristic, repetitive sequence rich in

thymine and guanine (TG-rich) on one strand (Blackburn 1991). While the majority of the

telomere is double-stranded, the TG-rich strand extends past the complementary cytosine and

adenine (CA)-rich strand to create a 3' overhang. Regeneration of this 3' overhang after each

round of DNA replication results in progressive sequence loss, but in cells that maintain telomere

length over successive generations, this end-replication problem is counterbalanced through the

extension of the 3' strand by telomerase (reviewed in Osterhage and Friedman 2009; Pfeiffer and

Lingner 2013; Bonnell et al. 2021). Telomerase uses an intrinsic RNA molecule as the template

for synthesis of the TG-rich strand (Greider and Blackburn 1989; Singer and Gottschling 1994)

while the lagging strand polymerase machinery fills in the complementary, CA-rich strand

(reviewed in Gilson and Géli 2007; Pfeiffer and Lingner 2013).

The idea of the telomere as a critical component of the chromosomes was proposed in

flies by Herman Mueller and maize by Barbara McClintock in the 1930s before DNA was

confirmed as the genetic material. Mueller and McClintock studied the behavior of chromosomes

in response to treatment with X-rays. They separately observed that natural chromosome ends were protected from the rearrangement and fusion experienced by broken chromosome ends, leading both to conclude that naturally occurring ends were different than broken ends. Based on these observations, Muller termed chromosome ends telomeres (McClintock 1938, Muller 1938, Blackburn *et al.* 2006). However, in 1938, when telomeres were first named by Mueller, the lack of knowledge about DNA and the lack of tools to study chromosomes at a molecular level prevented telomere research from making meaningful progress.

In the 1960s and 1970s, the tools to study DNA were developed and the availability of sequencing technology allowed for progress in the study of telomeres. In 1978, Elizabeth Blackburn isolated mini-chromosomes containing the ribosomal DNA from *Tetrahymena* and sequenced their native ends. The sequencing revealed nonprotein-coding 5'-T_2G_4-3' repeats (Blackburn and Gall 1978). In 1981, similar repeats were found by the Prescott lab in a ciliate class distinct from *Tetrahymena* (Klobutcher *et al.* 1981). These discoveries allowed researchers to begin to understand the nature of telomeres at the molecular level and, as more tools to study DNA became available, many research groups began to turn their attention to the study of telomeres (Zakian 2012). As the number of available sequences increased, it became clear that some features of telomeres are conserved or similar across multiple species. For example, telomeres are often TG rich on the 3' terminating strand, even if the exact sequence varies from species to species (Srinivas *et al.* 2020) . The telomeres of many different species have been studied to understand the structure and function of telomeres more generally. Particularly, the study of telomeres in the model organism *S. cerevisiae* has been instrumental in forming our current understanding of how telomeres function and what other factors are important for

facilitating telomere function. The remainder of this section will primarily discuss research regarding *S. cerevisiae* telomeres.

The sequence of the budding yeast telomere was first elucidated by the Blackburn lab in 1984 (Shampay *et al.* 1984). At this time, sequencing of the chromosome end with the existing technology was a challenge because telomeres are repetitive and vary in size. Furthermore, at the time, it was difficult to pinpoint where telomere began on the chromosome end. To circumvent these issues, a chromosome end from *S. cerevisiae* was cloned into a yeast plasmid and subcloned into an *E. coli* plasmid to generate DNA to sequence. To determine the sequence of the yeast telomeres, the DNA was mapped using restriction enzymes and sequenced with the Maxam-Gilbert method. This strategy revealed that the telomere sequence of *S. cerevisiae* is heterogenous with one T followed by 1 to 3 Gs (TG_{1-3}) (Shampay *et al.* 1984). The Zakian lab employed a similar strategy to sequence the yeast telomere in 1990 using a different cloning strategy wherein they utilized the T4 DNA polymerase to clone the chromosome end. This new strategy allowed for the majority of the chromosome end to be cloned and therefore the Zakian lab was able to sequence a longer region of the DNA. The authors sequenced many telomeres captured on linear plasmids and, when they compared the telomeres, they observed an internal region that was identical between clones and an external region that was more variable. From these observations, the Zakian lab inferred that the yeast telomere is composed of two regions; a centromere proximal region of about 120-150 bp that is protected from recombination, breakage, and degradation and a distal region that is more susceptible to those processes (Wang and Zakian 1990). We now know that this structure is due to the mechanism of telomerase action wherein the internal portion of the telomere is generally replicated by the canonical DNA replication machinery and is therefore identical on a particular chromosome end between closely related

cells. In contrast, the distal portion shortens and is re-lengthened by telomerase (Wellinger and Zakian 2012)

Together, these studies revealed that the *S. cerevisiae* telomere is heterogeneous (contains a variable pattern), unlike the *Tetrahymena* telomere that is homogenous and thus has a constant pattern. From an experimental standpoint, the variability of *S. cerevisiae* telomerase makes it possible to distinguish newly added telomeres from preexisting telomeres.

The discovery of the yeast telomeric sequence was an important milestone in the yeast telomere research field, allowing researchers to study different aspects of telomere biology. Along with the sequence of the yeast telomere, researchers were also interested in the factors that bound to telomeres or interacted with telomeres to affect telomere length. At this time, it was thought likely that yeast telomeres were extended by an enzyme, a prediction supported by the identification by Greider and Blackburn of an activity from *Tetrahymena* extract capable of adding nucleotides to a telomeric oligonucleotide (Greider and Blackburn 1987). Lundblad reasoned that mutations affecting this telomere lengthening activity would be defective in maintaining a linear plasmid.

To test this hypothesis the Lundblad lab employed a genetic screen to identify genes that, when mutated, cause defects in telomere maintenance, length, or structure (Lundblad and Szostak 1989). The screen made use of a *LEU2*-marked circular plasmid containing inverted repeats of the *Tetrahymena* telomeric sequence separated by a *URA3* marker. Rare spontaneous breakage events within the *URA3* gene allow formation of stable linear molecules through extension of the *Tetrahymena* telomeric repeats with yeast telomere sequence. Cells in which the plasmid underwent linearization were selected by monitoring for the loss of the *URA3* marker

and maintenance of the *LEU2* marker. *URA3* encodes an enzyme that converts 5-fluoroorotic acid (5-FOA) to the toxic compound 5-fluorouracil. Therefore, a wild-type strain transformed with the circular plasmid generates papillae on media lacking leucine and containing 5-FOA (-leu 5-FOA).

The cells described above were treated with ethyl methanesulfonate (EMS) to create mutations and clones were identified that failed to produce papillae on -leu 5-FOA. One candidate also resulted in telomere shortening and senescence, phenotypes predicted for mutations in a telomere-maintenance mechanism. This gene was termed *Ever-Shorter-Telomeres 1* (*EST1*) and the mutation identified from this screen was termed *est1-1*. Up to this point the function of Est1 was still unknown, researchers speculated that Est1 could be a component of an enzyme similar to the *Tetrahymena* telomerase but this would not be confirmed for a number of years (Lundblad and Szostak 1989).

When the Blackburn lab isolated the enzyme from *Tetrahymena* extract that would later be termed telomerase, they characterized it as a ribonucleoprotein complex. Purification of the RNA revealed a template region for the addition and elongation of telomeres (Greider and Blackburn 1989). In yeast, the discovery of the telomerase RNA was more fortuitous. The Gottschling lab discovered the yeast telomerase RNA, termed *TLC1,* when studying the phenomenon of telomeric silencing (Singer and Gottschling 1994). Telomeric silencing is a phenomenon in which genes placed in the vicinity of telomeres are transcriptionally repressed. It was suspected that such epigenetic repression was due to heterochromatin formation near telomeres and the Gottschling lab reasoned that heterochromatin formation might be sensitive to gene dosage. Therefore, they sought to identify genes that, when overexpressed, restored expression of silenced genes near telomeres. To overexpress genes, the yeast cells were

transformed with a high expression cDNA library. To monitor telomeric silencing, *URA3* and *ADE2* markers were integrated near a telomere. If silencing was overcome, then the cells would be able to grow on media lacking uracil and would be white (colonies formed by cells lacking *ADE2* expression are red). In contrast, if silencing was not disrupted, then the cells would die on media containing uracil and would generate red colonies. From this screen, the authors identified multiple plasmids containing the same DNA region that disrupted telomeric silencing but did not affect silencing in non-telomeric regions. Sequencing failed to identify a long open reading frame and the authors noted the presence of a short sequence that could function as the template for synthesis of the yeast telomeric repeats. Knocking out this region led to shortened telomeres, and lessened survival. To determine whether this region was truly the RNA template, the Gottschling lab introduced a mutation to the template and demonstrated that the corresponding nucleotide changes were incorporated into the yeast telomeres. They named the gene encoding the telomerase RNA *TLC1* (Singer and Gottschling 1994).

Once Est1 and *TLC1* were identified as factors that affected telomerase function, researchers hypothesized that Est1 was likely a component of telomerase or a positive regulator of telomerase because strains lacking either *EST1* or *TLC1* presented similar phenotypes of telomere shortening and senescence. To identify other components of the telomerase holoenzyme and other factors that affect telomere maintenance, Lundblad and colleagues conducted a more expansive screen to identify candidates with the same phenotype as the *tlc1* and *est1* mutant strains (Lendvay et al. 1996). Mutations for this screen were generated by EMS and a multitiered screen was performed to isolate mutations that caused telomere shortening and senescence. The *est1-1* strain originally identified by the Lundblad lab was found to exhibit chromosome loss, presumably as a consequence of telomere shortening. By incorporating an

initial screen for mutations inducing chromosome instability, the authors were able to greatly expand the number of mutant strains that could be screened (Lendvay et al. 1996).

From these experiments, three new *EST* genes were discovered (*EST2, EST3,* and *EST4*). Using epistasis analysis, *EST1-4* were determined to function in the same pathway suggesting that they are likely components of telomerase or positively regulate telomerase activity. The *est4-1* strain was later identified to contain an allele of *CDC13* (Lendvay et al. 1996). Sequencing of *EST2* did not reveal its function and the function of *EST3* was also not elucidated in this study (Lendvay et al. 1996). Later studies comparing the amino acid sequence of Est2 to that of other reverse transcriptases revealed conserved motifs that suggested Est2 played a catalytic role within the telomerase holoenzyme (Lingner *et al.* 1997).

The discovery that the *est4-1* mutation is an allele of *CDC13* was especially intriguing. Cdc13 was first identified by the Hartwell lab in a genetic screen seeking to identify either components of the cell cycle or factors that affect cell cycle components. When raised to the non-permissive temperature, *cdc13-1* cells arrest as large-budded cells. This cell cycle arrest was later shown to be due to activation of the Rad9 checkpoint at the restrictive temperature. Analysis of *cdc13-1* cells at the non-permissive temperature revealed single-stranded DNA that corresponded to the yeast telomere. This was the first indication that Cdc13 may play a role in the maintenance of telomeres (Garvik *et al.* 1995). When *CDC13* and *RAD9* were both mutated, cells no longer arrested in G2, suggesting that ssDNA could be a signal for the *RAD9* checkpoint (Weinert and Hartwell 1988). The *cdc13-1* allele suggested that Cdc13 somehow protects chromosome ends from nucleolytic resection of the 5' strand. The *est4-1* allele from the Lundblad screen, in contrast, didn't show a G2 arrest or evidence of single-stranded DNA generation but rather showed telomere shortening and senescence similar to that observed with

the *tlc1* mutant. This phenotype of the *est4-1* mutant suggested that Cdc13 was involved in telomere maintenance. However, the exact role of Cdc13 at the telomere was unclear (Lendvay *et al.* 1996). The role of Cdc13 was eventually revealed when the Lundblad lab used multiple alleles of Cdc13 to reveal a duel role for Cdc13 wherein Cdc13 acted in both end capping and telomere maintenance (Nugent *et al.*1996). Later, coimmunoprecipitation assays performed by the Lundblad lab would reveal that the telomerase holoenzyme is comprised of Est1, Est,2, and Est3. These same experiments revealed that although Cdc13 is not part of the telomerase holoenzyme it does interact with telomerase (Hughes *et al.* 2000).

To determine the role of Cdc13 at telomeres, the Lundblad lab created fusion proteins of Cdc13 and the Est proteins (Evans and Lundblad 1999). They found that fusion of full length Est1 and Cdc13 causes severe telomere overlengthening. However, they also found that fusing mutant versions of Cdc13 and Est1 would confer similar results suggesting that close proximity of Est1 and Cdc13 is enough to overcome mutations in either protein that inhibit telomerase activity. Together these data suggest that Est1 could be important for recruiting telomerase to Cdc13. To further test this hypothesis, researchers also fused Cdc13 to the catalytic core of telomerase in the absence of Est1 and found that this fusion was enough to promote telomere maintenance in the absence of Est1. This corroborates the idea that Est1 is the component of telomerase that mediates the recruitment of telomerase to the DNA via interaction with Cdc13 (Evans and Lundblad 1999).

Research up to this point had shown that Cdc13 has multiple interacting partners that could indicate multiple roles for Cdc13 at the telomere. The interacting partner Stn1 was identified through a screen for factors that could suppress the *cdc13-1* mutation and later it was found that Stn1 and Cdc13 also interact with Ten1 to form the CST complex which was

9

demonstrated to play a role in end capping (Grandin *et al.* 1997, 2001). Based on the interactions of Cdc13 with Est1 and with Stn1 and Ten1, the Lundblad lab hypothesized that these different interactions indicated distinct roles for Cdc13 at the telomere. To investigate this hypothesis, Lundblad and colleagues studied the effects of fusion proteins in a lethal cdc13 null strain (Pennock *et al.* 2001). The fusion of Stn1 to Cdc13 was enough to rescue the lethality of the *cdc13* null strain, however telomeres in this strain remained shortened. This result suggests that Stn1 plays a role in end protection but not telomere replication. The authors then simultaneously fused Est1 to the Cdc13 DNA binding domain and found that this restored normal telomere maintenance. From these experiments, they inferred that the end protection role of Cdc13 depends on the Cdc13/Stn1 interaction and telomerase recruitment depends on the Cdc13/Est1 interaction (Pennock *et al.* 2001).

Together, the early genetic analysis of yeast telomeres and telomerase has allowed for a fuller understanding of the structure and function of telomeres and telomerase. The understanding of the components of telomerase and of the proteins that interact with telomerase has allowed for continued study of yeast telomeres and their function. Furthermore, research in yeast paved the way for research on telomeres in other species that are more difficult to study, such as humans. Indeed, researchers have demonstrated that many components of yeast telomeres and telomerase such as the CST complex or Est1 are at least functionally conserved in humans (Snow *et al.* 2003; Wellinger and Zakian 2012). This conservation is significant because telomere dysfunction is associated with human diseases. Therefore, the more that we can learn about telomeres and telomerase in model organisms such as yeast, the more that knowledge can be applied when studying these processes in humans (Rossiello *et al.* 2022).

1.2 Telomere healing and disease

Because telomeres are, by definition, the end of a DNA molecule, they resemble a DNA double-strand break (DSB). Indeed, similar to telomeres, enzymatic resection at a DSB generates 3' overhangs that can serve as substrates for homologous recombination. The specific sequence of the 3' overhang at telomeres distinguishes it from 3' overhangs generated by resection at a double-strand break, thereby enforcing different outcomes at these otherwise similar structures (telomere elongation versus DNA repair, respectively; reviewed in Doksani and de Lange 2014; Casari et al. 2022. Rarely, the 3' overhang generated at a DSB is recognized by telomerase, resulting in the addition of a new or *de novo* telomere (reviewed in Pennaneach et al. 2006; Hoerr et al. 2023). *De novo* telomere addition (*dn*TA), also termed telomere healing, causes loss of sequences distal to the site at which the telomere is added but prevents additional resection that would ultimately be lethal.

In humans, telomere healing was first observed by sequencing the DNA of patients with diseases caused by terminal chromosome truncations (e.g., Phelan-McDermid syndrome and α-thalassemia). When researchers examined the sequencing data to determine the cause of the gene mutation, they found shortened chromosomes stabilized by a telomere in several patients (Lamb *et al.* 1993; Bonaglia *et al.* 2011; Guilherme *et al.* 2015). These *de novo* telomeres were often located within a small region of the gene suggesting that sequences associated with these diseases are unusually prone to telomerase action. Despite the knowledge that telomere healing likely occurs at these sites and results in disease, little is known about the mechanism of telomere healing and all the ways in which it can promote disease.

Recently, the availability of sequence databases and tools like CRISPR (Clustered Regularly Interspersed Shore Palindromic Repeats) have allowed for progress in the study of

telomere healing in humans. One study examining sequencing data from over 2000 cancer patients revealed that some patient sequences contained evidence of *dn*TA within cancer cells. However, the effect of these telomere addition events was unclear (Dentro *et al.* 2021). To determine the mechanism of *de novo* telomere addition in humans and whether *de novo* telomeres could possibly play a role in promoting cancer, the de Lange lab utilized CRiSPR Cas9 to generate a DSB in hTERT-immortalized retinal pigment epithelial cells (hTERT RPE-1 cells) and a HeLa stem cell line that could then be monitored for repair (Kinzig *et al.* 2022). Using this strategy, the de Lange lab found that *dn*TA in humans is likely inhibited by long-range resection and the activation of ATR signaling. Furthermore, their findings led them to speculate that *dn*TA promotes tumorigenesis by protecting the chromosome ends of cancer cells from breakage-fusion-bridge cycles that would be detrimental to the cancer cells. These findings helped elucidate additional consequences of *dn*TA in human cells and a possible regulatory mechanism that may act upon DSBs to prevent *dn*TA (Kinzig *et al.* 2022). Although recent developments in scientific technology have allowed for progress in the study of *dn*TA in humans, working with human cells to answer these questions still poses many challenges. Telomeres and telomere regulation in humans is complicated and these processes are not as well understood as they are in yeast. Additionally, the genetic tools available for studying humans are not as expansive as those that exist for model organisms. The use of a model organism like yeast to study the regulation of *dn*TA could elucidate some of the outstanding questions in this field by bypassing some of the challenges that are inherent when studying humans.

1.3 Regulation of telomerase action at double-strand breaks in yeast

As discussed above, in human cells expressing telomerase, a *de novo* telomere can be added to a DSB despite cellular mechanisms to negatively regulate telomere addition. These

events can lead to disease. This phenomenon is not limited to human cells and has been observed in other eukaryotic organisms including *Tetrahymena* (Yu and Blackburn 1991), mice (Sprung *et al.* 1999), and *S. cerevisiae* (Mangahas *et al.* 2001; Stellwagen *et al.* 2003; Ouenzar *et al.* 2017). In this thesis, I will primarily discuss telomere healing in the context of *S. cerevisiae.* Although *dn*TA events are generally very rare, the *S. cerevisiae* genome contains hotspots where *dn*TA occurs at frequencies estimated to be at least 200-fold above background (Obodo *et al.* 2016; Epum *et al.* 2020). These sequences, termed <u>Si</u>tes of <u>R</u>epair <u>A</u>ssociated <u>T</u>elomere <u>A</u>ddition (SiRTAs), present a unique opportunity to use yeast as a model to study the consequences of such sequences for genome stability and evolution.

To understand how telomere healing may occur at high frequencies at some sequences but not at others, it is important to understand how telomerase is regulated at DSBs. Given the potential for telomerase to compete with the DNA repair machinery at DSBs, it is not surprising that multiple mechanisms inhibit *dn*TA. These mechanisms fall into two classes: (1) mechanisms that spatially or temporally separate telomerase from DSBs and (2) mechanisms that alter the action of telomerase at a DSB. Examples of the first class include observations that telomerase is sequestered in the nucleolus in response to DSBs (Ouenzar *et al.* 2017) and that nuclear retention of Cdc13 requires association with DNA (most predominantly at telomeres), a property that may limit the concentration of free Cdc13 (Mersaoui and Wellinger 2019). Here, I concentrate on the second class of mechanisms whereby telomerase action at a DSB is distinguished from its action at a telomere.

Pif1 is a 5′–3′ helicase with roles in telomere length regulation, Okazaki fragment processing, unwinding of G-quadruplex structures, DNA repair, disassembly of stalled replication complexes, and 5′ end resection (reviewed in Dewar and Lydall 2012; Chung 2014;

Muellner and Schmidt 2020). Pif1 also facilitates mitochondrial DNA replication; yeast strains lacking Pif1 are respiration incompetent (Foury and Kolodynski 1983). The *pif1-m2* allele, which lacks the nuclear localization sequence, retains mitochondrial function but causes telomere overlengthening of ~100 bp and increases the association of telomerase with telomeres (Schulz and Zakian 1994; Boulé et al. 2005). *In vitro*, Pif1 preferentially unwinds DNA/RNA duplexes (Boule and Zakian 2007), suggesting that Pif1 removes telomerase from the telomere. Indeed, yeast telomerase is largely non-processive *in vitro* and remains bound to the primer following synthesis of a single telomeric repeat, but addition of Pif1 allows further rounds of elongation by facilitating telomerase release (Boulé *et al.* 2005). *In vivo*, the limited concentrations of telomerase [fewer than one telomerase complex per telomere (Mozdy and Cech 2006)] may mean that telomerase released by Pif1 action is unlikely to result in additional telomere elongation. While Pif1 preferentially binds long telomeres *in vivo* (Phillips et al. 2015), experiments analyzing telomere addition in a single cell cycle are consistent with Pif1 action independent of telomere length, suggesting that enrichment at longer telomeres may reflect roles of Pif1 during replication (Stinus *et al.* 2018).

Pif1 also inhibits *dn*TA at DSBs. In strains lacking nuclear Pif1, telomere addition frequencies are elevated in response to spontaneous breaks and after induction of cleavage by the homothallic switching (HO) endonuclease (200- to 1,000-fold, depending on the allele and assay) (Schulz and Zakian 1994; Myung *et al.* 2001). Remarkably, the roles of Pif1 at endogenous telomeres and in response to DSBs are genetically separable. Pif1 is phosphorylated in a Mec1-Rad53-Dun1-dependent manner following DNA damage, and a variant that cannot be phosphorylated at key residues (Pif1-4A) maintains normal telomere length but cannot repress

*dn*TA despite associating at normal (or increased) levels with DSBs (Makovets and Blackburn 2009). How phosphorylation alters Pif1 activity is unclear.

Using TG$_{1-3}$ sequences of varying lengths integrated adjacent to an HO cleavage site, the Durocher lab systematically probed how Pif1 function is influenced by the telomeric character of a DSB (Strecker et al. 2017). With TG$_{1-3}$ seeds of ≥34 bp, telomere addition to the broken end is observed in bulk culture and nearly 100% of cells survive HO cleavage, even in a strain expressing wild-type Pif1. In contrast, below this threshold, telomere addition is strongly reduced by the presence of Pif1. A phospho-mimetic allele of Pif1 (*pif1-4D*) does not affect the threshold, suggesting that phosphorylation cannot account for this distinction (Strecker et al. 2017). An exhaustive analysis uncovered Cdc13 as a mediator of differential Pif1 action on TG$_{1-3}$ tracts of differing lengths. Cdc13 variants predicted to reduce interaction with Est1 or decrease DNA binding increased the threshold of TG$_{1-3}$ sequence required for resistance to Pif1 negative regulation. These results suggest that the difference between a DSB and a short telomere is dictated by levels of Cdc13 association/function (Strecker et al. 2017). Interestingly, Hiraga and Sugimoto report that a telomeric seed sequence of 22 bp supports robust telomere addition and >90% survival following HO cleavage (Hirano and Sugimoto 2007). Neither group directly assessed the capacity of Cdc13 to bind the seed sequence *in vitro*, so the difference in threshold may be explained by differential affinity of Cdc13 for the sequences tested.

Regarding both *dn*TA and telomere maintenance at the end of the chromosome, work on Pif1 demonstrates that many of the mechanisms that regulate telomerase activity vary based on telomere length. Another protein that demonstrates a length dependence in the regulation of *dn*TA is the yeast Ku complex. Ku is a conserved protein that binds both broken DNA ends and telomeres. At broken ends, Ku facilitates nonhomologous end joining by acting as a scaffold to

recruit the proteins necessary for repair (Fell and Schild-Poulter 2015). In contrast, at telomeres,

Ku contributes to telomerase recruitment. In yeast, Ku is a heterodimer consisting of proteins

encoded by the genes *YKU70* and *YKU80*. Genetic analysis revealed that *TLC1* overexpression

has the same phenotype as a deletion of either *YKU70* or *YKU80*, and furthermore,

overexpression of *YKU70 and YKU80* together partially rescues the defects of *TLC1*

overexpression. These results suggest that Ku likely interacts with *TLC1* (Peterson *et al.* 2001).

The evidence that Ku interacts with *TLC1* led researchers to further investigate whether

Ku plays a role in the recruitment of telomerase. To study this, researchers did a random

mutagenesis screen looking for alleles of *yku70* or *yku80* that would allow for telomeric silencing

in the presence of *TLC1* overexpression (Stellwagen *et al.* 2003). From this screen they found an

allele of *yku80* (*yku80*-135i) and through gel shift assays they determined that this particular

allele was deficient for interaction with *TLC1* but could still bind DNA. Further characterization

of *yku80*-135i found that this mutant was deficient for yKu's role at telomeres as mutant cells

presented shortened telomeres and slowed telomere kinetics (Stellwagen *et al.* 2003). Once the

effect of this allele at normal chromosome ends was elucidated, the authors became interested in

the effect of this allele at DSBs. To study the role of Ku in the generation of spontaneous gross

chromosomal rearrangement (GCR) events, researchers utilized the a GCR assay originally

developed by the Kolodner laboratory that allows one to monitor the rate of GCR events that

occur randomly on chromosome V (Chen and Kolodner 1999, Stellwagen *et al.* 2003). In this

assay, the gene *HXT13* is replaced with the *URA3* marker (7.5 kb telomere proximal to *CAN1*)

(Chen and Kolodner 1999). Cells that simultaneously lose both markers are selected on medium

containing 5-FOA (lethal to cells expressing *URA3*) and canavanine (lethal to cells expressing

CAN1). The Gottschling lab adapted this assay to isolate and characterize clones that

experienced a GCR event in a *yku80-135i* strain. They found that *dn*TA was reduced by this mutation, but only at very short TG sequences (fewer than 4 bp). They proposed that the Ku complex may allow the recruitment of telomerase to a DSB when there is insufficient TG-rich sequence to bind Cdc13 (Stellwagen *et al.* 2003).

To identify other factors that promote telomere healing at DSBs, the Durocher lab utilized the Kolodner GCR assay in a *pif1Δ* background to identify candidates that reduce the high GCR rates characteristic of this strain background (Zhang and Durocher 2010). This screen was done using methyl methanesulfonate (MMS) to create DSBs and thereby further increase the GCR rate. Cells were mutagenized by random transposon insertion. Colonies that presented lower GCR rates than the control *pif1Δ* strain were selected and transposon insertion mapping was used to determine which gene was disrupted. One of the genes identified was *RRD1*. *RRD1* encodes a conserved activator of protein phosphatase 2a (PP2a) and further analysis revealed that Rrd1 promotes telomere healing at a DSB via activation of the yeast PP2a phosphatase, encoded by *PPH3*. This result led the authors to hypothesize that phosphorylation is important for the regulation of *dn*TA (Zhang and Durocher 2010). The Durocher lab became interested in determining which kinase was responsible for this negative regulation of telomere healing. Initially they focused on Mec1 and Tel1, the yeast orthologs of the ATM and Rad3-related (ATR) kinases, because both are activated by DSBs. Deletion of Mec1 but not Tel1 caused *dn*TA to increase (Zhang and Durocher 2010). In a series of elegant experiments to further investigate which targets of Mec1 would influence *dn*TA, the authors demonstrate that Mec1 negatively regulates telomere healing through its interaction with Cdc13. Specifically, Mec1 directly phosphorylates Cdc13 at serine 306, thereby preventing the accumulation of Cdc13 at DSBs (Zhang and Durocher 2010).

Mec1 action is opposed by the Pph3 phosphatase in a manner requiring the activator Rrd1. Remarkably, deletion of *RRD1* eliminates *dn*TA at TG tracts of fewer than 11 bp, consistent with a requirement for Cdc13 association at such sequences. While Mec1-dependent phosphorylation of Cdc13 is detected in response to DNA damage, bulk levels of phosphorylated Cdc13 do not increase upon deletion of *PPH3* or *RRD1*, suggesting that dephosphorylation may specifically occur at DSBs. Consistent with this idea, Pph3 accumulates at HO-induced breaks (Zhang and Durocher 2010).

The loss of *dn*TA events at sequences with fewer than 11 TG_{1-3} nucleotides is puzzling because Cdc13 binding requires 11 bases of TG-rich ssDNA. How can phosphorylation of Cdc13 influence its association with a sequence to which it is not predicted to directly bind? One possibility is that Cdc13 associates, albeit with lower affinity, to shorter TG tracts. While several positions of the 11-base Cdc13 binding site are critical (G1, G3, and T4), single mutations are tolerated in the rest of the binding site with minimal consequences for affinity (Eldridge et al. 2006; Lewis et al. 2014). Cdc13 associates with a resecting chromosome break, even in regions where "ideal" Cdc13 binding sites are not present (Oza *et al.* 2009), suggesting that Cdc13 binds with low affinity at multiple sites or that other interactions facilitate association with ssDNA. For example, proteins such as RPA and Rad51 influence the recruitment of Cdc13 at DNA ends and the outcome of DNA repair (Epum *et al.* 2020). These findings, along with work on Pif1, suggest that, in response to DNA damage, at least two proteins (Cdc13 and Pif1) are phosphorylated to reduce the probability of *dn*TA (Makovets and Blackburn 2009; Zhang and Durocher 2010). These mechanisms are additive, with both contributing to the extremely low rate of *dn*TA at most sequences (Zhang and Durocher 2010).

1.4 Sites of repair associated telomere addition (SiRTAs).

Sequences that promote telomerase action at DSBs and serve as hotspots of *dn*TA in

yeast were first observed as sites of telomere healing in response to an induced DSB on

chromosome VII (Mangahas *et al.* 2001). In these initial experiments, where hotspots of *dn*TA

were first observed in yeast, researchers examined how cells survived the sudden loss of the

telomere. To characterize the fate of these chromosomes, the authors utilized a yeast strain that

was haploid, but was disomic for chromosome VII. One copy of chromosome VII had a HO cut

site and the other did not. The chromosome containing the HO cleavage site was marked with

URA3 and *ADE3*. The HO cut site allowed for the creation of a break that could be monitored for

repair and the strategic placing of the genetic markers allowed researchers to distinguish between

various types of outcomes such as chromosome loss, *dn*TA, recombination, and non-homologous

end joining. Through this assay, the Zakian lab identified two sites on chromosome VII that

incurred *de novo* telomere addition in multiple independent clones (Mangahas *et al.* 2001). In

both cases, TG-rich sequences were located immediately proximal to the sites at which telomere

addition had occurred.

These sites on chromosome VII were not the only sites where higher instances of *dn*TA

were found. In section 1.3, I discussed the use of the Kolodner GCR assay to investigate the role

of Ku at telomeres and breaks. The Kolodner lab utilized this same assay to study the role of S-

phase checkpoint mutations in preventing chromosomal rearrangements (Myung *et al.* 2001). In

these experiments, the Kolodner lab determined if GCR rates were increased in the absence of

various checkpoint components. They then characterized the sites at which the rearrangement

occurred. Rarely, they observed a chromosome truncation that resulted from a *dn*TA in a specific

region on chromosome V (Myung *et al.* 2001). This same region was shown to stimulate *de*

novo telomere addition again when the Gottschling lab adapted the GCR assay to study Ku (Stellwagen *et al.* 2003). When these experiments were done in a strain background where the Ku and *TLC1* interaction was disrupted, the authors discovered that *dn*TA was eliminated at sequences with fewer than four TG_{1-3} residues. However, they noticed that *dn*TA was not reduced in a region of *NPR2* that contains TG-rich (telomere like) sequences (Stellwagen *et al.* 2003). The Durocher lab found *dn*TA events in this same region when investigating the *rrd1* deletion strain (Zhang and Durocher 2010).

Together, these results suggested the existence of sequences that act as hot spots for *dn*TA within the yeast genome. To characterize these sequences, the Friedman lab performed a structure/function analysis of the sequence on chromosome V and an additional hotspot on chromosome IX and named these sequences Sites of Repair-associated Telomere Addition or SiRTAs. Although the chromosome V SiRTA was identified as described above using a spontaneous GCR assay, the spontaneous events were very rare, making these events difficult to study. Generation of a chromosome break through induced cleavage with the HO endonuclease circumvents this issue. Furthermore, based only on the spontaneous assay, it was unclear whether SiRTAs are hotspots for breakage, repair, or both. By inducing the break several kb from the SiRTA, the Friedman lab demonstrated that these were hotspots of repair (Obodo *et al.* 2016). To perform the HO-cleavage assay, a recognition site for the HO endonuclease was placed ~3 kb distal to each SiRTA and a *URA3* marker was placed further distal to the cleavage site. The HO endonuclease is controlled by a galactose-inducible promoter such that growth on galactose creates a persistent DSB. Cell survival requires a repair event that mutates or removes the HO cleavage site. Counterselection for the *URA3* marker on medium containing 5-FOA allows selection of cells that have lost the end of the chromosome. Cells resistant to 5-FOA are screened

by PCR to identify those in which the truncation event occurred within the SiRTA and Southern blotting or PCR is used to determine whether *dn*TA had occurred. The percentage of total GCR events that occur in the SiRTA is used to measure the efficiency with which the SiRTA stimulates *dn*TA (Obodo *et al.* 2016).

To identify key *cis*-acting sequences, structure/function analysis was performed on the chromosome V SiRTA (Obodo et al. 2016). This analysis ultimately identified three separable regions: a 5' region, a middle region, and a 3' region. Initially the Friedman lab focused on the 3' region because of its TG-richness and because the majority of *dn*TA was observed in this region. The 5' region was then noted for its TG-richness and for the presence of a sequence predicted to bind Rap1, a transcription regulator with multiple roles including telomere length maintenance (Gajendra and Tomar 2016; Obodo et al. 2016) . The intervening middle sequence was not notably TG-rich. The 5' and 3' regions were individually replaced with a poly A sequence of equal size. A third mutation was created by changing each base of the middle sequence to its complement. The HO-cleavage assay was performed on each of these constructs to determine the effect of these mutations on SiRTA efficiency. Mutation of either the 5' or 3' sequence greatly reduced SiRTA efficiency while mutation of the middle sequence had no effect. As a result of this analysis, the 3' region that serves as the direct substrate for telomere addition by telomerase was named the "Core" sequence while the 5' region was named the "Stim" for its ability to stimulate *dn*TA at the Core sequence (Figure 1a). The intervening sequence (termed the spacer) does not contribute to *dn*TA, but deletion of the spacer greatly increases the frequency of *dn*TA, likely by allowing the Stim and the Core to be in closer proximity (Obodo *et al.* 2016).

The TG-rich nature of the Stim sequence suggested that it might bind either Rap1 (a protein with affinity for double-stranded telomeric DNA) or Cdc13 (a protein with affinity for

single-stranded telomeric DNA). To determine which factors may be important to stimulating

*dn*TA, Friedman and colleagues replaced the Stim sequence of chromosome V with two copies

of the Gal4 upstream activating sequence (2X-UAS), a sequence that is bound by the Gal4 DNA

binding domain (Gal4-DBD). Constructs were created that fused the Gal4-DBD to either *CDC13*

or *RAP1* to allow for the artificial recruitment of these proteins to the stim:2X-UAS sequence.

The *RAP1* fusion was not sufficient to stimulate *dn*TA, while the *CDC13* fusion supported robust

telomere addition that was dependent on the presence of the UAS sequence. Furthermore,

mutations of the Stim designed to favor binding by Cdc13 over Rap1 supported the conclusion

that Cdc13 binding is critical (Obodo *et al.* 2016). Together, these observations support a model

in which 5' to 3' resection at DSBs exposes TG-rich sequences that are bound by Cdc13, with

subsequent recruitment of telomerase (Figure 1).

(a)

(b) *de novo* telomere addition

(c) Cell death or GCR through alternative mechanism
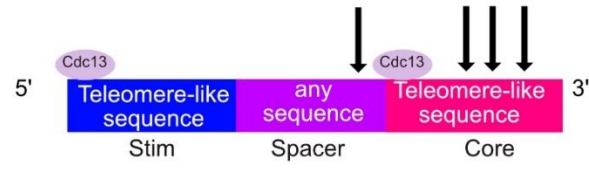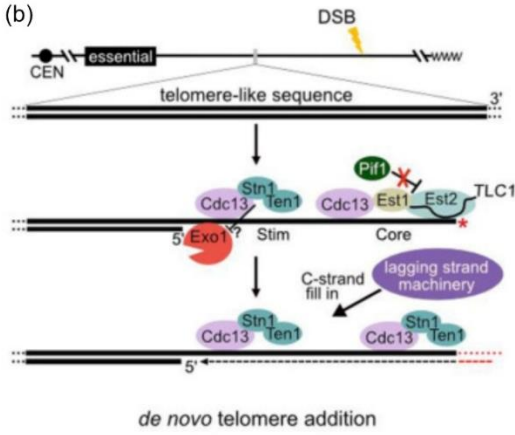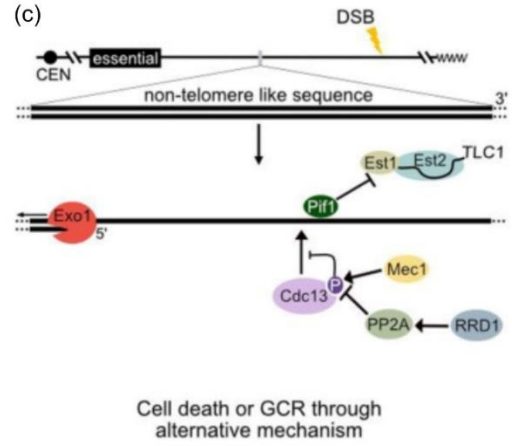
Figure 1. Models of telomerase regulation at a resecting break in the presence and absence of telomere-like sequences. a) Graphic depiction of sequence components of a general SiRTA. The blue and pink regions refer to the Stim and the Core which contain telomere-like TG rich sequence able to bind Cdc13. The Stim sequence is required to stimulate dnTA at the core sequence. The black arrows refer to possible telomere addition events which primarily occur in the core region but a few events can also occur in the other regions. The purple region refers to the spacer sequence which is the intervening sequence between the core Stim and the core; this sequence does not have to be TG rich (Obodo *et al.* 2016). b) Regulation of telomerase at Sites of Repair-associated Telomere Addition (SiRTAs). Following induction of a double-strand break (DSB), 5' end resection is initiated by the MRX complex (Mre11–Xrs2–Rad50) and Sae2. Multiple nucleases are involved in extensive resection following a DSB but this process requires the exonuclease Exo1 and helicase Sgs1 and creates single-stranded DNA (Zhu et al. 2008; Gravel et al. 2008; Mimitou and Symington 2008). Following resection through the TG-rich sequences, the resulting TG-rich ssDNA recruits Cdc13 binds to a "Core" sequence and recruits telomerase through interactions with Est1. Cdc13, in complex with Stn1 and Ten1 [likely as a hexamer (Ge et al. 2020)], also binds to a proximal "Stim" sequence to prevent further 5' resection. The limited generation of ssDNA inhibits Pif1 loading and removal of telomerase (see text). Both the Stim and the Core sequences are required to stimulate *de novo* telomere addition, but it is unclear whether Cdc13 complexes bound to each are functionally distinct (as depicted here). Telomerase must access a 3' terminus, which is generated through an unknown mechanism to prime telomere synthesis (depicted by a red *). Following *de novo* telomere addition by telomerase, the CST complex recruits the lagging strand machinery for C-strand fill-in. c) Regulation of telomerase at sequences lacking SiRTA sequences. In the absence of DSB repair, 5' resection proceeds unimpeded. Mec1 phosphorylation of Cdc13 at serine 306 inhibits Cdc13 accumulation at short TG_{1-3} sequences less than 11 bases (Zhang and Durocher 2010). Pph3 phosphatase (in a manner requiring the activator Rrd1) counteracts Cdc13 phosphorylation (Zhang and Durocher 2010), but Pif1 binds and inhibits telomerase action to strongly repress *de novo* telomere addition (Schulz and Zakian 1994). This figure was adapted from Hoerr *et al.* 2021 and was made by Remington Hoerr.

To determine whether the SiRTA identified on chromosome IX has a similar structure to the SiRTA on chromosome V, Friedman and colleagues utilized gel shift analysis to identify sequences within the chromosome IX that bind Cdc13 (Obodo *et al.* 2016). Two sequences with affinity for Cdc13 were identified (BS1 and BS2). Mutation of these sites to poly A revealed that BS1 plays a minor role in SiRTA activity, while BS2 is required for robust *dn*TA, consistent with a role for Cdc13 binding in SiRTA function. Interestingly BS2 is located closer to the Core sequence than BS1, which may explain why BS2 is the primary Stim sequence.

The characterization of the SiRTAs on chromosome IX and V revealed the importance of Cdc13 binding, which requires the generation of ssDNA. During DSB repair, Rad51 forms a nucleoprotein filament on the ssDNA in a manner that requires Rad52 (reviewed in Andriuskevicius et al. 2018). The Friedman lab considered the possibility that single-stranded

DNA binding proteins such as RPA and Rad51 might compete with Cdc13 for binding to the resected DNA and thereby influence the outcome of repair.

To address the potential competition between *dn*TA and homologous repair pathways, he Friedman lab initially examined the effect of *RAD51* deletion on the frequency of *dn*TA at SiRTAs (Epum *et al.* 2020). Interestingly, deletion of *RAD51* significantly reduced *dn*TA at the SiRTA. This result led the authors to hypothesize that a deletion of *RAD5*2 would have a similar affect because Rad52 is required to generate the Rad51 filament. Surprisingly, deletion of *RAD52* had no effect on the rate of *dn*TA. To gain additional insight, the Friedman lab performed epistasis analysis and discovered that the *RAD52* deletion phenotype is epistatic, consistent with a model in which Rad52 inhibits *dn*TA, but only in the absence of Rad51. To determine whether these observations were due changes in Cdc13 binding, the authors measured Cdc13 binding by chromatin immunoprecipitation (ChIP) at the chromosome IX SiRTA in *rad51Δ, rad52Δ and rad51Δrad52Δ* strain backgrounds. Consistent with the observed effect on SiRTA function, Cdc13 association with the SiRTA following induction of a break was only substantially reduced in the single *rad51Δ* strain. Additionally, artificial recruitment of Cdc13 to the chromosome V SiRTA via the GBD/UAS system (described above) suppressed the *rad51Δ* defect. These results suggest that the defect in the *rad51Δ* strain is due to lack of binding by Cdc13 (Epum *et al.* 2020).

Insights to the mechanism through which Rad52 might repress SiRTA function in the absence of Rad51 came from the Greene lab. While examining protein dynamics during presynaptic complex assembly, Greene and colleagues found that Rad52 stabilizes RPA on single-stranded DNA in a manner that is counteracted by Rad51 (Gibb *et al.* 2014). Indeed, Epum *et al.* found that the interaction between Rad51 and Rad52 (and not Rad51's ability to bind

ssDNA) is critically important to suppress the negative role of Rad52. Furthermore, mutations in

RPA that weaken its association with ssDNA suppress the *rad51Δ* phenotype. Together, these

results suggest that, in the absence of Rad 51 (and the presence of Rad52), tight association of

RPA with single-stranded DNA interferes with Cdc13 binding. However, Cdc13 is able to bind

the SiRTA in the context of the Rad51 nucleofilament (Epum et al. 2020).

Given the evidence that Cdc13 binding at a SiRTA is required for high levels of *dnTA*,

we hypothesize that SiRTAs are at least partially resistant to negative regulation by Pif1, Mec1,

and yKu. If a SiRTA is present when resection occurs following a break, robust binding by

Cdc13 is sufficient to recruit telomerase (Figure 1b). Additional recruitment of the Stn1 and

Ten1 proteins may also contribute to *dn*TA by preventing excessive resection and/or promoting

fill-in synthesis of the resected strand. At sequences lacking a SiRTA (Figure 1c), the negative-

regulatory pathway described above are expected to inhibit *dn*TA.

Overall, the characterization of the two SiRTAs on chromosome V and chromosome IX

suggests that telomere addition is favored at SiRTAs even when the initiating break is artificially

induced 2-3 kilobases distal to the eventual site of telomere addition, suggesting that SiRTAs

stimulate repair rather than serving as fragile sites *per se* (Obodo *et al.* 2016).

1.5 Characterization of SiRTAs in the yeast genome

Although progress has been made in the characterization of the *cis-* and *trans*-factors

facilitating *de novo* telomere addition, little is known about the distribution of SiRTAs and what

the purpose of SiRTAs might be. To address whether *dn*TA at SiRTAs provides an advantage to

cell survival that has driven the retention of these sequences in the yeast genome, it is important

to determine the genome-wide distribution of SiRTAs. Furthermore, once the distribution of

SiRTAs is understood, characterizing various types of SiRTAs either based on sequence or by location becomes possible.

In this thesis, I describe the development of two novel methods for testing putative sequences for their ability to stimulate *dn*TA. These methods simplify and expedite analysis of putative SiRTA sequences. In the first method, CRiSPR Cas9 is used to insert a sequence of interest onto a site on chromosome VII, facilitating the analysis of *dnTA* following HO cleavage. In the second method, high throughput sequencing is utilized to determine the number and location of telomere events in pooled samples of colonies that underwent a GCR event.

These new techniques allowed the rapid analysis of a much larger number of putative SiRTAs than previously possible and led to the development of a collection of sequences with a range of ability to stimulate *dnTA* (including a large number of non-functional sequences). This collection of sequences provided insight into the structure of a SiRTA and which sequences may or may not stimulate *dn*TA. From this initial collection of data, we determined that a simple accounting of T and G residues is alone insufficient. For example, a sequence with multiple poly T tracts is unlikely to stimulate *dn*TA because yeast telomeres do not contain multiple adjacent Ts. This insight allowed informed development of the Computational Algorithm for Telomere Hotspot Identification (CATHI). This thesis will describe the creation and validation of the CATHI which allowed, for the first time, development of a comprehensive map of putative SiRTAs in the yeast genome. With this map, I describe trends in the distribution of putative SiRTAs and find that sequences predicted to function as SiRTAs are more prevalent in the yeast genome than expected by chance.

Strategies for testing candidate sequences for SiRTA function

To study the effect of *dn*TA on genomic stability, the ability to efficiently test sequences

for their ability to stimulate *dn*TA is essential. Historically, the Friedman lab has tested putative

SiRTAs in their endogenous location using a multiplex PCR method to determine whether a

GCR event occurred within the SiRTA. Subsequently, Southern blotting was used to confirm

that the event was a *de novo* telomere. In this chapter, I describe the development of

methodologies to 1) rapidly generate yeast strains in which any sequence of interest can be tested

for SiRTA function at a defined location and 2) measure the efficiency of SiRTA function.

2.1 Introduction

One of the key functions of telomeres is to distinguish the single-stranded ends of

chromosomes from the single-stranded, 3' overhangs generated by DSBs (Glousker and Lingner

2021). SiRTAs are sequences that, when rendered single-stranded by a DSB, can be mistaken by

telomerase for a telomere, causing a *dn*TA event to occur instead of normal repair processes

(Obodo *et al.* 2016). *De novo* telomere addition results in terminal truncation of the chromosome

but can also protect the rest of the chromosome from further damage. Characterizing SiRTAs is

thus important for understanding the process of telomere healing and genomic instability more

generally. These studies require the ability to efficiently test putative sequences for their ability

to stimulate *dn*TA.

---

[3] This work is adapted from portions of the research article: Ngo K, Gittens TH, Gonzalez DI, Hatmaker EA, Plotkin S, Engle M, Friedman GA, Goldin M, Hoerr RE, Eichman BF, Rokas A, Benton, ML, and Friedman KL. 2023. A comprehensive map of hotspots of *de novo* telomere addition in Saccharomyces cerevisiae. *Genetics*, iyad076. https://doi.org/10.1093/genetics/iyad076

[4] This work is adapted from portions of the research article: Ngo K, Epum EA, Friedman KL. Emerging non-canonical roles for the Rad51-Rad52 interaction in response to double-strand breaks in yeast. Curr Genet. 2020 Oct;66(5):917-926. doi: 10.1007/s00294-020-01081-z.

Previously, SiRTAs were analyzed in their endogenous location by adding a recognition site for the HO endonuclease and a *URA3* marker telomere proximal to the sequence of interest. The HO site allowed for persistent cleavage by the HO endonuclease leading to a DSB that could be acted upon by telomerase. The *URA3* marker allowed the identification of GCR events because loss of the marker can be screened on medium containing 5-FOA. To determine whether the GCR events that occurred were *dn*TA, multiplex PCR was used to map the location of the event and Southern blotting was used to confirm whether GCR events involved *dn*TA. This strategy allowed for the detailed characterization of two SiRTAs, one on chromosome IX and the other on chromosome V (Obodo *et al.* 2016).

The strategy described above (testing each putative SiRTA at its endogenous chromosome location) is not feasible if the goals are to identify multiple SiRTAs throughout the yeast genome and to compare the ability of these sequences to stimulate *dnTA*. To test a SiRTA endogenously, an HO cut site and a *URA3* marker must be added at each putative SiRTA, a time-consuming process. Furthermore, the assay for *dn*TA assumes that telomere addition is compatible with viability, therefore requiring the sequence of interest to be located in a non-essential terminal region and to be oriented appropriately to stabilize the centromere-proximal side of the break. As a result, putative SiRTAs in regions of the genome that are essential (cannot tolerate chromosome loss) cannot be tested because *dn*TA would be lethal. Although SiRTAs in the nonessential regions can be tested endogenously, a comparison of SiRTA efficiencies is confounded by the varied chromosome context. This chapter describes the development of a common test site on chromosome VII where CRiSPR Cas9 is used to insert putative SiRTAs onto a site 2.1 kb away from an HO cut site with a *URA3* marker at the end of the chromosome.

Mapping the location of GCR events by PCR is time-consuming and inefficient. Each experimental trial requires the analysis of 30 individual clones using multiple PCR reactions and at least two experimental trials are required for each strain/condition tested. Furthermore, the precise location of telomere addition requires cloning and sequencing of individual telomere addition events. In this chapter, I describe the development and validation of a methodology that can analyze multiple GCR events (~30) in a single genomic DNA sample subjected to Illumina sequencing. Together, the novel methods described here offer more efficient and informative methods for testing putative SiRTAs, which will allow for the continued characterization and study of SiRTAs.

2.2 Results
2.2.1 Test site on Chromosome VII can be used to test sites for SiRTA activity

To circumvent the previously discussed limitations of testing a sequence endogenously, I developed a "test site" on the left arm of chromosome VII. CRISPR/Cas9 is used to insert sequences (typically 300 bp) oriented with the TG-rich sequence of interest on the bottom (3' to 5') strand. Sequences that have been tested during experiments described in this thesis are listed in Table 1 (see appendix A). A recognition site for the HO endonuclease is located ~2kb distal to the CRISPR/Cas9 integration site. A *URA3* marker located further distal to the HO cleavage site facilitates selection for cells carrying a truncated chromosome VII-L (Figure 2a). *RAD52* is deleted to prevent homology-directed repair between the sequence inserted on chromosome VII and that same sequence at its endogenous location. Oligos utilized for strain construction are listed in Table 2 (see appendix A).
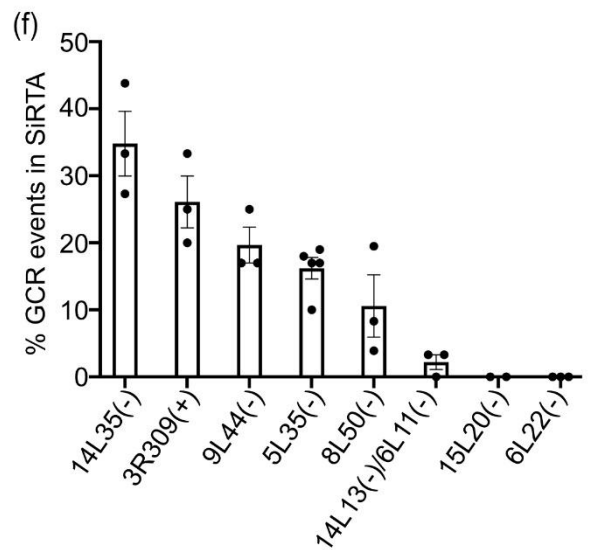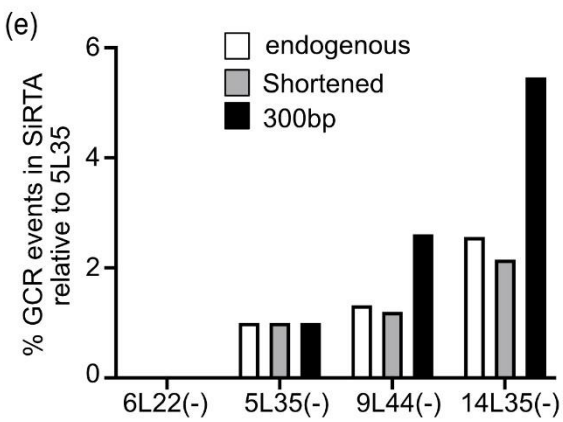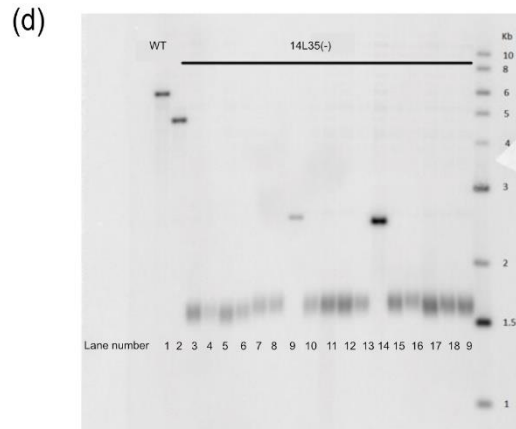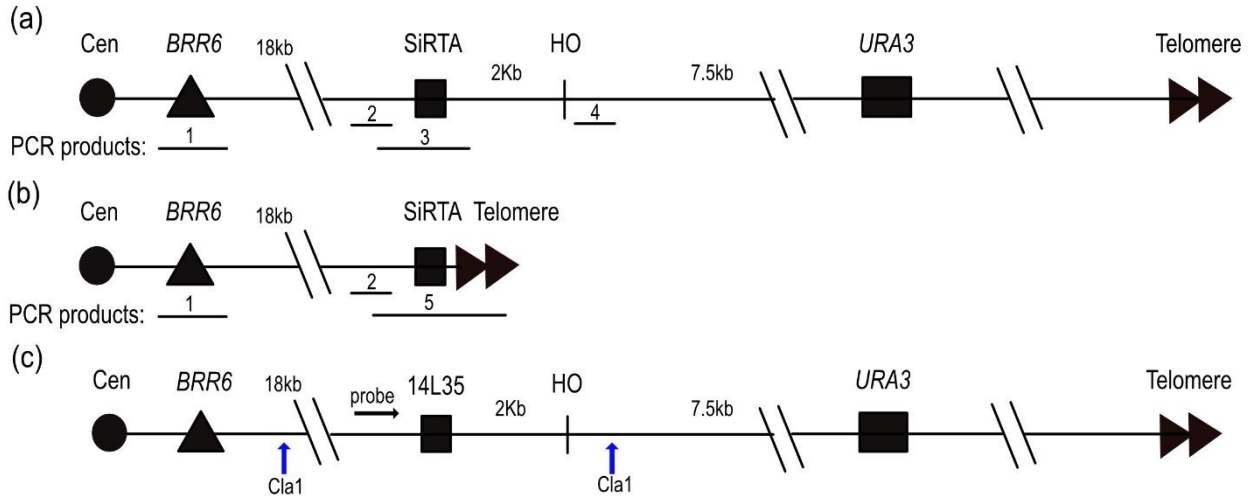
Figure 2. Development of chromosome VII site to test sequences for *de novo* telomere addition. a) Diagram depicting the structure of chromosome VII and the strategy for mapping GCR events resulting from a DSB generated by the HO endonuclease. Locations of the essential gene BRR6 and the sequence to be tested (SiRTA) are shown. A URA3 marker inserted distal to the HO cleavage sites allows selection for cells that lose the chromosome terminus following HO cleavage (see text for additional explanation). The approximate locations of PCR products utilized to map GCR events are shown. Product 1 amplifies sequences within BRR6 and is utilized as a positive control. Products 2 and 3 amplify regions internal to the putative SiRTA or across the SiRTA, respectively, and are used to identify clones in which a GCR event occurred within the SiRTA. Product 4 is used to verify loss of the chromosome terminus. b) Diagram of a GCR event in which telomere addition occurred in the putative SiRTA. Addition of telomeric DNA at the SiRTA is verified by the presence of PCR product 5, generated with a forward primer proximal to the SiRTA and a reverse primer complementary to the telomeric repeat. c) diagram of chromosome showing the probe (black arrow) and the restriction sites for Cla1 that were used for cleavage for the Southern blot d) Southern blot of 14L35(-) colonies that experienced GCR event at the SiRTA. WT in lane 1 refers to control that is the same strain as the other samples, but it did not undergo HO cleavage. Telomeres present as fuzzy bands around 1.5kb. e) Percent of GCR events in the SiRTA of the indicated SiRTAs tested on chromosome 7 and in their endogenous locations. Efficiencies are calculated as the percentage of GCR events that occur within the SiRTA. Individual data points represent individual experiments. Data for SiRTAs 5L-35 and 9L-44 in their endogenous locations are from Epum *et al.* 2020. All other data in this graph are from Ngo *et al.* 2020. f) Percent of GCR events in each of the indicated candidate SiRTAs as measured on chromosome 7. Sequences analyzed are in Table 1. Data corresponding to this figure are in Table 3. Data are from Ngo *et al.* 2020.

Cells are plated on solid medium containing galactose to induce expression of the HO endonuclease. Because correct repair restores the HO site and leads to recurrent cleavage, cells that survive to generate a colony have either incurred a mutation that blocks nuclease recognition or have lost the HO site completely through the formation of a gross chromosomal rearrangement (GCR), which includes *dn*TA. To identify GCR events, 100 clones arising on the galactose plate are screened for loss of the *URA3* marker via growth on a medium containing 5-fluoroorotic acid (5-FOA). Thirty 5-FOA-resistant clones are then analyzed to determine the nature of the resulting GCR event. Prior to the development of the high throughput sequencing strategy described later in this chapter, we utilized a clone-by-clone mapping strategy that employs multiple PCR reactions to identify the approximate location of each GCR event (Figure 2a). For colonies in which the event maps to the sequence of interest, Southern blotting or PCR utilizing a telomeric primer is utilized to determine if the event involved telomere addition

(Figure 2b; oligonucleotides used for PCR are listed in Table 2 (see appendix A)). By Southern

blot, 15 of 18 events observed at 14L35(-) on chromosome VII were *dn*TA (Figure 2c).

Furthermore, Southern blot also confirmed that 13 of 13 events observed at 5L35(-) on

chromosome VII involved telomere addition (Ngo and Friedman, unpublished data), suggesting

that most GCR events identified in this assay involve telomere addition. Unless specified

otherwise, we recorded the actual percentage of *dn*TA to use as an indicator of SiRTA

efficiency. However, because the majority of events at the SiRTA were confirmed as *dn*TA, in

instances where *dn*TA was not specifically tested for, the % of GCR events within the SiRTA is

recorded instead and used as a proxy for SiRTA efficiency. Throughout this thesis, sequences

tested are named using the following scheme: chromosome number, chromosome arm (L for left

and R for right), distance from the left telomere rounded to nearest kilobase, and the strand on

which the putative SiRTA is located [(+) for the forward strand and (–) for the reverse strand].

Unless specified, tested sequences are 300 bp in size.

To determine the effect of testing SiRTA function at a non-native location, we measured

telomere addition within four sequences at both their endogenous locations and after integration

at the test site on chromosome VII [published SiRTAs 5L35(-) and 9L44(-) (Obodo *et al.* 2016;

Epum *et al.* 2020), published SiRTA 14L35(-), and 6L22(-), a sequence found to be devoid of

SiRTA activity (Ngo *et al.* 2023)]. For these initial experiments, the sequences moved to

chromosome VII were ~100bp in size. For each active SiRTA, the efficiency measured on

chromosome VII was lower than in the endogenous location (for SiRTA 9L44(-) this

corresponds to efficiencies of 20% and 39%, respectively). When comparing the endogenous

rates to the chromosome VII test site rates we observe a consistent trend wherein 14L35(-) has a

higher SiRTA efficiency than 9L44(-) and 5L35(-) and 6L22(-). (Figure 2d). Furthermore, A TG-

rich sequence from chromosome VI (6L22(-)) does not stimulate GCR formation in either its endogenous location or on chromosome VII and serves as a negative control for the assay (Figure 2d, Figure 2e) (Ngo *et al.* 2020). Once I established that the chromosome VII test site could be used to monitor SiRTA activity, I used this strategy to test several additional sequences chosen for their TG-richness and proximity to the chromosome end for their ability to stimulate *dn*TA. From this work I identified several sequences that stimulate *dn*TA at varying frequencies including sequences that looked as though they should stimulate *dn*TA but did not (Figure 2e, Table 3 (see appendix A)).

I hypothesized that *dn*TA is reduced on chromosome VII relative to the endogenous location because the entirety of the SiRTA was not inserted into the test site. To ensure consistency in testing, a decision was made to insert 300 bp sequences at the Chr VII test site (Figure 2c; sequences tested in this manner are listed in Table 1 (see appendix A)). For 9L44(-), the 300 bp sequence supported *dn*TA at a frequency of 36% compared to 20% and 39% in the truncated SiRTA and the endogenous SiRTA, respectively. In contrast, a 300 bp sequence containing 5L35(-) had an efficiency of 13%, more similar to the frequency of events of the shortened sequence (16%) and less than the frequency of the endogenous sequence at 36%. Once again, we observe that 14L35(-) has the highest efficiency and 6L22(-) has no ability to stimulate *dn*TA. Interestingly the observation that the longer SiRTA did not increase *dn*TA to endogenous levels for 5L35(-) indicates that either the entire sequence was still not captured or there are other factors on chromosome V that contribute to *dn*TA that are not present on chromosome VII. This suggests that a caveat to this method is that there may be chromatin context at a SiRTA's endogenous location that is lost on the test site. Therefore, the chromosome 7 test site may be

used to identify SiRTAs, although the extent to which these sites are utilized in their endogenous locations may be underestimated or overestimated in some cases.

2.2.2 High throughput sequencing of pooled samples accurately measures de novo telomere addition

The creation of the chromosome VII test site was an important step in allowing for the efficient testing of putative SiRTA sequences. Even with this improvement in the testing method, the determination of SiRTA efficiency using the PCR method (Figure 2a and b) was the rate limiting step as multiple DNA extractions and PCR experiments are required. To feasibly test many sequences for their ability to stimulate *dn*TA, we sought to streamline the analysis of these sequences.

To increase the throughput of this analysis pipeline and to map *dn*TA events with nucleotide precision, we developed the Pool-Tel-seq (PT-seq) method (Chapter 4; *Materials and Methods*, and Figure 3a). We use our original method to identify thirty 5-FOA-resistant colonies. These clones are grown separately to saturation in liquid medium and equal volumes of each culture are pooled to generate a single genomic DNA sample for library construction and high throughput sequencing. The resulting sequence reads (>12 million) are analyzed for those that partially align to the 300 bp putative SiRTA and show evidence of telomere addition (TG_{1-3} or $C_{1-3}A$ sequence). To account for differences in read depth between experiments, the number of reads meeting these criteria is normalized to the number of reads mapping to a 300 bp sequence within *BRR6,* an essential gene on chromosome VII that lies centromere-proximal to the site at which the putative SiRTA is integrated (Figure 3a). At least two biological replicates are generated for each strain. To benchmark SiRTA efficiency based on our previous method, we

applied PCR-based mapping and PT-seq to multiple 30-colony samples derived from SiRTAs of

a range of efficiencies.
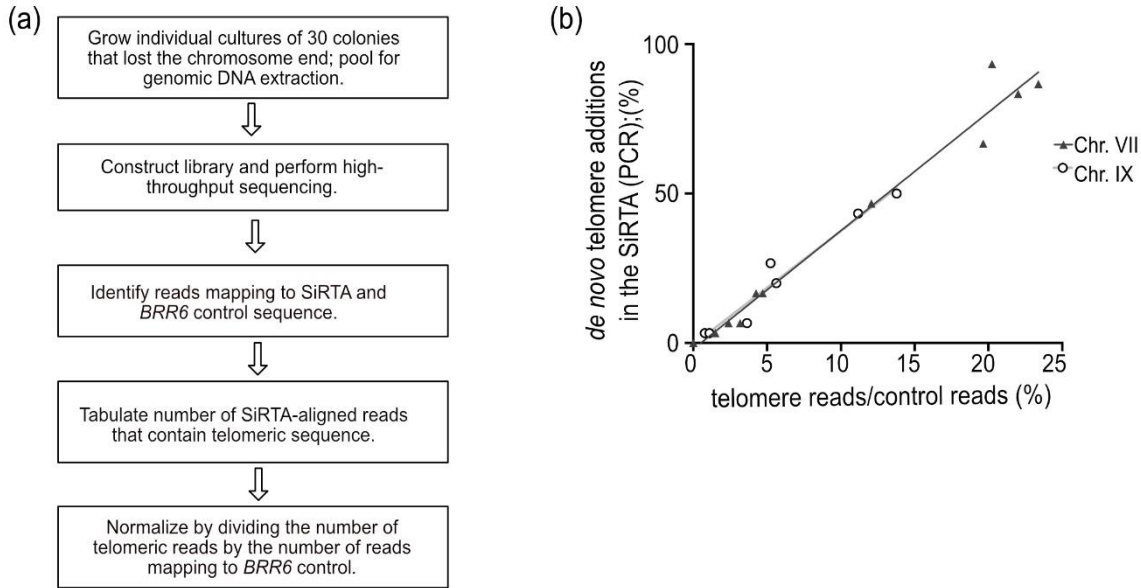
Figure 3. Validation of Pooled Telomere sequencing (PT-seq) as a method to quantify *de novo* telomere addition. a) Flow chart representing the PT-seq methodology. b) Correlation of the PCR-based and PT-seq methodologies (Table 3). Results on chromosome VII (black triangles; r^2=0.97) and chromosome IX (open circles; r^2=0.95) were analyzed by linear regression (lines are nearly superimposed). Line equations (VII: y=3.961x-2.039 and IX: y=3.8022x-0.5505) are statistically indistinguishable (p=0.94) by analysis of covariance.

Using linear regression, we find a strong correlation between the two methods (r^2=0.97), allowing us to use this standard curve to estimate the number of colonies within a 30-colony sample that underwent *dn*TA at the putative SiRTA (Figure 3b, closed triangles; Table 4 (see appendix A)).

This method also yields information about the relative frequency of telomere addition at each nucleotide position. To verify this method is applicable at other locations in the genome, we utilized PT-seq to test SiRTA 9L44(-) at its endogenous location on chromosome IX. *Cis*- and *trans*-acting mutations with effects on the efficiency of *dn*TA at SiRTA 9L44(-) were used to compare the PCR and PT-seq methodologies over a range of SiRTA efficiencies. The results of the two methods are strongly correlated (r^2=0.95; Figure 3b, open circles; Table 4 (see appendix A)). Slopes of the standard curves generated at both chromosome locations are statistically indistinguishable (p = 0.76 by analysis of covariance), suggesting that the percent of GCR events incurring *dn*TA (SiRTA efficiency) can be accurately estimated from PT-seq results regardless of chromosome location.

Once the methods for testing SiRTA efficiency were well established and it became clear we would be able to test many sequences for the ability to stimulate *dn*TA, we sought to define a minimal threshold of activity that a sequence must reach to be considered a SiRTA. The efficiency of SiRTA function is expressed as the percent of 5-FOA-resistant clones (from a total of 30) that contain a telomere-addition event within the sequence of interest at the insertion site on chromosome VII. Typically, average values of 2-3 biological replicates are reported. On chromosome VII, the 300 bp sequence analyzed represents ~1.4% of the 21,922 bp region within which a GCR event can be recovered [between the HO cleavage site and the first essential gene on VII-L (*BRR6*)]. To determine a threshold for SiRTA activity, we modeled the expectation for

random repair within this region (Figure 4). Assuming a random distribution of 30 GCR events in a single experiment, two or more GCR events would be expected to occur within the 300 bp test sequence in 6.2% of experiments, while three or more would be expected in 0.78% of experiments. Therefore, we chose to define a SiRTA as a sequence in which an average of greater than 2 out of 30 clones show evidence of *dn*TA within the 300 bp test sequence (SiRTA efficiency of 6.6%).

(a)
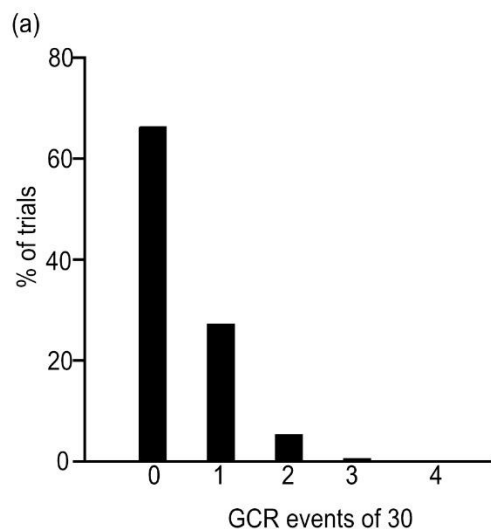
Figure 4. Modeling the random distribution of GCR events within a 300 bp sequence inserted on chromosome VII. Given a 21,922 bp region between the HO cleavage site and the first essential gene on VII-L (BRR6), the number of randomly distributed events (of 30) expected to occur in the 300 bp sequence was determined over 10,000 trials. See text for explanation.

2.3 Discussion

Together, these new methods allow for the efficient testing of putative SiRTAs for their ability to stimulate *dn*TA. Previous studies of SiRTAs have been limited to testing a small number of SiRTAs in the nonessential region of the genome (Obodo et al. 2016; Epum et al. 2020; Hoerr et al. 2023; Ngo et al. 2023). The labor and resources required to test SiRTAs were previously a limiting factor in the experimental workflow. Additionally, testing SiRTAs only at

their endogenous location resulted in a bias towards SiRTAs in the nonessential regions of the genome. The test site on chromosome VII addresses many of these limitations. Insertion of the sequence of interest using a validated CRISPR/Cas9 approach into a strain that already contains an HO cut site and the *URA3* marker streamlines the process of strain construction considerably. Furthermore, the test site allows any sequence to be tested regardless of location or orientation. Creation of the chromosome VII test site has also simplified the process of testing altered SiRTAs (e.g. SiRTAs lacking a Stim sequence or truncated/extended SiRTAs) and also allows for non-natural sequences to be tested (e.g. protein binding sites or perfect telomere repeats). The possibilities for what can be tested at this site are intriguing and could lead to many more exciting studies about SiRTA function and the mechanism behind SiRTA function. With that said, there are potential limitations regarding this method. There is a risk that we fail to include an important stimulatory sequence in choosing which piece of DNA to test. And there is also a possibility that chromatin context in the native location could influence SiRTA activity in a way that we will fail to capture in this assay. Additionally, we do not delete the endogenous sequence, so we do not fully understand how having this sequence of perfect homology may interfere with the assay. In the presence of *RAD52*, we do observe translocation events where the sequence on chromosome VII utilized the endogenous sequence to initiate repair, likely by break-induced replication. When testing 5L35(-) on chromosome VII, we observed 5 events out of 18 were translocations confirmed by Southern blot (unpublished observations). To address this last concern, we delete *RAD52* to prevent homology directed repair.

The PCR method for testing SiRTA efficiency was a limiting factor in our previous workflow and was not amenable to determining the precise location of telomere addition. To streamline the process of testing SiRTAs, we utilized Illumina sequencing of samples generated

from the pooling of 30 GCR events. This new strategy correlates strongly with the results from the PCR method and worked when testing on multiple chromosomes. Furthermore, the slopes for each chromosome tested were nearly identical suggesting that SiRTAs on other chromosomes can be tested without necessitating the generation of a new curve for each new chromosome. With that said, there are limitations of this strategy. One limitation is that this strategy does not easily facilitate detection of events other that dnTA (e.g. translocations) that have been observed in the past. In the future, we will use additional analysis to identify "split-reads" or will apply long read sequencing strategies to bypass this limitation. Additionally, because this strategy requires an initial alignment of reads to the putative SiRTA, it does not work well for highly repetitive sequences because a very large number of hits are returned and dnTA events are difficult to identify among this large number of partially aligned sequences. For sequences that are repetitive (for example, those found in X-elements) we revert back to the PCR method but we hope to develop a computational method to bypass this issue in the future. Despite these caveats, the development and confirmation of this new strategy has greatly increased the pace at which sequences can be tested for SiRTA function. These more efficient strategies for testing SiRTA activity are important tools for the continued study of SiRTAs.

Chapter 3[5]

A computational algorithm for telomere hotspot identification

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 (Pennaneach et al. 2006). We previously identified several sequences in the baker's yeast, *Saccharomyces cerevisiae,* that act as hotspots of *dn*TA (termed Sites of Repair-associated Telomere Addition or SiRTAs), but the distribution and functional relevance of SiRTAs is unclear. Here, I describe a computational algorithm that identifies SiRTA sequence motifs, allowing generation of the first comprehensive map of telomere-addition hotspots in yeast. In this chapter I will discuss the creation of this novel map and what we have learned about the distribution of SiRTAs from this map.

3.1 Introduction

Telomeres are nucleoprotein structures at the ends of most eukaryotic chromosomes that serve to protect chromosome ends, offer a solution to the end replication problem, and distinguish normal chromosome ends from ends created by DSBs (Blackburn 1991; Srinivas *et al.* 2020; Glousker and Lingner 2021). Telomerase elongates and maintains telomeres. To ensure proper telomere length and function, telomerase is regulated both positively and negatively. For example, the ssDNA binding protein Cdc13 recruits telomerase to telomeres, while the helicase Pif1 negatively regulates telomerase by removing it from DNA (Evans and Lundblad 1999; Li *et*

---

[5] This work is adapted from portions of the research article: Ngo K, Gittens TH, Gonzalez DI, Hatmaker EA, Plotkin S, Engle M, Friedman GA, Goldin M, Hoerr RE, Eichman BF, Rokas A, Benton, ML, and Friedman KL. 2023. A comprehensive map of hotspots of *de novo* telomere addition in Saccharomyces cerevisiae. *Genetics*, iyad076. https://doi.org/10.1093/genetics/iyad076

al. 2014). Instances where telomerase is not properly regulated and therefore improperly elongates telomeres have been implicated in cellular dysfunction and human disease. This work describes a type of telomerase error wherein, in response to a double-strand break, the 3' overhang created by that break is improperly recognized by telomerase and a *de novo telomere* is added instead of normal repair processes.

Hotspots of *dn*TA in response to an induced DSB were identified first on chromosome VII (Mangahas *et al.* 2001). Later, analysis of spontaneous truncations of chromosome V identified another hotspot of *dn*TA (Myung *et al.* 2001; Stellwagen *et al.* 2003; Pennaneach *et al.* 2006). Further characterization of this sequence on chromosome V and an additional hotspot on chromosome IX revealed a structure for these types of sequences. These sequences, termed Sites of Repair Associated Telomere Addition (SiRTAs), contain two TG-rich sequence tracts. The first tract (the Core) serves as the region where telomerase acts to add telomeres and the second tract (the Stim, located 5' to the Core on the TG-rich strand) is required for *dn*TA at the Core (Obodo *et al.* 2016). Previous work has demonstrated that artificial recruitment of Cdc13 to the Stim as a substitution for the normal Stim sequence is sufficient for SiRTA activity (Obodo *et al.* 2016). It has been hypothesized that SiRTA function requires Cdc13 association with the single-stranded DNA generated upon resection of a DSB. Association of Cdc13 with the Stim sequence facilitates the recruitment of telomerase, which in turn acts upon the Stim sequence to generate a *de novo* telomere. Ultimately, *de novo* telomere addition results in chromosome loss distal to site of telomerase action. With that said, the formation of a *de novo* telomere prevents additional resection from what would otherwise be an unrepaired break. Understanding the different factors that affect the process of *dn*TA will help elucidate whether SiRTAs are purely a source of genomic instability or if they offer utility in helping cells survive unrepaired breaks.

An important factor to consider when studying SiRTAs is the orientation of the SiRTA. When the TG-rich sequence of a SiRTA is on the same strand as the TG-rich 3' telomeric overhang on that chromosome arm, that SiRTA is said to be in the "TG-orientation." On the left arm of a chromosome, SiRTAs in this orientation are TG-rich on the bottom (3' to 5' or minus) strand while on the right arm, the TG-rich sequence is on the top (5' to 3' or plus) strand. Telomere addition at a SiRTA in the TG orientation requires a DSB distal to the SiRTA and stabilizes the centromere-containing side of the break. If the SiRTA is distal to all essential genes on that arm, the resulting terminal deletion is compatible with viability, even in a haploid strain. Interestingly, the Friedman lab recently described a SiRTA in the opposite or "CA-orientation" that promotes cell survival under sulfate-limiting conditions by facilitating the formation of an acentric fragment containing *SUL1*, encoding the primary sulfate transporter (Hoerr *et al.* 2023). Because this event is observed in diploid cells, the centromere-proximal side of the break may utilize the homolog for repair. Alternatively, *dn*TA may be initiated following a replication error that does not involve a DSB (Hoerr et al. 2023). Despite the identification of several additional SiRTAs (Ngo *et al.* 2020), an understanding of the genome-wide frequency and distribution of these sequences is lacking.

Here, we validate the use of a computational algorithm, the Computational Algorithm for Telomere Hotspot Identification (CATHI), to predict SiRTA function based on similarity with the TG_{1-3} pattern of the yeast telomeric repeat. As discussed in Chapter 2, we developed and validated a high-throughput sequencing method that dramatically increases the number of putative sequences that can be characterized while simultaneously yielding information about the site of telomerase action. Together, we use these approaches to determine the overall locations and orientations of SiRTAs on a genome-wide scale. All but one of the subtelomeric repetitive

regions (defined as X and Y' elements; Louis and Haber 1990, 1992) contain at least one SiRTA in the TG orientation. However, outside of the subtelomeric regions, there is no apparent bias in the location or orientation of predicted SiRTAs, although these sequences occur more frequently than expected by chance. SiRTA function correlates with the ability of a sequence to bind Cdc13, but overall binding affinity is insufficient to explain all variations in the frequency of *dn*TA. This work provides a foundation for developing a fuller understanding of how sites with a propensity to stimulate *dn*TA impact genomic stability and evolution.

3.2 Results
3.2.1 Putative SiRTAs are accurately identified using a computational method

Inspection of SiRTA sequences revealed similarity to yeast telomeres, consistent with prior work demonstrating that association of Cdc13 with the Stim sequence is required for *dn*TA (Obodo *et al.* 2016). The Core sequence is also TG-rich, likely reflecting the required complementarity to the *TLC1* template sequence and (perhaps) the ability to associate with Cdc13. We postulated that SiRTA function could be predicted by considering both the TG-richness of a sequence and its similarity to the pattern of the yeast telomeric repeat (TG_{1-3}). SiRTA function does not require a perfect match to the telomeric sequence, so we developed a strategy to score similarity to telomere sequence while allowing some divergence. The Computational Algorithm for Telomere Hotspot Identification (CATHI) identifies strings of consecutive Gs and Ts, awards one point for each base in that string, and subtracts 1.5 points for each instance of GGTGG, a sequence lacking in yeast telomeres (Figure 5a). Calculations are done in a sliding window that can be varied in size (see Chapter 4; *Materials and Methods*).

The algorithm was developed through an iterative process in which sequences were identified and tested for SiRTA function. This dataset included several previously published

SiRTAs, sequences identified in the work described here, and sequences that were not expected to function as SiRTAs. To standardize measurements of SiRTA efficiency, sequences were assayed on chromosome VII by inserting a 300 bp region encompassing the putative SiRTA. If boundaries of the SiRTA sequence were previously established, the SiRTA was centered within the 300 bp region. Sequences were tested at least twice by PT-seq (see Figure 3a). During initial testing, data were obtained for 37 sequences, seven of which had an average SiRTA efficiency above the 6.6% cutoff we previously established (Chapter 2). To optimize the algorithm, we calculated a score for each of the 37 sequences using varying window sizes (25 to 150 bp) and penalties (0 to 3) and chose the combination generating the best fit by linear regression (Table 5 (see appendix A)). A window size of 75 and a penalty of 1.5 for GGTGG sequences yielded the highest correlation between CATHI score and SiRTA efficiency ($r^2 = 0.63$ for all sequences and 0.69 for the seven sequences exceeding the 6.6% cutoff for SiRTA function).

Additional testing resulted in a final dataset of 47 sequences (13 active as SiRTAs), graphed relative to the CATHI score in Figure 5b (also see Table 3 (see appendix A)). A threshold CATHI score between 18.5 and 20 most effectively separates active and inactive sequences (Figure 4b) with a false positive rate of ~2% (1/47) and a false negative rate of ~4% (2/47). For those sequences with a CATHI score of 20 or greater, the score is moderately predictive of SiRTA efficiency ($r^2=0.43$; p=0.015). Sequences with the four highest CATHI scores are also the most efficient. Scores between 20 and 30 are less predictive of efficiency, suggesting that some aspects of SiRTA function are not captured by the algorithm (see Section 3.3; *Discussion*). Taken together, we conclude that the algorithm can be used to accurately identify sequences with a propensity to stimulate *dn*TA.

(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
[GTGGTGGTGTTTCTCCCTT**GGGT**TTGCCGAACAATGCCAT**GGGTGG**AGGAGGA**GGTGGTGGTGGTGG**AGGGGGAG]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)

(c)

Figure 5. Computational Algorithm for Telomere Hotspot Identification (CATHI) predicts SiRTA function. a) Summary of methodology used to generate CATHI score. An example corresponding to SiRTA 14R131(+) is shown (300 bp sequence beginning at chromosome XIV nucleotide 131308). Although multiple 75 bp windows within this sequence surpass the threshold score of 20, the calculation is shown only for the highest-scoring window, starting at nucleotide 131471 (bold, bracketed text). Underlined sequences correspond to strings of 4 or more guanine or thymine nucleotides conforming to the patterns described in the flowchart and in more detail in Chapter 4. Each underlined nucleotide is awarded one point. Subsequently, each occurrence of a GGTGG pentanucleotide incurs a 1.5-point penalty to generate the final score. b) Distribution of false positives and false negatives from the CATHI algorithm. The gray bars represent the sum of false positive and false negative sequences. The dotted black line represents false positives and the solid black line represents false negatives. c) Correlation of CATHI score and the percentage of GCR events that result from *de novo* telomere addition within the sequence of interest. Each value is the average of at least two experiments, each with 30 GCR events. The standard curve for chromosome VII (Figure 1d) is used to convert PT-seq values to the percentage of GCR events undergoing *dnTA* in the SiRTA. Horizontal dashed line indicates a minimum telomere-addition efficiency of 6.6% used to define an active SiRTA (see text for detail). Thirty-two of the sequences fall below this threshold and 15 are above this threshold. Vertical dashed line illustrates a CATHI score of 20 and a vertical gray bar indicates CATHI scores of 18.5 to 20 that effectively separate active and inactive sequences. Thirty-three sequences fall below this threshold and 14 are at or above this threshold (Table 5). Open circles are false negatives or false positives. Linear regression analysis on SiRTAs with a CATHI score of 20 or more yields a p-value of 0.01 (r^2 =0.43). This analysis excludes SiRTA 6R210(+) with a CATHI score >60, for which SiRTA efficiency underestimates the overall frequency of *dnTA* (see text).

3.2.2 Distribution of SiRTAs across the yeast genome

Using the algorithm parameters established above, the 16 chromosomes (excluding terminal TG_{1-3} telomeric sequences; see Table 6 for coordinates (see appendix A)) were scanned as a series of 75 bp sliding windows with a step size of 1. Given the lack of specific information for CATHI scores between 18.5 and 20 (Figure 5b), we chose to utilize the more stringent cut-off value (20) for our analysis. Overlapping windows with scores of 20 or greater were merged such that the starting and ending coordinates of a predicted SiRTA represent the maximum distance between the first and last window meeting the threshold value. The final score assigned to a set of overlapping windows is equivalent to the highest CATHI score in that set. The algorithm was separately applied to the top and bottom strands and strand information was retained. Overall, we

identified 728 sequences in the *S. cerevisiae* genome with a CATHI score of 20 or greater (Table 7 (see appendix A)).

The overall distribution of these 728 sequences within the 16 yeast chromosomes is shown in Figure 6. A diagram of the typical yeast chromosome is shown in Figure 6a. SiRTAs on the top strand (5' to 3'; 342) are blue and those on the bottom (3' to 5'; 386) strand are red (Figure 6b and c). The centromere of each chromosome is depicted as a black circle. The overall distribution between the two strands is not different from random (p=0.25 by chi-square test), but there is a minor but significant tendency for the SiRTAs to cluster on the same strand. This effect is quantified by measuring the number of times a SiRTA is found on the opposite strand from its neighbor (number of "strand switches"). We observe 322 strand switches across the genome, significantly fewer than the number expected by chance (351.8 ± 13.0; p=0.013; Figure 7a). This difference is driven almost entirely by the observation of fewer "singlet" SiRTAs compared to expectation (148 observed versus 187.7 ± 15.0 expected; p=0.0043; Figure 7b). In contrast, longer runs do not deviate significantly from expectation. We conclude that there is a minor tendency for SiRTAs to cluster on the same strand, but only with the nearest neighbor.

(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 and telomere

(b)

Distance from left end (kb)

(c)

Distance from left end (kb)

Distance from right end (kb)
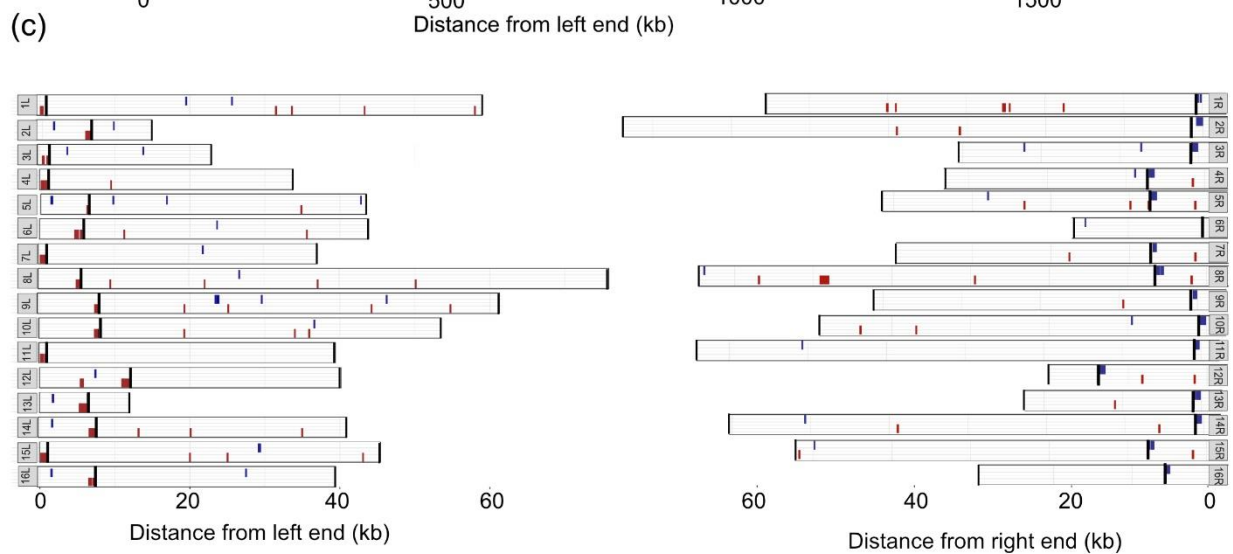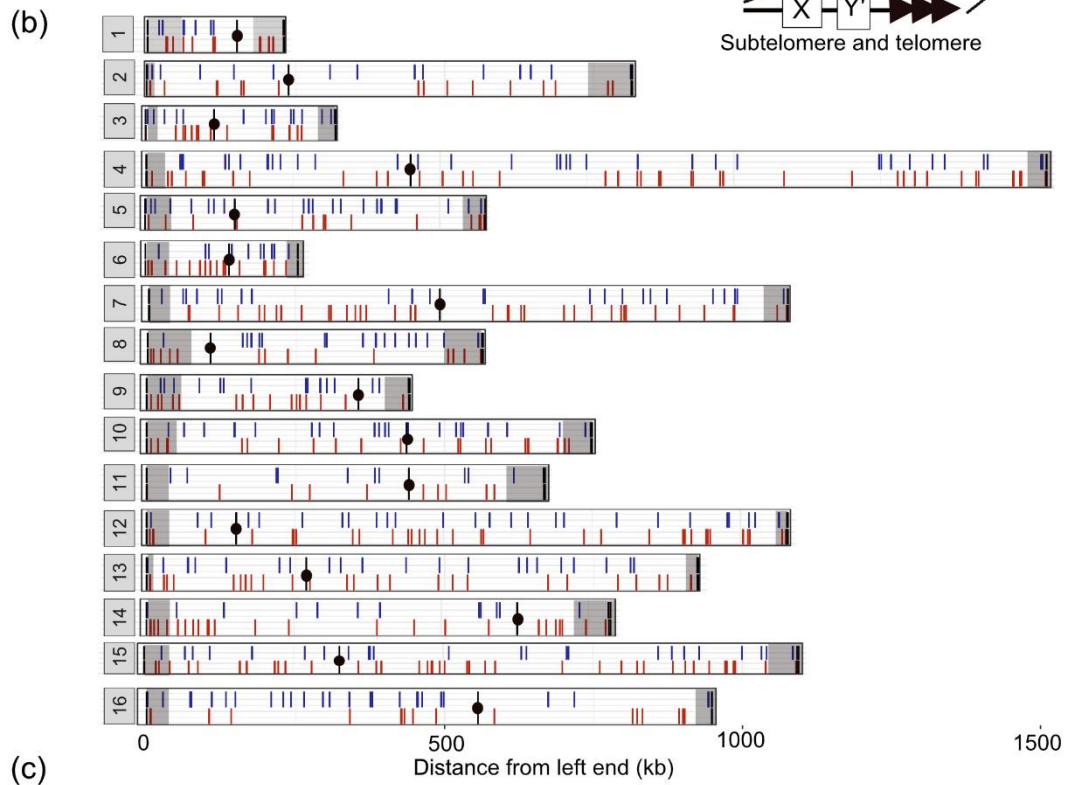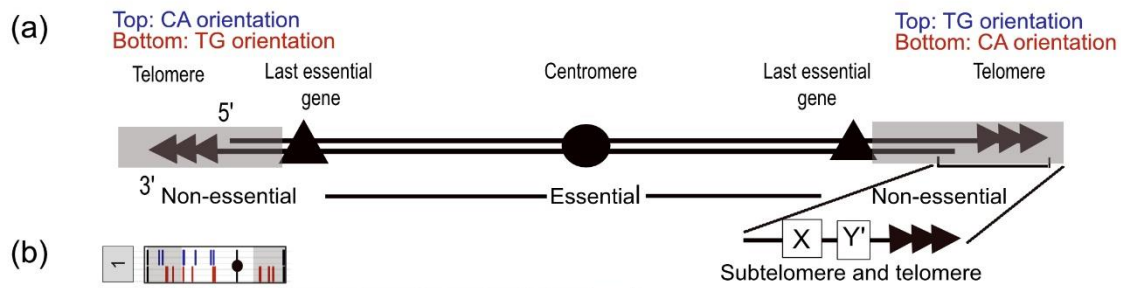
Figure 6. Summary of predicted SiRTAs across the S. cerevisiae genome. a) Diagram of chromosome landmarks. Predicted SiRTAs are listed in Table 7. Non-essential regions (grey boxes) are sequences located between the last essential gene (most distal gene on the chromosome arm that causes lethality in a haploid strain when deleted) and the telomere on each chromosome arm. Subtelomeres are located immediately adjacent to the telomeric repeats within the nonessential regions and are composed of a single complete or partial X element (all telomeres) and one or more Y' elements (a fraction of telomeres). Genomic coordinates are listed in Table 6 and Table 8. Location of the CA- or TG-oriented sequences are indicated for the left and right chromosome arms (see text). b) Distribution of SiRTAs on each of the 16 yeast chromosomes; distance in kilobases from thFigure 13. Association of Cdc13 DNA binding domain (Cdc13-DBD) with DNA substrates. a) Each fitted curve represents a separate experiment to detect binding of Cdc13-DBD to the labeled Tel11 sequence. Individual data points are the average of three technical replicates; error bars are standard deviation. b) Representative competition experiment conducted as described in Chapter 4. Tel11 is the labeled oligonucleotide. Results are shown for three different unlabeled competitor oligonucleotides (see Table 10 for sequences). Individual data points are the average of three technical replicates; error bars are standard deviation.

In a haploid cell, chromosome truncation proximal to the last essential gene on a chromosome arm will be lethal. Therefore, we examined whether the distribution of putative SiRTAs is different in essential versus nonessential regions. For our analysis, the nonessential region on each chromosome arm comprises sequences distal to the last essential gene (Figure 6a; Table 8 (see appendix A)). The last essential gene, in turn, is the most telomere-proximal gene for which single gene deletion was reported to cause lethality during the systematic knockout of each open reading frame in *S. cerevisiae* (Giaever *et al.* 2002). This definition does not account for synthetic lethality; some nonessential regions may be smaller than defined here if the combined loss of one or more genes results in cell death. In Figure 6b, nonessential regions are highlighted in gray; those same regions are shown in expanded form in Figure 6c. Nonessential regions are divided into unique sequences (in most cases) among the different chromosome arms and the highly repetitive subtelomeric X and Y' elements found immediately adjacent to the telomeric repeats (Figure 6a; Table 8(see appendix A)). All chromosome arms contain at least a portion of the X element while some also contain one or more Y' elements (~6 kb each; Louis and Haber 1990b; Zhu and Gustafsson 2009). The transition to subtelomeric sequence is shown

on each chromosome arm as a black line (Figure 6c). The diagrams in Figure 6 are based on the

published sequence of reference strain S288C. Recent long-read sequencing analyses confirm

that some subtelomeric regions contain additional terminal sequences (primarily Y' elements)

that were not included in the published reference genome (Yue *et al.* 2017), so our analysis likely

underestimates the number of subtelomeric SiRTAs.
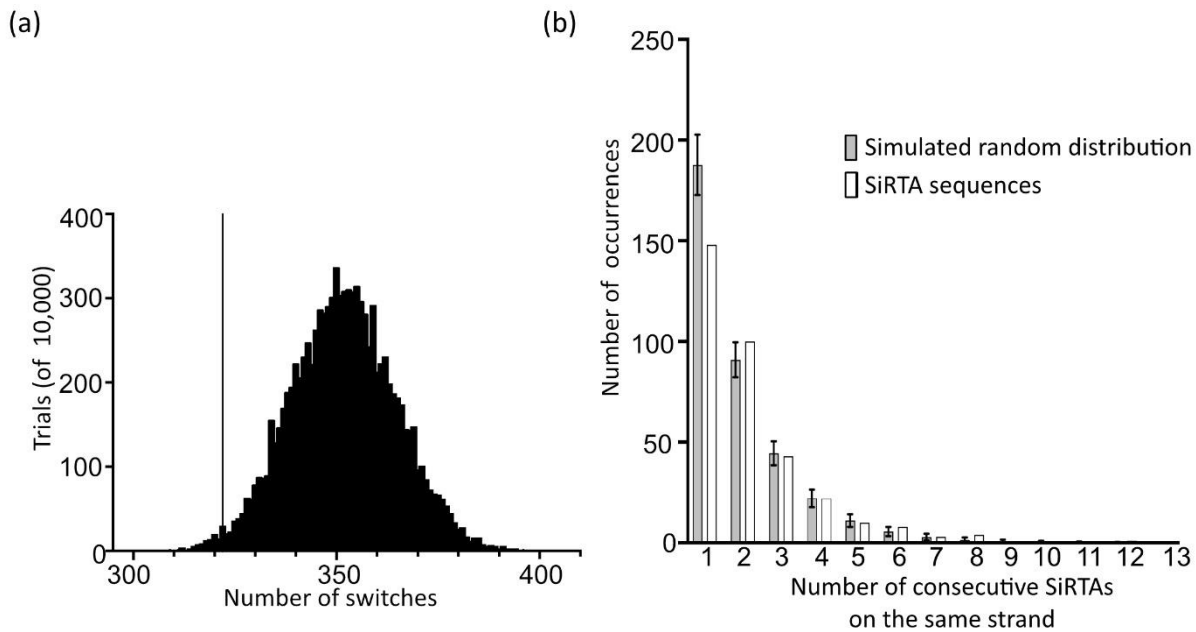
Figure 7. SiRTA distribution is mostly random, with a slight preference for neighboring SiRTAs to occur on the same strand. a) and b) Modeling the random strand distribution of SiRTAs. Using the number of predicted SiRTAs on the top and bottom strand of each chromosome (including subtelomeric SiRTAs; Table 7), 10,000 iterations were generated in which those SiRTAs were randomly distributed between strands. Distributions were analyzed in two ways. In (a), the number of times that neighboring SiRTAs are found on different strands (number of "strand switches") was calculated. The observed value of 322 (line) corresponds to a probability of 0.013. In (b), the number of adjacent SiRTAs on the same strand (the "run length") was tabulated and summed across all chromosomes for each of 10,000 iterations. Average and standard deviation is shown for the randomized trials (grey bars). The actual values are shown with the white bars. Only cases in which both flanking SiRTAs are on the opposite strand (runs of one) are statistically underrepresented relative to expectation ($p = 0.0043$).

To test the hypothesis that SiRTAs are preferentially located in nonessential terminal

regions, we utilized a permutation-based enrichment test to compare the distribution of predicted

SiRTAs between essential and non-essential regions to that of randomly shuffled sequences matched in number and length. This analysis shows significant enrichment for SiRTAs in the nonessential regions of the genome (p<0.01). However, enrichment disappears when the subtelomeric regions (X and Y' elements) are excluded from the analysis (Figure 8a). Seventy-five of the 728 putative SiRTAs lie within subtelomeric sequences. Nine of those 75 sequences consist of perfect telomeric (TG_{1-3}) repeats located between X and Y' elements and seven of these perfect repeats constitute the top-scoring sites in the genome (Figure 8b and Table 9(see appendix A)). The remaining predicted SiRTAs in subtelomeric regions are located within the X or Y' elements. Even when the nine perfect telomeric repeats are excluded, there remains significant enrichment for SiRTAs within the nonessential regions (p=0.001; Figure 8d). Notably, all chromosome ends [with one exception: 6R)] contain at least one region predicted to function as a SiRTA (Figure 6c). We conclude that SiRTAs are disproportionately found within the subtelomeric regions but are otherwise not significantly enriched within nonessential sequences.

(a)
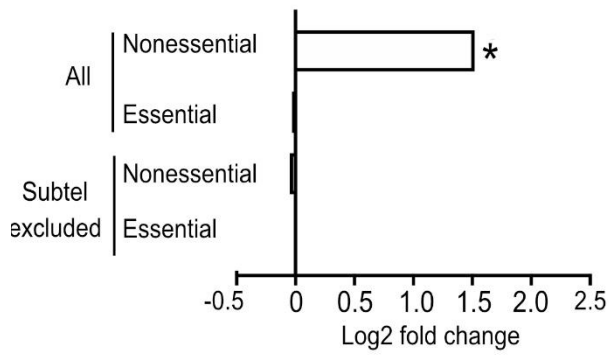
(b)
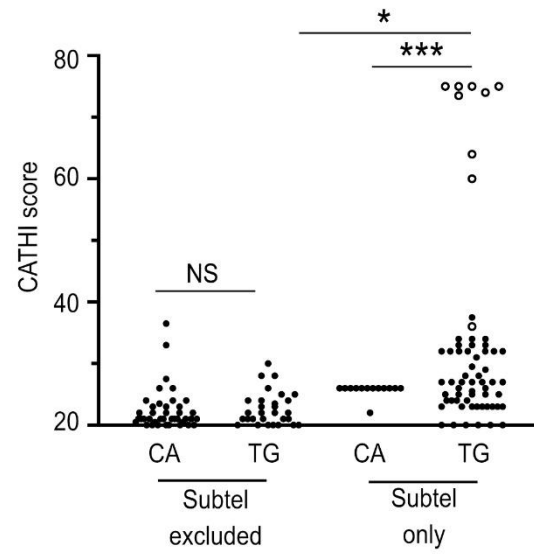
(c)
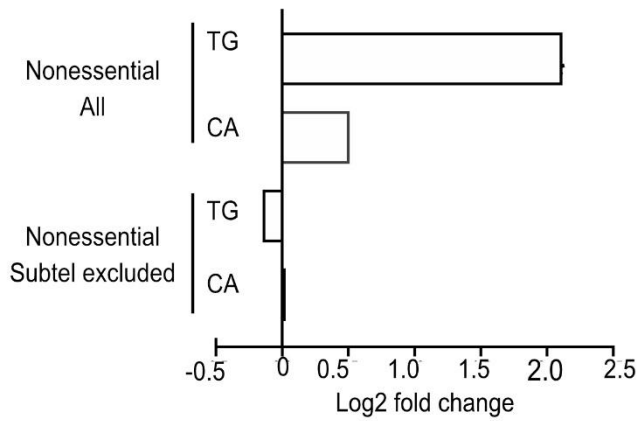
(d)
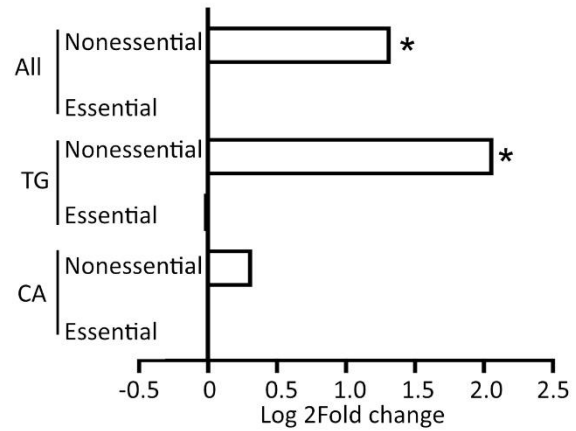
Figure 8. SiRTAs are enriched in subtelomeric regions. a) Using a permutation strategy (Chapter 4; *Materials and Methods*), enrichment of SiRTAs (Log_2-fold change) was determined for the nonessential and essential chromosome regions. Analysis utilized all genomic sequences (except the terminal telomeric repeats) or genomic sequences from which the subtelomeric regions were excluded, as indicated (coordinates in Table 8). *p-value <0.01 by chi-squared test with Bonferroni's correction. b) Distributions of CATHI scores (20 and greater) for putative SiRTAs in the TG- or CA-orientations. Analysis is presented separately for SiRTAs in nonessential regions (subtel excluded) versus those in the subtelomeric repeats (subtel only). Nine SiRTAs containing perfect telomeric (TG_{1-3}) repeats are indicated with open circles. Distributions of CATHI scores were compared by Kolmogorov-Smirnov test. The a value after Bonferroni's correction is 0.017. *p<0.017; ***p<0.0001; NS = not significant. Significant differences remain if the perfect telomeric repeats are excluded from the analysis. Coordinates of telomeric repeats are found in Table 9. c) Enrichment analysis was conducted separately for SiRTAs in the TG or CA orientation (see text and Figure 3a for definitions) as described in part (a). Results for the nonessential regions are shown. *p-value<0.01 by chi-squared test with Bonferroni's correction. d) SiRTAs are enriched in subtelomeric regions when excluding perfect telomeric repeats. This analysis was done as described in (a) and (c) except that nine subtelomeric SiRTAs containing perfect telomeric repeats (Table 9) were excluded. Analysis utilized all genomic sequences (except the terminal telomeric repeats). Essential and nonessential regions are defined in Table 8. *p-value <0.01 by chi-squared test with Bonferroni's correction.

Visual inspection of predicted SiRTAs (Figure 6) suggested that SiRTAs outside subtelomeric regions are randomly located along chromosome arms. To test this hypothesis, we modeled expected inter-SiRTA distances based on an assumption of random distribution (excluding subtelomeric regions and ignoring strand) as described in Chapter 4; *Materials and Methods*. As shown in Figure 9a, distances between the predicted SiRTAs match the expectation of random distribution. There is also no significant difference in the spatial distribution of predicted SiRTAs with the lowest and highest CATHI scores (Figure 9b and c).

(a)

(b)

(c)
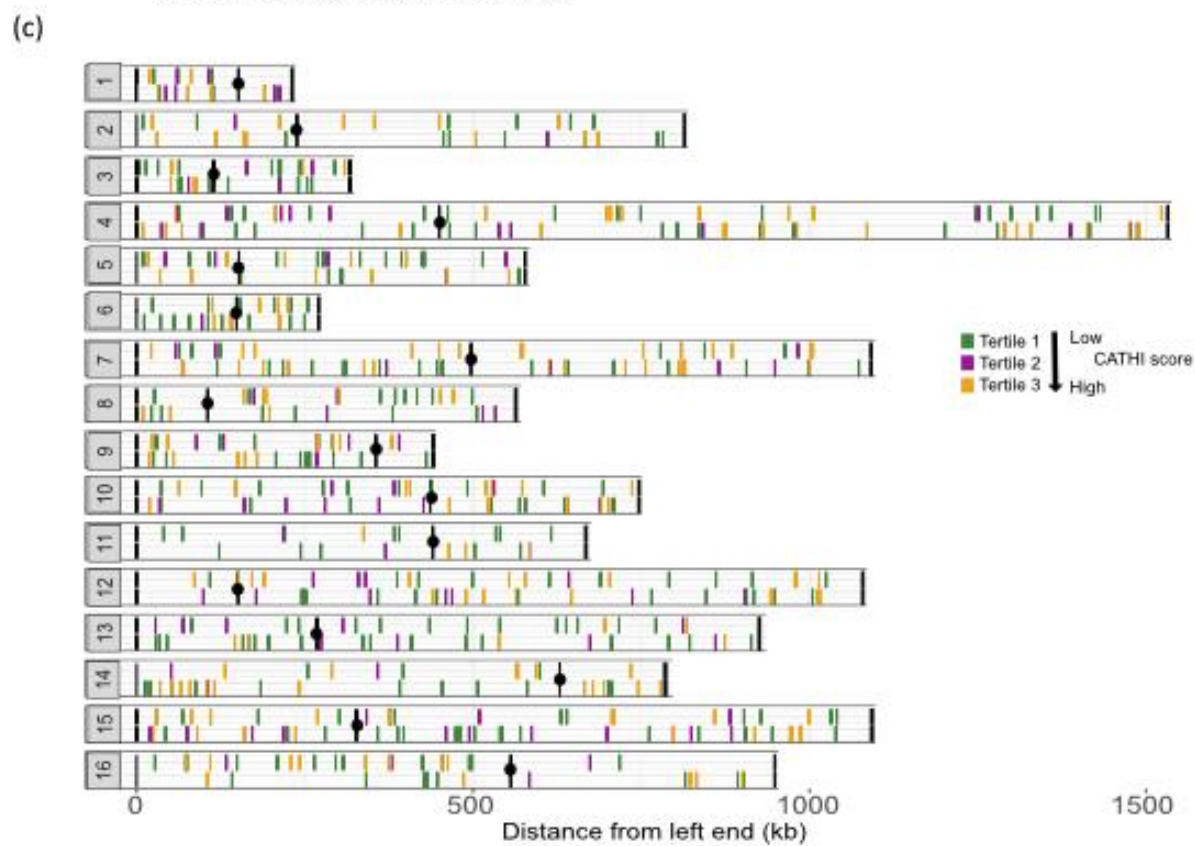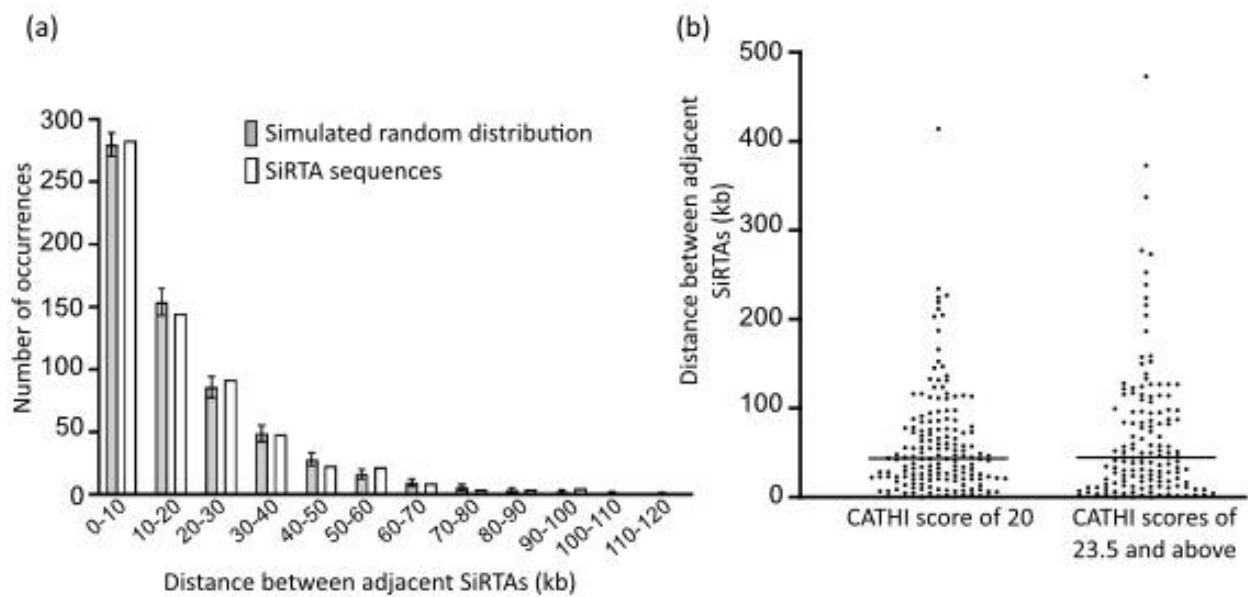
Figure 9. Analysis of the distance between adjacent SiRTAS. a) Modeling the random distribution of SiRTAs, measured as inter-SiRTA distance without regard to strand. Analysis excluded subtelomeric regions (Table 8). Using the number of predicted SiRTAs on each chromosome, 10,000 iterations were generated in which those SiRTAs were randomly distributed along the chromosome, inter-SiRTA distances were calculated, and binned in 10kb intervals. Average and standard deviation is shown for the randomized trials (grey bars). The actual values are shown with the white bars. b) Inter-SiRTA distance is not significantly different among predicted SiRTAs with the lowest and highest CATHI scores. Left: 178 predicted SiRTAs with a score of 20. Right: 157 highest CATHI scores (23.5 and up). Distributions were compared by Kolmogorov-Smirnov test (p=0.36). c) Genome-wide distribution of SiRTAs on non-subtelomeric sequences. Distance in kilobases from the left telomere is indicated at the bottom of the figure and corresponds to coordinates in the S288C reference genome. Black circles mark the centromere position. Vertical lines in the top half of each bar represents SiRTAs on the top (plus) strand and lines in the bottom half of each bar refer to SiRTAs on the bottom (minus) strand. SiRTAs are color-coded by CATHI score (see Chapter 4; *Materials and Methods*) with tertile 1 (green) representing the lowest subset of scores, tertile 2 (purple) representing intermediate scores and tertile 3 (orange) representing the highest subset of scores. Diagram generated using shinyChromosome (Yu *et al.* 2019).

As described in the Section 3.1 (*Introduction*), the strand on which a SiRTA is located has important implications for the consequence of *dn*TA. SiRTAs in the TG orientation (those oriented to stabilize the centromere-containing fragment when a break occurs distal to the site) are on the bottom strand for the left arm of a chromosome and on the top strand for the right arm. In the genome, predicted SiRTAs are not biased for the TG- versus CA-orientation (372 versus 356; p=0.675 by chi-square test). However, nonessential regions are enriched for SiRTAs in the TG orientation (p<0.01), but not the CA-orientation (Figure 8c). Excluding the nine perfect TG_{1-3} repeats does not alter the result (p=0.001; Figure 8d) but enrichment is no longer observed when all subtelomeric sequences are excluded (Figure 8c).

When the CATHI scores of individual putative SiRTAs are separated by CA or TG orientation and plotted, there is a striking difference in the strand bias and score distribution when comparing the sub-telomeric and non-subtelomeric regions (Figure 8b). In non-subtelomeric regions, there is no bias for the TG versus CA strand (p=0.18 by chi-square test) and the distributions of CATHI scores are indistinguishable. In contrast, SiRTAs in the X and Y'

elements are much more likely to be in the TG orientation than in the CA-orientation (p<0.0001 by chi-square test) and the distributions of CATHI scores are significantly different, even when perfect telomeric repeats are excluded. Among putative TG-orientated SiRTAs, the distribution of CATHI scores in the subtelomeric repeats is significantly different from that in non-subtelomeric regions (Figure 8b) regardless of whether the perfect telomeric repeats are included in the analysis. Striking enrichment of TG-oriented SiRTAs within subtelomeres is also apparent in the clustering of red symbols at or near the subtelomere junction on all left arms and of blue symbols on nearly all right arms (except 6R; Figure 6c).

We tested the ability of three sequences identified in subtelomeric regions to stimulate *dn*TA using the HO cleavage assay. Two of these sequences overlap with an X element [14L07(-) and 15R1084(+)] and one overlaps with a Y' element [7R1089(-)]. All three sequences function as SiRTAs, stimulating *dn*TA at frequencies of 10.0%, 13.3%, and 36.6%, respectively (Figure 10). Although 7R1089(-) functions well as a SiRTA in our assay, it is worth noting that it is in the CA orientation. Because it is part of a conserved Y' sequence, a very similar sequence occurs at multiple chromosome ends (also in the CA orientation). The X-element sequences are TG-oriented and are found in two distinct regions of the X-element on multiple chromosomes. Taken together, these results support the interesting possibility that sequences capable of functioning as SiRTAs have been retained near chromosome termini to facilitate chromosome rescue in the event of telomere loss.
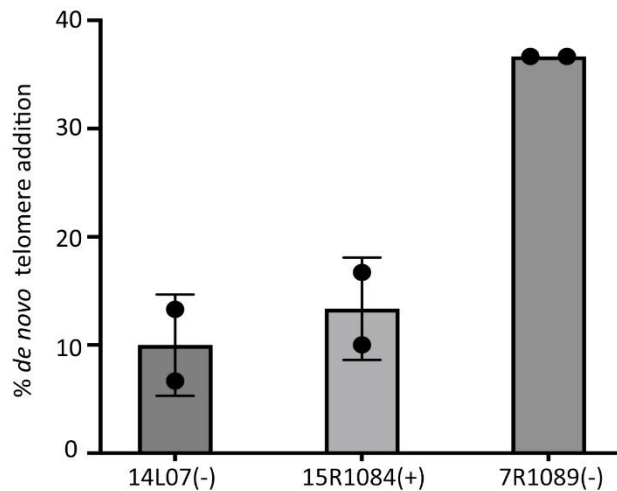
Figure 10. Representative predicted SiRTAs in the X and Y' elements stimulate *de novo* telomere addition. The indicated SiRTAs (300 bp) were integrated at the test site on chromosome VII. Each data point represents one PCR experiment in which 30 GCR events were analyzed as described in Chapter 4 (Materials and Methods). Average and standard deviation are shown. SiRTAs 14L07(-) and 15R1084(+) are contained within X elements while 7R1089(-) is found within a Y' element. These data are summarized in Table 3.

3.2.3 TG-rich sequences identified by the algorithm are overrepresented in the yeast genome

To address whether sequences predicted to stimulate *dn*TA are found in yeast at the expected frequency, we generated five scrambled genomes identical in sequence composition to the yeast genome. To avoid potential biases introduced by the subtelomeric regions, this analysis was done on sequences from which the subtelomeric X and Y' elements were excluded (see Chapter 4; *Materials and Methods*). The scrambled genomes contain, on average, 283.2 ± 18.3 sequences that score 20 or higher compared to 653 sequences observed in the yeast genome, an excess of 2.3-fold. The differential is increasingly 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 of 25. Among the scrambled genomes, an average of fewer than one sequence has a score of 30 or higher (range 0 to 2), while 21 sequences scoring 30 or higher are observed in the *S. cerevisiae* genome (Figure 11a). In Figure 11b, putative SiRTAs with scores of 25 or higher are shown to emphasize the strikingly

different distributions in the simulated versus actual genomes. To address whether the excess of

higher scores might be related to SiRTA function, we examined the predicted and actual

occurrence of CATHI scores below 20, which are unlikely to stimulate increased levels of *dn*TA

(see Figure 5b). For sequences with CATHI scores of 15-19, we still observe an excess in the

actual genome, although the excess is less pronounced (1.3-fold, with 2703 ± 16.3 predicted

compared to 3485 observed; Figure 11c). We conclude that TG-rich sequences matching the

pattern detected by the CATHI algorithm are significantly overrepresented in the yeast genome

compared to expectation, suggesting that sequences detected by the algorithm may provide a

selective advantage (see *Discussion*).
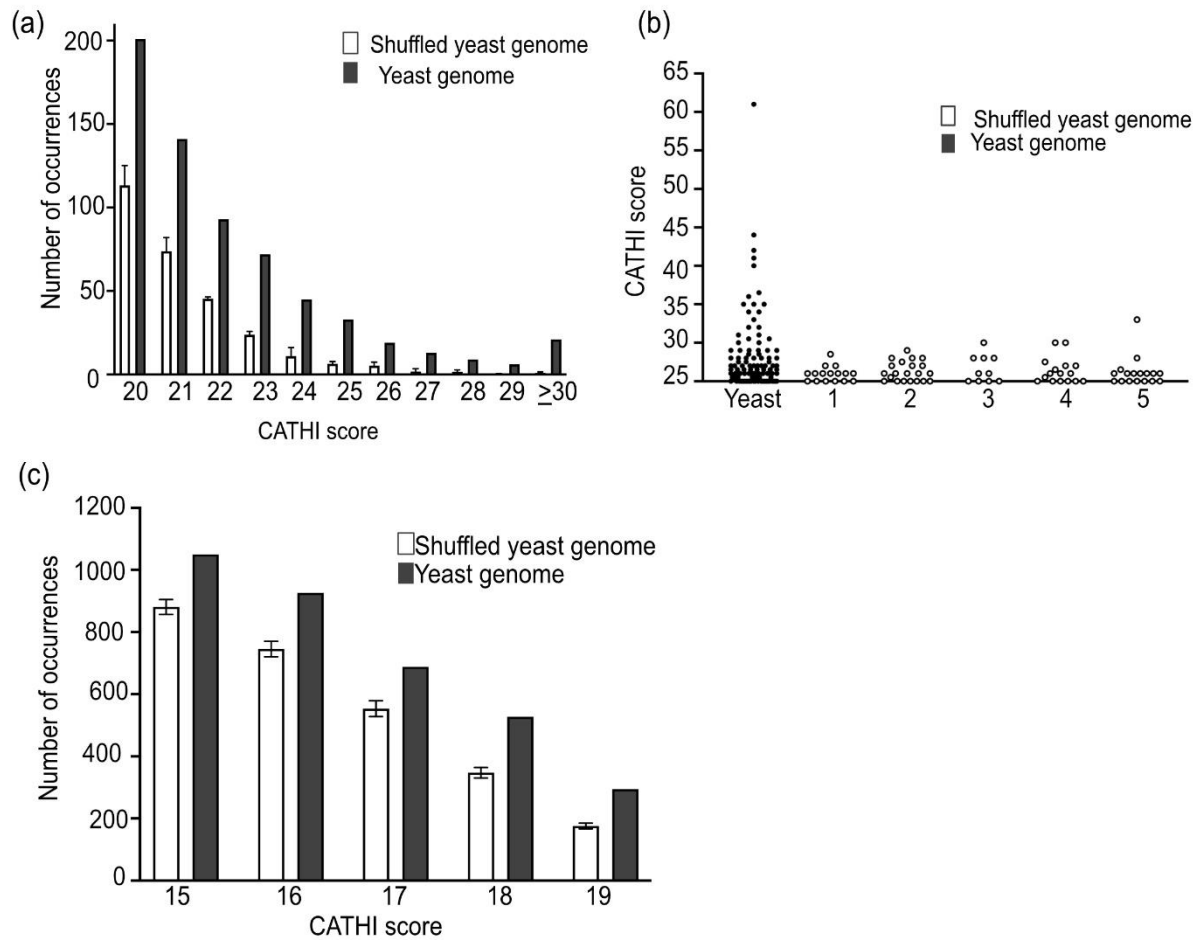
Figure 11. CATHI scores are significantly elevated in the S. cerevisiae genome relative to expectation. a) As described in Chapter 4, the algorithm was applied to the S. cerevisiae genome (excluding subtelomeric regions) and the number of sequences at each score (15 or higher; rounded down to the nearest integer) was graphed (solid bars). Genomic sequences (excluding subtelomeric regions) were scrambled five times and the identical procedure was applied. Data are presented as the average and standard deviation of the five trials (open bars). b) Distribution of CATHI scores of 25 and above in the S. cerevisiae genome (closed circles) and shuffled genomes (open circles). Subtelomeric sequences were excluded. c) As in (b), but data are shown for CATHI scores ranging from 15-19.

3.2.4 TG-dinucleotide repeats stimulate dnTA and are among the strongest SiRTAs in the genome

In analyzing predicted sites of *dn*TA, our attention was particularly drawn to SiRTA

6R210(+). With a score of 61, this sequence represents the strongest predicted site outside

subtelomeric regions (outlier in Figure 11b). SiRTA 6R210(+) contains a nearly perfect 62

nucleotide TG-dinucleotide repeat and is the longest TG-dinucleotide repeat in the *S. cerevisiae*

genome (the next longest is 41 nt; Table 9(see appendix A)). This sequence is in the TG

orientation but lies centromere-proximal to the last essential gene on the left arm of chromosome

VI, implying that repair by *dn*TA at this site in a haploid cell would be lethal.

To determine the efficiency of *dn*TA at SiRTA 6R210(+), we inserted the TG-

dinucleotide repeat (centered within a 300 bp region) at the test site on chromosome VII. Most

strains monitored for SiRTA function on chromosome VII generate GCR events at a frequency

of ~0.001%, equivalent to a negative control strain lacking a SiRTA. In contrast, a strain

containing the 62 bp TG-dinucleotide repeat generates 5-FOA resistant colonies at a 10-fold

higher frequency of ~0.01% (Figure 12a). By PT-seq, 86% of GCR events involve *dn*TA

addition within the inserted sequence (Figure 12b). Therefore, although SiRTA efficiency

(defined as the percentage of GCR events in which telomere addition occurred within the

sequence of interest) appears similar at the TG-dinucleotide repeat compared to SiRTA 14L35(-)

(86% versus 76%, respectively), the actual frequency with which *dn*TA occurs at the 62-nt

dinucleotide repeat is at least 10 times higher. We conclude that SiRTA efficiency alone

underestimates the propensity of this long TG-dinucleotide repeat to stimulate *dn*TA.

Using the PT-seq results, we mapped the sites at which *dn*TA occurred at SiRTA

6R210(+) (Figure 12c). Each arrow corresponds to the last nucleotide that aligns between the

chromosome and at least one PT-seq read, representing the 3'-most nucleotide at which

telomerase may have initiated synthesis. These results correspond to the mapping of 50-52

independent telomere addition events (86% of the 60 pooled strains). Sites identified in a larger

fraction of reads were likely targeted by telomerase in multiple independent clones. Consistent

with our prior observation that a 5' Stim sequence is required to stimulate telomere addition in a 3' Core sequence (polarity relative to the TG-rich strand), the vast majority of telomere addition events occur in the 3' half of the dinucleotide repeat or in sequences immediately downstream. Two events occurred at least 50 bases downstream of the TG-repeat, consistent with prior reports that TG-rich sequences can stimulate *dn*TA within neighboring sequences (Kramer and Haber 1993; Mangahas *et al.* 2001).

Excluding subtelomeric regions, TG-dinucleotide repeats comprise 11 of the 21 CATHI scores of 30 or greater (Figure 12d). A 34-nt perfect TG repeat [SiRTA 7L69(-); CATHI score = 34] was identified by the Zakian laboratory as a site capable of stimulating *dn*TA in response to a DSB induced more than 50kb distal to the eventual site of telomere addition (Mangahas *et al.* 2001). When integrated at the test site on chromosome VII, this 34-nt repeat stimulates *dn*TA with an efficiency of 47.9% by PT-seq (Table 3(see appendix A)). Together, these observations focus attention on TG-dinucleotide repeats as potential mediators of genome instability.
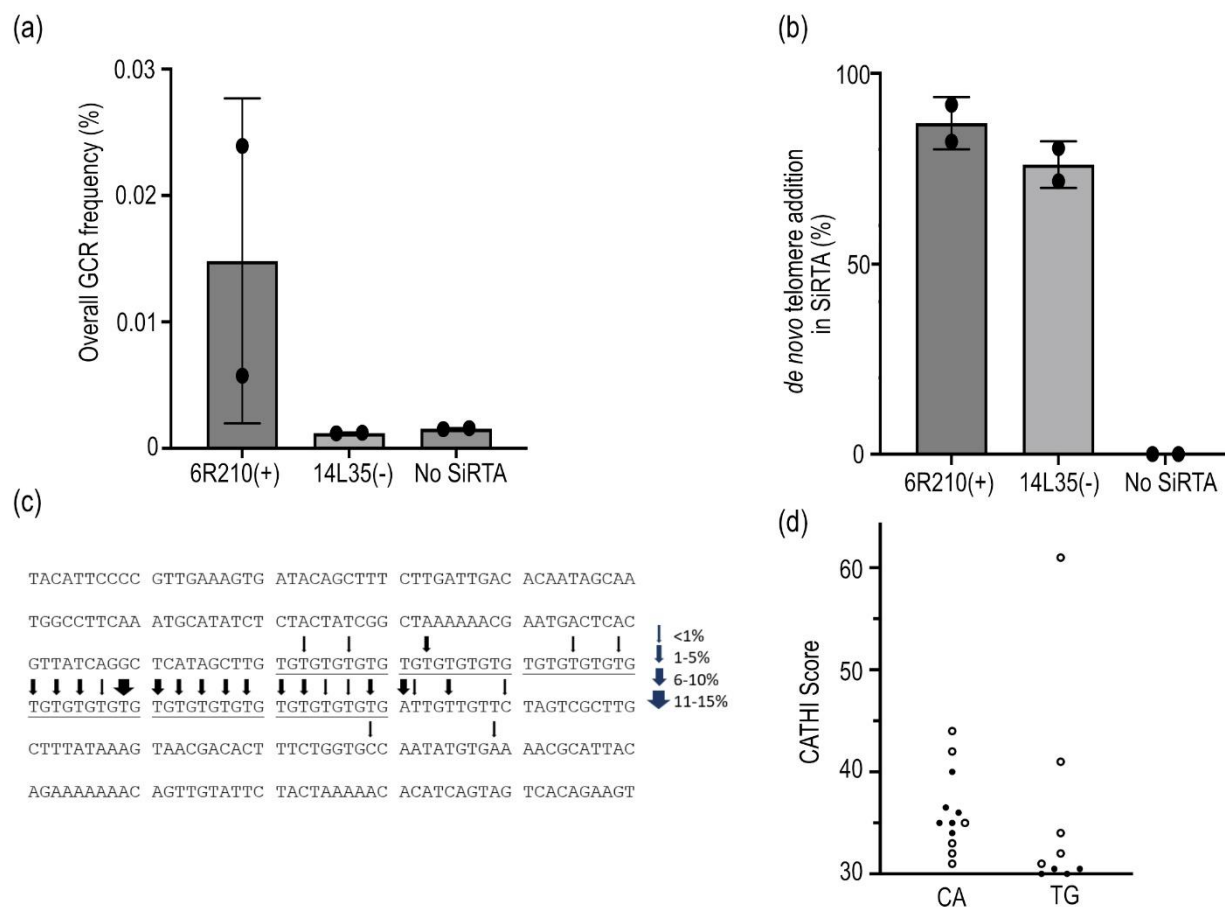
Figure 12. A 62 bp TG-dinucleotide repeat [SiRTA 6R210(+)] supports high levels of *dnTA*. a) Strains in which a 300 bp sequence encompassing SiRTA 6R210(+) or 14L35(-) was integrated on chromosome VII were subjected to the HO-cleavage assay as described in Chapter 4. The percent of cells that survived on galactose-containing medium and acquired 5-FOA resistance [cells containing a gross chromosomal rearrangement (GCR)] is shown. A strain lacking any insertion (No SiRTA) was utilized as a control. Error bars are standard deviation. b) The percent of GCR events that involve *de novo* telomere addition in the sequence of interest was determined by PT-seq for each strain described in (a). Each data point was generated by analysis of 30 GCR events. Average and standard deviation are shown. c) The 300 bp sequence encompassing 6R210(+) is shown. Sequence reads generated by PT-seq from a total of 60 GCR events [corresponding to the experiments shown in (b)] were filtered for those containing evidence of *de novo* telomere addition within the 300 bp sequence. Sites at which *de novo* telomere addition was observed are indicated (arrows). Arrow width indicates the percent of telomere-containing reads that map to that particular site. d) CATHI scores are shown for all non-subtelomeric SiRTAs with scores of 30 or higher, separated by CA- or TG-orientation. Open circles correspond to SiRTAs containing TG-dinucleotide repeats (also listed in Table 9).

3.2.5 Sequences that function to stimulate de novo *telomere addition bind Cdc13 in vitro*

Previous studies demonstrated that Cdc13 binding at the Stim sequence is required to promote *dn*TA. We hypothesized that sequences with CATHI scores of 20 or more will bind Cdc13 with greater affinity than sequences with lower scores. Additionally, we predicted that the two sequences identified as false negatives in our initial analysis (Figure 5b) should bind Cdc13 with higher affinity than the single sequence identified as a false positive. To test these predictions, we utilized fluorescence polarization to measure the ability of unlabeled, 75-base oligonucleotides to reduce the association of the Cdc13 DNA binding domain (Cdc13-DBD) with a 6-carboxyfluorescein (FAM) labeled 11-mer containing the canonical Cdc13 binding site (5'-GTGTGGGTGTG; referred to here as Tel11). Cdc13-DBD binds to Tel 11 with similar sequence specificity and affinity as full-length Cdc13 (Lewis *et al.* 2014). The goal of these analyses was not to identify individual Cdc13 binding sites but rather to measure the relative, cumulative ability of each sequence to bind Cdc13.

We first established that the FAM-labeled Tel11 oligonucleotide binds Cdc13-DBD (Figure 13a) and selected concentrations of 30 nM Cdc13-DBD and 25 nM labeled Tel11 for the competition analyses. The apparent inhibition constant (Ki,app) is the concentration of each competitor required to reduce binding to the FAM-labeled Tel11 by half. An unlabeled 75-mer containing the Tel11 sequence at the center of the oligonucleotide (Tel11-75) was included in each experiment and normalized Ki,app values are reported as fold change relative to this control (Ki,app of Tel11-75/Ki,app of experimental oligonucleotide). Sequences flanking the Cdc13 consensus binding site in Tel11-75 lack any TG or GG motifs to minimize additional association of Cdc13-DBD. Oligonucleotides used in these assays are found in Table 10 (see appendix A) and representative competition curves are shown in Figure 13b. To validate the method, we

64

determined the normalized Ki,app of a double-stranded version of Tel11-75 (0.5 +/- 0.3) and the inverse complement of the Tel11-75 sequence (0.4 +/-0.2), both of which show the expected reduction in binding relative to Tel11-75 (Figure 14a). A 75-mer sequence from chromosome VI previously shown to lack SiRTA activity (CATHI score=5) also competes very weakly for Cdc13-DBD association (normalized Ki,app = 0.5 +/- 0.2; Figure 14a). Finally, as expected, a 2xTel11-75 sequence that contains two adjacent Tel11 sequences competes twice as well as the Tel11-75 control oligonucleotide (normalized Ki,app = 2.2 +/- 0.2; Figure 15a).
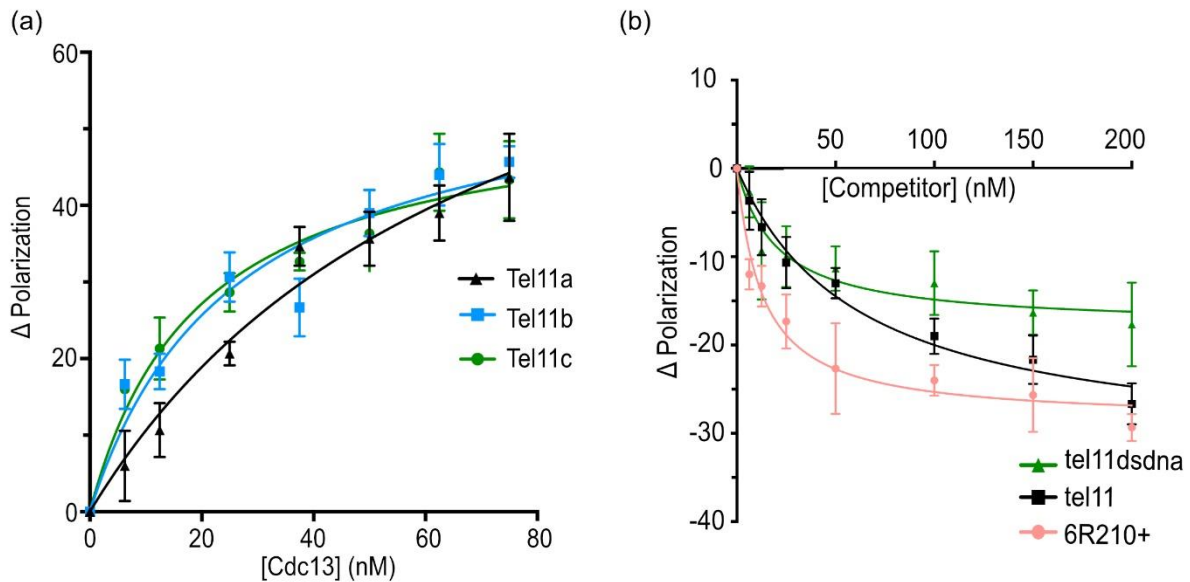
Figure 13. Association of Cdc13 DNA binding domain (Cdc13-DBD) with DNA substrates. a) Each fitted curve represents a separate experiment to detect binding of Cdc13-DBD to the labeled Tel11 sequence. Individual data points are the average of three technical replicates; error bars are standard deviation. b) Representative competition experiment conducted as described in Chapter 4. Tel11 is the labeled oligonucleotide. Results are shown for three different unlabeled competitor oligonucleotides (see Table 10 for sequences). Individual data points are the average of three technical replicates; error bars are standard deviation.

We tested several sequences with CATHI scores over 20 that were previously shown to stimulate *dn*TA. Oligonucleotides were designed to correspond to the 75 bases with the highest

CATHI score within the 300bp sequence tested for SiRTA function. Both 14L35(-) and 14R131(+) compete in a manner indistinguishable from the Tel11-75 control sequence (normalized Ki,app of 1.0 +/- 0.3 and 0.9 +/-0.2, respectively) and bind more robustly than the negative control sequences (Figure 14a). The two sequences identified as false negatives in Figure 2b [2R780(-) and 14R306(+)] both compete more effectively than the Tel11-75 control sequence (normalized Ki,app of 1.6 +/- 0.9 and 1.4 +/- 0.6, respectively; Figure 14a). This observation is consistent with the ability of these sequences to stimulate *dn*TA and suggests that the algorithm fails to predict Cdc13 binding in some cases. We also tested the false positive sequence [12R330(+)] with a CATHI score of 22 and an average *dn*TA frequency of 2.3% (below our cut-off for SiRTA function). This 75-base sequence has a normalized Ki,app of 0.7 +/- 0.2, intermediate to that of the Tel11-75 control sequence and the negative controls (Figure 14a).

Given the extremely high SiRTA activity of the 62-nt TG-dinucleotide repeat described above, we tested the ability of a 75-mer containing this repeat to compete for Cdc13-DBD binding. The normalized Ki,app of 4.9 +/- 2.9 measured for this sequence is considerably higher than any other sequence tested (Figure 14b). The first, second, and fourth base of the canonical Cdc13 binding site (5'-**GTGT**GGGTGTG) contributes most strongly to Cdc13 affinity, comprising a GxGT motif that recurs in the TG-dinucleotide motif (Anderson *et al.* 2003; Eldridge *et al.* 2006; Lewis *et al.* 2014). The 62-nt dinucleotide repeat is predicted to accommodate approximately five 11-mer binding sites, remarkably close to the observed 4.9-fold increase in competition compared to the Tel11-75 control oligonucleotide with a single binding site.

Our prior analysis of SiRTA 2R780(-) presented an additional opportunity to test the correlation between Cdc13 binding and SiRTA efficiency (Hoerr et al. 2023). Mutation of either one of two GxGT motifs within the Stim sequence of SiRTA 2R780(-) greatly diminishes SiRTA function, an effect that we attributed to reduced Cdc13 association (Hoerr et al. 2023). Consistent with this hypothesis, we find that mutation of one or both motifs significantly reduces the ability of the oligonucleotide to compete for Cdc13-DBD binding (Figure 14c).

The experiments described above provide evidence that sequences capable of stimulating *dn*TA associate more robustly with Cdc13 than sequences that do not function as SiRTAs, although our ability to distinguish borderline cases is limited. While there appears to be a threshold of binding required for SiRTA function, the cumulative "affinity" of a sequence measured in this assay is not fully predictive of SiRTA efficiency [e.g. 14L35(-) and 14R131(+) compete equivalently, but differ by a factor of two in SiRTA efficiency; 80.5% versus 30.7%]. This discrepancy may result in part from our choice of 75-mer sequence to test in each case, but also likely reflects the specific number, affinity, and distribution of Cdc13 binding sites within the sequence.
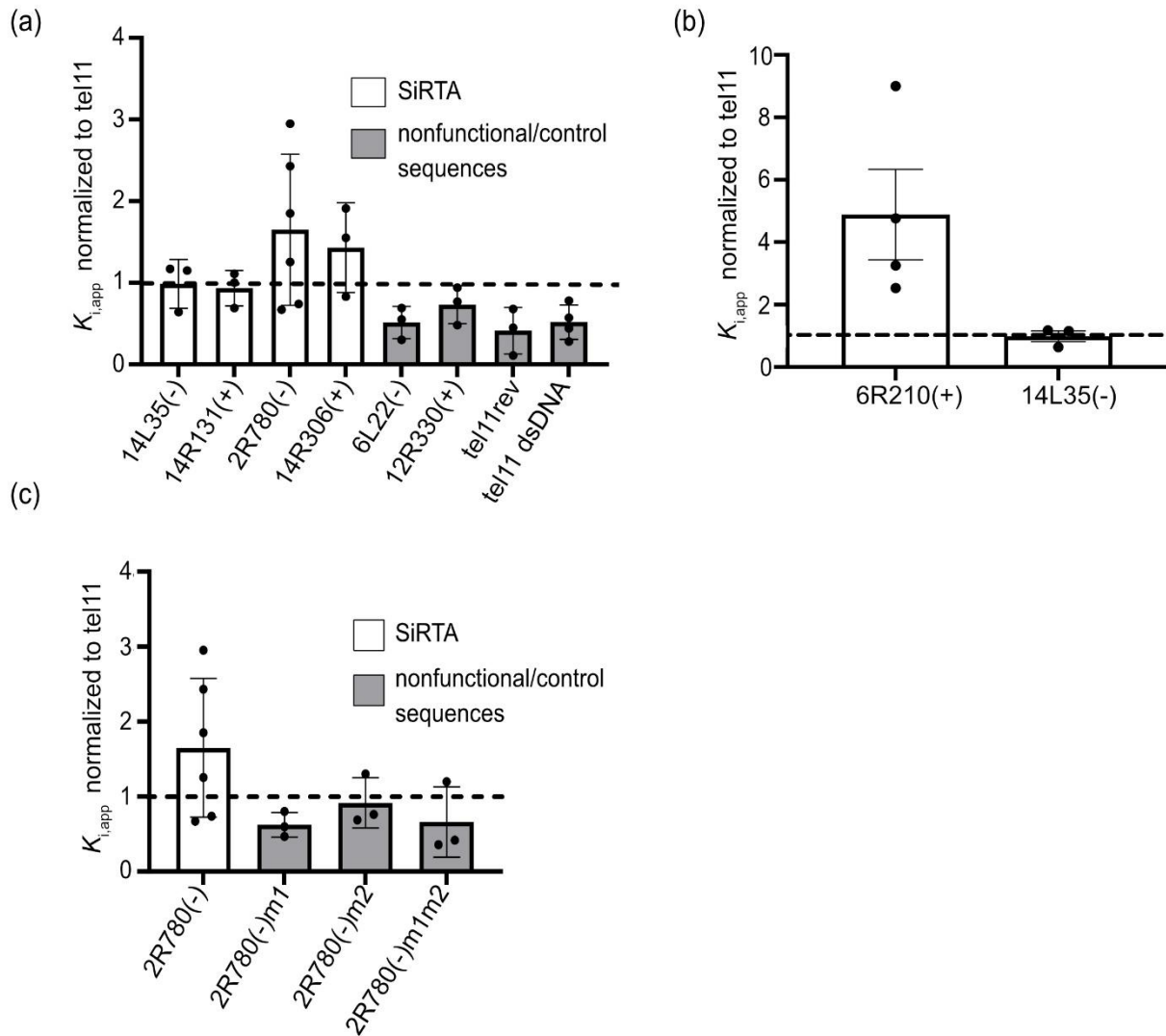
Figure 14. Sequences that function as SiRTAs bind Cdc13 *in vitro*. a) A competition fluorescence polarization assay was utilized to measure the relative association of the Cdc13 DNA binding domain with the indicated sequences (Table 10). Relative $K_{i,app}$ was determined as described in Chapter 4. Each point represents an independent measurement; error bars are standard deviation. The dotted line indicates normalization of values to the $K_{i,app}$ of a tel11-75 oligonucleotide included in each experiment. b) Same as in (a). $K_{i,app}$ of the TG-dinucleotide repeat analyzed in Figure 6 is shown [6R210(+)]. Data for 14L35(-) are repeated from (a) for comparison. c) Same as in (a). 2R780(-) and its mutated variants are described in Hoerr *et al.* (2023).

As another approach to benchmark the effect of high-affinity Cdc13 binding on SiRTA function, we tested the ability of either a single canonical Cdc13 binding site or two tandem sites to stimulate *dn*TA at the test site on chromosome VII. One copy of the Tel11 site stimulated

telomere addition in only four or five of 60 GCR events analyzed by PT-seq (7.5%; Figure 15).

Remarkably, adding a second Tel11 (2xTel11) site increases the frequency of GCR events

undergoing *dn*TA to 83.3% (Figure 15b). Together these results suggest that Cdc13 binding is

important for stimulating *dn*TA at the SiRTA however to efficiently stimulate *dn*TA more than

on Cdc13 binding site should be present. This is supports the model of a bipartite structure for

SiRTAs requiring at least one site that can bind Cdc13 in order to stimulate *dn*TA at another

Cdc13 binding site within the SiRTA.
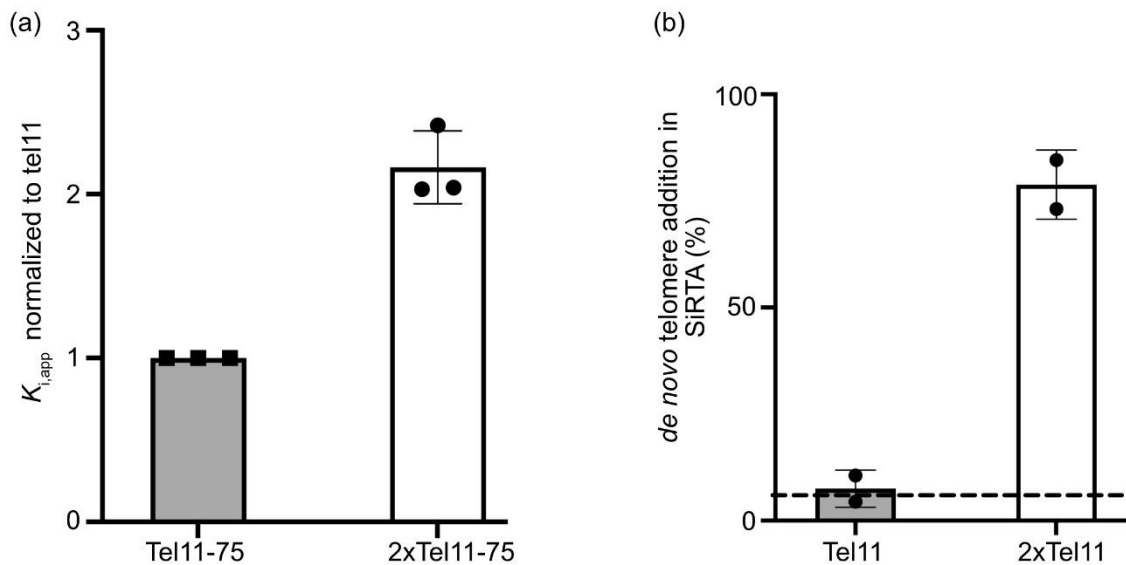
Figure 15. Integration of two canonical Cdc13 binding sites is sufficient to stimulate high levels of *dnTA*. a) A competition fluorescence polarization assay was utilized to measure the relative association of the Cdc13 DNA binding domain with the indicated sequences (Table 10). Relative $K_{i,app}$ was determined as described in Chapter 4 (*Materials and Methods*). Each point represents an independent measurement; error bars are standard deviation. b) The percent of GCR events involving *dnTA* within the indicated sequence was determined by PT-seq on chromosome VII. Tel11 contains a single canonical Cdc13 binding site; 2xTel11 contains two tandem sites. The dotted line represents the 6.6% threshold used to classify a sequence as a SiRTA. Data summarized in Table 5.

3.2.6 SiRTA distribution is not strongly associated with known protein binding sites or chromosome landmarks.

To gain insight into factors that may contribute to *dn*TA, we examined whether putative SiRTAs preferentially overlap with the binding sites of proteins related to telomere addition such as Est2 (a component of telomerase recently shown to associate with internal chromosome sites; Lendvay et al. 1996; Pandey et al. 2021), Rap1 (a transcription regulator that also binds telomeric repeats and affects telomere length homeostasis; Conrad et al. 1990; Rhee and Pugh 2011), and Pif1 (a helicase that negatively regulates telomerase at telomeres and DNA double-strand breaks; Schulz and Zakian 1994; Paeschke et al. 2011). We also examined the correlation between predicted SiRTAs and fragile sites, identified as regions that associate with үH2AX even in the absence of exogenous damage (Downs *et al.* 2000; Capra *et al.* 2010), and between SiRTAs and sequences predicted to form G-quadruplex structures (Capra *et al.* 2010). Using a permutation-based enrichment test under conditions that require an overlap with at least half of the predicted SiRTA sequence, we find statistically significant enrichment among SiRTAs for Est2, Rap1, үH2AX binding sites, and G4-forming sequences (Figure 16a and b). However, overlap never exceeds 20% of the putative SiRTAs, arguing against a strong functional relationship. Although the actual number of putative SiRTAs overlapping with a Pif1 binding site is the highest of all features tested (18.4%), this extent of overlap is not significant, likely because the regions reported to bind Pif1 by chromatin immunoprecipitation are relatively broad. To determine whether the predicted strength of the SiRTA affects these results, we divided putative SiRTAs into tertiles based on CATHI score but found no strong relationship between CATHI score and the significance of overlap (Figure 16c). Of the 14 SiRTAs confirmed to be active, only one overlaps with an Est2 binding site, and one overlaps with a Rap1 binding site.

Overall, these results fail to identify any overlapping binding sites with evidence of strong functional significance and suggest that fragile sites and G-quadruplex forming sequences are not strongly correlated with predicted hotspots of *dn*TA.
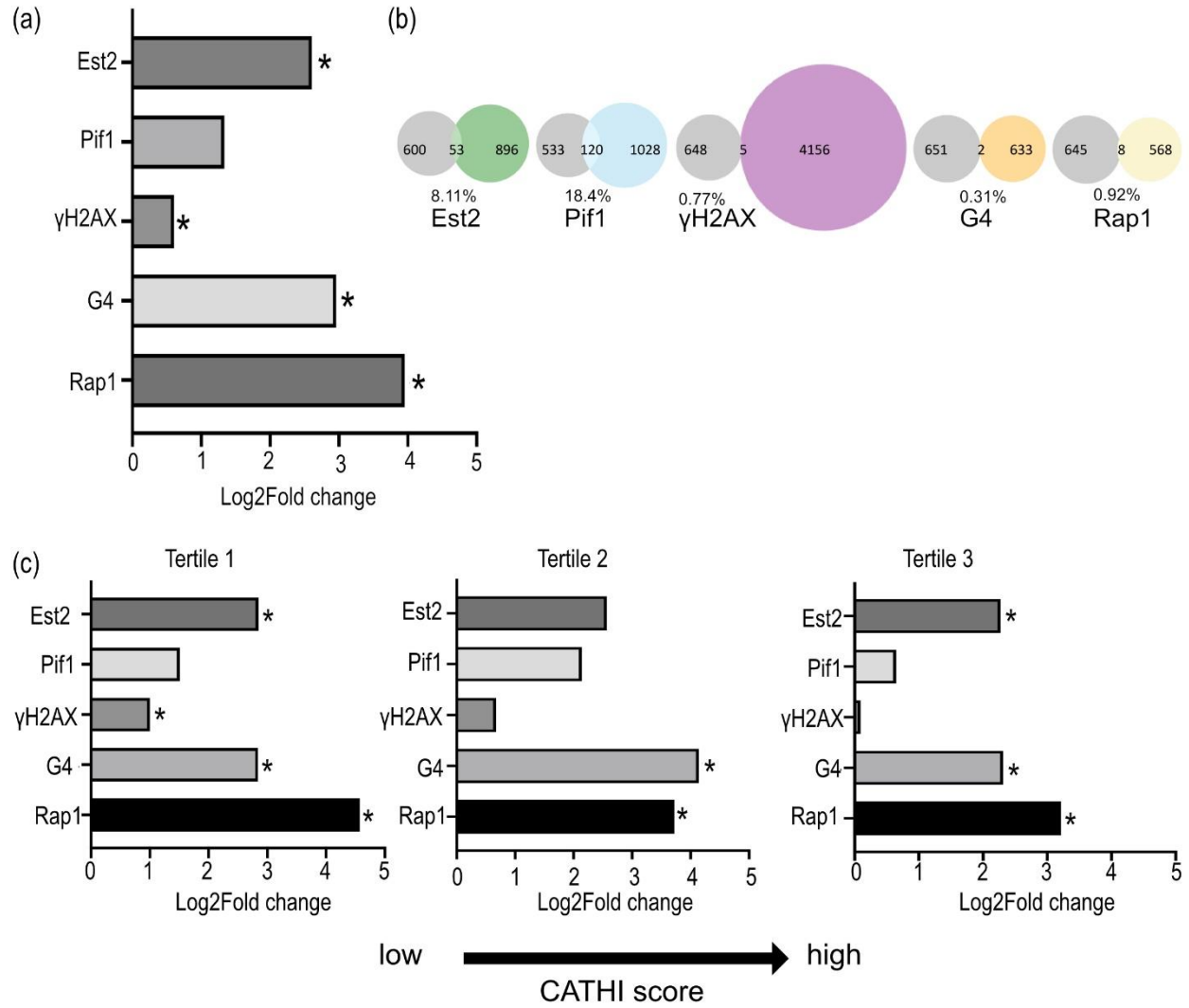
Figure 16. Overlap of SiRTAs with protein binding sites and G-quadruplex forming sequences. a) Using a permutation strategy, the enrichment of SiRTAs (Log2 fold change) for overlap with the indicated protein binding sites and with G-quadruplex forming sequences was determined (see Chapter 4; *Materials and Methods*). Analysis excluded sub-telomeric regions. *p-value<0.01 using Bonferroni's correction. b) Venn diagrams showing the number of SiRTAs that overlap with the indicated protein binding site or chromosome feature. c) As in (a), except that SiRTAs were divided into tertiles based on CATHI score prior to permutation analysis. Tertile 1 represents the lowest scores. *p-value<0.01 using Bonferroni's correction.

3.2.7 SiRTAs are predominantly found within coding regions

We were interested in determining whether SiRTAs are preferentially excluded from genic regions due to higher levels of evolutionary constraint. Each of the 728 SiRTAs was categorized as genic (any part of the SiRTA overlapped with a gene as defined by the start and stop codon of each annotated gene) or intergenic (Table 7 (see appendix A)). Seventy-eight percent of all SiRTAs overlap with coding regions and only 22% are exclusively found in intergenic regions (Figure 17). Given that approximately 30% of the yeast genome is intergenic (Hurowitz and Brown 2003; Lynch *et al.* 2008), we conclude that SiRTAs are not excluded from expressed regions. There are two exceptions. When a SiRTA contains long (>20 nt) TG-dinucleotide repeats, those repeats are virtually never found within a coding region. This result is not surprising since the expansion or contraction of a dinucleotide repeat is expected to disrupt the open reading frame. Second, intergenic SiRTAs are disproportionately found in the subtelomeric X and Y' elements. While only 8% of all SiRTAs are in the subtelomeric regions, 37% of intergenic SiRTAs are subtelomeric, consistent with the presence of a few transcribed regions in the subtelomeres (Table 7(see appendix A)). Interestingly, for those SiRTAs that overlap with open reading frames, 58% are located on the template strand, which is different from the expectation of random distribution (Figure 17; p<0.05 by chi-square test) and suggests that the presence of these sequences within genes may, in some cases, have consequences for cellular fitness.
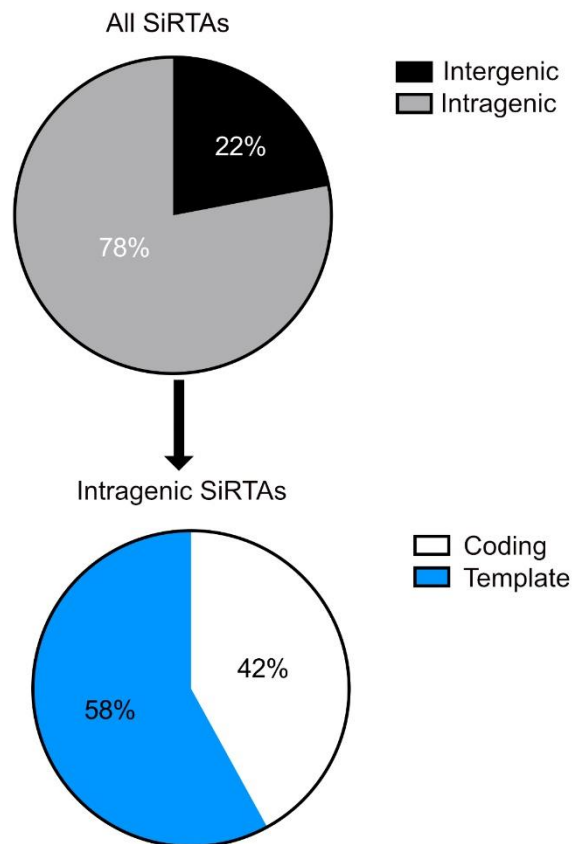
Figure 17. Analysis of SiRTA overlap with protein coding regions. SiRTAs were classified as intragenic or intergenic (see Chapter 4 (*Materials and Methods*) and Table 7). Intragenic SiRTAs were analyzed to determine if the TG-rich SiRTA sequence is located on the template or coding strand.

3.3 Discussion

3.3.1 Prediction of SiRTA function

In this work, we predict the distribution of SiRTAs in the yeast genome, an important

step in understanding the role of these sequences in genome stability and function. The Zakian

laboratory initially proposed that hotspots of *dn*TA addition contain tracts of 15 or more

nucleotides consisting exclusively of T and G in a "telomere-like" pattern (Mangahas *et al.*

2001). Our subsequent analysis of SiRTAs on chromosomes V and IX revealed that these

requirements are too strict. For example, SiRTA 5L35(-) (formerly called 5L-35) stimulates

*dn*TA in response to both spontaneous and induced DSBs (Stellwagen *et al.* 2003; Obodo *et al.*

2016; Ngo *et al.* 2020), but the longest uninterrupted string of TG sequence is 14 nucleotides, including several instances of a TT motif that never occurs within telomeric repeats.

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

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

74

3.3.2 Sequences that stimulate dnTA associate with Cdc13

Because prior work suggests that *dn*TA is stimulated by the association of Cdc13 with single-stranded DNA generated after a DSB, the CATHI algorithm likely identifies sequences with an affinity for Cdc13. To test this hypothesis, we developed a fluorescence polarization competition assay in which sequences are tested for their relative ability to compete with a labeled oligonucleotide for binding to the purified Cdc13 DNA binding domain. A 75-mer oligonucleotide containing two tandem copies of the canonical Cdc13 binding site competes twice as well as an oligonucleotide containing a single site, suggesting that the assay is sensitive to Cdc13 binding (Figure 15a). Our goal is to measure the overall association of Cdc13, which arises as a combination of the number and affinity of binding sites. We find that 75-mer oligonucleotides containing sequences that stimulate *dn*TA compete as well or better than a 75-mer containing a single match to the telomeric consensus Cdc13 binding sequence. In contrast, sequences that fail to support *dn*TA compete less well (Figure 15a). Importantly, mutations previously demonstrated to reduce SiRTA function also reduce Cdc13 binding (Figure 15c). Together with our previous demonstration that *dn*TA is stimulated through the artificial recruitment of Cdc13 to the Stim sequence of a SiRTA (Obodo et al. 2016; Epum et al. 2020; Hoerr et al. 2023), these results are consistent with the requirement for a threshold level of Cdc13 in stimulating *dn*TA.

Despite the observation at a single SiRTA that Cdc13 association correlates well with SiRTA efficiency, the apparent overall affinity for Cdc13 measured by fluorescence polarization poorly predicts SiRTA efficiency. For example, SiRTA 14L35(-) competes equivalently with the control sequence but stimulates *dn*TA more strongly than most other SiRTAs, including those that compete more effectively for Cdc13 binding. This apparent discrepancy may reflect the

effects of the distribution or spacing of Cdc13 binding sites on SiRTA function. In prior work, we observed that deletion of the ~30 nt spacer region between the Stim and Core sequences of a SiRTA dramatically increases SiRTA efficiency (Obodo *et al.* 2016). The highly efficient SiRTA 14L35(-) contains an unusually long region of TG-rich sequence that likely acts as both a Stim and Core region with little or no spacer, a property that may account for its ability to stimulate *dn*TA strongly despite an overall lower affinity for Cdc13.

3.3.3 Limitations to the predictive capacity of the CATHI algorithm

The well-characterized and functional SiRTA 2R780(-) has a CATHI score of only 13, despite stimulating *dn*TA with an efficiency of 31.1% (Hoerr et al. 2023). By fluorescence polarization, this sequence competes for Cdc13 binding more effectively than many sequences with higher CATHI scores, suggesting that the failure of the algorithm to predict SiRTA function (at least in this case) is primarily a failure to predict Cdc13 binding. One explanation is that the CATHI algorithm does not prioritize matches to the GxGT motif identified as particularly impactful for Cdc13 binding (Anderson *et al.* 2003; Eldridge *et al.* 2006). Indeed, we find that mutation of even one GxGT motif in SiRTA 2R780(-) strongly reduces Cdc13 binding and nearly eliminates SiRTA function (Figure 15c) (Hoerr et al. 2023). The 226 bp minimal sequence of SiRTA 2R780(-) contains seven GxGT sequences that may account for the ability of this sequence to stimulate *dn*TA despite lacking sufficiently long/abundant telomere-like tracts to be identified by the algorithm.

Although the presence of GxGT motifs is an attractive explanation for the activity of SiRTA 2R780(-), attempts to incorporate the motif into the algorithm did not improve the accuracy with which SiRTA function could be predicted and instead increased the number of false positive results. For example, the false negative SiRTA 14R306(+) (CATHI score = 15.5)

contains five GxGT motifs (two of which overlap), but the false positive SiRTA 12-330(+)

(CATHI score = 22) also contains five distinct GxGT motifs. There are at least three (non-

exclusive) explanations for the remaining discrepancies between the predictive algorithm and

measured rates of *dn*TA. First, it remains unclear how Cdc13 affinity is affected by deviation

from the consensus telomere binding site. Although extensive mutagenesis has been conducted

in vitro, these studies either altered single nucleotide sites (showing that positions 2 and 5-11 are

tolerant of single changes) or simultaneously mutated the seven 3'-most nucleotides (showing

that the GxGT motif is insufficient; Anderson et al. 2003; Eldridge et al. 2006; Lewis et al. 2014;

Glustrom et al. 2018). Neither approach fully recapitulates sequences that Cdc13 will encounter

at internal sites exposed by resection. Second, as described above, the distance between Cdc13

binding sites likely contributes strongly to SiRTA function. We have attempted to account for

this property by using a window size of 75. However, some effects on SiRTA efficiency likely

arise from varied distributions of Cdc13 binding sites that are not fully captured with the

algorithm. Third, we suspect that *dn*TA can be stimulated either by a small number of high-

affinity Cdc13 binding sites or by a larger number of low-affinity sites. SiRTAs at either extreme

of this continuum may be difficult to identify using the current strategy.

*3.3.4 SiRTAs do not colocalize strongly with binding sites for other telomere/telomerase-
associated proteins*

Our co-localization analysis failed to identify additional proteins that strongly impact

SiRTA function. Overall, we consider it unlikely that the observed enrichment represents a

functional relationship. For example, Rap1 binding sites are overrepresented among SiRTAs, but

this result is not surprising given that the consensus binding site for Rap1 is also TG-rich. In

previous work, we showed that Rap1 association *per se* is not required for SiRTA activity

(Obodo *et al.* 2016). Our results suggest that Est2 binding in undamaged conditions is not

required for a sequence to function as a SiRTA. Only 8.1% of predicted SiRTAs overlap with experimentally determined sites of Est2 enrichment and only one of the fourteen active SiRTAs is also an internal Est2 binding site.

Consistent with our observation that SiRTAs stimulate *dn*TA even when located several kilobases from an induced DSB, we observe only a modest correlation between sequences that function as SiRTAs and sites enriched for phosphorylated H2A (gH2AX), a marker of DNA damage. Since enrichment was determined in undamaged cells, these sites represent regions of the genome that are prone to spontaneous damage, likely due to difficulties encountered during DNA replication. It will be interesting to determine whether SiRTAs that overlap with fragile sites are more likely to stimulate the spontaneous formation of gross chromosomal rearrangements. For example, we have proposed that generation of acentric fragments through *dn*TA on chromosome II is facilitated by an unusually high density of inverted repeats in this region combined with errors in the resolution of stalled replication forks (Hoerr et al. 2023). In this light, it is interesting to note that sequences required to stimulate *dn*TA on chromosome II [SiRTA 2R780(-), coordinates 779784 to 780009] overlap with a region of enhanced yH2AX association (779987-780040; (Capra *et al.* 2010)).

Interestingly, we find a statistically significant tendency for predicted SiRTAs, if found within an open reading frame, to be oriented with the TG-rich sequence on the template strand. We propose that the presence of the TG-rich sequence within the exposed strand of the transcription bubble may be deleterious. Interestingly, this bias is opposite to that observed for G-quadruplex forming sequences in mammalian cells, which are more likely to be found on the coding strand (Agarwal *et al.* 2014; Rhodes and Lipps 2015; Kim 2019). In yeast, Replication protein A (RPA)-bound single-stranded DNA at stalled transcription complexes has been

implicated as a major signal of DNA damage (Wang and Haber 2004; Tapias *et al.* 2004; Fanning 2006). Conceivably, competition for binding to single-stranded DNA by Cdc13 could interfere with this process, leading to selection against Cdc13 target sequences on the exposed coding strand. Despite the frequent presence of SiRTAs within genes, transcription does not appear to be required for SiRTA function since the test site that we developed on chromosome VII is contained within an intragenic region, more than 1.5 kilobases from the 3' end of the *ADH4* gene.

3.3.5 Implications of SiRTA distribution in the yeast genome

The compact and well-annotated yeast genome presents an opportunity to assess evidence of selective pressures that might act upon sequences with a propensity to stimulate *dn*TA. Based on the hypothesis that *dn*TA within the nonessential terminal region of a chromosome arm might provide a selective advantage by allowing a cell to survive a persistent DSB, we examined the distribution of predicted SiRTAs in essential and nonessential chromosome regions. As predicted, we found a significant enrichment of putative SiRTAs in nonessential terminal regions. Furthermore, as expected for a role in chromosome stabilization, only SiRTAs in the TG orientation are overrepresented. However, both effects disappear when the subtelomeric X and Y' elements are removed from the analysis (Figure 8a and c).

Nine of the TG-oriented SiRTAs in the subtelomeric regions correspond to stretches of TG$_{1-3}$ (perfectly "telomeric") sequence that is located predominantly between tandem Y' elements. However, the vast majority, while TG-rich, deviate substantially from the TG$_{1-3}$ pattern. Whether these sequences are vestiges of ancient telomeric repeats is unclear. Because the Y' and X elements are similar between chromosome ends, many of the subtelomeric SiRTAs have similar or identical CATHI scores and represent a small number of unique sequences.

Given the near ubiquity of TG-oriented SiRTAs within subtelomeric regions (identified at 31 of 32 chromosome ends), we speculate that these sequences are conserved, at least in part, due to an ability to stimulate dnTA in the event of catastrophic telomere loss, most likely due to replication errors within the telomeric repeats. The subtelomeric region on the right arm of chromosome VI contains a truncated X element followed immediately by TG_{1-3} telomeric repeats and therefore lacks sequences predicted to function as a SiRTA (Figure 6c). This subtelomeric structure may have resulted from a prior dnTA event within the X element. In the future, it would be interesting to determine whether the right arm of chromosome VI is more sensitive to catastrophic loss of telomeric repeats than other chromosome termini that contain intact X element repeats.

While the spatial distribution and orientation of putative SiRTAs outside the subtelomeres are not strongly skewed, the number of sequences with the potential to act as SiRTAs is significantly higher than predicted by chance (Figure 11). This excess is observed at both low and high scores but is increasingly pronounced at higher CATHI scores. Scores of 30 or greater are approximately 20-fold overrepresented in the yeast genome compared to the random expectation (Figure 11b). Our data do not provide evidence that SiRTA function *per se* is driving this excess, particularly because we also observe an excess of scores below 20, representing sequences that are not likely to stimulate dnTA at unusually high levels (Figure 11c). Since many SiRTAs are located within coding regions, we considered the possibility that codon bias might explain this pattern. However, codons consisting of only G and T (or only C and A; TTT and AAA excluded) collectively are not overrepresented among all codons (Table 11 (see appendix A); Nakamura *et al.* 2000). It is possible that particular amino acid repeats could result in this effect. For example, poly-proline tracts consisting of CCA and CCC codons can generate a

SiRTA signature. However, only a small fraction of poly-proline tracts are also identified as potential sites of *dn*TA.

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

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

This work provides, for the first time, a genome-wide map of sites predicted to stimulate *dn*TA. With the exception of sites clustered in subtelomeric regions, the largely random distribution and orientation of SiRTAs throughout the genome stands in interesting contrast to

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

Chapter 4[6]

Materials and Methods

4.1 Strain construction

Strains were constructed in the S288C background as described (Hoerr et al. 2023; Ngo et al. 2023). The parental strain contains a *URA3* marker distal to an HO recognition site on chromosome VII (YKF1975 *MATa::ΔHOcs::hisG hmlαΔ::hisG HMRa::NAT URA3Δ851 trp1Δ63 leu2Δ::KANR ade3::GAL10::HO* Chr VII, 15828-16027 (*adh4*)::HOcs::*HYGR pau11::URA3*). 300 bp sequences to be tested for SiRTA function were inserted using the CRISPR/Cas9 system as described in Anand *et al.* (2017) using a guide sequence of 5'-TGCGGCAAGTTCATCTTCCA located ~2kb centromere-proximal to the HO recognition site. PCR products for recombinational insertion were generated as follows. Forward primers were designed by including 40 bases upstream of the gRNA recognition site (5'TTTCTTTGGAAAACGTTGAAAATGAGGTTCTATGATCTAC) followed by the first ~20 bases on the 5' end of the sequence of interest. Reverse primers were constructed by taking the reverse complement of the 40 bases downstream of the gRNA site (5'-AGAACATAGAATAAATTTGGTACTGGAACGTTGATTAACT) followed by the last ~20 bases of the sequence of interest. The sequences tested are listed in Table 1(see appendix A). The DNA fragment needed to insert SiRTA 6R210(+) onto chromosome VII using the CRISPR system was synthesized and inserted into the pMX plasmid using Invitrogen GeneArt Gene Synthesis services (Thermo Fisher Scientific). The 6R210(+) DNA fragment was amplified from the plasmid using PCR primers designed as described above. *RAD52* was replaced by one-step gene replacement using template DNA from pFA6a-TRP1 (Longtine *et al.* 1998).

---

[6] This work is adapted from portions of the research article: Ngo K, Gittens TH, Gonzalez DI, Hatmaker EA, Plotkin S, Engle M, Friedman GA, Goldin M, Hoerr RE, Eichman BF, Rokas A, Benton, ML, and Friedman KL. 2023. A comprehensive map of hotspots of *de novo* telomere addition in Saccharomyces cerevisiae. *Genetics*, iyad076. https://doi.org/10.1093/genetics/iyad076

For testing on chromosome IX, the *URA3* marker and HO cleavage site were integrated on chromosome IX to create strain YKF1752 (*MATa::ΔHOcs::hisG hmlαΔ::hisG HMRa::NAT URA3Δ851 trp1Δ63 leu2Δ::KANR ade3::GAL10::HO* Chr9;35050-41450::HOcs::HPHR *soa1::URA3*) as described in (Obodo *et al.* 2016). Yeast strains containing the BS Mut1 and BS Mut2 mutations are described in (Obodo *et al.* 2016) as YFK1610 and YFK1652, respectively. Strains for testing sequences endogenously were created as described in Obodo et. al 2016 with primers for strain construction listed in Table 2(see appendix A).

For the yeast strains in which the SiRTA test site is directly adjacent to the HO cut site (close HO assay; Chapter 5), the HO cut site and *URA3* markers were constructed as described in Hoerr et al. 2023; Ngo et al. 2023 for the assays where the test site is located ~2kb away from the HO cut site except the putative SiRTA sequence is added adjacent to the HO cut site and transformed with the HO cut site. Primers used to generate PCR products for integration are listed in Table 2. *Pif1m2* mutations were made as described in Schulz and Zakian 1994.

4.2 HO cleavage assay

The HO cleavage assay was performed as described (Hoerr et al. 2023; Ngo et al. 2023). Briefly, cells were grown in synthetic dropout media lacking uracil (SD-Ura) + 2% raffinose to an optical density at 600nm (OD600) of 0.6-1.0. Cells were serially diluted and plated on yeast extract peptone medium with either 2% dextrose (YEPD) or 2% galactose (YEPgal). After incubation at 30˚C for three days, colony number was determined on at least two plates of each condition. The frequency of survival on YEPgal was calculated as: (average number of colonies per plate on YEPgal x dilution factor)/(average number of colonies per plate on YEPD x dilution factor). At least 100 colonies surviving on YEPgal were patched to a medium containing 1 mg/mL 5-fluoroorotic acid (5-FOA) to select for cells in which the *URA3* marker was lost [gross chromosomal rearrangement (GCR) events]. The overall frequency of GCR events was

determined as: (the frequency of survival on YEPgal*frequency of clones demonstrating 5-FOA resistance). Thirty clones that displayed growth on a medium containing 5-FOA were selected and inoculated in liquid YEPD for genomic DNA extraction using the MasterPure™ Yeast DNA Purification Kit (Lucigen). Multiplex PCR was used as described in Ngo et al. 2020 to map the approximate site of *dn*TA in relationship to the sequence of interest. Primers for chromosomes VII and IX are listed in Table 2 (see appendix A). Colonies where the DNA loss event mapped within the sequence of interest were tested for telomere addition using one primer centromere proximal to the putative telomere addition site and a second primer complementary to the telomeric repeat (Table 2(see appendix A)). SiRTA efficiency is defined as the percent of GCR events involving *dn*TA in the SiRTA. This value is a reproducible measure of SiRTA function when the frequency of GCR formation is similar between strains (Epum *et al.* 2020). Southern blotting to confirm *dn*TA was performed as described in Obodo *et al.* 2016 except the restriction enzyme used in these experiments was *Cla1* and the primer used as a probe was chr7presirta5for which can be found in Table 2 (see appendix A).

For the *pif1m2* "close" assays described in Chapter 5, the assay is done as described above except that the over frequency of GCR events (the frequency of survival on YEPgal*frequency of clones demonstrating 5-FOA resistance) is used as the end point of the analysis.

4.3 Pooled Tel-seq
Thirty 5-FOA-resistant clones isolated as described above were separately inoculated in 200 μL of YEPD in a 96-well culture plate and incubated overnight at 30˚C to reach saturation. Equal volumes (at least 30 μL) of each culture were pooled and DNA was extracted using the YeaStarTM genomic DNA kit (ZYMO research).

Libraries were prepared using 50 ng of genomic DNA and a modified protocol using the

Twist Library Preparation kit (Twist Bioscience 106543). Denatured DNA templates in a 96-well

plate were randomly primed with 5' barcoded adapters. Samples were pooled, captured on

streptavidin-coated magnetic beads, and washed to remove excess reactants. A second 5'

adapter-tailed primer with a strand-displacing polymerase was utilized to convert the captured

templates into dual adapter libraries. Beads were washed to remove excess reactants. Four cycles

of PCR were utilized to amplify the library and incorporate the plate barcode in the index read

position. Libraries were sequenced using the NovaSeq 6000 with 150 bp paired-end reads

targeting 13 to 15 million reads per sample. Real-Time Analysis software (version 2.4.11;

Illumina) was used for base calling and data quality control was completed using MultiQC v1.7.

Sequencing data are available from the NIH Sequence Read Archive (SRA) under

BioProject ID PRJNA939836. Reads mapping to a 300 bp control sequence located in the

essential gene *BRR6* (Chr VII: 36933 to 37233) or to the 300 bp sequence of interest [inserted on

chromosome VII or at the endogenous location of SiRTA 9L44(-)] were identified using

Bowtie2 (Galaxy Version 2.5.0+galaxy0) with the sensitive local setting (Langmead *et al.* 2009;

Langmead and Salzberg 2012). Any remaining library primer sequences were removed using the

Trimmomatic tool (Galaxy Version 0.38.0; Bolger et al. 2014) and reads mapping to the putative

SiRTA that also contains telomere sequence (match to 5'-GGGTGTGG or 5'-CCACACCC)

were identified and tabulated. The number of individual reads with evidence of telomere addition

was normalized by expressing telomere reads as a percentage of control reads mapping to *BRR6*.

Where applicable, sites of telomere addition were mapped to the original SiRTA sequence to

determine the location of the event.

4.4 Purification of Cdc13-DBD

Cdc13-DBD was expressed in *E. coli* using pET21a-Cdc13-DBD-His6, a gift from the

Wuttke lab. Purification was done as described (Anderson *et al.* 2002; Obodo *et al.* 2016).

4.5 Fluorescence Polarization binding assays

Binding assays were conducted using a fixed concentration of a 5'-6-carboxyfluorescein

(FAM) labeled tel-11 oligonucleotide (25 nM). Cdc13-DBD was added at final concentrations of

0, 6.25, 12.5, 25, 37.5, 50, 62.5, and 75 nM. Competition binding assays were conducted at fixed

concentrations of Cdc13-DBD (30 nM) and 5'-6-FAM labeled tel-11 oligonucleotide (25 nM).

Unlabeled oligonucleotides of 75 bases each were used at final concentrations of 0, 6.25, 12.5,

25, 50, 150, and 200 nM. Oligonucleotide sequences are listed in Table 10 (see appendix A) .

Labeled and unlabeled oligonucleotides and protein were mixed in binding buffer (50 µM Tris

pH 8, 1 µM EDTA pH 8, 15% glycerol, 75 µM NaCl, 75 µM KCl) in a final volume of 80 µL

and incubated at 4°C for 30 minutes. Each reaction was measured in triplicate (25 µl per

measurement) in a Corning 384 well assay plate using the BioTek Synergy H1 hybrid reader.

This procedure was repeated at least three times for each competitor. The relative polarization

(ΔP) was determined using the following equation: $\Delta P = P0 - Px$, where P0 is the polarization value

at a competitor concentration of 0 and Px represents the polarization value at x competitor

concentration. For each experiment, technical replicates were averaged and the averaged data

were fit to the following equation: $\Delta P = (Pmax[competitor])/(K_{i,app}[competitor])$, where Pmax is

the maximal polarization value and $K_{i,app}$ is the apparent inhibition constant (Anderson *et al.*

2008; Vaasa *et al.* 2009). An unlabeled 75mer containing the Tel11 sequence at the center of the

oligonucleotide (Tel11-75) was included in each experiment and normalized $K_{i,app}$ values are

reported as the fold change relative to this control ($K_{i,app}$ of Tel11-75/$K_{i,app}$ of experimental

oligonucleotide)

4.6 Implementation of the CATHI algorithm

Initially, the program generates a series of sliding windows to be utilized in the score calculation. The window and step size of the sliding windows can be customized using the –window and –step options. For each window, the program searches for strings of at least 4 characters that begin with a G and consist of only Gs and Ts. These become the set of candidate scoring regions. From these candidates, regions that consist only of Gs are removed. The program then scans candidate regions for any consecutive Ts, or four or more consecutive Gs. If either are encountered, that candidate region is truncated after the first T or the third G, respectively. Once the set of candidates has been filtered, the number of nucleotides remaining in the candidate set is counted and any applicable scoring penalties are subtracted. There are no penalties applied by default, but users can choose to apply them. The –penalty option imposes score deductions for any GGTGG sequences, and the –ttpenalty imposes score deductions for Ts that flank the candidate regions.

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

telomeres were trimmed prior to analysis (if present). The coordinates used for each chromosome

are in Table 6 (see appendix A).

Overlapping and adjacent windows meeting or exceeding the threshold value can be

merged into a single region using the –cluster option, where the beginning is the start coordinate

of the most upstream window and the ending is the end coordinate of the most downstream

window. The score of the merged region is the maximum CATHI score across all merged

windows.

4.7 Generation of a shuffled yeast genome

Five randomized versions of the *S. cerevisiae* (sacCer3) genome were generated to

evaluate the number of SiRTAs expected when applying the CATHI algorithm to a null model.

A set of five was chosen to evaluate the reproducibility of CATHI score distribution across

randomized genomes (expressed as standard deviation).

The DNA sequence was downloaded from the sacCer3 reference genome using the

BedTools (version 2.30.0) 'getfasta' command (Quinlan and Hall 2010) after adjusting the start

and end coordinates of each chromosome to exclude subtelomeric regions (Table 8 (see appendix

A)). Nucleotides were randomly shuffled within each adjusted chromosome using Python's built-

in randomization library. This procedure maintains the nucleotide composition and length of

each chromosome while randomizing the actual DNA sequence.

4.8 Enrichment for genomic annotations in putative SiRTAs

Overlap between putative SiRTAs and other genomic annotations was determined using a

permutation-based enrichment test. Enrichment for SiRTAs with several different genomic

annotations was determined: (1) essential and non-essential genomic regions (Giaever *et al.*

2002); (2) Est2 binding sites (Pandey *et al.* 2021); (3) Pif1 binding sites (Paeschke *et al.* 2011);

(4) yH2AX binding sites (Capra *et al.* 2010); (5) G-quadruplex regions (Capra *et al.* 2010) and

(6) Rap1 binding sites (Rhee and Pugh 2011). When the original dataset included strand information (as in the case of G4-sites) that information was considered in the analysis.

Enrichment between the SiRTAs and the annotations was calculated as the fold change between the observed and expected overlap. To ensure meaningful overlaps between the SiRTAs and the genomic annotations, at least 50% of the binding site was required to overlap with the SiRTA, or at least 50% of the SiRTA was required to overlap with the essential/non-essential region. To create the distribution of expected overlap, 1000 permutations were performed by randomly shuffling regions throughout the genome and calculating the amount of SiRTA overlap. Shuffled regions are non-overlapping, length- and strand-matched (G4 sequences only) with the annotations. When specified, telomeric and/or subtelomeric regions were excluded. Subtelomeric regions are defined in (Table 8 (see appendix A)) For G4 sites, overlap was only recorded if the G4-forming sequence and SiRTA are on the same strand. An empirical p-value is calculated for the overlap using the expected distribution; where relevant, p-values are corrected for multiple comparisons using the Bonferroni method.

To determine whether there was a relationship between the enrichment for functional annotations and CATHI score, all enrichment analyses were performed on putative SiRTAs divided by tertile. For all SiRTAs with a score above the threshold (CATHI score > 20), regions were divided into bins based on the quantile of the score using the qcut function in Pandas (v1.4.2) (McKinney 2010). Regions with the same score will belong to the same bin. Due to the number of scored regions and the distribution of scores, three bins (tertiles) were chosen to represent low, medium, and high-scoring SiRTAs.

4.9 Determining overlap with genes

The location of predicted SiRTAs was compared to the location of genes within the *S. cerevisiae* genome to determine the number of predicted SiRTAs in both inter- and intragenic

regions. Coordinates for genes and subtelomeres (defined as X and/or Y' elements) were obtained from the *S. cerevisiae* S288C annotation available from NCBI (accession numbers NC_001133.9, NC_001134.8, NC_001135.5, NC_001136.10, NC_001137.3, NC_001138.5, NC_001139.9, NC_001140.6, NC_001141.2, NC_001142.9, NC_001143.9, NC_001144.5, NC_001145.3, NC_001146.8, NC_001147.6, NC_001148.4); FASTA and GFF3 files for the reference assembly of strain S288C (GCF_000146045.2) were downloaded from NCBI's RefSeq database. The RefSeq genome annotation is identical to that in the *Saccharomyces* Genome Database (SGD).

Coordinates for predicted SiRTAs were obtained from the CATHI program and manually converted into GFF3 files, one for each chromosome. Overlap between predicted SiRTAs and genes was calculated using the "intersect" function within bedtools v2.30.0 (Quinlan 2014) for each chromosome. Predicted SiRTAs within annotated genes were manually assigned to the template or coding strand using chromosome visualization in Geneious Prime v2020.1.2.

4.10 Modeling SiRTA distribution

Python programs to model the expected distribution of telomere addition events within a region, to model the random distribution of SiRTAs between the forward and reverse strand, and to model the random spacing of SiRTAs are available at https://github.com/geofreyfriedman/sirta. For the strand analyses, random strand distributions were generated for each chromosome based on the observed number of SiRTAs on each strand. The expected distribution of SiRTAs between the forward and reverse strands was quantified by 1) determining the number of consecutive SiRTAs on the same strand (run length) or 2) summing the number of times that consecutive SiRTAs are found on different strands (number of strand switches). To avoid "edge effects" generated at the ends of each chromosome, 10,000

iterations were generated for each chromosome, and run lengths or strand switches were summed across the 16 chromosomes (sum of iteration 1, the sum of iteration 2, etc). In each case, the observed value was compared to the random distribution generated from 10,000 iterations.

To model the random spacing of SiRTAs, locations were randomly chosen along the length of each chromosome (ignoring strand). Because the number of SiRTAs is small compared to the length of each chromosome, each location was assigned to a single nucleotide. For each iteration, distances between adjacent locations were determined on each of the 16 chromosomes. The average and standard deviation generated from five iterations was compared with the inter-SiRTA distances observed in the genome (determined as the distance between the 5'-most nucleotide of each SiRTA on the top DNA strand, regardless of orientation.

4.11 Assay to determine GCR rate in the absence of cleavage by the HO endonuclease

To construct a strain with the putative SiRTA sequence and the *URA3* marker but no functional HO cut site, cells from the strain that has the HO cut site and the *URA3* marker were plated onto galactose media to induce cleavage by the HO endonuclease and then 30 of the surviving colonies were streaked onto -URA media to select for colonies that maintained the end of the chromosome but survived HO cleavage. Mutation of the HO cut site was confirmed via PCR and Sanger sequencing. This was done for twice the 6R210(-) strain and the resulting strains were termed (YKF2440 and YKF2441 respectively). A strain (termed YKF2127) with no SiRTA added to the chr. VII test site and with an *URA3* marker but no HO cut site was used as a control. Two strains lacking a functional HO site were chosen for testing for both 6R210(-) and WT. A single colony for each strain was grown in liquid medium lacking uracil to an OD of ~0.6 then 10µl of liquid culture diluted by 10^{-3} was plated on YPD and 50µl of undiluted liquid

culture was plated onto media containing 5-FOA. The growth rate was then calculated by dividing growth on 5-FOA by growth on YPD media.

4.12 Analysis of conservation of TG dinucleotide repeats between closely related yeast species.

The *S. cerevisiae* (strain S288C) and *S. paradoxus* (strain CBS432) genome FASTA sequences were retrieved from the *Saccharomyces* Genome Database (SGD) at Stanford University (Stanford, CA) and the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (Bethesda, MD), respectively. Conservation of the TG dinucleotide repeats across yeast species relative to *S. cerevisiae* S288C was assessed via the Multiz Alignment & Conservation (7 Yeasts) track on the University of California Santa Cruz (UCSC)'s Genome Browser. With this tool, the previously mentioned six *Saccharomyces* species were aligned to *S. cerevisiae* S288C per input of chromosomal locus (*i.e.*, ChrVI:210345-210406). Conservation was quantified with the following equation: Ratio of conservation= (number of bases conserved)/(Total number of bases in the TG-dinucleotide repeat).

Chapter 5

Concluding remarks and future directions

Telomeres, the TG-rich ends of most eukaryotic chromosomes, function to protect the chromosome end and to distinguish natural chromosome ends from ends created by a DSB. The importance of telomeres and proper telomere maintenance has been well established within scientific literature. The effects of telomere dysfunction have been linked to mutation and cell death. In humans, telomere dysfunction is associated with cancer and other diseases (reviewed in Doksani and de Lange 2014; Casari et al. 2022). Telomere healing, a phenomenon where telomerase adds a *de novo* telomere to a non-telomeric sequence in response to a DSB, has been observed in multiple species including both humans and *S. cerevisiae.* In *S. cerevisiae*, hotspots for telomere healing have been identified and termed SiRTAs (reviewed in Pennaneach et al. 2006; Hoerr et al. 2023). Prior to the work described in this thesis, research on SiRTAs was limited to sequences in regions that could tolerate genomic loss and the process of testing these sequences was laborious and time-consuming (Obodo *et al.* 2016; Epum *et al.* 2020; Ngo *et al.* 2020). Furthermore, we lacked a comprehensive description of the number and distribution of SiRTAs within the yeast genome. In this work, I have described new techniques and strategies for identifying and testing sequences for their ability to stimulate *dn*TA. The development of the chromosome VII test site facilitates the analysis of mutant variants of SiRTAs and the map of SiRTAs offers a robust pool of SiRTAs which could be further analyzed to get a more complete understanding of SiRTA function.

In this chapter, I present preliminary data from ongoing projects that aim to address lingering questions regarding the mechanism and impact of *dn*TA at different types of SiRTAs. The work presented in this thesis offers insight into how SiRTAs are distributed in yeast and

describes the development of tools for the study of *de novo* telomere addition. In this chapter, I discuss some remaining questions and how they could be addressed.

5.1 Preliminary results and future directions
5.1.1 Negative regulation of telomerase by Pif1 at SiRTAs

The regulation of telomerase in yeast has been well studied in the context of perfect telomeric sequence. As discussed in Chapter 1, mechanisms regulating telomere addition at DSBs are well known. However, no published studies explicitly address whether these regulatory mechanisms are counteracted at SiRTAs. One known inhibitor of telomerase at DSBs is the Pif1 helicase. Pif1 is a 5' to 3' DNA helicase that is thought to negatively regulate *dn*TA by removing telomerase from the DNA (Boulé *et al.* 2005; Li *et al.* 2014). The absence of Pif1 results in a ~600-fold increase in *dn*TA in response to DSBs (Schulz and Zakian 1994). Work from the Durocher lab revealed a threshold for Pif1 activity wherein Pif1 only effectively inhibits telomerase action at telomeric tracts that are shorter than 34 base pairs (Strecker et al.2017; Boulé et al. 2005). Durocher and colleagues proposed that longer telomeric tracts recruit sufficient Cdc13 to outcompete Pif1 in binding to the DNA. This work was done specifically in a model system with either perfect telomeric tracts or with TG dinucleotide repeats and offered great insights regarding telomere function, but this analysis leaves questions regarding the importance of this pathway at naturally occurring telomere-like sequence.

SiRTAs while telomere-like, are for the most part not perfect telomere repeats and often have non-TG sequences interspersed between the telomere-like sequence. This poses the question of whether Pif1 regulation at SiRTAs is similar to Pif1 regulation at perfect telomeric repeats. This question is composed of two distinct issues. The first issue is whether Pif1 functions differently when a break has been resected a considerable distance prior to *dn*TA. The

second issue is whether SiRTAs bind sufficient Cdc13 to resist Pif1 negative regulation. In this Chapter, I specifically address the second issue.

Given the results from the Durocher laboratory and the demonstrated relationship between SiRTA efficiency and Cdc13 binding (Chapter 3), I hypothesized that SiRTAs with high amounts of *dn*TA would be less sensitive to negative regulation by Pif1. In this model, Cdc13 binding to the Stim and the Core sequence of the SiRTA prevents Pif1 from binding and knocking telomerase off the DNA; at a sequence that does not bind Cdc13, Pif1 is able to bind and prevent *dn*TA. To test this model, I made the *pif1m2* mutation, which impairs the nuclear function of Pif1, and inserted SiRTAs of interest adjacent to the HO cleavage site on chromosome VII (close HO assay) (Figure 18a). For these experiments, the sequence of interest must be tested next to the HO cleavage site because the lack of nuclear Pif1 function in the *pif1m2* mutations causes telomere addition in the region between the putative SiRTA and the HO cleavage site, complicating evaluation of *dn*TA at the SiRTA. Moving the putative SiRTA directly next to the HO cleavage site solves this issue. In this assay, a sequence subject to negative regulation by Pif1 would show increased frequency of *dn*TA when Pif1 is mutated. In contrast, a sequence that is resistant to Pif1 negative regulation would display a frequency of *dn*TA equal to wild type when Pif1 is mutated. In the case of SiRTAs, we speculate that we may observe a range of sensitivity to Pif1 negative regulation, rather than a strict threshold. Therefore, we might find a correlation between the efficiency of a SiRTA and the degree of resistance to Pif1 negative regulation.

To begin to address this hypothesis, we have tested 5L35(-) (a SiRTA with a lower rate of *dn*TA), 6L22(-) (a negative control that does not stimulate *dn*TA), 1Cdc13 site (a sequence that stimulates very little *dn*TA) and 2Cdc13 (a sequence that stimulates high amounts of *dn*TA).

For these experiments, the short sequence versions from Figure 2 were used for 5L35(-) and 6L22(-). To test these sequences, I performed the close HO assay in both WT and *pif1m2* backgrounds and compared the *dn*TA rates of both backgrounds. This assay was completed at least three times per tested sequence. For each sequence tested, the frequency of cells that survived the DSB as measured by growth on galactose was multiplied by the frequency of cells that lost the chromosome end as measured by growth on 5-FOA. Due to the proximity of the HO cut site to the SiRTA, all GCR events likely involve *dn*TA within or telomere proximal to the SiRTA. This result was confirmed through PCR analysis of ~120 clones using one primer that anneals to telomeric DNA and a second primer proximal to the cleavage site. Based on this result, subsequent analyses assume that all GCR events are due to telomere addition in the SiRTA. The ratio of the overall GCR frequency in the WT *PIF1* background to the overall GCR frequency of the *pif1m2* background is reported as a measure of a sequence's ability to escape negative regulation by Pif1. If a sequence is subject to negative regulation by Pif1 then the overall GCR frequency of the *pif1m2* background will be much higher than that of the WT (ratio of WT to *pif1m2* of much less than one) whereas if the sequence can escape negative regulation, then the ratio will be closer to one.

In these preliminary assays, we found that 1Cdc13 and 6L22(-), both sequences that stimulate low levels of *dn*TA, had ratios that were very close to zero. In contrast, the weak SiRTA 5L35(-) had a ratio that was low but still higher than both negative control sequences. This result suggests that 5L35(-) is subject to negative regulation by Pif1 but not as much as sequences that do not stimulate any *dn*TA or stimulate very low levels of *dn*TA. The highly active SiRTA 2Cdc13 had a higher ratio than the other sequences, supporting the hypothesis that sequences that stimulate higher amounts of *dn*TA circumvent negative regulation by Pif1 more

(Figure 18b and Table 12 (see appendix A)). Additional replicates are needed to confirm these results.

A future goal of this project is to test these SiRTAs using the longer 300bp sequence to see if the extra sequence affects the sensitivity to negative regulation by Pif1 similarly to how it may affect the frequency of *dn*TA. It is also important to test more SiRTAs with varying rates of *dn*TA and determine how closely the *pif1m2* ratio correlates with the frequency of *dn*TA. SiRTAs of immediate interest include 14L35(-) and 6R210(+) because of their rates of *dn*TA and 2R780(-), which was a false negative in the CATHI analysis.

Another future direction for this project is to address whether the resistance of a SiRTA to Pif1 negative regulation can be measured in the standard assay where the sequence of interest is located several kilobases from the induced chromosome break. To do this, we will replace the sequence between the test site and the HO cut site with an artificial sequence of the same size that does not contain any adjacent T and G nucleotides. The goal is to create a region in which telomerase is unable to act, even when lacking negative regulation by Pif1. This artificial sequence has been designed and cloned into a plasmid and is awaiting transformation into yeast.

The preliminary experiments described here offer insight into the role of Pif1 in regulating telomerase at potential SiRTA sites. The results suggest that SiRTA function and sensitivity to negative regulation by Pif1 are correlated but there not seem to be a strict threshold like that reported by the Durocher lab (Strecker et al. 2017). One explanation for this difference is that the Durocher lab utilized sequences comprised of strong Cdc13 binding sites while SiRTAs likely contain multiple, weaker Cdc13 binding sites. Therefore, the SiRTAs may represent a more nuanced range of Cdc13 binding than sequences that bind Cdc13 strongly,

creating a more gradual transition between sequences that are sensitive or resistant to negative regulation.
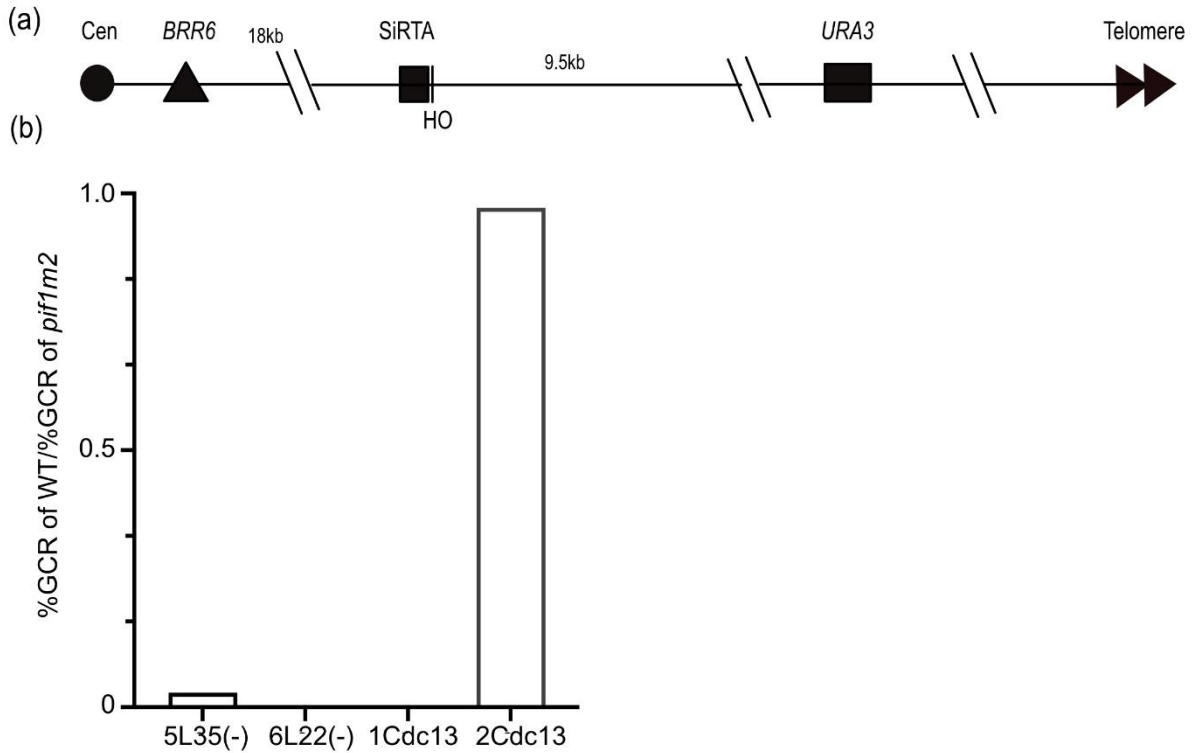
Figure 18. SiRTAs circumvent negative regulation by Pif1. a) Diagram depicting the structure of chromosome VII and the strategy for mapping GCR events resulting from a DSB generated by the HO endonuclease adjacent to the inserted SiRTA/putative SiRTA sequence. Locations of the essential gene BRR6 and the sequence to be tested (SiRTA) are shown as is the HO cut site and URA3 marker. This assay is the same as the assay depicted in Figure 2a except the HO cleavage site is located next to the SiRTA rather than 2kb away. b) Sequences that stimulate *dnTA* are subject to less negative regulation by Pif1 than sequences that do not stimulate *dnTA*. The ratio of the overall GCR frequency of the WT PIF1 strain to the overall GCR frequency of the pif1m2 strain are shown for the indicated SiRTAs. A ratio of one indicates that the sequence is fully resistant to Pif1 negative regulation Data of GCR Frequencies can be found in table 12.

5.1.2 Naturally occurring GCR rates at TG-repeat SiRTAs

The work done to characterize SiRTAs currently primarily characterizes SiRTAs in a manner that induces a break near a SiRTA and then selects specifically for events that would occur within the SiRTA. While these experiments are useful for answering questions about how

SiRTAs may function, the question of the extent to which these sites increase genomic instability remains. Initial evidence that SiRTAs may influence genomic instability comes from the Kolodner lab where they observed genomic instability at 5L35(-) (Myung *et al.* 2001; Stellwagen *et al.* 2003; Zhang and Durocher 2010) (see Chapter 1). The Friedman lab also demonstrated that SiRTA 2R780(-) also likely contributes to genomic instability (Hoerr *et al.* 2023). Currently, it is unknown whether other SiRTAs contribute to genomic instability. However, considering that 5L35(-) presents as a weaker SiRTA in our assays, we hypothesize that other SiRTAs (particularly those that stimulate high levels of *dn*TA) contribute to genomic instability. To address this issue, we have begun investigating the spontaneous GCR rate of SiRTA 6R210(+). This SiRTA was chosen because it has the highest CATHI score of all predicted SiRTAs and the highest rate of *dn*TA of those SiRTAs we have assayed. We hypothesize that a yeast strain with 6R210(+) at the chromosome VII test site will experience a higher frequency of spontaneous terminal truncation than a strain with no SiRTA added to the chromosome VII test site (WT). To measure the spontaneous frequency of GCR events, we sought to eliminate the HO cleavage site in the chromosome VII test strain. To identify such strains, cells surviving HO cleavage were patched on medium lacking uracil to identify strains in which the HO cleavage site was mutated but the chromosome end was retained. PCR amplification and Sanger sequencing confirmed that the HO cut site was mutated. For both 6R210(+) and WT strains, two candidates that mutated the HO cut site from the galactose assay were chosen (labeled A and B). For the WT strain, only variant A was tested due to time constraints.

To test whether insertion of 6R210(+) causes high frequencies of GCR at the chromosome VII test site, we initially grew both yeast strains in liquid medium lacking uracil to

ensure each strain retained *URA3* function. Initially, cells were plated at high density on rich medium for two days to allow cells an opportunity to lose the *URA3* marker and were subsequently replica plated to medium containing 5-FOA. However, selection on 5-FOA was insufficient to obtain unambiguous colony counts by this method. As an alternative approach, cultures grown in medium lacking uracil were separately plated on medium containing 5-FOA (to select for truncation events) and on rich medium (YPD) to determine the total number of colony-forming units. The frequency of mutation at *URA3* was calculated by dividing the number of colonies that grew on 5-FOA by the number of colonies that grew on rich medium after adjusting for dilution. In these experiments, performed by undergraduate Honors student Tristen Gittens under my supervision, there was no significant difference in the generation of 5-FOA-resistant colonies between the two strains (data not shown). However, this experiment is not ideal because cells transferred immediately from medium lacking uracil onto medium containing 5-FOA may not have time to dilute or degrade the Ura3 protein product before encountering 5-FOA.

In the future, we will perform the Luria-Delbrück Fluctuation assay to determine mutation rates. Using this method, cultures will undergo an initial period of growth in rich medium to allow cells to lose the *URA3* marker before being plated onto 5-FOA. We will adapt the method as described (Luria and Delbrück 1943; Lang 2018) which involves analysis of many cultures in parallel (at least 72 colonies per strain). Once established, we can examine the effect of other SiRTAs or of a mutated version of 6R210(+) that reduces activity.

Determining the extent to which SiRTAs cause increased frequencies of GCR in the absence of an induced DSB will give insight into the possible evolutionary forces acting on these sites as well as the impact of these sites for genome stability. If these sites do incur higher

frequencies of GCR, it would beg the question why they have not been selected against. If they do not incur higher rates, which is highly unlikely due to what is observed in SiRTA 5L35(-), then it would suggest these events are indeed very rare and have not been selected against for that reason. Either way, this information will be useful for understanding the impact of SiRTAs in the genome.

5.1.3 Improving the CATHI Algorithm

In its current iteration, the CATHI algorithm predicted whether the tested sequences could stimulate *dn*TA in response to a break with ~95% accuracy. This provides a novel tool for predicting the distribution of SiRTAs within the genome and allows for insights into where SiRTAs are located and where they might have the most impact on genomic instability. However, the CATHI does not accurately predict how well a sequence will function as a SiRTA, especially between the scores of 20 and 25. It would be useful to have the ability to predict how well a SiRTA can function because this additional information would allow one to prioritize which sequences to test or use sequences with varied predicted efficiencies to ask questions about the mechanism of *dn*TA. The current version of the algorithm only considers similarity to yeast telomere sequence when scoring. However, past work by the Friedman lab and work I've presented here suggest that both Cdc13 binding and distance between potential Stim and Core sites influence *dn*TA. Updating the CATHI to award bonus points for sequences matching the GxGT sequence, which is known to stimulate binding of Cdc13, could allow the algorithm to be more predictive (Anderson *et al.* 2003; Obodo *et al.* 2016). Accounting for the distance between TG-rich sequences is another possibility. For example, bonus points could be awarded for strings of telomere-like sequences that are 20 bases or longer so that sequences lacking spacers are favored.

5.1.4 Investigating the role of Cdc13 at SiRTAs in non-subtelomeric regions

One surprising observation from the analysis done on the CATHI results was that the distribution of SiRTAs appears random, apart from subtelomeric regions. We hypothesized that SiRTAs would be enriched in nonessential regions and depleted in essential regions because SiRTAs act as hotspots for gross chromosomal rearrangements and would thus likely cause cell death in the essential regions of the genome. Due to the potential deleterious effects of SiRTAs in essential regions, we hypothesized that they would be selected against. We did not observe depletion of SiRTAs in essential regions, and we also found that SiRTAs are more abundant in the yeast genome than what would be expected by chance. Together, these observations beg the question: why aren't SiRTAs selected against? Are events at SiRTAs rare enough that they simply do not sufficiently affect fitness or do SiRTAs serve a purpose outside of *dn*TA?

Work from the Friedman lab has consistently identified Cdc13 as an important factor in stimulating *dn*TA in response to a DSB (Obodo et al. 2016; Hoerr et al. 2023; Ngo et al. 2023) One possibility is that the instability caused by SiRTAs is rare enough to be counterbalanced by another function that is under selection. Conceivably, this function may not involve *dn*TA but could require binding by Cdc13. Cdc13 has multiple roles at the telomere. It not only facilitates the recruitment of telomerase but it also facilitates the recruitment of the CST complex to cap the telomere and prevent resection ongoing resection (Rice and Skordalakes 2016; Ge *et al.* 2020). At telomeres, Cdc13 is also important for synthesis of the C-rich strand following telomerase elongation of the 3' end (Chandra *et al.* 2001). Perhaps the abilities of Cdc13 to limit resection and/or promote resynthesis of the resected strand make SiRTAs useful to the cell. These various roles of Cdc13 are well understood at the chromosome end but not at SiRTAs. Parsing out the various roles of Cdc13 at SiRTAs would help elucidate what the true roles of SiRTAs may be.

I hypothesize that SiRTAs may help facilitate fill-in synthesis. One way to approach this question is to utilize already known temperature sensitive mutations of Cdc13. One allele of Cdc13 that is of interest is the temperature-sensitive allele *cdc13-1*. This allele is of interest because in a study looking at the role of Cdc13 in preventing resection, Resnick and colleagues found that at restrictive temperature, *cdc13-1* mutants experience extensive resection at telomeres (estimated to extend 15 kb). However, when cells were shifted to the permissive temperature, telomeres were rapidly (<1 hour) restored to their normal structure (Westmoreland *et al.* 2018). After the extensive resection that occurs at the non-permissive temperature, it is difficult to imagine that once Cdc13 function was restored it would only act on telomeres to regenerate the resected strand.   The ability to restore telomeres after extensive resection could be interesting to study in the context of SiRTAs as we could look to see if there is a difference in resection or restoration at SiRTAs that could make a difference in survival. More specifically, we could examine different chromosome termini and determine whether fill-in rate or efficiency is correlated to the number or strength of the predicted SiRTAs. Or we could mutate SiRTAs on a single arm and determine if fill-in synthesis is reduced.

The *cdc13-1* point mutation presents a useful tool for stimulating resection and observing the ability of cells to recover (Garvik *et al.* 1995). With this tool we could elucidate whether SiRTAs affect C-strand fill in as this allele could be used to measure the accumulation of C-strand ssDNA, which would indicate a lack of fill-in synthesis, at SiRTAs.

One process that may also be active at SiRTAs is the break induced replication (BIR) repair pathway. BIR is a repair pathway that utilizes wherein a DSB invades a homologous sequence that is then used as a template for unidirectional replication to repair a one-ended DSB (Kramara *et al.* 2018). BIR has been observed to act at telomeres and allow survival when

telomerase is non-functional (Reviewed in Bonnell *et al.* 2021). If SiRTAs are similar enough to telomeres to bind Cdc13, I hypothesize that SiRTAs are also similar enough to promote the strand invasion required for BIR and are therefore favored sites at which BIR may initiate. It is possible that Cdc13 may also help facilitate this process via its role in fill in synthesis (Figure 19).
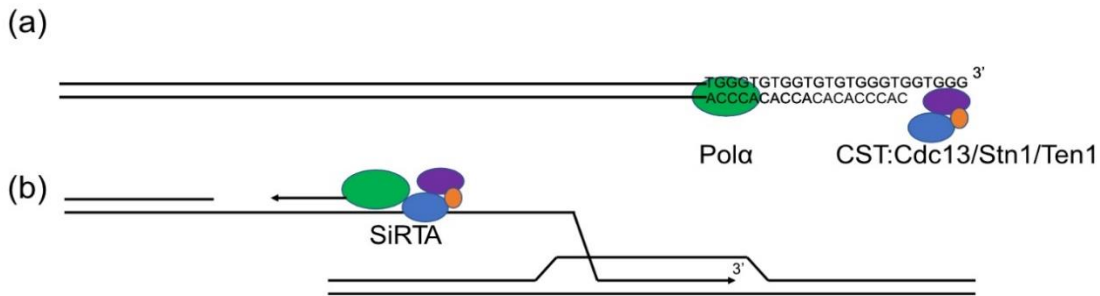
Figure 19. Model of Cdc13 facilitation of break induced repair at the SiRTA. a) Cd13 (purple oval) binding to the telomere in complex with Stn1 and Ten1 (blue and orange ovals) facilitates C-strand fill-in via polymerase α. b) At a SiRTA when the DNA is made single stranded the CST complex (blue purple and orange ovals) are able to bind and facilitate fill-in via polymerase α during BIR.

I also hypothesize that the presence of SiRTAs in the region of the chromosome that must be replicated by BIR increases the efficiency of repair. To address whether the efficiency of BIR is affected by SiRTAs, we could utilize the modified chromosome fragmentation vector (CFV) described by the Symington lab. The pCF2/D8B vector contains a unique sequence to target recombination, a *URA3* marker, and telomeric sequence. When the vector is linearized and transformed, researchers observed recombination between the sequences at the ends of the CNV and duplications of the telomeric sequences which allowed for the formation of a stable chromosome fragment (Davis and Symington 2004). For our purposes, we could replace the non-telomeric sequence with a SiRTA and see if BIR still occurs to create a stable fragment (Figure 20). If so, we could then either mutate SiRTA sequences or use various alleles of Cdc13 to determine what factors play a role in BIR. We could also add Cdc13 sites to determine if that is

enough to facilitate BIR.
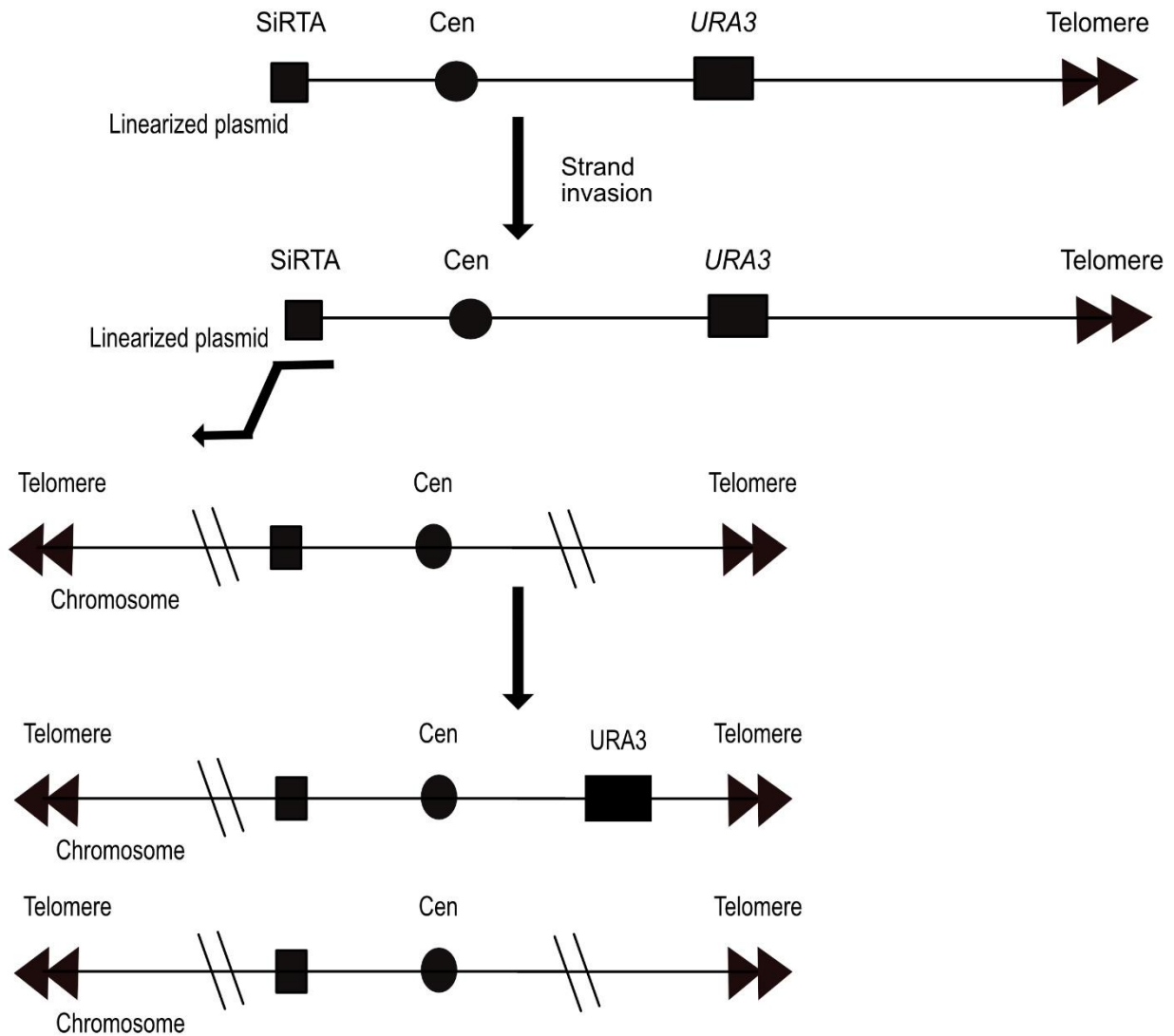
Figure 20. Chromosome fragmentation assay to assess rates of break induced repair. When a plasmid containing a telomere a SiRTA and a *URA3* marker is transformed into the yeast and linearized the fragment will be lost unless stabilized through BIR via a strand invasion event with a chromosome that contains homology with the SiRTA. If BIR occurs the corresponding chromosome will be stabilized with a telomere and can be selected for using the *URA3* marker.

Another way to address this question is to identify variants of Cdc13 or other telomere-related proteins that are able to maintain telomeres but cannot bind SiRTAs. One way to do this

is to develop a random mutagenesis screen to identify mutant strains in which telomere addition can occur at endogenous telomeres, but not at SiRTAs. One possible screen could utilize a modified version of the plasmid developed by the Lundblad lab to identify *EST1,* described in section 1.1. The original circular plasmid contained a *LEU2*-marker and inverted repeats of the *Tetrahymena* telomeric sequence separated by a *URA3* marker. A stable linear chromosome would form when rare spontaneous breakage events within the *URA3* gene linearized the plasmid and then the subsequent extension of the *Tetrahymena* telomeric repeats with yeast telomere sequence would stabilize the mini-chromosome. We could modify this plasmid by replacing the one of the *Tetrahymena* telomeric sequences with yeast telomeric sequence and replacing the other *Tetrahymena* sequence with a sequence containing a high activity SiRTA such as 14L35(-), followed by a selectable marker like *TRP1* and yeast telomeric repeat sequence. This latter cassette should be oriented such that the SiRTA is adjacent to the *URA3* marker (Figure 21). When the WT strain transformed with this plasmid is plated on media lacking leucine and tryptophan and containing 5-FOA, we would expect a small number of colonies to arise where spontaneous breakage within *URA3* leads to telomere formation at the telomeric seed sequence on one end and the SiRTA on the other end. Mutants that generate colonies at a reduced rate might be completely deficient in telomere addition. To identify those mutations that selectively affect telomere addition at a SiRTA, candidates that fail to generate colonies in the initial screen could be further screened for those that still generate colonies on medium lacking leucine and containing 5-FOA (i.e. telomere addition fails at the SiRTA, but can occur at the more internal telomeric tract). Initially we could rely on spontaneous breakage to linearize the plasmid. However, if too few colonies grow, we could also add a cut site to one either side of the *URA3* marker or into an artificial intron within *URA3* to increase linearization of the plasmid. This

experiment would allow for the identification of mutants that can extend telomeres but are ineffective at SiRTAs which would offer insights into what factors are specific for SiRTA function. This screen could also be used to specifically target mutagenesis of Cdc13 and concentrate on the identification of mutations in Cdc13 that alter its DNA binding specificity. Such mutations may separate endogenous telomere maintenance from SiRTA function. Simultaneously, while the screen is being developed, we could also determine whether any natural variants of Cdc13 (from other species) might show variable specificity. The caveat with this approach would be that we would first have to elucidate whether there are differences in the telomeric repeat and/or the sequence of Cdc13 DBD might be useful to explore so we could only do this with species where this information is readily available.
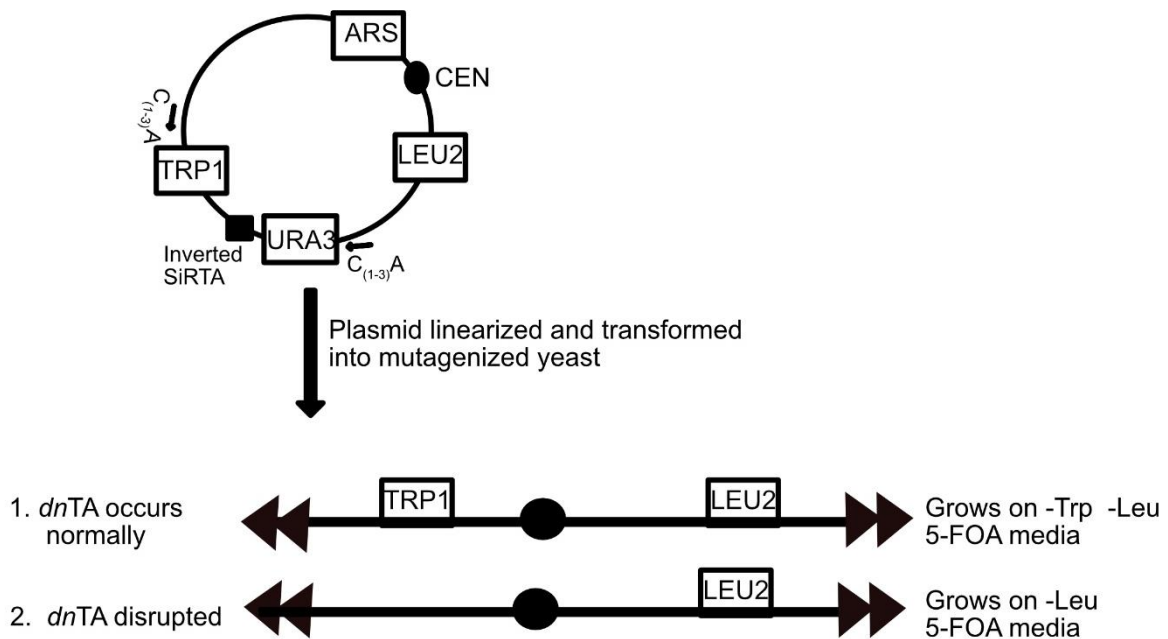
Figure 21. Screen to identify factors affecting *de novo* telomere addition. A plasmid containing inverted telomere repeats, a SiRTA and various selectable markers would be transformed into randomly mutagenized yeast. Rarely, once transformed the plasmid may linearize within the URA3 marker. If *dn*TA is able to occur normally at the SiRTA, the chromosome is stabilized and the selectable markers (excluding *URA3*) are maintained. If *dn*TA does not occur but telomerase can still act at the internalized telomeric repeat the *TRP1* marker is lost but the *LEU2* marker is maintained. Selecting for candidates that grow on -Leu 5-FOA media but not -Trp -Leu 5-FOA media could reveal candidates that are deficient for *dn*TA but able to maintain normal telomeres.

Overall, the work presented in this thesis suggests that the overrepresentation of sequences in the genome that can function as SiRTAs is likely not due to their ability to stimulate *dn*TA *per se*. Perhaps *dn*TA is simply a consequence of these sequences that we were able to detect but not the main feature or function of these sequences. In the future elucidating what other roles these sequences play in the genome would be important for fully understanding the impact of these sites.

5.1.5 Testing Cdc13 binding at a variety of SiRTAs in vivo

Although it is known that Cdc13 binding is important for SiRTA function, the extent of binding required is unknown. This thesis presented the use of fluorescence polarization to test association of the DNA binding domain of Cdc13 with a variety of sequences. Work from the Friedman lab and my results shown in Chapter 3 demonstrated that sequences that stimulate *dn*TA also bind Cdc13 *in vitro* (Obodo *et al.* 2016). While this insight was valuable, there were a few limitations to the assay. One limitation is that only 75 bases of a sequence could be tested at a time. Sequences tested in the HO cleavage assay are typically 300 bp so we may fail to identify the full scope of Cdc13 binding in our assay. Additionally, because full length Cdc13 is more difficult to purify, I monitored binding by the Cdc13-DBD, which has been shown to bind with an affinity and specificity very similar to the full-length protein. However, this assay almost certainly cannot fully recapitulate the behavior of the full-length protein in a cellular context. This means that we are likely not getting the full picture of how much Cdc13 binding is required for *dn*TA and we may be underestimating Cdc13 binding for some of the SiRTAs that we tested.

To get a more accurate picture of Cdc13 binding at these sites, we could do chromatin Immunoprecipitation (ChIP) assays to assess Cdc13 binding at the SiRTA in response to a DSB.

ChIP is an *in vivo* method of assessing protein binding wherein cells are lysed and protein-DNA interactions are fixed with formaldehyde. Following shearing to produce short chromatin fragments, an antibody is used to isolate a protein of interest. To identify DNA bound to the protein of interest, one can use techniques like qPCR to quantify the amount of a specific sequence that is immunoprecipitated. ChIP seq a strategy where the immunoprecipitated sequence is sequenced is a strategy to identify all sequences that are enriched. In the past, ChIP directed against Myc-tagged Cdc13 has shown that Cdc13 associates with sequences surrounding a DSB when homologous recombination is inhibited (Oza *et al.* 2009; Epum *et al.* 2020). For example, ChIP has been used to examine the association of Cdc13-Myc with SiRTA 9L44(-) following induction of HO cleavage (Epum *et al.* 2020). These experiments were done in the context of deleting Rad51 and/or Rad52 and the extent of binding by Cdc13 was shown to correlate with SiRTA efficiency in the different genetic backgrounds. While this prior work demonstrated that ChIP is a viable method for determining the relative amounts of Cdc13 binding *in vivo* at a SiRTA in response to a DSB, the extent to which Cdc13-Myc association depended on the presence of SiRTA 9L44(-) was not determined.

A study utilizing ChIP to survey the Cdc13 binding of many different SiRTAs with varying CATHI scores and ability to stimulate *dn*TA will elucidate how well the extent of Cdc13 association *in vivo* correlates with SiRTA efficiency. Furthermore, ChIP can be applied to determine genetic backgrounds that might affect Cdc13 binding. For example, in section 4.1.1, I discuss the possibility that sequences functioning as SiRTAs do so by binding sufficient Cdc13 to bypass negative regulation by Pif1. I hypothesize that sites with low sensitivity to negative regulation by Pif1 would bind more Cdc13. To determine whether this is the case, ChIP could be used to determine *in vivo* Cdc13 binding in a variety of sequences with varying sensitivities to

negative regulation by Pif1. Currently, ChIP has only been used to determine Cdc13 binding at one unmodified SiRTA (9L44). Using ChIP to study Cdc13 binding at multiple SiRTAs with varying efficiencies would further elucidate whether a sequence's affinity to bind Cdc13 is sufficient to confer resistance to negative regulation by Pif1 and more generally if it is enough to predict SiRTA function.

Questions regarding the mechanism of *dn*TA at SiRTAs could also be answered by combining ChIP experiments with mutations that alter SiRTA function. In chapter 1, I discussed how the structure of a SiRTA consists of a Stim, a Spacer and a Core sequence. Removal of the spacer sequence increases rates of *dn*TA. Increased activity as a result of moving the Stim and Core sequence in closer proximity might reflect an increased ability to bind Cdc13. This model is supported by experiments that suggest that Cdc13 may bind DNA as a dimer (Rice and Skordalakes 2016). Alternatively, the relative location of Cdc13 binding sites might affect the efficiency of *dn*TA independent of the absolute amount of Cdc13 bound. If the first model is correct, then Cdc13-Myc binding measured by ChIP should increase when the length of the spacer sequence is decreased. Overall, the use of ChIP to study Cdc13 binding on varying SiRTAs would be a useful tool for understanding telomere healing and this contributor to genomic instability.

5.1.6 Utilizing the algorithm strategy to identify SiRTAs in other species

The focus of my work has been to elucidate the distribution of SiRTAs in the *S. cerevisiae* genome. I have previously discussed ways we can continue to study SiRTAs in *S. cerevisiae* and the mechanisms behind SiRTA function. However, telomere healing as a response to DSBs has been observed in other species and SiRTAs are likely present in species other than *S. cerevisiae* (Hannes *et al.* 2010; Ouenzar *et al.* 2017; Nevado *et al.* 2022). The CATHI

algorithm is unlikely to be useful in species distantly related to budding yeast. While most eukaryotes maintain TG-rich telomeres with telomerase, the characteristics of telomeres vary between species. Telomere length, sequence, and even the proteins that maintain the telomeres are all characteristics that vary between the telomeres of different species. For example, *S. cerevisiae* telomeres contain a heterogeneous TG_{1-3} pattern, whereas many mammalian telomeres (including those of humans) are composed of perfect TTAGGG repeats (Moyzis *et al.* 1988). This variability poses a challenge to the study of *dn*TA because, while the basic concept of *dn*TA may be similar from species to species, the specific mechanisms and details may differ. The current technology to study *dn*TA in species with more complicated telomeres and telomere maintenance, like mammals is lacking. Long read sequencing after telomere isolation has made it possible to identify *dn*TA in cancer cells (Kinzig *et al.* 2022) however it is still unclear how factors like the interstitial telomeric repeats (ITS) may affect *dn*TA. Furthermore, most somatic cells will not encounter this issue because telomerase is not expressed. Here I focus primarily on applying the CATHI to closely related yeast species to get a better understanding of whether other species also have SiRTAs.

In chapter 3, I discussed the observation that, outside of the sub-telomere, the distribution of SiRTAs is random, but the number of SiRTAs is greater than what would be expected by chance. These observations suggest that the role of SiRTAs in *dn*TA is unlikely to drive their abundance in the yeast genome. One question that remains is whether sequences with the capacity to stimulate *dnTA* are also overrepresented in other yeast species. Another question is whether the location of SiRTAs is conserved across yeast species. To answer these questions, a map of putative SiRTAs in other yeast species is essential. The challenge with creating this map in other species is that the CATHI relies specifically on the nucleotide pattern of telomeres in *S.*

cerevisiae. Creating a similar map in other yeast species may require adjustments to the algorithm if telomeric nucleotide patterns are different. To accomplish this, one would have to adjust the algorithm to match the pattern of telomeres in the species being mapped and then the penalties and bonuses of the program would have to be adjusted based on *in vivo* testing. It is unclear in other yeast species how much a sequence could differ from the telomere sequence and still be recognized, so the endeavor to adjust CATHI to be able to test other yeast species would contribute to answering the question: what are the minimum requirements to be a telomere?

Yeast species that are closely related to *S. cerevisiae* like *S. paradoxus* would be good candidates to start this mapping because the CATHI would likely not have to be adjusted very much and many of the tools used to validate the CATHI could likely be used in these species as well (Naumov *et al.* 1992). Any effort to adapt the CATHI to identify SiRTAs in other species would be useful in the study of telomere healing and the study of genomic instability as a whole.

In the meantime, while efforts are made to develop the CATHI in other closely related species, another strategy for identifying SiRTAs is to see if any of the *S. cerevisiae* SiRTAs are conserved in other species. If so, then these sequences could possibly be used to help validate the CATHI in other species as we could check to see if these conserved sequences are identified in the adapted versions of the CATHI. In this section I describe preliminary work to see if a specific type of SiRTA (TG-dinucleotide repeats) are conserved in other species of yeast.

The CATHI program identified 20 SiRTAs containing TG-dinucleotide repeats in non-subtelomeric regions. SiRTAs consisting primarily of TG-dinucleotide repeats may have a higher rate of *dn*TA than other SiRTAs because the repeats contain multiple copies of the Cdc13 GxGT

binding site and these SiRTAs also lack a spacer sequence. In Chapter 3, I discussed the characterization of 6R210(+), the longest TG-dinucleotide repeat and the sequence with the highest CATHI score outside of the subtelomeric region. This characterization revealed that 6R210(+) stimulates *dn*TA at a high rate. *In vitro* fluorescence anisotropy assays showed that this sequence binds an amount of Cdc13 consistent with each molecule of Cdc13-DBD occupying ~12 nucleotides (6 TG repeats). We hypothesize that the other TG-repeats in the genome with a CATHI score above 20 will also stimulate *dn*TA. One question that remains is whether the TG-repeats are conserved in other species and whether these repeats also stimulate *dn*TA. Due to the high rate of *dn*TA, studying these sites will also allow us to investigate the potential role of these sequences in creating GCRs and whether this role causes negative selection. In general, it is known that dinucleotide repeats are subject to expansion and contraction as a result of replication slippage and unequal (Sinden and Wells 1992; Pâques *et al.* 1998). If we find that the longer repeats are conserved between multiple yeast species, it might suggest that these sequences are advantageous in some way. If we find great variability, that might suggest that these sequences are unstable but may be relatively neutral in terms of cellular fitness.

To study these possibilities and determine whether TG-dinucleotide repeats are conserved between different species of yeast, Liraz Stilman, an undergraduate in the lab, selected six different *Saccharomyces* species to compare to *S. cerevisiae.* These species were chosen based on the availability of their relatively complete genomes. The chosen species and strains used for this study are as follows and species are listed in order of relatedness to *S. cerevisiae: S. cerevisiae* (strain S288C), *S. paradoxus* (strain CBS432)*, S. mikatae* (strain IFO1815)*, S. kudriavzevii* (strain IFO1802)*, S. bayanus* (strain FM677)*, S. castelli* (strain NRRL Y-

12630)*, and *S. kluyveri* (strain NRRL Y-12651) (Naumov *et al.* 1992; Cliften *et al.* 2003; Hittinger 2013). For this analysis, 16 of the TG-dinucleotide repeats were identified as a SiRTA by the CATHI and one TG repeat that was not identified by the CATHI. This sequence was not identified CATHI because the CATHI required that a string starts with a G and not a T so when this 20-base repeat was scored it was given a score of 19 instead of 20 but we decided to include it because it is a 20 bp repeat (an interesting caveat of the CATHI that may be addressed in later iterations of the program).

The TG dinucleotide repeats were compared across each species to determine the conservation of the sequence. To measure conservation, the following equation was used: Ratio of conservation= (number of bases conserved)/(Total number of bases in the TG-dinucleotide repeat). The closer to one the ratio is, the more conserved the repeat. We found that the most closely related species to *cerevisiae* (*paradoxus)* has the highest mean conservation ratio (0.4105), and the species that are the least related to *cerevisiae* (*castellii, and kluyveri),* have the lowest mean ratios (0.06, 0.26, respectively) (Figure 22). Overall, each species shows at least partial (>50%) conservation of at least one of the TG-repeat sequences. This result suggests that there is some level of conservation of TG repeats. To determine whether this result is what would be expected for this type of repetitive sequence, we would need to compare the variability of the TG repeats we observed here to that of other dinucleotide repeats. knowing of the variability of other dinucleotide repeats. Another caveat of this work is that because we first identified long TG-dinucleotide repeats in *S. cerevisiae*, it is perhaps not surprising that we only observed reductions in the repeat length in other species. It will be important to start with long repeats in other species to determine whether *S. cervisiae* has a propensity for long TG-repeats or whether this result is biased by the experimental approach. In the future we plan to do further

analysis to elucidate these issues and determine whether the TG dinucleotide conservation is significant.

Interestingly, 16L486(-), a 26 bp TG repeat, is the only TG repeat analyzed that is at least somewhat conserved in each species. An interesting future experiment would be to mutate the *cerevisiae* 16L486(-) to the sequences in the other species and then test to see the extent to which these changes affect the ability of the sequence to stimulate *dn*TA. Although this would not test exactly whether the sequence changes affect *dn*TA in the context of each species, it could give some insight into whether the changes in the TG repeats from species to species make a difference in that sequence's ability to stimulate *dn*TA.

Overall, determining whether SiRTAs are evolutionarily conserved will require analysis of SiRTAs and the genomes of other species. This preliminary work suggests that at least some of the TG-repeat sequences are conserved, although there is still not enough evidence to determine whether this is different than what would normally be expected. One thing to consider when answering this question is the varying telomere sequence from species to species so an important caveat of this work is distinguishing between sequence changes that could be due to the deleterious effects of *dn*TA and changes that may improve *dn*TA rates from species to species. Grappling with these questions and challenges will be an interesting field of study in the

future that could offer insight into the true impact of SiRTAs on genomic instability.
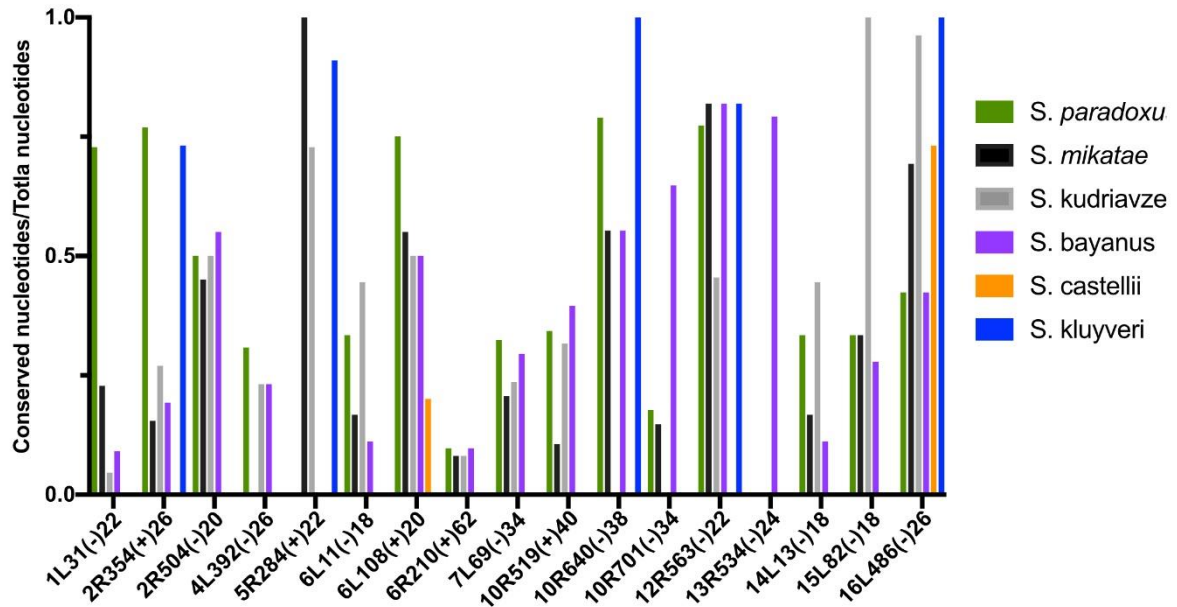
Figure 22. The conservation of S. cerevisiae TG dinucleotide repeat sequences in closely related Saccharomyces yeast species. The amount of conservation is expressed as the number of bases conserved in each TG repeat for the respective species divided by the original number of bases in the repeat in S. cerevisiae. The closer the ratio is to one the more conserved the sequence is in that species. The TG repeat is labeled by the naming convention used throughout the paper except that the number following the parentheses corresponds to the number of bases of that TG repeat. Data for Figure was compiled by Liraz Stilman.

5.2 Concluding remarks

The work presented in this thesis answers important questions regarding telomere healing in *S. cerevisiae* by creating the first comprehensive map of putative SiRTAs within the yeast genome. This map and the new tools used to generate and validate this map demonstrate novel insights into how SiRTAs are distributed. SiRTAs are disproportionately found in the subtelomeric X and Y' elements, however outside the subtelomeres, SiRTAs are randomly distributed. This random distribution argues that SiRTAs outside subtelomeric regions have likely not evolved to rescue a chromosome when a break fails to be repaired. Intriguingly,

117

SiRTAs are more common in the yeast genome than would be expected by chance, raising

questions about the potential role of these TG-rich sequences. One exciting possibility is that the

prevalence of sequences predicted to function as SiRTAs is tied to their ability to bind Cdc13.

One way to accomplish this would be to produce strains wherein Cdc13 retains function at

endogenous telomeres but is no longer able to bind at SiRTAs and then we could possibly ask

whether there is a fitness advantage to these sites, perhaps in situations where cells are subjected

to chronic stress. Overall, the work presented in this thesis and the future work made possible by

these insights will lead to an increased understanding of this source of genomic instability and

the mechanisms that govern it.

APPENDIX

Appendix A: Tables

Table 1. Sequences tested for the ability to stimulate *de novo* telomere addition

SiRTA	Used in CATHI calibration	CATHI score	Sequence 5' to 3'
12R1009(+)	yes	4	ATTTTCTAATGACGCAGGATTGAAAAGCAAATAAGGAGT TGGAGATTCGTTTTCTTGACCTGGGTCCTCATTTTGATTTT GGCCTCGATTGTCATTTTGGCCTTGTTGCTGATTTTGATGT TGTTGATTTGCTGAAGAATAAAGATTGACTTGAGGCAAAT CAGTTATCTCTTCAAATTGATAGAAATCCTTCCATTGGAC TTTCCGGTTGAATCCCATTTGCTTCAAGGTCATGATATGC GAATTTAAATTTGGATTCATGTCTAAGGCCTTGGATATAG AATTTATAGAGTCATTTAT
12R130(-)	yes	14	TTTCACTGGCTTGAGAGCTGCTACTTACCCTACTCAAATTC ATATCTTGGATAAGATGGTCAGACGATACTTGTACAGTAG GCCCTCGCTGCTGCGACTGCGACTGCGACTGTGACTGTGA CTGTGACTGGGGTTGTGGTTGTGGTTGTGACTGTTGTTGTT GTTGCTGCTGTTGTTTTTTCTGCCTGGAAGTATCTTTCACA GCATCTCCTTTTTCCTGTACAATGCCACGACTGCGACGGT TATCTTCATCGTCGATCTCTTCATCTTCGTCAACCTCGTAT TTTTCCTCGTAAAGAT
12L23(-)	yes	11	CAGAGTAAAAATGTTCAGCACGTTGATCTTGAAAGCAGA CGGAAAGTCTTATTAGATGAAATTAATGCTGAAATCTCGA AACTGTTCGCTCAGCTACCATGTGCATGTGTGTGGTGTTG GTATATTTTAGTATACTTCTTGATAAAGGTACCAAAGTAC TGTATACACATGATAATAACTGGTTAGTCATCCTGGTGTT TTACTTCTTGGCATTCGGCTCGATCATGAGGATTTCGGGC CGAACCTGCTCCGAAGAAATGAAGAAATCTTTCGGCCAG GCATCATATCACAATGGGTTGC
12R291(+)	yes	11	GTTTGGTATGAAACATAAGGATGGTACCCACTATTTGTCA AACGTCAGAACAGGTTTAATCGTCGCCATTTTCAATATTG GCTGTGCCTTTGGTGGTATTATACTTTCCAAAGGTGGAGA TATGTATGGCCGTAAAAAGGGTCTTTCGATTGTCGTCTCG GTTTATATAGTTGGTATTATCATTCAAATTGCCTCTATCAA CAAGTGGTACCAATATTTCATTGGTAGAATCATATCTGGT TTGGGTGTCGGCGGCATCGCCGTCTTATGTCCTATGTTGA TCTCTGAAATTGCTCCAAA
12R330(+)	yes	22	TTGATATATATTTTACTATATACTTTTGCTGTGTATTTCTA TATGAGGCTAGTTGGTTGGTGGGAACTGGAAGTGAAAGT TCTGTTCTTTAGGGTGCAGCGGTTGCTGTTGCTGTTGCTGT TGCTGTTGCTGCTGTGGTTGGATTGATGCTTGTGATGTTTG TGGTGCCTGTGGTGCTTGTGGTACAGTTTTGCTGCCAGTTT GCTGCGAGTATGTAATTATGGGGTGAGCGGCATTTACATT AGGTGCTGGATTGATGGCTGGATTTGTTTTTTCGTAAGCG GCGATACTAGGCGTTGT
12R500(-)	yes	9	GAAAGCAGTCCCACTTTCTACAGAATATTGTGGAACATAC GTGGGAGTAGGCCGCGCACTTTGCGGAGAAGGTCTATAA TCTTGCATTTCATAATCATAGTCGTCATCATTTCCCCCATC GTTGCCGTTGTTGTTGTTGTTACTGTTATTATTATTATTATT

			ATTGTTATTATTGTTGTTAGTACCACCAGAGTGATGATGTGAAGGTTGTGATGAAGAGCGTCCTTCGTCGGATGACTGCTTATTAGATTGTTTCCTAGAACTATCATGTGTATGAACAGATTCACTGGACAGAAAGTC
12R555(-)	yes	5	CAAAGGTTGTTGAAAATCAGCGGGCTGTTGTTGTTGTTGTTGCATATTAGCGGCTGTTTGTTGCTGTTGCAAGGCAGAAGCGGCGGCCGCAGCTTGTTGTTGCTGCTGAAAATATTGTTGTTCAGCAAGCTGCTGTTGTTCTTGCTGCTGGTCCAGCCATCTTTGCTGAGCCATTGCTTGTTCCTGGTCCTGTTGGTACTGATACTGTAAGTATTCATCTTGGGAGATTGGATTACCGAAAATGTCGTAGTACGCTGGTTGTTGTTGTGGTTGTTGTTGTTGTAAAGGAGCTTGAAATTG
12R700(+)	yes	8	AGGGTCATTCAAAAGGCTCGCGGATTGGCAGATAGCATTATTGTTGTTTATGTTGTTCGTACTGGCCGATGTATCGTTAGTGGCCGCATTAACGTTGCTGCCGTTATTGTTTGTGTTCTCCAGGGGTTTGTTGGAAGGAGATGGCAGCGGCAGTATGACCATATTGCTGCTGTTCGGTCTAGAGTAATCAGAAGTGTTGTTACTGAACAAATGAGAAACCTTCTTGTACTGGGATAATTCGTAGTTGTACTGGTTGCATCCCGGCGATAAATCGAGCAAAGTCTCCAAACTTTTTAGTTT
12R874(-)	yes	11	ACGGGCTCATACGCAGTGTGTTGTGATCCTATTGATGGCTCCTCAAATTTGGACGCCGGTGTCTCCGTTGGAACTATCGCGTCTATATTCAGACTGCTACCAGACTCATCAGGTACTATAAACGACGTACTGAGATGTGGTAAAGAAATGGTAGCCGCTTGCTATGCCATGTACGGATCCTCTACGCATCTAGTATTGACATTGGGTGATGGAGTTGATGGGTTTACCTTAGACACAAACTTGGGCGAATTCATCTTGACTCATCCTAACTTAAGAATTCCGCCTCAAAAGGCCATCTAC
14R131(+)	yes	25	TTTAATTTCTTGTGTGGACGAGGATACTTTTCAAATAATGAGGGCGATTGAGGAAGTAATGGTGATGTTACAGCTGATTTGATTTGCGATGCCGGCATCATAGGTGGAGCTGGCGGAATGACGCCATCAGTAGATGAAGATAATACAGATGGTAGCGGCGGCGGTGGTGGTGTTTCTCCCTTGGGTTTGCCGAACAATGCCATGGGTGGAGGAGGAGGTGGTGGTGGTGGAGGGGGAGGAGCAGGAGAATCCCCAGTAGTTTCTTGCTTGACAGTGTCGTCTTCCGACTTTTTCTCTTTCT
14R235(+)	yes	9	ATATTGTAGCCCGTTGCGCCGGTGGTAACAATGCCGGGCATACCATTGTTGTAGACGGTGTTAAGTATGATTTCCATATGTTACCATCTGGTTTAGTCAACCCAAACTGCCAAAACCTTTTGGGTAATGGTGTTGTTATTCATGTTCCATCTTTTTTCAAAGAGTTGGAAACCTTGGAAGCTAAAGGTTTGAAGAACGCAAGGAGTAGATTATTTGTTTCTTCCAGAGCACATTTAGTCTTTGACTTTCATCAGGTGACTGACAAGCTAAGAGAATTGGAGTTATCAGGTCGTTCTAAAGA
14R250(+)	yes	6	TCTTGGTCATTAAAGGGCATTTTAAATCAAGATGGTAGTAATAATTATCCTTACAGAATCCGTAGGTACAATGCTGATGTTGTTGCTGATAATTTTGAATTTGGTTCAGACAAAAAGGTTATCGTTGCCTTGAAAGACGATGTCTCTCTATCGAAAGTGAAATCCTTTAAACAGCCAAGCAAAGGAGTCTTGATGCCAAGTGCTTCCCTCCAAGATTTTTATGGCTAATTCTTATTTTCAAGGTTTTAAAGTTGTATTCCGAGTGTATTATATAGTTGTTTCCCTTCCGGAAGTTGTGTAA
14R265(+)	yes	4	CAAATCCTCCTGAGCCAGGACTGGTTGGCTGTGTATAATATCCTGTCAACGAACGAGAATTGACATCTTGTGATAGCGTTTGCAGTCGTTTTTGTAACTGTTGAGACAATGATGGATCAA

			AAAGCATATTGGTACCTGGGTTGCTTATTGTTGTATTACT GAATGCACTCGTATCTGTAATGCCGTTTCTCTCTTTATTAC TATTTTTATTGTTTTTCCCATTTTTAGCATTTTTAGCCACCC TGGTGCCTCTTTTTTTCGTTCGGTTGTTTGGCATGTTTTTAT TTCCATAGTTCTTAA
14R269(+)	yes	18	GATTCCACAGACAAATCTTGATGTGAATGGTTAGAATGTA ACCCAATAACCTGGCCTTGAGAAGATAACGGTATTGAAG AATTCACTTGTTGTTGATTACATTGTTGTGGTGAATGTTTT TGCGTAGAGTGTTGTTGTGGCTGTTGTGGCTGTTGCTGCT GCTGTTGAGATTGCTGTTGGCTTAAAGGAGAATCATTGGA AAATGGATCGCTGAATGAAAACCTGGATCTTTGTGATAAC AGACTTGGCCTTTTCATTCCCGAGTTCGATTTATTCAATCT TGAGGTGTCATTGAAAATGG
14R306(+)	yes	15.5	TGAAATGAAGTTGCAGATATGGAAGACAAGTATGTGTAA CTAGGGGATATTACGGTGGGTGTTGATGGTGAGAAAATA GAAGATCTGCTGTCGAACGAAAACAATTTAGTGTCTGAAT ATGATGAAGATAGTGATAATGACAACGAGGACGATACCA GTAACGAAGATGGTGATGATAACGACGACCATGATGATG ACAGCGATGTAGCTGATGACGATAACGTTGGTGTTGTTGA TGTTAACGATGACCATTCCGACGTTATTGACGACGATGAT GACATTGGTGACGTCGATGAGAAA
14R321(+)	yes	11	AGGAGTAGAGGTGGTAATCTTTTCGACCCACCAGGTGTGA GAATAGGCGGAGATATTATAGTAGGTCTTTGTTGGCCTGG TAAAAAACTAGGTGAGTCTTTTTTTACGTCATTATGATTA CTATTATCTGTAACGGTTGAATTCCCATTTTCTTCTGGAAT ACTGTTCAAACTATGAGTGGAACGTTGGGAATGTTGTGTT ATCGTTGAGGTAGAATTGTTTCGTGAAAGCTCTTTTGGTA GAGACGGATGGGAGATCGTGCCTGGAAGTGACAAAGCCT TCTTGGAAGAGGCGAAGCTT
14R332(-)	yes	12	GGTATTGATCATCTGATTTGAAGCTGCGTAATTATTATGT GGTGGTGGTGTGGCTGGATAATTCAGGGCTGGAACGTCTG TAAAACCGCTGCCAAGGTTACCTGTCTCACGTTGTAAATG CTCTTGTAACGATTGATGAGAGCCTTGCTCTTTCATGTAGT TGCTACTAAATCCTGCAGCAGGCGGTCTAGGAATTTGCTG ATACGGATTCATTTCGTTGGCGTATTGTTGTTGTTGTTGAT GCTGTTGGTCTTGTTGTTGGGACATATAATAGCCATGTCC TGAAGTAGGTGCACCCTCA
14R333(-)	yes	0	GCTGCTGCTTTATCTTGCGTTCCTTTAGTCAACAAATCTGG TCGTCTTTCAAAGTACATATAGTTTGATATGGTACCATTAT TAAAGCCGCTCTGAACGGGAGATTGTGATTGCTGCTGCTG TTGTTGCTGCTGTTGTTGTTGTTGTATTTGCATATGTTGCT GTTGTTGTTGTTGTTGCTGCTGCTGGTATTGAAGTTGT TGCTGTTGTTGCCGGGAATTTTGTTGCTGTGCTTGGGAAA GTTGTTGCTGAACTTGGACATTTTCCTGTTCTTGTTGTAAC CTCTGGAAGCTTTTCA
14R343(+)	yes	10	TCATACGAAGTTGGATCAAGCTTCGACGCCTCTCTTTCTTC GTCATAGTGCGAGTCATCTGGTTGTAATTGATAACGATTG GCCTGTTGCTGGGCCTGGGCTTGTTGCAATTGTTGTTGCTG CACTTTGGCTTGTGCTTGAGCCTGTGCTTGAGCCTGAGCC TGCTGTTGTTGTTGCTGTTGATTGTAATGATGGTGATGATC GGGTGATGAATATTGATGGTTTTTACTTCTATGCCTTTTTG ATTTTTCATTTGGTGGATTTGGGTGACCATAACCATGTAA GTTCGGCTTTTTGTTTA
14R409(+)	yes	6	ACCTCGATAATTTCCATGGCGCCAGCGGCTACAACGAGA AATATAGTCGGTGGTGCAGATGGATCTACTATTGTTAACA

			ACAGTCAGGAAATGTATAAGAATTTACGTCACTTGATATATGCTGCAAACCAGCCAAATGGGACCGAAATTTTACATTTGGATTTACCTGCAACAAGTGCAGAGGAGTCAAACAATATGTTCAATGTAGATGAGGTCACTTTGAAGCAAAGGAAAGATAAGCATGGTCTTTTTAGTATAAGACTAACCCCATTTATTGATAGTTCGTCCACGACAAATCAAG
14R416(+)	yes	10	TCCCGAATTATTGAAATACGACAGGAGATCTTATGGTGGCGGTCACCCAAGATACGGTGGTGGTCGTGGTGGTCGTGGTGGCTATGGCCGTAGAGGTGGTTACGGTGGTGGCCGTGGTGGTTACGGCGGTAACAGGCAGAGAGATGGTGGCTGGGGTAACAGAGGTCGTTCAAACTATTGAAGTCAAAGTGATTCAATTCATTCATCTCTATTAATAATAATTATATAAGTTGACTGCACCTTATTTCTTCAATGCATGATAAAATTTTATTTGCTGTATACTACATTTTTAATATTTTAA
14R487(+)	yes	4	AAGACGAACTTAACGGAAAGGAAGGATCTAGCATGTCCATGGTTAAATCTTCTTCTTCGCTGTCACTTGATTCTATTTGGAACTTGGGATTCTTGTGCATGTTTGCAACAGGGTCAACGTTTGATAAAATGGATTGTCTTTGAAATAGTATCGAAGATGACCGGTTTTTTGGGGATTTTGCATTGTTATTAGCATTATTTGAATATTGTCTCGCAGATTGCTGCGGAGATACATAAGGCGAAGTAGCTGGTAATTCTGTAGTTTTGATGAAGGGGTTGGAAAATGTCGTTGTTGCTGACG
14R527(-)	yes	4	TTGTGCTTATAGTCGCGATTGTGTTATTCAATATGGCACCAGTGCTAATATGTCTTGAATTTTTCAGTCGGCTTAATAATTTTTTGTTATTGTTTAGCAATGGGGCGCTTATCTTTGAGGTTATTCCATGATTTTGTTGCTGTTGCTGTTGTTTTGGAGGTTGCTGTTGCTGCTGCTGCTGCTGGCTGGAGGAGTCAGGAAATATTAGGTTTTTAGTCGAATTAGCTGCCGATTCATAAACAAATGTTTCCTCAGAATCACTAATTTCATTTTCACCTACAGCTGTAGCCAGCCTTGCAGC
14R530(-)	yes	18	AGAGCGTGGAGTAGTGGGCATCCTTACGGCCATCGGTGTTCCATGATGTGATGACTTATTATATGGACTTGTGGCTAAATACATTAAGAGATCAGCGCCCGCAGTGTTGTCGCTGTCCTGTGGAGCATCCTTAGAGCCTGCTGCCTTCTTCTTGATATTATTATTGTTATTGCTATGGTTACTGCTGTTGTTATTATTACGTTTTTGGCGTCATGAAATTCTTATCATGGGGGCGCTGGTAGTGTGACGAAGCGTGTGGTAAGATTATAGGTGTAGCAGAAGCCCTCGGAGTAGATGGAGTG
14R566(+)	yes	12	GTCATCGCTTAAGTATTGTGATGAAACGGTTGATTCATTAATGTGATCATTGTTCACCCTTGAGAATTTACTAGATGTAGCAATATTGTGAGCTTTTGCTGCTCTTGGTTTTGAACTATTTGTTCTCTGATAATTACGAGTTGTAACCTTTGTGTTAGTCGTGTTTGGAGATGTGCGAGAAACATTTTTGTATTCGCTGTGTTGTTTGTATTGGGACGGTTTATTTAGATTCATGATTTTATTTTCGTCATATATTAACTGAATTTCAGCGAATAATTGGGAAAGTTTCATGTTCGTTTTG
14R567(+)	yes	22.5	GTGGTGGGATATACCGTGGCAACGGATAGTGTGGGTCCGGTTGATAATACGCAGGTGACGTTTGATAATGTGAACACGTCATCATAGATGGTTGTTGGTAATGTGCCCAGTTAGAGGCCATAGCTTTGTTTGGGGTCATCATGCCGTGCTGTTGGTACTGTCCATTCTGTGGAGGTGGTACTGAAGCAGGTTGAGGAGAGACATGATGATGGTTCTCTGGAACAGCTGATGTCCCAGGTGTTGTCTCTTGTTGAGAATTAACCTTAGTGGAATCTCTATCAAATTCCGGTAAATTGGAAG

14R580(-)	yes	10	GATCTTGGAGATAAAAACTGGTTTGTGAGGGATGGCCGATGTTGATTATCAGCAGGTGTTAAGATATCGTCCCACTGAAATGTATCAGATTGTTCGACTTTTAAAGTTGGTGTTTGCTTCTGCTGTTGCTGTTGTTGTTGCTGTTGCTGTTGCTGTTGCTGTTGTAACTGTTGCTGTTGTAACTGTTGCTGTGGAAGATCTGGTTTGTATAGACTTTGGTAATTCAAATCTGCAGAATAAGAACCGCTTGACGGAGACAGCAAGCTGTTCACGTTAGAATCAAATGAGTCCGGGATGCCAG
14R596(+)	yes	10	TAACACATTGTAATAAAGGTTCTTTGAACATGTTACCAGCCTCTTCAGAGATCGATCCGGTTCTGTTGAGCCAAATACTCACTTATCCAAGGATAACTTTCGTCCCGCGAGGAGCAGTGGTTTTGGAGGAGGTTTAGGCTTTAAATGTGAAGGCTTAGGTGGTGGGGTAGGTTTTGTTTTTTTGTCTTTTGAAATCGGGCCACTTGATTTCAGCATTGTTGGCTTATCTATCTCCCCTTTCTCATGCAGCTTGACCTTGGACATTTTCCTAGCGTTGAATGACTCCTCAGAAGAGTTGAGT
14R604(-)	yes	8	TTTTTTTTTTAAAAATAGTATAATATAATGAATGGATTCTTGTCGTTCTTTTTTTATTTTATCTTGGGATTTGTAGGTTGCCTCTCTTTATTCTTTCTTTTTGTTTCATTCCACTTTTCTTCATAATATTTATTGTTGTTGCTGCTGTTGCTGCTGTTGCTGTTGTTGCTCAGACATAACCGGTTGTTGTTGGTCATTCATTATAAGTGGCTGACCGCCCTGTTGCTGCTGCTGTGGCATAAAATTTGATCTTTCCTTACCCCAACCGGTTCTCAAGTTTCTGCCTTGAAATGGGAAGTTT
14L07(-)	no	23	CATAATGATGTGGGTGCATTTGGTACTGATTTAGTGAGAGATGGGCCATGGAGTGGAGTGGAATGTTATGGCAGGGTAAGTTGAGATGGTATATACTGTAGCATCCGTGTACGTATGACCGATCAGAATACAAGTGAAGGTGAGTATGGCATGTGGTAGTGGGATTAGAGTGGTAGGGTAAGTATATGTGTATTATTTACGATTATTTGTTAACGTTTCAATATGGAGGGTAGAACAACAGTATAGTGAGTAGCAGATGGTCGATGGTAGGGTAATGGTAGTAGAGTTGGATTTGGGTAAT
14R719(+)	yes	11	TTTTAAACTTTTTTTTTTAAAATTGTGTATACATATAGTACATAAATGATCATTGGTCCCCATCAATACCGTATATGACACCTTGGTACTTGGCGAAATCCGTACCGTTTGGGAACTTATGCATAGACAGCTTGCTCAATTGGCAGAAGGATAGAAGCATCAATAGACTCTGGCCGCCTGTAGTAGTAAATATAGAGAATCTTGGTAGCCCTAACTCATCGGCAAGTGTGGTTAGTGAATATTGCTGCTGCTGTTGCTGTTGTGAAGATTGTTGCAATTGAGGGTTCTTGAACTCCTGGAT
14R720(+)	yes	5	GCTGCTGTTGCTGCTGTTGTTGTTGCTGCTGTTGTTGTTGCTGCAGAACGTTCAGTCCGGGCGGCAAACTAGCCACCGAAACAGGCGTTCCTATATTAAATGGCTGCTGTTGTTGTTGCTGTTGCTGTTGTTGCTGAGTTAACAGTGGCTCGTTCTTCTGCTTCAAAAGATAGGCATTGTTCATCTCCTGATTAACAGAGCCATTAATTGTATTTACAGGCATCCCTTGAGGAAACATTCTTGCTCTATTAACCTTTGGCGAATACACCTGAGGTCCCATGTTTTGCTGCTGGTGTTGTTC
14R724(+)	yes	0	ATTTTACCAGTTGCCTTAGTTTTTCTGTTTTTAAATCCAGTGAAGTGGTGGACGAAGCTTGTTCTTCTACTTCTCGATCGAGCTTTTCTCCATTCTCTTTATCTGGTTGTTCTACATCCTCTTTTTCTTCTTGTTTTTGCTCTTCTTCTATGTCTGACTCGTTTTCACTTTCGATGTCTTCTACCTTCTTACTGAAGATATCATCCTTACTAGAAATTTTTTTCTTGGTCTGAATTAAGACGTTATGTTGATCGGTTCCTCATCTGAGGAAAAGTCTTCAAAGTCACTATTTACTTTTT

14R773(+)	yes	13	CTTCTTGTGGCTATGTTCAGTATTGGATGTGCCATAGGTGGTCTTATTTTTGCCCGTCTTGCTGATACTTTAGGTAGAAGGCTGGCAATTGTAATCGTGGTGTTGGTATATATGGTTGGTGCAATTATTCAGATCAGTTCGAATCACAAATGGTACCAATACTTCGTTGGTAAGATTATCTATGGTCTTGGTGCTGGTGGCTGTTCGGTGTTGTGTCCAATGCTTTTGTCTGAAATAGCTCCTACAGATTTGAGAGGTGGACTTGTCTCATTGTACCAACTTAACATGACCTTCGGTATTT
14R779(-)	yes	22	AAAAGCGATAAAAATTAATTGAATGTCCACCCTGCATGATTAGTCCTTTTTATCTGGCTGGACAAATGAAGCCTATTATAAGTATTAATATATTAGCTGCTGCAGACTAGAAACGAAAGCAGTGATAAATCATTGATAGACACACATTTCCTCATCGTTGTTCTTCATTATTTACAGGAAAGAGCAGAGAAAGGAAGAAAGAAATTGCAAAATATATGACAATTAACACAAAGAACATAGTTGTTGTTGGTGCTGGTGTGTTTGGTGTGTCTGTGGCAAATCACTTGTACAGGGAACTGGG
14L35(-)	yes	30.5	TCCTCGTCATCATTGTCATCATGAGGTAATTCCCGGATAGATGAATGCATTCGTTTCGGAGAGCTGGGAGTAGCTGTATTTAAGTCTGGCACAGAGTGCGACTTGGAGGATAGCGACTCTAGTGTGGGCGGCAAGTGTTGTTGGTGCTCGTGGTGGTGCAGCGTGTTGTTGCTGCTATGTACGATTGGTGATTGCGAGCTTACAGATCGCTTGGATTTATTGGCATGCAGGGCGAACTTCAACCCTTCATGCAGACCCTCAAGATACCTGATATCGTTTTGGATCTCCTTTGGATCATCGT
15R1084(+)	no	23	AGGGTAATAGTAGGGTAAGTGGTGGTGGAGTTGGATATGGGTAATTGGAGGGTAACGGTTATGATGGGCGGTGGATGGTAGTAGTAAGTAGAGAGATGGATGGTGGTTGGGAGTGGTATGGTTGAGTGAGACAGGGTAACGAGTGGAGAGGTAGGGTAATGGAGGGTAAGTTGAGAGACAGGTTCATCATATATATGTCACTGTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGTAACCTTGTACGAAAATAGGCAATATTTCCTGTTTAGGCGAT
3R309(+)	no	30	GTATCTGCGGATCTGATTTTCATATAGCCGTTGGTAATTGGGGTCCAGTCCCAGAAAATCAAATCCTTGGACATGAAATAATTGGCCGCGTGGTGAAGGTTGGATCCAAGTGCCACACTGGGGTAAAAATCGGTGACCGTGTTGGTGTTGGTGCCCAAGCCTTGGCGTGTTTTGAGTGTGAACGTTGCAAAAGTGACAACGAGCAATACTGTACCAATGACCACGTTTTGACTATGTGGACTCCTTACAAGGACGGCTACATTTCACAAGGAGGCTTTGCCTCCCACGTGAGGCTTCATGA
10R733(+)/4L13(-)	no	17	CGTGGCCATGTTCAGTGTAGGATGTTCCATTGGCGGTGTTGCTTTTGCGAGACTTGCTGATACTTTAGGTAGAAGGCTAGCAATTGTAATCGTGGTTTTGGTATATATGGTTGGTGCAATTATTCAGATCAGTTCGAATCACAAATGGTACCAATACTTTGTCGGTAAGATCATCTACGGTCTTGGTGCTGGTGGCTGTTCGGTGTTGTGTCCAATGCTTTTATCTGAAATAGCCCCCACAGATTTGAGAGGTGGACTTGTCTCATTGTACCAACTTAACATGACCTTCGGTATTTTCTTG
9L44(-)	no	20	AGGAAGAAAGAAGACCATTTGGTAAAGGTGGAGCTGGAGGAGTAACTAAATCGTGATTGGAGGGGCCTTTGTTCATCGACAAGGACTCGGGTAAAGGCGGTGGCGGTGGTGGTGGTGGAACGGCGCCACAAGTTTTAGTTTTAAATAGATCTGGAAGGGGTGGCGGTGCAGGTGCTGCAATACAGGAAACACCATCCGGTTTAGCTTCTGCTTCAGTCAAAAGAGACTGTGGAAgtggaggaggggggaggtggaggaggggggcggtggaAGTTGAGGTAGTTTAACAACTTCAGGTACTA

124

5L35(-)	no	28	ATTCTCATCAATAAATCTTGTATGGAAAATGCGGATTGCG GAAGGCCCAAATCCTTGATAAGGTCACTGTATCTGCTGCT TCTCATATCCTTTGTATCCTTTTCGGATGTAGGATGAGTTG GTGTGGTGTTACTACTAGGATTGGCGTGGATGAAGGACC TGCAGTGGAGGGTGTTGTTGTGGAGTTTTCCAACACCTTC AAATCGAAGAAAACTGGATCTCTTTCCGATTTGGATAGCA ATTGATTTTGTTCTTCAAGAACTTTGAACATTTTGCCCAGC CTAGTAATAGCTGGTTCGT
6R210(+)	no	61	GTACATTCCCCGTTGAAAGTGATACAGCTTTCTTGATTGA CACAATAGCAATGGCCTTCAAATGCATATCTCTACTATCG GCTAAAAAACGAATGACTCACGTTATCAGGCTCATAGCTT GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG TGTGTGTGTGTGTGTGTGATTGTTGTTCTAGTCGCTTGC TTTATAAAGTAACGACACTTTCTGGTGCCAATATGTGAAA ACGCATTACAGAAAAAAACAGTTGTATTCTACTAAAAAC ACATCAGTAGTCACAGAAGT
6L22(-)	no	5	GAAGACGCGGATATAATGACATTTCCTAACTTTTGGGCAA AAATTCGCTATCATATGCGAGAACCGTTTGCGGAGTTTCT CGGGACACTAGTTCTTGTCATTTTTGGTGTTGGTGGTAATC TTCAAGCAACTGTAACAAAAGGTAGTGGTGGTTCCTATGA ATCCCTATCATTTGCATGGGGGTTCGGTTGTATGCTTGGT GTTTACGTCGCAGGCGGTATTAGTGGTGGTCATATTAACC CTGCTGTTACGATTTCAATGGCAATTTTTCGAAAATTCCCC TGGAAAAAGGTGCCCGTA
7R1089(-)	no	26	ATACGACTCTTTGTTAATGTCGGTGACTGGATGGAATCTA TTATCCTCAGCATTGCCATCTTTATTGGCGTCCTCCTTGGC ACTAGCGTTGGTACTTTCAGTGGTAGTGGCATTAGTGCTG GAGTTGGTGCTAGCAGTGGTAGTGGCATTAGTGCTGGAGT TGGTGCTAGCAGTGGTAGTAGCACTAGTGTTGGAGTCGGT ACTTTCGGTGGTAGTAGCACTAGTGTTGGAGTCGGTACTT TCGGTGGTAGTAGCACTAGTGTTGGAGTTGGTACTTTCAG TGGTAGTCGCACTAGTCCTG
7L67(-)	yes	28	ACGTATGATGGCTGTTTTTTATTCTGTCGGTTTTTGGTTGC CGTAGTAAATATAATTTGGTAAGCGTGTTGTGCTGTGTTG TGTTGTGTTGTGTTGTGTTGTGTTCTACTAATTGGATTGTG CGAATAAGTGAAACGGTTGCTTCACCCAGCGGGAAAAGG ACAACCTGGTATTTGGTGTGAGATAAAGTTATAGACAATC TGGTCTTTTGTTATTAAGTTCTCTTTTCTTCTATACCGTCCA GTTTACTGCTCTCTTGCTTTTCGATAAAGCTAAAATAAAA TAAAAATATGTCACTAT
7L69(-)	yes	34	ATATTATATTCTTATATGGCTATATATGTTTATATATAAAG GTAAAGACAAACAATAAAGAACATACATATATATATAGA TAGATATTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG TATAGCAAGCGCATGAATTTCATCACAATGTCATGGTAAT GGTTGTGGTGATGGTCGAAGTAGAGGTGTAGGATACGTCT TCTTCGTCTTCTTCTTCCTGTTGTCGTAATAGCTCCTGCTG CCTTTGGAACTCTTCTGCTGCAGCTTTTTCACGCTCTTTTC TTTCACGTTCCTTTCTT
2R780(-)	no	13	CTTGACCAATTTGGGGATGTAAGTTGTGTGGTGCTGGAGG AGGAGGAGGGGTAGAAGTCGCTGATTGATTAGCAATGGG TTGAGATTGTTGCGGAGGAGGGGGTGGAGATTGATGCGT TTGCGGTGGCGGTGAAGTTTGTTGTTGTTGTTGTTGTTGTT GTTGTAGACGACTCCTCAGTAACTGTTGGTGTCGTTGCTG AATCATCTGCAATTGTTGTGGTGTGAG

7L37(+)	n/a	4	ATGGAACTCCGTAGTTTTTCTAGACAGCCTGATGGCATAC TTGCTAACCCAAGATTGGGAAGGGAGGAAGTATTGGAAG GTGAACATCCGCAAGATGCTAGATTAGCCCGACAAAGCA TTTGGTTGAGCCCAAGTTTAATTGCGGAGTATATACAGCT TTTTTTCAATTTTATTATAGGAACTATAGGACTGTCTCTTG CTATTAAATTCATTTTGATGATAAGAAACGATGTTAACTT AAAACTGGAGCATAACGTGAGGGAAGAATTAGATAAAAT TGCAACTTGCAAATCAAGGTATT
Tel-11	n/a	11	GTGTGGGTGTG
2xTel11	n/a	22	GTGTGGGTGTGGTGTGGGTGTG
9L44(-)short	n/a	12	GGGCCTTTGTTCATCGACAAGGACTCGGGTAAAGGCGGT GGCGGTGGTGGTGGTGGAACGGCGCCACAAG
5L35(-)short	n/a	28	GGATGAGTTGGTGTGGTGTTACTACTAGGATTTGGCGTGG ATGAAGGACCTGCAGTGGAGGGTGTTGTTGTGGAGTTTT
14L35(-)Short	n/a	26.5	GTGTGGGCGGCAAGTGTTGTTGGTGCTCGTGGTGGTGCAG CGTGTTGTTGCTGCTATGTACGATTGGTGATT
3R309(+)Short	n/a	26	AAAATCGGTGACCGTGTTGGTGTTGGTGCCCAAGCCTTGG CGTGTTTTGAGTGTGAACGTTGCAAAAGTGACAACG
8L50(-)short	n/a	13	TGTGTCTCACAGTCAACGTCAGGTCCGTACAGAACGGTGT CTCGAGCCTTATCGCTGCACGCCATATACTGTGGATATAA TCCG
14L13(-)/6L11(-)short	n/a	17	TAGTTTTATATATGTGTGTGTGTGTGTGAAATTCATTCC ACA
15L20(-)short	n/a	17	CTTTGTAGTCGCTGGGTCGCTTCTTGGTGTCATCGGAATG GGTCTTTTGTACAAATACAGGGTGGTATATGACGGGATTT CGGG
6L22(-)short	n/a	5	GTCATTTTTGGTGTTGGTGGTAATCTTCAAGCAACTGTAA CAAAAGGTAGTGGTGGTTCCTATGAATCCCTATCATTTGC ATGGG

Table 2. Oligos used for strain construction and PCR

Oligo Name	Sequence 5' to 3'	Length	Description
14L35URAfor	AAGATAGTACATATCTTCTGTGATTG GCAGCAAATTTTTTCgtactgagagtgcaccacg c	60	Used to add *URA3* marker to XIV
14L35URArev	CTGTTAATTGCTGTTGTATTTGTTATT GTAGGAGGGTGCActccttacgcatctgtgcgg	60	Used to add *URA3* marker to chromosome XIV
14L35HOfor	TCATCCTCCAGTGTAAAGACGACCAC ATCGACAGGGAGCGccagctgaagcttcgtac gc	60	Used to add HO cut site to chromosome XIV
14L35HOrev	TACGTACTGGGAGGAGAAACGCTGG ACAAAAGTAGTTTGCcgcataggccactagtg gatctg	60	Used to add HO cut site to chromosome XIV
6L22HOFor	AAGGATTGATACTTTGGCTAATCCAA GGCCAAAGCAGAAccagctgaagcttcgtacgc	60	Used to add HO cut site to chromosome VI
6L22HOrev	AATAACGCCAAGAACATCCGAAGAA GTCCTGGTCGCGCTTcgcataggccactagtgg atctg	60	Used to add HO cut site to chromosome VI

6L21*URA3*rev	TGGGCCTAATATCATTGATGAATGTG AGCGTTGGAGCTGCctccttacgcatctgtgcgg	60	Used to add *URA3* marker to VI
6L21*URA3*for	TTCCAACAGGCAGAAGGATATAAGC ATCAACAAAATTGCCgtactgagagtgcacca cgc	60	Used to add *URA3* marker to chromosome VI
14L35PreSiRT AFor	TCGTTACCAGTCCTCCGCCGCCTAG	25	Used for multiplex PCR for chromosome XIV
14L35PreSiRT Arev	ACGAATGCATTCATCTATCCGGGAA	25	Used for multiplex PCR for chromosome XIV
14L35PostSiRT Arev	ATCAGGTATCTTGAGGGTCTGCATG	25	Used for multiplex PCR for chromosome XIV
6L22PreSiRTA For	CATCGAGACGAGGGTCGACCGATAT	25	Used for multiplex PCR for chromosome VI
6L22PreSiRTA rev	AATTGACATTATGCGAAGTACGCTT	25	Used for multiplex PCR for chromosome VI
6L22PostSiRT Arev	CGTAACAGCAGGGTTAATATGACCA	25	Used for multiplex PCR for chromosome VI
chr7PreSiRTAt estfor	CGGAAAGCTATATGATATAAACTAG CC	27	Used for multiplex PCR for chromosome VII
chr7PreSiRTAt estrev	GCTCGTAGACATAGGTTAAGATCA	24	Used for multiplex PCR for chromosome VII
chr7PostSiRTA testrev	CTTCTATAACTAAACCATAATAATG	25	Used for multiplex PCR for chromosome VII
UBP10 KO FOR	CTACAACTCATAGTATCTGTAAATCC GTCCTATTGTCATATCACAATCACCG GATCCCCGGGTTAATTAA	70	Used to construct strain for Chr. IX standard curve in Figure 1
UBP10 KO REV	CGCTTCTCCAACTCTGTATACTGAGA GTTGTTTTATAAGTCAATAAAGACCG GGAATTCGAGCTCGTTTAAAC	73	Used to construct strain for Chr. IX standard curve in Figure 1
UBP10 KO CHECK FOR	GAGCTATTAAGCATCATATAC	21	Used to construct strain for Chr. IX standard curve in Figure 1
UBP10 KO CHECK REV	GTTAGCAAAGCTCGGTAG	18	Used to construct strain for Chr. IX standard curve in Figure 1
HIM1 KO FOR	GACAGCTTTTTATACATATAGACCCT TTGAAGAATATTCCAAACTAGAAAG GCGGATCCCCGGGTTAATTAA	72	Used to construct strain for Chr. IX standard curve in Figure 1
HIM1 KO REV	GTATTAGGCTACTATTTATTTAATAA CTTCGCATCTATCAAAATAAAAGACT GGAATTCGAGCTCGTTTAAAC	73	Used to construct strain for Chr. IX standard curve in Figure 1
HIM1 KO CHECK FOR	CAATACATATAATGATGGGTGAC	23	Used to construct strain for Chr. IX standard curve in Figure 1
HIM1 KO CHECK REV	GAATTGTCGGCCATTACTGGCG	22	Used to construct strain for Chr. IX standard curve in Figure 1
NUP170::Trp Rev	GGCCTCTTTATTACATTAAAATATAC ACGTACATTACCCTGCTATCTATATG GAATTCGAGCTCGTTTAAAC	72	Used to construct strain for Chr. IX standard curve in Figure 1
NUP170::Trp For	CCTTTACATCAAATAAGCACCGCAAG ATATCCTAAAATCGACATCCCGGATC CCCGGGTTAATTAA	23	Used to construct strain for Chr. IX standard curve in Figure 1

NUP170 Check For	GGATGTGTTCTTTGTGACACGC	22	Used to construct strain for Chr. IX standard curve in Figure 1
NUP170 Check Rev	CTGATTCTCTCTGACCTGTTCG	22	Used to construct strain for Chr. IX standard curve in Figure 1
Rad52koR	GATGCAAATTTTTTATTTGTTTCGGCCAGGAAGCGTTTCAGAATTCGAGCTCgtttaaac	60	Utilized to generate RAD52 deletion construct
Rad52koF	CGAAAAATATAGCGGCGGGCGGGTTACGCGACCGGTATCGCGGATCCCCGGGTTAATTAA	60	Utilized to generate RAD52 deletion construct
chr7presirta5testfor	CGGAAAGCTATATGATATAAACTAGCC	27	PCR primer to amplify centromere proximal region or SiRTA region in HO cleavage assay on chr VII
chr7presirta5testrev	GCTCGTAGACATAGGTTAAGATCA	24	PCR primer to amplify centromere proximal region in HO cleavage assay on chr VII
c7closechkrev	GTTAATCAACGTTCCAGTACCA	22	PCR primer to amplify SiRTA region in HO cleavage assay on chr VII
chr7LEGfor	TCTAATTCTTCCCTCACGTTATGCT	25	PCR primer for positive control in HO cleavage assay
chr7LEGrev	CGTTAAGTCAACACCTGCTAA	21	PCR primer for positive control in HO cleavage assay
c7newendchkfor	GAACCAGTTCAAAGTAGCAGC	21	PCR primer for negative control in HO cleavage assay
c7newendchkrev	AGTTAATATACGGAAAGAGC	20	PCR primer for negative control in HO cleavage assay
telcheckblueR	CCCACACACCACACCCACA	19	PCR primer to amplify a fragment containing telomeric repeats
rt9L44 HS for 2	AACGTGTCTTCCCAAAGTGTGTC	23	PCR primer to amplify telomere region in HO cleavage assay on chr IX
Telcheckgreen R	CACCCACACCACACCCACA	19	PCR primer to amplify a fragment containing telomeric repeats
Int BNR1 for	GGACAATGTCGTTATCACAG	20	PCR primer to amplify centromere proximal region or SiRTA region in HO cleavage assay on chr IX
HS check PSirta1rev2	CGATGAACAAAGGCCCCTCC	19	PCR primer to amplify centromere proximal region or SiRTA region in HO cleavage assay on chr IX
Core 2 check	CTAGGAACACAACCTAATTACC	22	PCR primer to amplify SiRTA region in HO cleavage assay on chr IX
MCM10 F	CCTGAAGATGACTTAAAACGGG	22	PCR primer for positive control in HO cleavage assay
MCM10 R	CCTCGCCCTGAGGGTCTCCCATG	23	PCR primer for positive control in HO cleavage assay
GencLhostitchF2	tttcagctttCCGCAACAGTATAATTTTATAAACCCTGGTTTTGGTT	47	PCR primer to insert close ho cut site

Chr7HOrev	CAAATCAATACCCCTTCAAAGCACAG GCATCGGTGATTGGGTTAGAGGCGG	50	PCR primer to insert close ho cut site
Chr7HOinsshort F	TAATCTCTTCTTGGCCTTTGGAC	23	PCR primer to insert close ho cut site
Chr7HOinsshort R	CCCAATCACCGATGCCTGTG	20	PCR primer to insert close ho cut site
1cdc13cLhostit chF1	TAATCTCTTCTTGGCCTTTGGACATTG CATGTTGGCCTCTGTGTGGGTGTGtttca g	57	PCR primer to insert 1cdc13 sequence plus adjacent ho cut site
1cdc13cLhostit chR1	ATACTGTTGCGGaaagctgaaaCACACCC ACACAGAG	37	PCR primer to insert 1cdc13 sequence plus adjacent ho cut site
2cdc13honewF	TAATCTCTTCTTGGCCTTTGGACATTG CATGTTGGCCTCTGTGTGGGTGTGGT GTGGGTGTGtttca	67	PCR primer to insert 2cdc13 sequence plus adjacent ho cut site
2cdc13honewR	tactgttgcggaaagctgaaaCACACCCACACCA CACCCACACAGAGGC	49	PCR primer to insert 2cdc13 sequence plus adjacent ho cut site
2cdc13sizechec kF	TAATCTCTTCTTGGCCTTTGGA	22	PCR primer to check 2cdc13 sequence plus adjacent ho cut site
2cdc13sizechec kR	Tactgttgcggaaagctgaaa	21	PCR primer to check 2cdc13 sequence plus adjacent ho cut site
6L22clhostf1	TAATCTCTTCTTGGCCTTTGGACATTG CATGTTGGCCTCTGTCATTTTTGGTGT TGGTGGTAATCTTCAAGCAACTGTAA C	81	PCR primer to insert 6L22(-) sequence plus adjacent ho cut site
6L22clhostr1	GATAGGGATTCATAGGAACCACCAC TACCTTTTGTTACAGTTGCTTGAAGA TTAC	55	PCR primer to insert 6L22(-) sequence plus adjacent ho cut site
6L22clhostR2	TAAAATTATACTGTTGCGGAAAGCTG AAACCCATGCAAATGATAGGGATTC ATAGGAAC	59	PCR primer to insert 6L22(-) sequence plus adjacent ho cut site
5L35CLhof1stit ch	taatctcttcttggcctttggacattgcatgttggcctctggatg agttggtgtggtGTTACTACTAGGATTTGGC G	77	PCR primer to insert 5L35(-) sequence plus adjacent ho cut site
5L35CLHo10R 1stitch	GAAAGTAAAAGAAAAAAACTCCACA ACAACACCCTCCACTGCAGGTCCTTC ATCCACGCCAAATCCTAGTAGTAACA C	78	PCR primer to insert 5L35(-) sequence plus adjacent ho cut site
5L35CLHOR1s titch	AAAGCTGAAAAAAACTCCACAACAA CACCCTCCACTGCAGGTCCTTCATCC ACGCCAAATCCTAGTAGTAACAC	74	PCR primer to insert 5L35(-) sequence plus adjacent ho cut site
HO Check F	CCGATGTACTTAGTCATATCG	26	PCR primer to check HO cut site insertion
HO Check R	CTAATCTACTATGAAAGTTCCTAATT	26	PCR primer to check HO cut site insertion
pif1m2chkfor	GAATCATATTATACCAAGAAGGC	23	PCR primer to check *pif1m2* mutation
pif1m2checkrev	GCTTCTCAGTTTCTAACTGTATGC	24	PCR primer to check *pif1m2* mutation

Table 3. Results of testing SiRTA ability to stimulate *de novo* telomere addition

SiRTA	CATHI	avg % *dnTA*	Method
9L44(-)endogenous	NA	33	PCR (Obodo 2016)
5L35(-)endogenous	NA	25	PCR (Obodo 2016)
14L35(-)endogenous	NA	64	PCR(Ngo2020)
6L22(-)endogenous	NA	0	PCR(Ngo2020)
9L44(-)short	12	19.67	PCR(Ngo2020)
5L35(-)short	28	16.2	PCR(Ngo2020)
14L35(-)Short	26.5	34.8	PCR(Ngo2020)
3R309(+)Short	26	27.61	PCR(Ngo2020)
8L50(-)short	13	7.57	PCR(Ngo2020)
14L13(-)/6L11(-)short	17	3.33	PCR(Ngo2020)
15L20(-)short	17	0	PCR(Ngo2020)
6L22(-)short	5	0	PCR(Ngo2020)
14R333(-)	0	0	PT-seq
14R724(+)	0	0	PT-seq
14R265(+)	4	0	PT-seq
14R487(+)	4	0	PT-seq
14R527(-)	4	0	PT-seq
12R1009(+)	4	0.13	PT-seq
14R720(+)	5	0	PT-seq
6L22(-)	5	0	PT-seq
12R555(-)	5	0.18	PT-seq
14R409(+)	6	0	PT-seq
14R604(-)	8	0	PT-seq
14R250(+)	8	0	PT-seq
12R700(+)	8	1.41	PT-seq
14R235(+)	9	0	PT-seq
12R500(-)	9	0.11	PT-seq
14R580(-)	10	0	PT-seq
14R596(+)	10	0	PT-seq
14R416(+)	10	0.77	PT-seq
14R343(+)	10	5.93	PT-seq
14R719(+)	11	0	PT-seq
14R321(+)	11	0.61	PT-seq
12R874(-)	11	1.56	PT-seq
12R291(+)	11	1.64	PT-seq
12L23(-)	11	2.22	PT-seq
14R332(-)	12	0	PT-seq
14R556(+)	12	0	PT-seq

2R780(-)	13	31.11	PCR (Hoerr *et al.* 2023)
14R773(+)	13	3.16	PT-seq
12R130(-)	14	4	PT-seq
14R306(+)	15.5	17.05	PT-seq
10R733(+)/4L13(-)	17	3.64	PT-seq
14R530(-)	18	0	PT-seq
14R269(+)	18	4.27	PT-seq
9L44(-)	20	36.36	PT-seq
14R779(-)	22	2.34	PT-seq
12R330(+)	22	32.22	PT-seq
14R567(+)	22.5	9.60	PT-seq
15R1084(+)	23	13.35	PCR
14L07(-)	23	9.99	PCR
14R131(+)	25	30.05	PT-seq
7R1089(-)	26	36.60	PCR
7L67(-)	28	49.14	PT-seq
5L35(-)	28	13.92	PT-seq
3R309(+)	30	53.37	PT-seq
14L35(-)	30.5	76.00	PT-seq
7L69(-)	34	47.88	PT-seq
6R210(+)	61	86.00	PT-seq

Table 4. Results of PCR VS PT-seq experiments

Chromosome VII		
Sequence	%reads (PTseq)	*%dnTA* PCR
14R269(+)	2.39	6.67
14R306(+)	4.71	16.67
14R332(-)	0	0
14L35(-)a	22.02	83.33
14L35(-)b	19.65	66.67
14R131(+)	4.26	16.67
7L69(-)	12.07	46.67
Tel11a	3.17	6.67
Tel11b	1.49	3.33
2xTel11a	23.34	86.67
2xTel11b	20.25	93.33
Chromosome IX		
Sequence	%reads (PTseq)	*%dnTA* PCR
9L44(-)A	11.17	43.33
9L44(-)B	5.24	26.70

9L44(-)BS2A	0.78	3.30
9L44(-)BS2B	1.09	3.30
UBP10	3.64	6.67
HIM1	5.63	20.00
NUP170	13.78	50.00

Table 5. Results of the calibration of the CATHI program

Window size	GGTGG penalty	R^2
25	0	0.4957
25	0.5	0.493
25	1	0.4847
25	1.5	0.4743
25	2	0.4665
25	2.5	0.4688
25	3	0.4691
50	0	0.6192
50	0.5	0.6209
50	1	0.6182
50	1.5	0.6128
50	2	0.6087
50	2.5	0.6022
50	3	0.5927
75	0	0.6165
75	0.5	0.6277
75	1	0.6339
75	1.5	0.6347
75	2	0.6298
75	2.5	0.6188
75	3	0.6013
100	0	0.5754
100	0.5	0.5787
100	1	0.5776
100	1.5	0.5721
100	2	0.5624
100	2.5	0.5482
100	3	0.5295
125	0	0.51
125	0.5	0.5098
125	1	0.5063
125	1.5	0.5

125	2	0.4906
125	2.5	0.4782
125	3	0.4628
150	0	0.493
150	0.5	0.4948
150	1	0.4935
150	1.5	0.4858
150	2	0.4744
150	2.5	0.4592
150	3	0.4404

Table 6. Clip coordinates for the yeast chromosome used to run the CATHI program

Chr.	left clip	right clip
1	62	230118
2	1	813135
3	365	316520
4	156	1531933
5	1	576874
6	1	270112
7	36	1090940
8	36	562455
9	78	439814
10	61	745662
11	69	666603
12	76	1078177
13	53	924304
14	1	784037
15	118	1091273
16	1	948066

Table 7 List of putative SiRTAs from the CATHI program

chr	start	end	score	strand	gene
chrI	61	457	37.5	-	None X element
chrI	19447	19593	26	+	None
chrI	25574	25684	20.5	+	FLO9

chrI	31423	31568	26	-	None
chrI	33569	33696	20	-	BDH2
chrI	43270	43373	22	-	ACS1
chrI	57985	58101	22	-	GCV3
chrI	59749	59866	22	+	PTA1
chrI	61611	61757	20.5	+	ERV46
chrI	75900	76024	23.5	-	RBG1
chrI	81100	81273	23	+	FUN19
chrI	107194	107323	21.5	+	PMT2
chrI	110910	111045	23	-	CCR4
chrI	111753	111887	20	+	CCR4
chrI	114009	114149	21	-	ATS1
chrI	190063	190267	36.5	-	YAT1
chrI	191193	191321	21	-	YAT1
chrI	204823	205171	22	-	FLO1
chrI	205687	205797	20.5	-	FLO1
chrI	212532	212666	22	-	None
chrI	229579	229752	23	+	None X element
chrI	229949	230118	24	+	None
chrII	1464	1636	26	+	YBL113C
chrII	5787	6626	32	-	None X element
chrII	9515	9618	21	+	None
chrII	23349	23494	25	+	RTG3
chrII	30253	30399	24.5	-	Transposon YBLWTy2-1
chrII	89730	89864	21	+	None
chrII	117630	117753	25	-	TOD6
chrII	119167	119303	23	-	None
chrII	146499	146661	21.5	+	MRPL16
chrII	159023	159169	29	-	RIB1
chrII	163058	163206	29	-	PET9
chrII	212719	212865	27	+	SLA1
chrII	221666	221794	20	-	Transposon YBL005W-B
chrII	307779	307923	25	+	TLC1
chrII	353576	353721	26	+	UBP14
chrII	449883	450030	24	+	YMC2
chrII	456062	456225	20	-	AIM3
chrII	463592	463735	21	+	CYC8
chrII	465006	465125	21	-	CYC8
chrII	504291	504443	29	-	Dubious ORF YBR134W
chrII	547199	547334	20	-	SPP381
chrII	565362	565472	20.5	+	YSY6

chrII	610651	610749	22	-	MSI1
chrII	627054	627179	25	+	MCM7
chrII	644742	644859	20.5	+	None
chrII	665707	665844	21	-	PDB1
chrII	665917	666075	24	-	PDB1
chrII	679445	679577	21	+	None
chrII	685995	686109	23	-	ARC40
chrII	774178	774311	20	-	YBR285W
chrII	782155	782351	21	-	SNF5
chrII	812368	812557	25	+	None X element
chrII	812671	813136	33	+	None X element
chrIII	362	842	25.5	-	None X element
chrIII	907	1116	32	-	None ARS300
chrIII	3612	3721	20	+	Transposon YCLWTy5-1
chrIII	13715	13861	21	+	HMLALPHA1
chrIII	31760	31869	20	+	SPB1
chrIII	50759	50884	25.5	-	GLK1 and YCL042W
chrIII	52237	52383	26	+	GLK1
chrIII	62818	62942	21	-	MXR2
chrIII	63033	63148	21	+	MXR2
chrIII	66101	66247	20	-	HIS4
chrIII	77139	77441	22	-	AGP1
chrIII	85420	85566	26.5	-	Transposon YCL019W
chrIII	88529	88630	23	-	Transposon YCL019W
chrIII	108614	108699	20	-	IDB16
chrIII	108955	109081	21	-	None ARS307
chrIII	135914	136050	21	-	ADP1
chrIII	163613	163729	22	+	Dubious ORF YCR025C
chrIII	200875	201021	21	+	MATALPHA1 and Dubious ORF YCR041W
chrIII	210636	210772	21	+	None
chrIII	211302	211401	21	-	BUD23 and Dubious ORF YCR047W-A
chrIII	213169	213304	21.5	-	ARE1
chrIII	214261	214545	20.5	+	YCR051W
chrIII	239797	239932	20	-	CPR4
chrIII	240777	240915	20.5	-	RSA4
chrIII	242300	242420	20.5	+	RSA4
chrIII	247121	247255	23	+	SOL2
chrIII	253594	253716	20	-	PTC6
chrIII	260340	260466	20	-	TUP1
chrIII	261140	261296	22.5	+	TUP1
chrIII	294403	294549	21	+	HMR, Dubious ORF YCR097W-A, and ARS318

chrIII	309272	309419	30	+	ADH7
chrIII	315815	315919	23	+	ARS319 X element None
chrIII	315944	316521	27	+	YCR108C and ARS319
chrIV	155	612	33	-	None ARS400
chrIV	726	915	25	-	None ARS 400
chrIV	9434	9569	25	-	SOR2
chrIV	36801	36998	22	-	YDL233W
chrIV	43232	43371	23	-	PTP1
chrIV	57560	57707	23	+	HBT1
chrIV	59730	59929	22	+	HBT1
chrIV	61084	61229	21	+	FMP45
chrIV	62694	62943	20	+	CDC13
chrIV	67175	67305	28.5	-	YDL218W
chrIV	95624	95763	22	-	RTN2
chrIV	98833	98939	21	-	MRPL11
chrIV	134119	134224	22.5	+	LYS20
chrIV	140792	140893	20.5	+	DLD2
chrIV	147612	147757	21	-	None
chrIV	160070	160162	20	+	SFA1
chrIV	175789	175913	20	-	YDL156W
chrIV	205746	205934	24	+	RPO21
chrIV	207261	207396	20	+	RPO21
chrIV	214757	214837	22	+	RGT2
chrIV	228117	228222	22.5	+	LYS21
chrIV	257515	257645	20	+	ATG20
chrIV	287335	287473	21.5	+	PMT1
chrIV	335238	335384	20	-	IDP1
chrIV	392081	392233	27	-	YDL034W
chrIV	410974	411092	21	-	None
chrIV	427383	427673	21	+	NOP1
chrIV	462377	462506	21	+	TRP1 and Dubious ORF YDR008C
chrIV	465145	465253	21	-	Dubious ORF YDR010C
chrIV	504337	504466	20.5	-	PST2
chrIV	518892	519038	26.5	+	Transposon YDR034C-D
chrIV	538791	538911	21.5	-	None
chrIV	556172	556277	22	-	TPI1
chrIV	601181	601575	25	-	SED1
chrIV	621748	621863	21	+	None
chrIV	698816	699008	30	+	INO2
chrIV	704749	704895	25.5	+	ARO1
chrIV	714914	715034	21	+	SAC6 and YDR129C_mRNA

chrIV	721201	721428	24.5	+	Dubious ORF YDR133C
chrIV	749139	749259	20	+	SWI5
chrIV	781369	781507	20	-	CWC15
chrIV	802823	802961	21	-	None
chrIV	803531	803659	20	-	Transposon YDR170W-A
chrIV	835901	836029	21	-	None
chrIV	836276	836420	25	+	Dubious ORF YDR187C
chrIV	841803	841928	22	-	RVB1
chrIV	872430	872576	24.5	-	Transposon YDRWTy2-2
chrIV	875539	875640	23	-	Transposon YDR210W-B
chrIV	927461	927659	20	-	HEM1
chrIV	929506	929636	20	+	RTN1
chrIV	930869	931001	26	-	None
chrIV	969123	969298	23	+	CTA1
chrIV	976599	976705	20	-	None
chrIV	981780	981926	24.5	-	Transposon YDRWTy2-3
chrIV	1005907	1006023	24	+	CCC2 and Dubious ORF YDR269C
chrIV	1085423	1085570	28	-	TFB1
chrIV	1201180	1201296	20	-	None
chrIV	1247215	1247330	20	+	MUS81
chrIV	1251045	1251180	22	+	RVS167
chrIV	1267633	1267739	20	+	UTP5
chrIV	1279113	1279244	20	-	None
chrIV	1289196	1289310	22	-	None
chrIV	1289776	1289877	24	-	SIZ1
chrIV	1299639	1299736	20	+	SYF5
chrIV	1307640	1308625	25	-	HKR1
chrIV	1328831	1329052	27	-	NPL3
chrIV	1338273	1338409	20.5	+	THI74
chrIV	1358912	1358995	20	+	None
chrIV	1387891	1388015	22	-	STP1
chrIV	1412329	1412469	27.5	-	SNF1
chrIV	1416672	1416749	20	-	PEX29
chrIV	1425257	1425392	20	+	VPS72
chrIV	1432003	1432137	20	+	PKH1
chrIV	1474976	1475063	23	-	EMI2
chrIV	1476142	1476305	21.5	-	EMI2
chrIV	1476640	1476780	22	-	None
chrIV	1486876	1486980	20	-	SPS1
chrIV	1489576	1489707	23	-	YDR524C-B
chrIV	1523017	1523163	23	+	None

chrIV	1524670	1524761	20	+	Dubious ORF YDR543C
chrIV	1524921	1525237	27	+	None X element
chrIV	1525254	1525532	73.5	+	None telomeric repeats
chrIV	1530345	1530527	26	-	YRF1-1
chrIX	7281	7739	34	-	None ARS902
chrIX	19196	19342	24	-	None
chrIX	23347	23492	24	+	YIL169C
chrIX	23560	23877	26	+	YIL169C
chrIX	25046	25187	20	-	YIL169C
chrIX	29514	29648	20.5	+	None
chrIX	44157	44297	20	-	BNR1
chrIX	46181	46297	23	+	AIM20
chrIX	54672	54813	24	-	IMP2
chrIX	88820	88926	22	+	None
chrIX	123535	123647	20	+	KGD1
chrIX	123853	123964	21	+	KGD1
chrIX	129448	129565	22	+	SIM1
chrIX	150506	150652	24	-	SDP1
chrIX	161073	161220	27	-	YIL108W
chrIX	175494	175714	21	+	XBP1
chrIX	178980	179160	23	-	SGA1
chrIX	207142	207256	21	-	Retrotransposon YIL082W-A
chrIX	243375	243491	21	-	YRB2
chrIX	251720	251854	21	-	VHR1
chrIX	257469	257590	21	-	None
chrIX	266854	266975	22	+	SYG1
chrIX	267999	268145	22	-	None
chrIX	268322	268505	40	+	YIL046W-A unknown function
chrIX	270452	270598	20	+	MET30
chrIX	291308	291404	20	+	BCY1
chrIX	292010	292147	25.5	+	Dubious ORF YIL032C
chrIX	292520	292629	20	-	None
chrIX	302336	302578	23	+	Dubious ORF YIL028W
chrIX	315582	315685	22	+	FAF1
chrIX	333921	334030	20	-	TIR3
chrIX	379241	379375	24	+	SQT1
chrIX	390084	390197	22	+	MUC1
chrIX	390525	390657	22	+	MUC1
chrIX	430386	430517	21	-	None
chrIX	439344	439815	28	+	None X element
chrV	1463	1728	22	+	YEL077C

chrV	6269	6441	36	-	None X element
chrV	9726	9848	20	+	None ARS504
chrV	16884	16982	23	+	DLD3
chrV	34795	34923	28	-	NPR2
chrV	42769	42873	22	+	HHY1
chrV	78471	78584	20	+	UTR2
chrV	81625	81747	23	-	RAD23
chrV	107699	107802	20	+	RIP1
chrV	116753	116853	22	+	*URA3*
chrV	134153	134294	23.5	+	GLC3
chrV	155574	155685	20	-	MNN1
chrV	208360	208497	21	+	CHO1
chrV	220166	220347	25	+	ZRG8
chrV	266528	266724	25	-	FCY2
chrV	269014	269154	20	+	None
chrV	277323	277431	20.5	+	FCY22
chrV	277971	278112	20	+	FCY22
chrV	284369	284510	22	+	None
chrV	285264	285410	20	-	ICL1
chrV	302761	302890	20	-	VTC1
chrV	305933	306065	21	-	None
chrV	318493	318632	24	+	None
chrV	332024	332178	21	+	AIM10 and Dubious ORF YER087C-A
chrV	349685	349819	23	-	None
chrV	370496	370633	20	+	NUP157
chrV	393614	393711	20	+	BOI2
chrV	400207	400284	20	+	AVT6
chrV	400995	401141	24	+	Dubious ORF YER119C-A
chrV	424758	424884	20	+	SNORNA SNR4
chrV	426641	426744	20	+	PMD1
chrV	427565	427678	20	+	PMD1
chrV	460938	461084	22.5	-	FTR1
chrV	461503	461640	23	-	FTR1
chrV	514230	514341	20	+	DNF1
chrV	548814	548961	22	+	DMC1
chrV	553456	553591	24	-	FAU1
chrV	566930	567070	21	-	None
chrV	569248	569349	21	-	None ARS523 X element
chrV	569581	569871	23	+	YER188C-A ARS523 X element
chrV	569892	570341	31	+	None ARS523 X element
chrV	575162	575344	26	-	YRF1-2

chrVI	4618	5167	75	-	None X element
chrVI	5394	5548	23	-	Dubious ORF YFL063W and ARS600 X element
chrVI	11181	11322	21	-	None
chrVI	23589	23673	21	+	DAK2
chrVI	35548	35651	20	-	ALR2
chrVI	55546	55667	21	-	YPT1
chrVI	78414	78525	20	-	CAK1
chrVI	96645	96761	21	-	GAT1
chrVI	97157	97303	22	-	GAT1
chrVI	106133	106270	20	-	MDJ1
chrVI	106625	106762	21	+	Dubious ORF YFL015W-A and Dubious ORF YFL015C
chrVI	112948	113088	25	+	HXT10
chrVI	115238	115373	30.5	-	AUA1 and WWM1
chrVI	126905	127027	20	-	BLM10
chrVI	138522	138668	26.5	-	Transposon YFL002W-A
chrVI	141631	141732	23	-	Transposon YFL002W-A
chrVI	153523	153630	20	+	RPN11
chrVI	167181	167299	20	-	AIM13
chrVI	182776	182961	32	+	YFR017C
chrVI	204667	204805	20	+	HIS2
chrVI	210286	210471	61	+	None
chrVI	210825	210959	23	-	PTR3
chrVI	211461	211590	23	-	PTR3
chrVI	213016	213130	24	-	None
chrVI	224738	224878	26	+	QCR6
chrVI	228113	228235	20.5	-	RSC8
chrVI	229042	229181	20.5	+	RSC8
chrVI	249560	249688	20	-	PRE4
chrVI	254656	254763	21	+	HXK1
chrVII	35	153	28	-	None X element
chrVII	177	579	27	-	None X element
chrVII	645	749	20	-	None X element
chrVII	21809	21941	24	+	ZRT1
chrVII	58149	58271	22.5	+	ADE5,7
chrVII	63521	63598	20	+	EMC4
chrVII	67089	67235	28	-	None
chrVII	69281	69446	34	-	SHE10
chrVII	81932	82071	20	+	NIF3
chrVII	117299	117436	22	+	ARO8
chrVII	119659	119785	20	-	MCM6
chrVII	124140	124265	20	+	YIP4

chrVII	152218	152298	24	-	TPN1
chrVII	157983	158117	24	+	GTS1
chrVII	175778	175934	25	+	KEM1
chrVII	188359	188460	24	-	PMR1
chrVII	197164	197302	20	-	None
chrVII	217679	217789	20	-	NUT1
chrVII	225646	225786	25.5	-	Dubious ORF YGL149W
chrVII	260638	260739	20	-	ITC1
chrVII	307540	307636	20	-	ARC1
chrVII	310631	310726	21	-	None
chrVII	337889	338026	23	-	NUP145
chrVII	352383	352502	24	-	None
chrVII	361154	361231	20.5	-	DBP3
chrVII	370992	371147	21.5	-	HSF1 and Dubious ORF YGL072C
chrVII	371299	371438	20	-	Dubious ORF YGL072C
chrVII	409137	409270	25	+	TIF4632
chrVII	420238	420359	20	-	HEM2
chrVII	446786	446884	21	-	TRP5
chrVII	449434	449583	23.5	+	PGD1
chrVII	454298	454389	20	-	STT3
chrVII	455433	455567	20	-	ALK1
chrVII	479573	479721	31	+	None
chrVII	570880	570981	23	+	Transposon YGRCTy2-1
chrVII	573945	574091	26.5	+	Transposon YGRCTy2-1
chrVII	587125	587242	20	-	TFC4
chrVII	612437	612572	27	-	ADE6
chrVII	614819	614934	22	-	ADE6
chrVII	635779	635861	23	-	SMD1 and Dubious ORF YGR073C
chrVII	640943	641059	20.5	-	YGR079W
chrVII	709039	709153	21	-	Transposon YGR109W-B
chrVII	726533	726678	25	-	None
chrVII	753190	753288	23	+	YGR130C
chrVII	756241	756388	24.5	-	PHB1
chrVII	778523	778653	21	+	None
chrVII	789929	790016	20	-	YGR149W
chrVII	806491	806638	27	-	NSR1
chrVII	808769	808870	21	+	RTS3
chrVII	808909	809056	23	+	RTS3
chrVII	812056	812202	26.5	-	Transposon YGR161W-B
chrVII	815165	815266	23	-	Retrotransposon YGR161W-B
chrVII	844149	844271	21	+	RBG2

chrVII	856674	856934	27.5	+	None
chrVII	865274	865410	22	-	UBR1
chrVII	884770	884906	24	+	PDX1
chrVII	906167	906278	20	-	ADE3
chrVII	948553	948747	22	-	DIE2
chrVII	963385	963541	20	+	YGR237C
chrVII	982873	982999	22	+	BRF1
chrVII	998718	998836	20	-	PUP2
chrVII	1000423	1000569	25.5	-	None
chrVII	1000967	1001155	27	+	ENO1
chrVII	1004958	1005084	23.5	+	GND2
chrVII	1073222	1073353	20	-	None
chrVII	1083924	1084344	34	+	None ARS736
chrVII	1089228	1089410	26	-	YRF1-3
chrVIII	5005	5344	32	-	None X element
chrVIII	5369	5523	23	-	None X element ARS801
chrVIII	9504	9613	23	-	ARN2
chrVIII	22047	22148	20	-	EFM1
chrVIII	26629	26763	21	+	MUP3
chrVIII	37066	37187	21	-	GUT1
chrVIII	50103	50242	28	-	WSC4
chrVIII	159715	159861	29	+	THR1
chrVIII	167202	167330	20	+	DAP2
chrVIII	173948	174105	21.5	+	ERC1
chrVIII	174946	175153	20	+	ERC1
chrVIII	187252	187365	21	-	VMA10
chrVIII	187864	188015	26	+	BCD1
chrVIII	192427	192561	23	+	NCP1
chrVIII	197487	197650	35	-	INM1
chrVIII	235952	236092	21	-	TRM5
chrVIII	282790	282917	21.5	-	GAR1
chrVIII	282943	283081	21	-	GAR1
chrVIII	298180	298267	22	+	YHR097C
chrVIII	301299	301417	23	+	SFB3
chrVIII	362279	362414	20	+	FUR1
chrVIII	380565	380682	21	-	YHR140W
chrVIII	383857	383965	20	+	CHS7
chrVIII	384391	384531	20	+	CHS7
chrVIII	398807	398940	21	+	PEX28
chrVIII	416567	416700	20	+	KEL1
chrVIII	439732	439856	20	+	THP2

chrVIII	451367	451526	23	+	ENO2
chrVIII	470628	470766	26	+	None
chrVIII	498701	498833	21	+	YHR199C-A_mRNA and NBL1
chrVIII	505654	505792	20	-	None
chrVIII	514493	514606	22	-	SKN7
chrVIII	533134	533268	22	-	None
chrVIII	556150	556268	24	+	None ARS824 X element
chrVIII	556399	556727	27	+	None X element
chrVIII	556732	557051	75	+	None telomeric repeats
chrVIII	560681	560863	26	-	YHR219W
chrX	7264	7722	34	-	None ARS1002
chrX	19178	19324	24	-	None
chrX	33944	34053	22	-	OPT1
chrX	35833	35978	20	-	OPT1
chrX	36533	36660	20	+	None
chrX	62819	62961	23	+	PHO90
chrX	96501	96613	21	+	RFA3
chrX	147156	147306	23.5	+	TIM17
chrX	147408	147541	25	+	TIM17
chrX	148283	148380	21	+	YAK1
chrX	160110	160214	22	-	None
chrX	169338	169440	20	-	URA2
chrX	182780	182896	20	+	SPT10
chrX	221921	222015	22	-	IME2
chrX	277245	277371	20	+	ALY2
chrX	280007	280118	22	-	TAX4
chrX	290296	290428	22	+	PRY1
chrX	313924	314063	20	+	None
chrX	317717	317837	20	-	LAS21
chrX	360332	360446	22	-	YJL043W
chrX	382534	382650	22	+	KAR2
chrX	389935	390012	20	+	VPS53
chrX	400076	400374	34	+	BBC1
chrX	406141	406280	24	+	YJL016W
chrX	426323	426455	21.5	-	CYR1
chrX	438513	438639	21	+	AVT1
chrX	464831	464978	23	-	ILV3
chrX	491598	491743	21	+	CPR7
chrX	519196	519361	41	+	None
chrX	522706	522851	25	-	TAH11
chrX	526247	526478	20	-	WITH CYC1

chrX	527558	527660	23	+	UTR1
chrX	531223	531346	22	+	OSM1
chrX	569121	569251	20	-	HAM1
chrX	572987	573127	23	+	MOG1
chrX	577893	578054	20	-	MIR1
chrX	604965	605110	21	+	IME1
chrX	635820	635961	21	-	YJR111C
chrX	639927	640088	42	-	YJR115W unknown function
chrX	640955	641165	20	-	YJR116W
chrX	689438	689557	22	-	HOM6
chrX	690643	690790	24	-	HIR3
chrX	693257	693378	20	+	HIR3
chrX	701686	701839	33	-	MGM101
chrX	708807	708923	21	-	DAN1
chrX	736231	736366	25	+	SOR1
chrX	744891	745513	26	+	None ARS1025
chrX	745517	745663	32	+	None X element
chrXI	68	519	29	-	None X element
chrXI	660	808	23	-	Dubious ORF YKL225W X element
chrXI	40171	40306	20	+	UBA1
chrXI	68288	68425	20	+	PTK1
chrXI	122698	122836	21	-	SNU114
chrXI	217828	217936	22	+	OAC1
chrXI	220760	220852	20	+	SBA1
chrXI	243839	243970	20	-	GFA1
chrXI	273770	273868	20	-	MIF2
chrXI	337450	337649	35	+	DEF1
chrXI	369843	369982	22	-	Dubious ORF YKL036C and UGP1
chrXI	382891	383019	21	+	MAE1 and YKL030W
chrXI	390172	390287	20	+	None
chrXI	464301	464488	24	-	PRY2
chrXI	488948	489114	28	-	GCN3
chrXI	502583	502727	20	-	SPO14
chrXI	533693	533826	21	+	None
chrXI	539985	540109	21	+	DYN1
chrXI	570398	570533	21	-	BET3
chrXI	583257	583393	23	-	ECM4
chrXI	583623	583740	22	-	MSA2
chrXI	616040	616163	20	+	PTR2
chrXI	665891	666082	25	+	None X element
chrXI	666105	666604	24	+	None X element

chrXII	5386	5522	20	-	None
chrXII	5595	5830	64	-	None telomeric repeats
chrXII	7304	7486	26	+	YLL066C
chrXII	10921	11057	20	-	None
chrXII	11130	11829	75	-	None telomeric repeats
chrXII	11920	12067	32	-	None X element
chrXII	85972	86103	24	+	TPO1
chrXII	99271	99401	21.5	-	HIF1
chrXII	108944	109057	21	+	COX 19
chrXII	151451	151614	25	+	YLR001C
chrXII	171540	171671	25	+	GAT3
chrXII	177772	177945	21.5	-	MEU1
chrXII	189561	189701	24	+	UBR2
chrXII	245559	245663	21	-	YLR050C
chrXII	247346	247483	20	-	IES3
chrXII	251591	251729	21	-	SPT8
chrXII	262431	262569	22	+	FRS1
chrXII	329720	329867	22	+	IOC2
chrXII	340198	340316	21.5	+	ICT1
chrXII	347144	347290	22	-	SEN2
chrXII	358289	358380	20	-	MDN1
chrXII	387205	387325	21	+	YPS1
chrXII	405393	405579	24	+	ACE2
chrXII	414619	414709	21	-	SLX4
chrXII	419156	419302	21	+	NHA1
chrXII	440673	440798	23	-	STM1
chrXII	446310	446456	20	-	ACS2
chrXII	459565	459678	22.5	-	NTS2-1
chrXII	468702	468815	22.5	-	NTS2-2
chrXII	489457	489654	27	-	YLR162W and RDN5-6
chrXII	499297	499400	20	+	RPS31
chrXII	515739	515848	23	-	SAM1
chrXII	553577	553770	23	+	HMX1
chrXII	563417	563574	32	-	None
chrXII	567238	567317	20.5	-	CRR1
chrXII	577669	577822	25.5	+	CCC1
chrXII	613647	613764	20	+	THI7
chrXII	642309	642442	22	+	None
chrXII	645884	646022	24.5	-	Dubious ORF YLR255C
chrXII	689354	689468	20	+	PIG1
chrXII	702946	703064	23.5	+	YLR278C

chrXII	736674	736785	22	-	ACO1
chrXII	765355	765451	20	-	TAD3
chrXII	791304	791436	20	+	MID3
chrXII	846438	846584	20	-	VPS38
chrXII	860914	861038	20	+	SSQ1
chrXII	903306	903438	22.5	-	ECM19
chrXII	906505	906642	21	-	None
chrXII	915012	915153	21	+	None
chrXII	917200	917331	21	-	SKI2
chrXII	941801	941947	26.5	-	Transposon YLR410W-B
chrXII	944910	945011	23	-	Transposon YLR410W-B
chrXII	948793	948906	20	-	BER1
chrXII	977881	977982	23	+	Transposon YLRCTy2-2
chrXII	980945	981091	24.5	+	LTR Transposon YLRCTy2-2
chrXII	1004462	1004596	20	-	CNA1
chrXII	1012401	1012564	28	-	CAR2
chrXII	1014131	1014278	28	+	ISM3
chrXII	1015453	1015583	25	-	SEC39
chrXII	1024634	1024744	20	+	YLR445W
chrXII	1064313	1064417	23	+	None X element
chrXII	1064577	1065221	75	+	None X element
chrXII	1069869	1070051	26	-	YRF1-4
chrXII	1076532	1076714	26	-	YRF1-5
chrXIII	1645	1827	26	+	YML133C
chrXIII	5262	5398	20	-	None
chrXIII	5471	6052	60	-	None X element
chrXIII	6166	6355	25	-	None X element
chrXIII	28082	28228	22	+	None
chrXIII	29029	29172	21	-	NDI1
chrXIII	34343	34468	20	-	NAB6
chrXIII	45472	45604	21	-	CTK3
chrXIII	69096	69197	22	+	CAC2
chrXIII	70060	70201	21	+	CUE4 and YML100W-A
chrXIII	81748	81879	21	+	RAD10
chrXIII	133614	133704	22	+	DAK1
chrXIII	146035	146181	23	-	None
chrXIII	158004	158143	20	-	NTE1
chrXIII	165721	165898	28.5	-	CYB2
chrXIII	166397	166543	20	-	CYB2
chrXIII	175776	175884	21	-	RSE1
chrXIII	196120	196234	20	-	TRNA (tP(UGG)M)

chrXIII	222976	223089	20.5	+	RPS18B
chrXIII	241046	241153	20	+	PPZ1
chrXIII	244656	244779	21	-	UBX2
chrXIII	274096	274183	22	-	MVP1
chrXIII	306647	306776	22	+	None
chrXIII	325785	325892	20	+	None
chrXIII	336274	336419	20	-	HOF1
chrXIII	347523	347649	21	-	CCS1
chrXIII	362163	362291	20	+	LTR Transposon YMRCTy1-3
chrXIII	387644	387771	22	-	AAC1
chrXIII	408334	408540	21	-	NAT4
chrXIII	435329	435485	20	+	ADH3
chrXIII	489687	489823	20	-	MYO5
chrXIII	491419	491559	21	+	HFD1
chrXIII	513645	513741	20	-	PKR1
chrXIII	538547	538773	44	-	GID8
chrXIII	539734	539867	20	+	GID8
chrXIII	624584	624728	21	+	RGM1
chrXIII	638454	638553	21	+	GCV2
chrXIII	655274	655399	21	+	YMR196W
chrXIII	673820	673944	21.5	-	PFK2
chrXIII	695916	696054	24.5	+	SCJ1
chrXIII	705725	705862	20	-	TRS130
chrXIII	716831	716952	21	+	UBP8
chrXIII	771707	771800	21	+	GAD1
chrXIII	790750	790870	21	-	TPS3
chrXIII	812134	812278	22	+	ZDS1
chrXIII	817737	817883	27	+	BUL1
chrXIII	821830	821984	20	-	FCP1
chrXIII	860290	860424	22	-	LCB1
chrXIII	874058	874181	23	-	ADH2
chrXIII	913370	913508	21	-	FET4
chrXIII	923584	923717	20	+	Dubious ORF YMR326C
chrXIII	923810	924183	33	+	None X element
chrXIII	924200	924305	32	+	None X element
chrXIV	1552	1734	26	+	YRF1-6 and YNL339W-A
chrXIV	6551	6819	74	-	None telomeric repeats
chrXIV	6923	7446	27	-	None X element
chrXIV	13085	13226	21	-	None
chrXIV	20040	20167	20.5	-	PEX6
chrXIV	34893	35046	30.5	-	VNX1

chrXIV	51256	51385	22	+	SKP2
chrXIV	52997	53127	21	-	STB1
chrXIV	53491	53629	23.5	-	STB1
chrXIV	65841	65987	25	-	TOS6
chrXIV	79298	79459	23	-	RIM21
chrXIV	88032	88165	20	-	PCL1
chrXIV	103756	103878	23	-	MRPL10
chrXIV	105647	105772	22	-	WSC2
chrXIV	115669	115815	26	-	CAF120
chrXIV	131440	131588	25	+	BNI1
chrXIV	184442	184543	20	-	YNL247W
chrXIV	241500	241646	23	-	None
chrXIV	254474	254655	21	+	YNL208W
chrXIV	289985	290131	24	+	UBP10
chrXIV	358303	358449	22	+	MEP2
chrXIV	391049	391182	20.5	-	ESBP6
chrXIV	395950	396027	20	+	NMA111
chrXIV	453917	454033	20.5	-	NST1
chrXIV	506790	506912	21	-	YDJ1
chrXIV	564172	564273	23	+	TransposonYNLCTy2-1
chrXIV	567236	567382	26.5	+	Transposon YNLCTy2-1
chrXIV	580495	580609	21	-	CRZ1
chrXIV	594028	594175	23	+	HDA1
chrXIV	599364	599487	21	+	None
chrXIV	665421	665565	24	-	ARE2
chrXIV	678011	678127	23	-	YNR029C
chrXIV	694366	694506	29	-	MRPS12
chrXIV	701213	701343	21	-	COQ2 AND Dubious ORF YNR042W
chrXIV	705462	705608	21	-	AGA1
chrXIV	734411	734536	23.5	+	BIO3
chrXIV	745327	745495	23.5	-	YNR062C
chrXIV	778614	778761	27.5	-	AIF1
chrXIV	783296	784038	32	+	YNR077C
chrXV	117	646	29.5	-	None X element
chrXV	711	865	24	-	YOL166W-A X element
chrXV	19945	20079	22.5	-	ENB1
chrXV	24952	25098	23	-	None
chrXV	29124	29244	21	+	HPF1
chrXV	29284	29417	23	+	HPF1
chrXV	43035	43137	20	-	None
chrXV	68465	68598	20	+	PFK27

chrXV	74752	74851	20	-	ALR1
chrXV	74948	75074	22	-	ALR1
chrXV	81582	81731	35	+	None
chrXV	90116	90256	24	-	SMF1
chrXV	109585	109725	25.5	+	SHR5
chrXV	109836	109973	20	+	SHR5
chrXV	160315	160438	23	-	ADH1
chrXV	171825	171960	22	-	IRA2
chrXV	180683	180782	20	+	REX4
chrXV	218622	218739	22	-	None
chrXV	224584	224685	20	-	THI20
chrXV	236363	236484	24	-	GAL11
chrXV	268813	268962	23	+	GAS5
chrXV	280228	280339	20	-	TSR4
chrXV	301242	301381	20	+	None
chrXV	341688	341819	22	+	SLG1
chrXV	358326	358472	21	-	RTS1
chrXV	375658	375786	23	+	YOR022C
chrXV	376474	376613	22	+	AHC1
chrXV	377916	378049	20	+	Dubious ORF YOR024W
chrXV	383874	384005	20	+	CIN5
chrXV	388254	388358	20	-	DFG16
chrXV	396817	396953	20	-	AKR2
chrXV	459877	459996	22	-	NRT1
chrXV	474034	474163	21	-	ATX2
chrXV	480226	480410	21	-	WHI5
chrXV	481926	482051	20	-	None
chrXV	494456	494594	22	-	TMA46
chrXV	502423	502544	20	-	YOR093C
chrXV	509442	509563	22	+	NUP1
chrXV	539996	540115	21	-	RPO31
chrXV	543601	543745	20	-	RPO31
chrXV	570351	570483	20	-	ORT1
chrXV	586771	586917	22	-	None
chrXV	630630	630761	20	+	None
chrXV	639184	639388	21	+	YOR161C-C
chrXV	699168	699287	22	-	THI72
chrXV	706204	706305	23	+	Transposon YOR192C-B
chrXV	709268	709414	24.5	+	LTR Transposon YORCTy2-1
chrXV	761800	761974	21	-	ISU2
chrXV	797788	797901	23	-	SRL1

chrXV	798149	798303	22	-	SRL1 and Dubious ORF YOR248W
chrXV	824130	824263	21.5	-	HRK1
chrXV	835320	835410	20	-	TPO4
chrXV	859201	859347	25	+	SNF2
chrXV	881929	882057	20	+	RAX1
chrXV	882142	882288	22	+	RAX1
chrXV	883871	883995	22	-	CPA1
chrXV	902738	902866	21	+	SPS4
chrXV	904762	904858	20	-	SFG1
chrXV	918636	918777	24	-	LDB19
chrXV	927759	927848	20	+	MYO2
chrXV	944642	944770	21	-	MRS2 and Dubious ORF YOR333C
chrXV	970896	971152	24.5	-	LTR Transposon YORWTy2-2
chrXV	974005	974125	24	-	Retrotransposon YOR343W-B
chrXV	985409	985545	27	-	PYK2
chrXV	987332	987465	26	-	PUT4
chrXV	988086	988205	20.5	-	PUT4
chrXV	999175	999300	23	+	SOG2
chrXV	1032043	1032157	21	+	GPB1
chrXV	1039177	1039312	20	-	NUD1
chrXV	1040504	1040665	21	+	ALD4
chrXV	1083954	1084100	23	+	None X ELEMENT
chrXV	1084261	1084640	27	+	YOR394C-A (YOR394C-A)
chrXV	1089497	1089679	26	-	YRF1-8
chrXVI	1461	1643	26	+	YRF1-7
chrXVI	6505	6884	27	-	None X element
chrXVI	7045	7191	23	-	None X element
chrXVI	27421	27553	20	+	YPL272C
chrXVI	72980	73123	20	+	VIK1
chrXVI	73720	73866	22.5	+	YAH1 and Dubious ORF YPL251W
chrXVI	74877	75051	24	+	None
chrXVI	75324	75470	21	+	None
chrXVI	105053	105199	25	-	VMA11
chrXVI	109517	109654	23	+	FAS2
chrXVI	133239	133343	22	+	FLC1
chrXVI	141903	142006	21	-	BMS1
chrXVI	149113	149251	20	+	THI6
chrXVI	209076	209181	20	+	PPQ1
chrXVI	229267	229398	23.5	+	MEX67
chrXVI	242321	242474	26.5	+	SVS1
chrXVI	264106	264243	20	+	RAD53

chrXVI	295825	295936	21	+	GIP3
chrXVI	307091	307230	20	+	TBF1
chrXVI	340177	340334	23	+	CAR1
chrXVI	341064	341258	20	-	GDE1
chrXVI	375627	375766	24	+	GLR1
chrXVI	378614	378753	22	+	None
chrXVI	424859	424992	21	+	YPL068C
chrXVI	426899	427037	20	-	YPL066W
chrXVI	428046	428152	20	-	VPS28
chrXVI	432535	432642	21	-	ALD6
chrXVI	447037	447136	21	-	PDR12
chrXVI	453594	453743	25.5	+	LCL1
chrXVI	455226	455371	21	+	None
chrXVI	462336	462508	36	+	DIG1
chrXVI	485860	486015	31	-	None
chrXVI	494209	494345	20	+	TRM44
chrXVI	498506	498618	20	+	ERG10
chrXVI	583575	583707	22.5	-	YPR011C
chrXVI	673954	674088	22	+	YMC1
chrXVI	717765	717855	20	+	YPR091C
chrXVI	815352	815456	20	-	TAZ1
chrXVI	823111	823235	23.5	-	ASN1
chrXVI	832078	832270	25	-	URN1
chrXVI	893332	893446	21	-	PRP4
chrXVI	900087	900239	29	-	SMX3
chrXVI	902709	902803	21	-	GDB1
chrXVI	942415	942559	26	+	None ARS 1632
chrXVI	942659	942840	25	+	None ARS1632

Table 8. Coordinates of nonessential and subtelomeric regions

Chr.	left border of Nonessential region	Cen left	Cen right	right border of Nonessential region	left subtel start	left subtel end	right subtel start	right subtel end
1	58695	151465	151582	175135	1	801	229411	230218
2	14241	238207	238323	739544	1	6608	812379	813184
3	22429	114385	114501	286443	1	1098	315783	316620
4	33415	449711	449821	1499335	1	904	1524625	1531933
5	43252	151987	152104	536021	1	6473	569599	576874
6	43628	148510	148627	253329	1	5530	269731	270161
7	36933	496920	497038	1051725	1	781	1083635	1090940

8	75412	105586	105703	498422	1	5505	556105	562643
9	61013	355629	355745	399236	1	7784	439068	439888
10	53341	436307	436425	696949	1	7767	744902	745751
11	39163	440129	440246	603072	1	807	665904	666816
12	39804	150828	150947	1058518	1	12085	1064281	1078177
13	11483	268031	268149	902413	1	6344	923541	924431
14	40619	628758	628875	724306	1	7428	783278	784333
15	44938	326584	326702	1039389	1	847	1083922	1091291
16	39121	555957	556073	919041	1	7223	942396	948010

Table 9. List of perfect telomere and TG repeats

chromosome	start	End	CATH I score	orientation	Subtelomere	Type	Conservation analysis
chrI	31423	31568	26	TG	no	TG repeat	y
chrII	353576	353721	26	TG	no	TG repeat	y
chrII	504291	504443	29	CA	no	TG repeat	y
chrIV	147612	147757	21	TG	no	TG repeat	N
chrIV	392081	392233	27	TG	no	TG repeat	Y
chrV	284369	284510	22	TG	no	TG repeat	Y
chrVI	11181	11322	21	TG	no	TG repeat	Y
chrVI	210286	210471	61	TG	no	TG repeat	Y
chrVII	69281	69446	34	TG	no	TG repeat	Y
chrVII	479573	479721	31	CA	no	TG repeat	N
chrX	313924	314063	20	CA	no	TG repeat	N
chrX	519196	519361	41	TG	no	TG repeat	Y
chrX	639927	640088	42	CA	no	TG repeat	Y
chrX	701686	701839	33	CA	no	TG repeat	Y
chrXII	563417	563574	32	CA	no	TG repeat	Y
chrXII	915012	915153	21	TG	no	TG repeat	N

chrXIII	538547	538773	44	CA	no	TG repeat	Y
chrXIV	13085	13226	21	TG	no	TG repeat	Y
chrXV	81582	81731	35	CA	no	TG repeat	N
chrXVI	485860	486015	31	TG	no	TG repeat	Y
chrVI	4618	5167	75	TG	yes	tel repeat	N/A
chrVIII	556732	557051	75	TG	yes	tel repeat	N/A
chrXII	11130	11829	75	TG	yes	tel repeat	N/A
chrXII	1064577	1065221	75	TG	yes	tel repeat	N/A
chrXIV	6551	6819	74	TG	yes	tel repeat	N/A
chrIV	1525254	1525532	73.5	TG	yes	tel repeat	N/A
chrXII	5595	5830	64	TG	yes	tel repeat	N/A
chrXIII	5471	6052	60	TG	yes	tel repeat	N/A
chrV	6269	6441	36	TG	yes	tel repeat	N/A
chrIV	1525254	1525532	73.5	TG	yes	TG repeat	N/A
chrVI	4618	5167	75	TG	yes	TG repeat	N/A
chrVIII	556732	557051	75	TG	yes	TG repeat	N/A
chrXII	5595	5830	64	TG	yes	TG repeat	N/A
chrXII	11130	11829	75	TG	yes	TG repeat	N/A
chrXII	1064577	1065221	75	TG	yes	TG repeat	N/A
chrXIV	6551	6819	74	TG	yes	TG repeat	N/A

Table 10. Oligos used for Fluorescence polarization experiments

Oligo Name	Sequence 5' to 3'	Length	Description
6L22(-)75	AACTGTAACAAAAGGTAGTGGTGGTTCCTATGAATCCCTATCATTTGCATGGGGGGTTCGGTTGTATGCTTGGTGT	75	polarization oligo negative control
14L35(-)75	GTGTGGGCGGCAAGTGTTGTTGGTGCTCGTGGTGGTGCAGCGTGTTGTTGCTGCTATGTACGATTGGTGATTGCG	75	polarization oligo SiRTA high efficiency
2R780(-)75	TGGGGATGTAAGTTGTGTGGTGCTGGAGGAGGAGGAGGGGTAGAAGTCGCTGATTGATTAGCAATGGGTTGAGAT	75	polarization oligo SiRTA false negative

14R306(+)75	GGTGGGTGTTGATGGTGAGAAAATAGAAGATCTGCTGTCGAACGAAAACAATTTAGTGTCTGAATATGATGAAGA	75	polarization oligo false negative
12R330(+)75	TGTTGCTGTTGCTGCTGTGGTTGGATTGATGCTTGTGATGTTTGTGGTGCCTGTGGTGCTTGTGGTACAGTTTTG	75	polarization oligo false positive
14R131(+)75	GTGGTGGTGTTTCTCCCTTGGGTTTGCCGAACAATGCCATGGGTGGAGGAGGAGGTGGTGGTGGTGGAGGGGGAG	75	polarization oligo SiRTA
2R780(-)75m1	TGGGGATGTAAGTTCACAGGTGCTGGAGGAGGAGGAGGGGTAGAAGTCGCTGATTGATTAGCAATGGGTTGAGAT	75	polarization oligo mutated SiRTA
2R780(-)75m2	TGGGGATGTAAGTTGTGTGGTGCTGGAGGAGGAGGAGCCCAAGAAGTCGCTGATTGATTAGCAATGGGTTGAGAT	75	polarization oligo mutated SiRTA
2R780(-)75m1m2	TGGGGATGTAAGTTCACAGGTGCTGGAGGAGGAGGAGCCCAAGAAGTCGCTGATTGATTAGCAATGGGTTGAGAT	75	polarization oligo mutated SiRTA
6R210(+)75	CATAGCTTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGATTGTT	75	polarization oligo TG repeat SiRTA
tel11-75	GCTAGATCGAGCGCTCCTTTTATCACATTTCCGTGTGGGTGTGCTCGACGACCGCCGCCTTCGCTCTTTCGCTAG	75	polarization oligo unlabeled control
tel11-75rev	CTAGCGAAAGAGCGAAGGCGGCGGTCGTCGAGCACACCCACACGGAAATGTGATAAAAGGAGCGCTCGATCTAGC	75	polarization oligo negative control
2xtel11-75	TCGAGCGCTCCTTTTATCACATTTCCGTGTGGGTGTGGTGTGGGTGTGCTCGACGACCGCCGCCTTCGCTCTTTC	75	polarization oligo 2xTel11
Tel 11 6-FAM™ (fluorescein) 5'	GTGTGGGTGTG	11	polarization oligo labeled Tel11

Table 11. Frequencies of the usage of TG- and CA-rich codons

Codon	Amino acid	Percent	Codon	Amino acid	Percent
CAA	Gln	2.73	TTG	Leu	2.72
AAC	Asn	2.48	GTT	Val	2.21

ACC	Thr	1.27	GGT	Gly	2.39
CCA	Pro	1.83	TGG	Trp	1.04
CCC	Pro	0.68	GGG	Gly	0.60
CAC	His	0.78	GTG	Val	1.08
ACA	Thr	1.78	TGT	Cys	0.81
	CA Total	11.55		TG Total	10.85

Table 12. GCR frequencies of the close HO assay

Strain and trial	GCR rate(%)
5L35wta	1.0780
5L35wtb	1.0800
5L35wtc	0.7600
5L35pif1m2a	23.7000
5L35pif1m2b	47.1000
5L35pif1m2c	46.0000
5L35pif1m2d	22.7000
6L22pif1m2a	13.0680
6L22pif1m2b	0.0497
6L22pif1m2c	0.0192
6L22howta	0.0170
6L22howtb	0.0006
6L22howtc	0.0000
1cdc13WT	0.0089
1cdc13closeWT 6/3 *	0.0178
1cdc13closeWT 6/3	0.0090
1cdc13closepif1m2 4/26	78.8700
1cdc13closepif1m2 4/2	41.5800
1cdc13closepif1m2 4/2	45.5400
2cdc13closeWTA	40.0000
2cdc13closeWTB	37.5000
2cdc13closeWTC	19.7000
2cdc13closepif1m2A	77.0000
2cdc13closepif1m2B	10.0000
2cdc13closepif1m2B	12.7500

Author Contributions:

Katrina Ngo: Wrote initial draft, formulated research questions, performed experiments and curated/managed data collection.

Tristen Gittens: specifically performed the experiments that contributed to Figures 2, 5, and 15.

She also contributed to the making of Figure 5. Performed initial experiments discussed in

section 5.1.2.

David Gonzalez: Performed experiments and contributed data to figures 2 and 5. Performed data

visualization for Figure 16.

E. Hatmaker: Implemented computer code and supporting algorithms for Figure 17 and Table 7.

Simcha Plotkin, Mason Engle: Contributed data to Figures 2 and 5.

Geofrey Friedman: Designed and implemented implementation of the computer code and

supporting algorithms for Figures 7 and 9.

Melissa Goldin: Contributed data to Figures 2 and 15.

Liraz Stilman: Data collection for Figure 22.

Remington Hoerr: Preformed the experiments for Figures 2 and 5. Figure creation of Figure 1.

Brandt Eichman: Gave guidance and materials for the fluorescence polarization experiments

(Figures 13, 14, and 15).

Antonis Rokas: Gave guidance regarding computational methods and approaches.

Mary Lauren Benton: Contributed to the development of the CATHI and specifically performed

the experiments regarding the implementation of the CATHI and the analysis of CATHI hits.

Katherine Friedman: Supervised research, formulated research questions and helped with

commentary and revision.

References

Anderson E. M., Wayne A Halsey, and Deborah S. Wuttke, 2002 Delineation of the high-affinity single-stranded telomeric DNA-binding domain of Saccharomyces cerevisiae Cdc13. Nucleic Acids Res 30: 4305–4313. https://doi.org/10.1093/nar/gkf554

Anderson E. M., W. A. Halsey, and D. S. Wuttke, 2003 Site-Directed Mutagenesis Reveals the Thermodynamic Requirements for Single-Stranded DNA Recognition by the Telomere-Binding Protein Cdc13. Biochemistry 42: 3751–3758. https://doi.org/10.1021/bi027047c

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

Andriuskevicius T., O. Kotenko, and S. Makovets, 2018 Putting together and taking apart: assembly and disassembly of the Rad51 nucleoprotein filament in DNA repair and genome stability. Cell Stress 2: 96–112. https://doi.org/10.15698/cst2018.05.134

Azad Kumar Gajendra, and R. S. Tomar, 2016 The multifunctional transcription factor Rap1 a regulator of yeast physiology. Frontiers in Bioscience 21: 4429. https://doi.org/10.2741/4429

Blackburn E. H., and J. G. Gall, 1978 A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol 120: 33–53. https://doi.org/10.1016/0022-2836(78)90294-2

Blackburn E. H., 1991 Structure and function of telomeres. Nature 350: 569–573. https://doi.org/10.1038/350569a0

Blackburn E. H., C. W. Greider, and J. W. Szostak, 2006 Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nat Med 12: 1133–1138. https://doi.org/10.1038/nm1006-1133

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

Bonaglia M. C., R. Giorda, S. Beri, C. De Agostini, F. Novara, *et al.*, 2011 Molecular Mechanisms Generating and Stabilizing Terminal 22q13 Deletions in 44 Subjects with Phelan/McDermid Syndrome. PLoS Genet 7: e1002173. https://doi.org/10.1371/journal.pgen.1002173

Bonnell E., E. Pasquier, and R. J. Wellinger, 2021 Telomere Replication: Solving Multiple End Replication Problems. Front Cell Dev Biol 9. https://doi.org/10.3389/fcell.2021.668171

Boulé J.-B., L. R. Vega, and V. A. Zakian, 2005 The yeast Pif1p helicase removes telomerase from telomeric DNA. Nature 438: 57–61. https://doi.org/10.1038/nature04091

Boule J.-B., and V. A. Zakian, 2007 The yeast Pif1p DNA helicase preferentially unwinds RNA DNA substrates. Nucleic Acids Res 35: 5809–5818. https://doi.org/10.1093/nar/gkm613

Capra J. A., K. Paeschke, M. Singh, and V. A. Zakian, 2010 G-Quadruplex DNA Sequences Are Evolutionarily Conserved and Associated with Distinct Genomic Features in Saccharomyces cerevisiae. PLoS Comput Biol 6: e1000861. https://doi.org/10.1371/journal.pcbi.1000861

Casari E., M. Gnugnoli, C. Rinaldi, P. Pizzul, C. V. Colombo, *et al.*, 2022 To Fix or Not to Fix: Maintenance of Chromosome Ends Versus Repair of DNA Double-Strand Breaks. Cells 11: 3224. https://doi.org/10.3390/cells11203224

Chandra A., T. R. Hughes, C. I. Nugent, and V. Lundblad, 2001 Cdc13 both positively and negatively regulates telomere replication. Genes Dev 15: 404–414. https://doi.org/10.1101/gad.861001

Chen C., and R. D. Kolodner, 1999 Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. Nat Genet 23: 81–85. https://doi.org/10.1038/12687

Chung W.-H., 2014 To peep into Pif1 helicase: Multifaceted all the way from genome stability to repair-associated DNA synthesis. Journal of Microbiology 52: 89–98. https://doi.org/10.1007/s12275-014-3524-3

Cliften P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, *et al.*, 2003 Finding Functional Features in *Saccharomyces* Genomes by Phylogenetic Footprinting. Science (1979) 301: 71–76. https://doi.org/10.1126/science.1084337

Cock P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, *et al.*, 2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Conrad M. N., J. H. Wright, A. J. Wolf, and V. A. Zakian, 1990 RAP1 protein interacts with yeast telomeres in vivo: Overproduction alters telomere structure and decreases chromosome stability. Cell 63: 739–750. https://doi.org/10.1016/0092-8674(90)90140-A

Dale R. K., B. S. Pedersen, and A. R. Quinlan, 2011 Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27: 3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Davis A. P., and L. S. Symington, 2004 *RAD51* -Dependent Break-Induced Replication in Yeast. Mol Cell Biol 24: 2344–2351. https://doi.org/10.1128/MCB.24.6.2344-2351.2004

Dentro S. C., I. Leshchiner, K. Haase, M. Tarabichi, J. Wintersinger, *et al.*, 2021 Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell 184: 2239-2254.e39. https://doi.org/10.1016/j.cell.2021.03.009

Dewar J. M., and D. Lydall, 2012 Similarities and differences between "uncapped" telomeres and DNA double-strand breaks. Chromosoma 121: 117–130. https://doi.org/10.1007/s00412-011-0357-2

Doksani Y., and T. de Lange, 2014 The Role of Double-Strand Break Repair Pathways at Functional and Dysfunctional Telomeres. Cold Spring Harb Perspect Biol 6: a016576–a016576. https://doi.org/10.1101/cshperspect.a016576

Downs J. A., N. F. Lowndes, and S. P. Jackson, 2000 A role for Saccharomyces cerevisiae histone H2A in DNA repair. Nature 408: 1001–1004. https://doi.org/10.1038/35050000

Eldridge A. M., W. A. Halsey, and D. S. Wuttke, 2006 Identification of the Determinants for the Specific Recognition of Single-Strand Telomeric DNA by Cdc13. Biochemistry 45: 871–879. https://doi.org/10.1021/bi0512703

Epum E. A., M. J. Mohan, N. P. Ruppe, and K. L. Friedman, 2020 Interaction of yeast Rad51 and Rad52 relieves Rad52-mediated inhibition of de novo telomere addition. PLoS Genet 16: e1008608. https://doi.org/10.1371/journal.pgen.1008608

Evans S. K., and V. Lundblad, 1999 Est1 and Cdc13 as Comediators of Telomerase Access. Science (1979) 286: 117–120. https://doi.org/10.1126/science.286.5437.117

Fell V. L., and C. Schild-Poulter, 2015 The Ku heterodimer: Function in DNA repair and beyond. Mutation Research/Reviews in Mutation Research 763: 15–29. https://doi.org/10.1016/j.mrrev.2014.06.002

Foury F., and J. Kolodynski, 1983 pif mutation blocks recombination between mitochondrial rho+ and rho- genomes having tandemly arrayed repeat units in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 80: 5345–5349. https://doi.org/10.1073/pnas.80.17.5345

Garvik B., M. Carson, and L. Hartwell, 1995 Single-Stranded DNA Arising at Telomeres in *cdc13* Mutants May Constitute a Specific Signal for the *RAD9* Checkpoint. Mol Cell Biol 15: 6128–6138. https://doi.org/10.1128/MCB.15.11.6128

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

Giaever G., A. M. Chu, L. Ni, C. Connelly, L. Riles, *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387–391. https://doi.org/10.1038/nature00935

Gibb B., L. F. Ye, Y. Kwon, H. Niu, P. Sung, *et al.*, 2014 Protein dynamics during presynaptic-complex assembly on individual single-stranded DNA molecules. Nat Struct Mol Biol 21: 893–900. https://doi.org/10.1038/nsmb.2886

Gilson E., and V. Géli, 2007 How telomeres are replicated. Nat Rev Mol Cell Biol 8: 825–838. https://doi.org/10.1038/nrm2259

Glousker G., and J. Lingner, 2021 Challenging endings: How telomeres prevent fragility. BioEssays 43: 2100157. https://doi.org/10.1002/bies.202100157

Glustrom L. W., K. R. Lyon, M. Paschini, C. M. Reyes, N. V. Parsonnet, *et al.*, 2018 Single-stranded telomere-binding protein employs a dual rheostat for binding affinity and specificity that drives function. Proceedings of the National Academy of Sciences 115: 10315–10320. https://doi.org/10.1073/pnas.1722147115

Grandin N., S. I. Reed, and M. Charbonneau, 1997 Stn1, a new Saccharomyces cerevisiae protein, is implicated in telomere size regulation in association with Cdc13. Genes Dev 11: 512–527. https://doi.org/10.1101/gad.11.4.512

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

Gravel S., J. R. Chapman, C. Magill, and S. P. Jackson, 2008 DNA helicases Sgs1 and BLM promote DNA double-strand break resection. Genes Dev 22: 2767–2772. https://doi.org/10.1101/gad.503108

Greider C. W., and E. H. Blackburn, 1987 The telomere terminal transferase of tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51: 887–898. https://doi.org/10.1016/0092-8674(87)90576-9

Greider C. W., and E. H. Blackburn, 1989 A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. Nature 337: 331–337. https://doi.org/10.1038/337331a0

Guilherme R. S., K. E. Hermetz, P. T. Varela, A. B. A. Perez, V. A. Meloni, *et al.*, 2015 Terminal 18q deletions are stabilized by neotelomeres. Mol Cytogenet 8: 32. https://doi.org/10.1186/s13039-015-0135-6

Hannes F., J. Van Houdt, O. W. Quarrell, M. Poot, R. Hochstenbach, *et al.*, 2010 Telomere healing following DNA polymerase arrest-induced breakages is likely the main mechanism generating chromosome 4p terminal deletions. Hum Mutat 31: 1343–1351. https://doi.org/10.1002/humu.21368

Harris C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, *et al.*, 2020 Array programming with NumPy. Nature 585: 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hirano Y., and K. Sugimoto, 2007 Cdc13 Telomere Capping Decreases Mec1 Association but Does Not Affect Tel1 Association with DNA Ends. Mol Biol Cell 18: 2026–2036. https://doi.org/10.1091/mbc.e06-12-1074

Hittinger C. T., 2013 Saccharomyces diversity and evolution: a budding model genus. Trends in Genetics 29: 309–317. https://doi.org/10.1016/j.tig.2013.01.002

Hoerr R. E., K. Ngo, and K. L. Friedman, 2021 When the Ends Justify the Means: Regulation of Telomere Addition at Double-Strand Breaks in Yeast. Front Cell Dev Biol 9. https://doi.org/10.3389/fcell.2021.655377

Hoerr R. E., A. Eng, C. Payen, S. C. Di Rienzi, M. K. Raghuraman, *et al.*, 2023 Hotspot of *de novo* telomere addition stabilizes linear amplicons in yeast grown in sulfate-limiting conditions. Genetics. https://doi.org/10.1093/genetics/iyad010

Hughes T. R., R. G. Weilbaecher, M. Walterscheid, and V. Lundblad, 2000 Identification of the single-strand telomeric DNA binding domain of the *Saccharomyces cerevisiae* Cdc13 protein. Proceedings of the National Academy of Sciences 97: 6457–6462. https://doi.org/10.1073/pnas.97.12.6457

Hurowitz E. H., and P. O. Brown, 2003 Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. Genome Biol 5: R2. https://doi.org/10.1186/gb-2003-5-1-r2

Kinzig C. G., G. Zakusilo, K. K. Takai, and T. de Lange, 2022 Neotelomere formation by human telomerase. bioRxiv 2022.10.31.514589. https://doi.org/10.1101/2022.10.31.514589

Klobutcher L. A., M. T. Swanton, P. Donini, and D. M. Prescott, 1981 All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus. Proceedings of the National Academy of Sciences 78: 3015–3019. https://doi.org/10.1073/pnas.78.5.3015

Kramara J., B. Osia, and A. Malkova, 2018 Break-Induced Replication: The Where, The Why, and The How. Trends in Genetics 34: 518–531. https://doi.org/10.1016/j.tig.2018.04.002

Kramer K. M., and J. E. Haber, 1993 New telomeres in yeast are initiated with a highly selected subset of TG1-3 repeats. Genes Dev 7: 2345–2356. https://doi.org/10.1101/gad.7.12a.2345

Lamb J., P. C. Harris, A. O. Wilkie, W. G. Wood, J. G. Dauwerse, *et al.*, 1993 De novo truncation of chromosome 16p and healing with (TTAGGG)n in the alpha-thalassemia/mental retardation syndrome (ATR-16). Am J Hum Genet 52: 668–76.

Lang G. I., 2018 Measuring Mutation Rates Using the Luria-Delbrück Fluctuation Assay, pp. 21–31 in.

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

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

Lendvay T. S., D. K. Morris, J. Sah, B. Balasubramanian, and V. Lundblad, 1996 Senescence Mutants of *Saccharomyces cerevisiae* With a Defect in Telomere Replication Identify Three Additional *EST* Genes. Genetics 144: 1399–1412. https://doi.org/10.1093/genetics/144.4.1399

Lewis K. A., D. A. Pfaff, J. N. Earley, S. E. Altschuler, and D. S. Wuttke, 2014 The tenacious recognition of yeast telomere sequence by Cdc13 is fully exerted by a single OB-fold domain. Nucleic Acids Res 42: 475–484. https://doi.org/10.1093/nar/gkt843

Li J.-R., T.-Y. Yu, I.-C. Chien, C.-Y. Lu, J.-J. Lin, *et al.*, 2014 Pif1 regulates telomere length by preferentially removing telomerase from long telomere ends. Nucleic Acids Res 42: 8527–8536. https://doi.org/10.1093/nar/gku541

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

Longtine M. S., A. Mckenzie III, D. J. Demarini, N. G. Shah, A. Wach, *et al.*, 1998 Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast 14: 953–961. https://doi.org/10.1002/(SICI)1097-0061(199807)14:10<953::AID-YEA293>3.0.CO;2-U

Louis E. J., and J. E. Haber, 1990a The subtelomeric Y' repeat family in Saccharomyces cerevisiae: an experimental system for repeated sequence evolution. Genetics 124: 533–545. https://doi.org/10.1093/genetics/124.3.533

Louis E. J., and J. E. Haber, 1990b Mitotic recombination among subtelomeric Y' repeats in Saccharomyces cerevisiae. Genetics 124: 547–559. https://doi.org/10.1093/genetics/124.3.547

Louis E. J., and J. E. Haber, 1992 The structure and evolution of subtelomeric Y' repeats in Saccharomyces cerevisiae. Genetics 131: 559–574. https://doi.org/10.1093/genetics/131.3.559

Lundblad V., and J. W. Szostak, 1989 A mutant with a defect in telomere elongation leads to senescence in yeast. Cell 57: 633–643. https://doi.org/10.1016/0092-8674(89)90132-3

Luria S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virua resistance. Genetics 28: 491–511. https://doi.org/10.1093/genetics/28.6.491

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

Makovets S., and E. H. Blackburn, 2009 DNA damage signalling prevents deleterious telomere addition at DNA breaks. Nat Cell Biol 11: 1383–1386. https://doi.org/10.1038/ncb1985

Mangahas J. L., M. K. Alexander, L. L. Sandell, and V. A. Zakian, 2001 Repair of Chromosome Ends after Telomere Loss in *Saccharomyces*. Mol Biol Cell 12: 4078–4089. https://doi.org/10.1091/mbc.12.12.4078

McClintock B. 1938. The Production of Homozygous Deficient Tissues with Mutant Characteristics by Means of the Aberrant Mitotic Behavior of Ring-Shaped Chromosomes. Genetics. 23(4):315-76. doi: 10.1093/genetics/23.4.315

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

Mersaoui S. Y., and R. J. Wellinger, 2019 Fine tuning the level of the Cdc13 telomere-capping protein for maximal chromosome stability performance. Curr Genet 65: 109–118.

Mimitou E. P., and L. S. Symington, 2008 Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature 455: 770–774. https://doi.org/10.1038/nature07312

Moyzis R. K., J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, *et al.*, 1988 A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proceedings of the National Academy of Sciences 85: 6622–6626. https://doi.org/10.1073/pnas.85.18.6622

Mozdy A. D., and T. R. Cech, 2006 Low abundance of telomerase in yeast: Implications for telomerase haploinsufficiency. RNA 12: 1721–1737. https://doi.org/10.1261/rna.134706

Muller HJ, 1938 The remaking of chromosomes. The Collect. Nat., 13.

Muellner J., and K. H. Schmidt, 2020 Yeast Genome Maintenance by the Multifunctional PIF1 DNA Helicase Family. Genes (Basel) 11: 224. https://doi.org/10.3390/genes11020224

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

Naumov G. I., E. S. Naumova, R. A. Lantto, E. J. Louis, and M. Korhola, 1992 Genetic homology between Saccharomyces cerevisiae and its sibling species S. paradoxus and S. bayanus: Electrophoretic karyotypes. Yeast 8: 599–612. https://doi.org/10.1002/yea.320080804

Nevado J., S. García-Miñaúr, M. Palomares-Bralo, E. Vallespín, E. Guillén-Navarro, *et al.*, 2022 Variability in Phelan-McDermid Syndrome in a Cohort of 210 Individuals. Front Genet 13. https://doi.org/10.3389/fgene.2022.652454

Ngo K., E. A. Epum, and K. L. Friedman, 2020 Emerging non-canonical roles for the Rad51–Rad52 interaction in response to double-strand breaks in yeast. Curr Genet 66: 917–926.

Ngo K., T. H. Gittens, D. I. Gonzalez, E. A. Hatmaker, S. Plotkin, *et al.*, 2023 A comprehensive map of hotspots of de novo telomere addition in Saccharomyces cerevisiae. Genetics iyad076. https://doi.org/10.1093/genetics/iyad076

Nugent C. I., T. R. Hughes , N. F. Lue, V. Lundblad, 1996 Cdc13p: a single-strand telomeric DNA-binding protein with a dual role in yeast telomere maintenance. Science.;274(5285):249-52. doi: 10.1126/science.274.5285.249.

Obodo U. C., E. A. Epum, M. H. Platts, J. Seloff, N. A. Dahlson, *et al.*, 2016 Endogenous Hot Spots of De Novo Telomere Addition in the Yeast Genome Contain Proximal Enhancers That Bind Cdc13. Mol Cell Biol 36: 1750–1763. https://doi.org/10.1128/MCB.00095-16

Osterhage J. L., and K. L. Friedman, 2009 Chromosome End Maintenance by Telomerase. Journal of Biological Chemistry 284: 16061–16065. https://doi.org/10.1074/jbc.R900011200

Ouenzar F., M. Lalonde, H. Laprade, G. Morin, F. Gallardo, *et al.*, 2017 Cell cycle–dependent spatial segregation of telomerase from sites of DNA damage. Journal of Cell Biology 216: 2355–2371. https://doi.org/10.1083/jcb.201610071

Oza P., S. L. Jaspersen, A. Miele, J. Dekker, and C. L. Peterson, 2009 Mechanisms that regulate localization of a DNA double-strand break to the nuclear periphery. Genes Dev 23: 912–927. https://doi.org/10.1101/gad.1782209

Paeschke K., J. A. Capra, and V. A. Zakian, 2011 DNA Replication through G-Quadruplex Motifs Is Promoted by the Saccharomyces cerevisiae Pif1 DNA Helicase. Cell 145: 678–691. https://doi.org/10.1016/j.cell.2011.04.015

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

Pâques F., W.-Y. Leung, and J. E. Haber, 1998 Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. Mol Cell Biol 18: 2045–2054. https://doi.org/10.1128/MCB.18.4.2045

Pennaneach V., C. D. Putnam, and R. D. Kolodner, 2006 Chromosome healing by *de novo* telomere addition in *Saccharomyces cerevisiae*. Mol Microbiol 59: 1357–1368. https://doi.org/10.1111/j.1365-2958.2006.05026.x

Pennock E., K. Buckley, and V. Lundblad, 2001 Cdc13 Delivers Separate Complexes to the Telomere for End Protection and Replication. Cell 104: 387–396. https://doi.org/10.1016/S0092-8674(01)00226-4

Peterson S. E., A. E. Stellwagen, S. J. Diede, M. S. Singer, Z. W. Haimberger, *et al.*, 2001 The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku. Nat Genet 27: 64–67. https://doi.org/10.1038/83778

Pfeiffer V., and J. Lingner, 2013 Replication of Telomeres and the Regulation of Telomerase. Cold Spring Harb Perspect Biol 5: a010405–a010405. https://doi.org/10.1101/cshperspect.a010405

Phillips J. A., A. Chan, K. Paeschke, and V. A. Zakian, 2015 The Pif1 Helicase, a Negative Regulator of Telomerase, Acts Preferentially at Long Telomeres. PLoS Genet 11: e1005186. https://doi.org/10.1371/journal.pgen.1005186

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinlan A. R., 2014 BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics 47. https://doi.org/10.1002/0471250953.bi1112s47

Rhee H. S., and B. F. Pugh, 2011 Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147: 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013

Rice C., and E. Skordalakes, 2016 Structure and function of the telomeric CST complex. Comput Struct Biotechnol J 14: 161–167.

Rossiello F., D. Jurk, J. F. Passos, and F. d'Adda di Fagagna, 2022 Telomere dysfunction in ageing and age-related diseases. Nat Cell Biol 24: 135–147. https://doi.org/10.1038/s41556-022-00842-x

Schulz V. P., and V. A. Zakian, 1994 The saccharomyces PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation. Cell 76: 145–155. https://doi.org/10.1016/0092-8674(94)90179-1

Shampay J., J. W. Szostak, and E. H. Blackburn, 1984 DNA sequences of telomeres maintained in yeast. Nature 310: 154–157. https://doi.org/10.1038/310154a0

Sinden R. R., and R. D. Wells, 1992 DNA structure, mutations, and human genetic disease. Curr Opin Biotechnol 3: 612–622. https://doi.org/10.1016/0958-1669(92)90005-4

Singer M. S., and D. E. Gottschling, 1994 *TLC1*: Template RNA Component of *Saccharomyces cerevisiae* Telomerase. Science (1979) 266: 404–409. https://doi.org/10.1126/science.7545955

Snow B. E., N. Erdmann, J. Cruickshank, H. Goldman, R. M. Gill, *et al.*, 2003 Functional Conservation of the Telomerase Protein Est1p in Humans. Current Biology 13: 698–704. https://doi.org/10.1016/S0960-9822(03)00210-0

Sprung C.N., G. E.Reynolds, M. Jasin, J.P. Murnane, 1999 Chromosome healing in mouse embryonic stem cells. Proc Natl Acad Sci U S A.;96(12):6781-6786. doi:10.1073/pnas.96.12.6781

Srinivas N., S. Rachakonda, and R. Kumar, 2020 Telomeres and Telomere Length: A General Overview. Cancers (Basel) 12: 558. https://doi.org/10.3390/cancers12030558

Stellwagen A. E., Z. W. Haimberger, J. R. Veatch, and D. E. Gottschling, 2003 Ku interacts with telomerase RNA to promote telomere addition at native and broken chromosome ends. Genes Dev 17: 2384–2395. https://doi.org/10.1101/gad.1125903

Stinus S., K. Paeschke, and M. Chang, 2018 Telomerase regulation by the Pif1 helicase: a length-dependent effect? Curr Genet 64: 509–513. https://doi.org/10.1007/s00294-017-0768-6

Strecker J., S. Stinus, M. Pliego Caballero, R. K. Szilard, M. Chang, *et al.*, A sharp Pif1-dependent threshold separates DNA double-strand breaks from critically short telomeres. https://doi.org/10.7554/eLife.23783.001

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

Wang S. S., and V. A. Zakian, 1990 Sequencing of Saccharomyces telomeres cloned using T4 DNA polymerase reveals two domains. Mol Cell Biol 10: 4415–4419. https://doi.org/10.1128/MCB.10.8.4415

Weinert T. A., and L. H. Hartwell, 1988 The *RAD9* Gene Controls the Cell Cycle Response to DNA Damage in *Saccharomyces cerevisiae*. Science (1979) 241: 317–322. https://doi.org/10.1126/science.3291120

Wellinger R. J., and V. A. Zakian, 2012 Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: Beginning to end. Genetics 191: 1073–1105. https://doi.org/10.1534/genetics.111.137851

Westmoreland J. W., M. J. Mihalevic, K. A. Bernstein, and M. A. Resnick, 2018 The global role for Cdc13 and Yku70 in preventing telomere resection across the genome. DNA Repair (Amst) 62: 8–17. https://doi.org/10.1016/j.dnarep.2017.11.010

Yu G.-L., and E. H. Blackburn, 1991 Developmentally programmed healing of chromosomes by telomerase in tetrahymena. Cell 67: 823–832. https://doi.org/10.1016/0092-8674(91)90077-C

Yue J.-X., J. Li, L. Aigrain, J. Hallin, K. Persson, *et al.*, 2017 Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet 49: 913–924. https://doi.org/10.1038/ng.3847

Zakian V. A., 2012 Telomeres: The beginnings and ends of eukaryotic chromosomes. Exp Cell Res 318: 1456–1460. https://doi.org/10.1016/j.yexcr.2012.02.015

Zhang W., and D. Durocher, 2010 De novo telomere formation is suppressed by the Mec1-dependent inhibition of Cdc13 accumulation at DNA breaks. Genes Dev 24: 502–515. https://doi.org/10.1101/gad.1869110

Zhu Z., W.-H. Chung, E. Y. Shim, S. E. Lee, and G. Ira, 2008 Sgs1 Helicase and Two Nucleases Dna2 and Exo1 Resect DNA Double-Strand Break Ends. Cell 134: 981–994. https://doi.org/10.1016/j.cell.2008.08.037

Zhu X., and C. M. Gustafsson, 2009 Distinct Differences in Chromatin Structure at Subtelomeric X and Y' Elements in Budding Yeast. PLoS One 4: e6363. https://doi.org/10.1371/journal.pone.0006363