IDENTIFICATION AND CHARACTERIZATION OF TELOMERASE RNA GENES IN YEAST GENOMES

by

LEILEI GUO

(Under the Direction of Michael J. McEachern)

ABSTRACT

The majority of eukaryotes use the enzyme telomerase to ensure that their chromosome ends are completely replicated. The core components of telomerase are the catalytic reverse transcriptase (TERT) and telomerase RNA (TR). The first part of this thesis describes a novel approach to predict the core secondary structure of telomerase RNAs, using TRFolder, a program we developed. With TRFolder, we confirmed and improved the previously studied TR structures in *Kluyveromyces* and *Saccharomyces*, and made novel structural predictions of core elements of the TRs from *Schizosaccharomyces pombe*, *Candida albicans*, and several other yeast species. In the second study, we showed that the presence of two types of telomeric repeats (A-repeat and C-repeat) in *Candida tropicalis* strain B-4414 was associated with the presence of a highly divergent allele of telomerase RNA gene *TER1*. To our knowledge, this is the first identification of a natural organism that has dimorphic telomeric repeats as a result of having two different telomerase RNA gene alleles. Our study also showed that the high degree of allelic sequence divergence was not

confined to *TER1* but included DNA sequences from across the genome. This striking high sequence divergence may suggest that mating or recombination can occur in this "asexual" diploid yeast. Together, the studies in this dissertation can help us to get a better understanding of the rapid evolution and functional maintenance of the telomerase RNA template and structure.

INDEX WORDS: telomere, telomerase, telomerase RNA gene, conserved secondary core structure, TRFolder, *Candida tropicalis*, high allelic divergence

IDENTIFICATION AND CHARACTERIZATION OF TELOMERASE RNA GENES IN YEAST GENOMES

by

LEILEI GUO

B.S., Sun Yat-sen University, P. R. China, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Paritial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2011

IDENTIFICATION AND CHARACTERIZATION OF TELOMERASE RNA GENES IN YEAST GENOMES

by

LEILEI GUO

Major professor:	Michael J. McEachern
Committee:	Edward Kipreos
	Sidney Kushner
	Russell Malmberg
	Michael Terns

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

DEDICATION

To my beloved parents, Baoquan Guo and Xiaoqing Zhang

&

To my precious daughter Emma and son Evan

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest appreciation to my major professor Dr. Michael J. McEachern. He has been constantly helping me during the whole process of our research, from big pictures to small details. Without his invaluable advice, timeless support and kind encouragements on my research and other aspects of life, the completion of this dissertation would not have been possible. His patience, confidence, research mindsets as well as his attitude toward professionalism have influenced me a lot during my Ph.D. training, and will keep guiding me in my future life. It is a truly privilege and pleasure to work with him.

I would like to extend my gratitude to Dr. Edward Kipreos, Dr. Sidney Kushner, Dr. Russell Malmberg and Dr. Michael Terns for enhancing this dissertation, offering valuable suggestions, and serving on my committee. Special thanks to Dr. Liming Cai for being a great collaborator and supervisor for my first project.

I am deeply grateful for being given the chance to pursue my Ph.D. study in Department of Genetics, the University of Georgia. My thanks are due to all the professors, staff and friends who are always ready to help and easy to work with. UGA provides wonderful atmosphere for academic accomplishments, and Athens has become my hometown in the States.

I would like to thank all the past and present members of the McEachern lab, who have been great friends and colleagues and a joy to work with. Especially

Dr-to-be Evelina Basenko, thanks so much for giving me comfort and encouragement, and walking together with me through the most difficult time in my life.

Moreover, many thanks to my dearest friends who were with me through this entire journey, Dr. Bing Xu, Dr. Zhiqian Hao, Dr-to-be Shan Huang, Dr. Xuehai Liang and Dr. Jing Xu. With them I shared more than five years of my life, the best and worst of me, the happiest and saddest moments, and the unforgettable memories in my life.

Last but not least, I would like to express my love to my family. My parents, Baoquan Guo and Xiaoqing Zhang, for your unconditional caring and love. I feel blessed to be your daughter. And my beautiful angels Emma and Evan, the most precious gifts given to me. I will make every effort to become a better mom and friend. There is nothing in the world I would trade for you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

TABLE OF FIGURES

Page

CHAPTER 1

LITERATURE REVIEW AND INTRODUCTION

Discovery of Telomere and Telomerase

Telomeres are regions of repetitive DNA sequences at the ends of eukaryotic chromosomes (Cech 2004; Donate and Blasco 2011). Telomeres serve as a “cap” on chromosome ends, preventing them from eliciting DNA damage responses like end-to-end fusion or exonucleolytic degradation. In the 1930s, Barbara McClintock and Herman Muller’s independent studies in maize and fruit flies showed that a broken chromosome end usually becomes unstable, while natural chromosome ends were relatively stable (McClintock 1931; Muller 1938). In 1938, Muller named the “natural chromosome end” the telomere, which is derived from the Greek nouns telos (*τέλος*) “end” and meros (*μέρος*) "part" (Muller 1938).

After the discovery of semi-conservative DNA replication, the existence of the “end-replication problem” was first predicted by Watson (Watson 1972) and Olovnikov (Olovnikov 1973). DNA replication by DNA polymerases starts within a RNA primer and always moves in the 5’ to 3’ direction. The consequence of the gradual shortening of DNA ends caused by the “end-replication problem” was first described as the inability of DNA polymerases to complete the replication of the

lagging strand. Although DNA polymerase makes a complementary DNA continuously on the leading strand, the replication of the lagging strand requires a new RNA primer for the synthesis of every Okazaki fragment. These RNA primers will then be deleted by DNA polymerases and gaps between the Okazaki fragments will be sealed by DNA ligase. Upon the completion of the lagging-strand synthesis during DNA replication, the final RNA primer will be removed, and result in a short overhang at the end. During each cell division, a portion of the DNA end will be lost and telomere will shorten gradually over time.

However, while the above is true only for DNA molecules with blunt ends, it is not necessarily true for all ends. Studies on telomere structures in multiple organisms (Klobutcher *et al.* 1981; Henderson and Blackburn 1989; Wellinger *et al.* 1993) showed that, instead of blunt ends, native telomeres have a single-stranded 3' overhang of 12 nucleotides or even longer. The problem of sequence loss at the telomeres during replication is caused instead by the inability of DNA polymerases to completely replicate the 3' overhang, which is produced by the leading strand synthesis (Figure 1.1). As long as the RNA primer (typically 8 to 14 nt) hybridizes to the 3' overhang only, no DNA will be lost on the lagging strand upon RNA primer removal after DNA replication. In contrast, the newly synthesized DNA on the leading strand will lose the 3' overhang since it lacks the complementary template from the parent strand (Lingner *et al.* 1995).

The nature of telomeres as tandem DNA repeats was first revealed in the ciliated protozoan *Tetrahymena thermophila* by Elizabeth Blackburn in 1970s (Blackburn and Gall 1978), and later in a distantly related eukaryote, the budding yeast *Saccharomyces cerevisiae*, by the collaboration between Blackburn and Jack Szostak (Szostak and Blackburn 1982). In 1984, the enzyme "telomerase", which is used in the majority of eukaryotes for telomere length maintenance, was discovered in *Tetrahymena thermophila* by Blackburn and her graduate student Carol Greider (Greider and Blackburn 1985), and later in a wide variety of organisms, including vertebrates, yeast and plants (Greider and Blackburn 1989; Morin 1989; Prowse *et al.* 1993; Mantell and Greider 1994; Cohn and Blackburn 1995; Fitzgerald *et al.* 1996). Elizabeth Blackburn, Carol Greider and Jack Szostak were awarded the 2009 Nobel Prize in Physiology or Medicine for the discovery of telomere and telomerase.

Senescence, Cancer and Telomerase

Telomere maintenance in humans is important for the studies of both aging and cancer (Shay and Wright 2010). In human cells, telomerase is expressed during early development and remains fully active in specific germline cells, but it is undetectable in most normal somatic cells except for proliferative cells of renewal tissues. The lack of telomerase activity leads to progressive telomere shortening (Flores and Blasco 2010). Progressive telomere shortening is thought to eventually truncate telomeres to a critically short length. The chromosome ends will then become recognized as double strand breaks. This will consequently lead to the activation of a DNA damage

signaling pathway called “replicative senescence” or “cell aging”, which acts as a tumor suppressor pathway (Kuilman *et al.* 2010). Cells that lose critical cell cycle checkpoint functions can bypass the initial growth arrest (replicative senescence) and continue to divide. Cells that bypass senescence eventually enter a second growth arrest (crisis), where many chromosomes with shortened telomeres start to form bridge-breakage fusion, and cells almost universally go to apoptosis. In human cells, senescence and crisis are two mechanisms that restrict cell growth and provide at least initial anticancer protection (Wright and Shay 1992).

Most human cells die in crisis phase, while cells can rarely acquire the ability to maintain telomeric sequences and escape from crisis, normally by reactivation of telomerase activity (Shay and Roninson 2004; Shay and Wright 2005). This is generally believed to be a critical step in cell immortalization and cancer progression (Hanahan and Weinberg 2000). This has led to much interest in the possibility that inhibition of telomere maintenance pathways might be an efficient target for anti-cancer drugs.

Multiple telomerase targeting approaches have currently been tested in preclinical or clinical trials. These therapies can lead to several possible outcomes. First, it is predicted that there will be a lag phase until the telomeres of the cancer cells achieve critical shortened length after telomerase inhibition. Based on such predictions, it has been suggested that the telomerase inhibition therapy should be used together with

other conventional treatments for more effective results or afterwards to prevent cancer reoccurrence (Chen *et al.* 2003). However, it has been observed that the telomeres in tumor cells are significantly heterogeneous in length (Holt and Shay 1999). Telomerase works preferentially on short telomeres (Ouellette *et al.* 2000; Hemann *et al.* 2001), and a few dysfunctional telomeres can trigger growth arrest or apoptosis (Kim *et al.* 2001; d'Adda di Fagagna *et al.* 2003; Zou *et al.* 2004). In contrast, telomeres in human gametes and stem cells where telomerase is active (White *et al.* 2001; Forsyth *et al.* 2002; Granger *et al.* 2002) are much longer in length than tumor cells, so the cell death caused by telomerase inhibition should occur much earlier in tumor cells than normal stem cells. As long as the telomerase inhibitors are safely used in the "therapeutic window", these should be specifically selective and universal antitumor drugs and leave only minimal if any effects on stem cells with longer telomeres (Zvereva *et al.* 2010).

Clinical trials with telomerase immunotherapy approaches have not found stem cell effects to be a serious problem (Minev *et al.* 2000; Nair *et al.* 2000; Vonderheide 2002). However, it is worth noting that only embryonic stem cells have fully active telomerase, while normal tissue stem cells also show progressive telomere shortening with increased age, due to an insufficient level of telomerase activity to fully maintain telomere length (Shay and Wright 2010). Better understanding in basic science and careful clinical trials in animal models are necessary for telomerase inhibition therapy.

Overview of Telomeres

Telomeres contain regions of double-stranded (ds) DNA and single-stranded (ss) DNA. The majority of telomeric DNA is double-stranded, with a 3' single-stranded overhang of the G-rich strand at the extreme tip of the DNA (Zhao *et al.* 2009). The telomeric overhangs result from both the DNA end-replication problem and further processing after replication (Wellinger *et al.* 1993; Lingner *et al.* 1995; Dionne and Wellinger 1998). The G-overhang is crucial to the formation of a structure called a t-loop (telomeric-loop), which was proposed to mask the telomere terminus from double-stranded break (DSB) repair machinery and to limit access to telomerase (Griffith *et al.* 1999; Stewart *et al.* 2011). This t-loop structure at the telomeric end has been observed in a wide variety of species, including human, plants, mice and multiple ciliates (Griffith *et al.* 1999; Murti and Prescott 1999; Munoz-Jordan *et al.* 2001; Cesare *et al.* 2003). However, it is not likely to be a universal feature of telomeres as telomeres in some ciliates are too short to be able to form them (Klobutcher *et al.* 1981; Murti and Prescott 1999).

The length of telomeres varies among different organisms. In the hypotrichous ciliate *Euplotes crassus*, the length of telomeres is only 28 bp of double-stranded telomeric sequence and 42 nt on the G-rich strand (Klobutcher *et al.* 1981). The telomeres in yeast are typically several hundreds base pairs in length (Shampay *et al.* 1984). In humans, telomeres are heterogeneous in length and generally in the 5-15 kilobases (kb)

length range (de Lange *et al.* 1990; Lansdorp *et al.* 1996). Telomeres in strains of lab mouse (*Mus musculus*) are even longer, usually 40-50 kb (Starling *et al.* 1990).

While telomeres are most commonly composed of tandem repeats, the sequences of these repeats sometimes vary appreciably among organisms. Human and other vertebrates have telomeres made up of repeats of the sequence TTAGGG (Moyzis *et al.* 1988; Meyne *et al.* 1989). *T. thermophila* telomeres are composed of TTGGGG repeats (Blackburn and Gall 1978), and the plant *Arabidopsis thaliana* has homologous telomeric repeats composed of the sequence TTTAGGG (Richards and Ausubel 1988). However, the telomeric repeat units can be much longer in yeast species, for example, 23 bp in *Candida albicans* (McEachern and Hicks 1993) and 25 bp in *Kluyveromyces lactis* (McEachern and Blackburn 1994). A list of telomeric repeat sequences in multiple organisms is shown in Table 1.1.

Although the telomeric repeats of *S. cerevisiae*, *K. lactis* and *C. albicans* are quite different in both sequence and length, they are all capable of binding the double-stranded DNA-binding protein Rap1 (Berman *et al.* 1986; Krauskopf and Blackburn 1996; Cohn *et al.* 1998). Rap1 is a component of shelterin, a six-protein complex bound to the telomeres in mammalian cells. The Rap1 of vertebrates is divergent from that of yeast in apparently being unable to bind DNA directly (Li *et al.* 2000). Rap1 is involved in telomere length regulation and also affects telomere length heterogeneity (Conrad *et al.* 1990; Lustig *et al.* 1990; Krauskopf and Blackburn 1996;

Marcand *et al.* 1997; Li and de Lange 2003). Analysis of telomeres cloned from a wide variety of yeast species showed that the highly divergent yeast telomeric repeats share a reiterated conserved core sequence which preserves a binding site for Rap1 (McEachern and Blackburn 1994; Cohn *et al.* 1998) (Figure 1.2).

Overview of Telomerase

The vast majority of all eukaryotes use the enzyme telomerase to ensure that the chromosome ends are completely replicated (Legassie and Jarstfer 2006). The telomerase enzyme is a ribonucleoprotein (RNP) that consists of multiple components including two essential core components: the telomerase reverse transcriptase catalytic protein subunit (commonly called TERT); and an RNA component, the telomerase RNA (commonly called TR). The TERT protein has regions homologous to the catalytic motifs of reverse transcriptase (Lingner *et al.* 1997). The RNA component contains a short region that is complementary to the telomeric repeat sequence and serves as a template for telomerase to reverse transcribe telomere DNA repeats (Greider and Blackburn 1989).

While TERT and TR are the minimum requirements for telomerase activity *in vitro*, additional proteins may be required for the activity of the telomerase holoenzyme complex *in vivo* (Lingner and Cech 1996; Greene and Shippen 1998; Schnapp *et al.* 1998). In budding yeast *S. cerevisiae*, after the discovery of the RNA component Tlc1 (Singer and Gottschling 1994), a yeast genetic screen for the ever shorter telomere

(EST) mutants identified several genes, *EST1-EST4*, that are essential for telomerase activity *in vivo* (Lendvay *et al.* 1996). The Est2 protein emerged as the yeast TERT (Lingner *et al.* 1997). Co-immunoprecipitation experiments suggested that Est1p binds directly to Tlc1 (Seto *et al.* 2002) and recruits telomerase to telomeres through the interaction between Cdc13p (also known as Est4p) (Steiner *et al.* 1996; Zhou *et al.* 2000). Cdc13p is part of an RPA-like heterotrimeric protein complex that binds telomeric 3' overhangs (Lin and Zakian 1996; Nugent *et al.* 1996; Hughes *et al.* 2000). The orthlogs of Est1p has been identified in some other organisms including human and fission yeast (*Schizosaccharomyces pombe*), suggesting that it is a common telomerase subunit (Beernink *et al.* 2003; Reichenbach *et al.* 2003; Snow *et al.* 2003). Est3p associates with the holoenzyme complex through the amino-terminus of Est2p and this association is Est1p-dependent (Hughes *et al.* 2000; Livengood *et al.* 2002; Friedman *et al.* 2003).

Telomerase activity is low or undetectable in human somatic cells, but is upregulated in the majority (80-90%) of human cancer cells (Shay and Bacchetti 1997). The active human telomerase complex is composed of two protein components, hTERT and dyskerin (DKC1), together with the RNA component hTR (Cohen *et al.* 2007). Dyskerin binds and stabilizes hTR. Mutations in dyskerin resulted in low levels of telomerase activity and poor telomere length maintenance (Mitchell *et al.* 1999) and are the cause of X-linked Dyskeratosis Congenita (DC) (Connor *et al.* 1986; Heiss *et al.* 1998; Hassock *et al.* 1999).

TERT was firstly identified in the ciliate *E. aediculatus* by biochemical purification of telomerase (Lingner and Cech 1996) and in *S. cerevisiae* by the genetic screen mentioned above that identified EST proteins (Lingner *et al.* 1997). Later, TERT orthologs were identified in human and mouse, ciliates, fission yeast and others (Counter *et al.* 1997; Harrington *et al.* 1997; Kilian *et al.* 1997; Meyerson *et al.* 1997; Nakamura *et al.* 1997; Bryan *et al.* 1998; Collins and Gandhi 1998). The alignment and analysis of TERT sequences revealed at least four conserved domains (Legassie and Jarstfer 2006). The central reverse transcriptase (RT) domain contains seven RT motifs that are all universally conserved. Mutations in these domains were shown to result in disruption of telomerase activity (Counter *et al.* 1997; Harrington *et al.* 1997; Lingner *et al.* 1997; Weinrich *et al.* 1997). The carboxy-terminal extension (CTE) is essential for telomerase activity in human and *Tetrahymena* (Bachand and Autexier 2001; Lai *et al.* 2001), but is dispensable in *S. cerevisiae* (Friedman and Cech 1999). This domain has been implicated to function in telomerase processivity (Huard *et al.* 2003). The amino-terminal half of TERT contains several conserved motifs that form TR-binding domains (Legassie and Jarstfer 2006). RID2 is a TR-binding domain that interacts with the CR4-CR5 domain of TR and has been identified in hTERT and ciliate TERT. RID1 binds the pseudoknot-template region of TR.

TRs have been identified in ciliates (Greider and Blackburn 1989; Shippen-Lentz and Blackburn 1990; Romero and Blackburn 1991), vertebrates (Blasco *et al.* 1995; Feng

et al. 1995; Chen *et al.* 2000), and various yeast species (Singer and Gottschling 1994; Feng *et al.* 1995; McEachern and Blackburn 1995; Hsu *et al.* 2007; Leonardi *et al.* 2008; Webb and Zakian 2008; Gunisova *et al.* 2009; Kachouri-Lafond *et al.* 2009) by different methods. Telomerase RNAs in ciliates were the first to be identified, by hybridizing macronuclear RNA with a synthetic DNA oligonucleotide, which was complementary to the telomeric repeat sequence (Shippen-Lentz and Blackburn 1990). The *Schizosaccharomyces pombe* TR gene was independently identified recently by two groups using almost identical methods involving the copurification of the RNA component with the TERT protein subunit of the enzyme (Leonardi *et al.* 2008; Webb and Zakian 2008).

These methods require sufficient enzyme abundance in the cell and are labor consuming. In *Kluyveromyces lactis* and *Candida albicans*, TR genes were identified by direct probing of the whole genome with the telomeric repeat sequence, taking the advantage of the large and homogenous telomeric repeats which are 25 bp and 23 bp, respectively (McEachern and Blackburn 1995; Hsu *et al.* 2007). However, this strategy is limited to organisms with long telomeric repeats. The human TR was identified using the predicted TR template as a probe to hybridize the cDNA library (Feng *et al.* 1995). Twelve TR candidates had to be examined individually in order to find the real one. *TLC1*, the gene encoding TR in *Saccharomyces cerevisiae*, was identified by screening for highly expressed genes that suppress telomeric silencing (Singer and Gottschling 1994). The strategy found *TLC1* by serendipity and could not

be a general method for TR identification. The most powerful method to date is to find TRs by phylogenetic comparison with known TRs (Romero and Blackburn 1991; Chen *et al.* 2000; Tzfati *et al.* 2003; Lin *et al.* 2004). However, TR sequences can be aligned with confidence only among closely related organisms, since the primary sequences have diverged quickly.

Universal Core Structure of Telomerase RNA

Telomerase RNAs are surprisingly divergent among eukaryotes. The size of telomerase RNAs varies from ~150 nucleotides in ciliates (Ye and Romero 2002) to ~500 nucleotides in vertebrates (Xie *et al.* 2008) and lengths of 779-1817 nucleotides in yeast (Gunisova *et al.* 2009). The largest telomerase RNA identified to date is 2.2 kb, in the malarial parasite *Plasmodium falciparum* (Chakrabarti *et al.* 2007). Additionally, the primary sequences of telomerase RNAs also show great diversity, especially when comparing with the generally more conserved TERT gene length and sequence.

Although telomerase RNA sequences can only be aligned among closely related organisms with confidence, secondary structure models were derived for the different groups using phylogenetic comparisons (Romero and Blackburn 1991; Lingner *et al.* 1994; Chen *et al.* 2000; Dandjinou *et al.* 2004; Zappulla and Cech 2004). In most species, the proposed models of telomerase RNAs share a common core secondary structure, which contains four functional structural elements that are conserved among

eukaryotic telomerase RNAs: the template region; the 5' boundary element upstream of the template, a downstream pseudoknot structure which binds TERT; and a long-range base-pairing element that encloses the template and pseudoknot. These common core features have persisted across the wide phylogenetic diversity represented by ciliates, vertebrates and yeasts (Figure 2.1) (Lin *et al.* 2004).

The telomerase template is complimentary to the telomeric tandem repeat on the G-rich strand of telomeres. In all telomerase RNAs, the template is defined by terminal repeats of ~3-10 nucleotides (nt) that are present at both ends (Yu *et al.* 1990; Wang *et al.* 2009). These repeats are thought to be involved in the alignment with the last telomeric repeat before a new round of telomeric repeat synthesis. The sequence synthesized based on the 5' terminal repeat of the template can align with 3' terminal repeat after translocation, providing 3-10 nt of complementarity between the telomere and the template. This appears to account for the uniformity in telomeric repeat sequences within most species.

The template regions in vertebrate and ciliate TRs mostly have perfect terminal repeats, usually 3-5 nt. Mutational analysis suggested that the terminal repeats in the *K. lactis* TR template are 9-nt long and contain a mismatch (Wang *et al.* 2009). A variety of yeast species were examined for the possible presence of an extended base-pairing potential between the 3′side of the RNA template and the telomeric DNA, as exists in *K. lactis*. As shown in Figure 1.2, most species examined were found to

have template terminal repeats of 7 to 10 nt in size. In three species, *Saccharomyces kluyveri*, *Ashbya gossypii*, and *Debaryomyces hansenii*, the repeats were perfect matches of 8 to 10 nt. Eight other species, including five of six *Kluyveromyces* species as well as *Candida albicans*, *Candida tropicalis*, and *Candida glabrata*, had repeats that differed from one another by a single nucleotide near their middles.

Kluyveromyces nonfermentans and *Kluyveromyces aestuarii* had repeats that differed from one another by a single extra nucleotide in the 5' repeat. The templates of *Candida guilliermondii* and *Pichia stipitis* could be described as being bordered either by perfect 7- to 8-nt repeats or by imperfect 9- to 10-nt repeats. *S. cerevisiae*, as well as some other closely related *Saccharomyces* species, has a template that can be drawn having a few slightly different pairs of 7- to 9-nt terminal repeats. Among the yeasts examined, only *Saccharomyces castellii*, with 4-nt terminal repeats, and *S. pombe* could not be reasonably drawn to have extended repeats. It was suggested that the long, often imperfect telomere base-pairing domain at the 3′ end of the telomerase RNA template is a feature common to many yeast telomerase RNAs that may be related to those species having long perfect telomeric repeats (Wang *et al.* 2009). Notably, *S. cerevisiae*, *S. castellii*, and *S. pombe* are species that have telomeric repeats that are highly heterogeneous in size and sequence.

The template boundary element of the RNA functions as a barrier for reverse transcription beyond the end of the template. Mutations in this element can lead to the

synthesis of a telomeric repeat that is lengthened by sequence copied from beyond the normal template (Autexier and Greider 1995; Prescott and Blackburn 1997; Miller *et al.* 2000; Tzfati *et al.* 2000). The structure of this boundary element differs in different phylogenetic groups. A stem-loop structure upstream of the template region defines the 5' boundary in both *Kluyveromyces* and *Saccharomyces* (Tzfati *et al.* 2000; Seto *et al.* 2003). In ciliates, the boundary element is defined by a conserved sequence motif upstream of the template region, which is bound by the TERT protein (Autexier and Greider 1995; Lai *et al.* 2002). In human telomerase RNA, the structure and position of the long base-pairing region that encloses the pseudoknot and template is important for boundary definition. However, mouse telomerase RNA lacks that helix, and the boundary element is established by its 5' end located only two nucleotides upstream of the template boundary (Chen and Greider 2003). These divergent boundary elements suggest that divergent mechanisms exist for template boundary definition.

The conserved pseudoknot element within the telomerase RNA core structure has been identified in ciliates, vertebrates and yeasts (Autexier and Triki 1999; Gilley and Blackburn 1999; Chen *et al.* 2000; Ly *et al.* 2003; Tzfati *et al.* 2003). In *T. tetrahymena*, disruption of this pseudoknot base-pairing region in vivo affects the stable association with TERT, while compensatory mutations restore telomerase activity (Gilley and Blackburn 1999). Pseudoknot structures have also been proposed for telomerase RNAs from *Kluyveromyces* (Tzfati *et al.* 2003) and in multiple

alternative forms in *S. cerevisiae*, with different base pairing possibilities in the putative pseudoknot region (Chappell and Lundblad 2004; Dandjinou *et al.* 2004; Lin *et al.* 2004; Zappulla and Cech 2004). Mutational analysis based on the different models of the telomerase RNA core region suggest that the actual structure fit best the model of Lin *et al*. (Lin *et al*., 2004; Qiao and Cech 2008).

A triple-helix motif within the pseudoknot has been identified in human telomerase RNAs (Theimer *et al.* 2005), *Kluyveromyces* (Shefer *et al.* 2007) and *Saccharomyces* (Qiao and Cech 2008). It has also been predicted in ciliates by phylogenetic comparison (Ulyanov *et al.* 2007) and in additional yeast species including *S. pombe* by RNA folding program (Guo *et al.* 2011), suggesting it is a structural motif conserved across the highly divergent yeast, ciliates, and vertebrate telomerase RNAs. The three-dimensional triple-helix formation is essential for the function of telomerase *in vivo* as demonstrated by disruption and compensatory mutations in *Kluyveromyces* (Shefer *et al.* 2007). Disruption of the triple helix in *S. cerevisiae* results in lower level of telomerase activity *in vitro* and telomere shortening *in vivo* (Qiao and Cech 2008). One or more 2'-OH groups in and adjacent to the triple helix regions of yeast and human telomerase RNAs participate in telomerase activity, possibly by directing the primer-template to the catalytic active site (Qiao and Cech 2008; Huang and Yu 2010).

Multiple mechanisms for synthesis of heterogenous telomeric repeats

Most of the known species of eukaryotes have telomeric repeats that are uniform in size and sequence. However, several organisms including some protozoa, fungi, slime molds and plants, have irregular telomere repeat sequences (Podlevsky *et al.* 2008). This is most pronounced in *Saccharomyces cerevisia*e and *Schizosaccharomyces pombe* with the telomere consensus sequences $(TG)_{1\text{-}4}G_{2\text{-}3}$ (Cohn *et al.* 1998) and $GGTTACA(G)_{1\text{-}4}$ (Hiraoka *et al.* 1998), respectively. Previous studies have also shown that some strains of *C. tropicalis* have telomeres containing two forms of telomeric repeats differing by a single nucleotide (A-repeat and C-repeat) (McEachern and Blackburn 1994). One chapter of this dissertation focuses on studying the mechanism of how these two forms of telomeric repeats are synthesized. We here review the existing hypotheses for synthesis of heterogeneous telomeric repeats in various organisms.

Different models have been put forward to explain the synthesis of variable telomeric repeats with only one RNA template. In *S. cerevisiae*, two possible mechanisms may account for the heterogeneous telomeric repeat synthesis: redundant alignment possibilities within the template region and abortive reverse transcription (Forstemann and Lingner 2001). The telomere template region of *TLC1* (CACCACACCCACACAC) suggests that the telomerase RNA may be able to align with a telomere terminus at a number of different points within the RNA, especially if just CAC can be sufficient for annealing. It is also possible that that the telomerase

could abort a round of reverse-transcription at several different positions along the RNA. If a terminal DNA sequence such as GTG is left, then alignment with the CAC of the template in the next round of elongation can readily occur (Singer and Gottschling 1994). In *Tetrahymena thermophila*, macronuclear telomeres are 250-350 bp, while the "germ-line" micronuclear telomeres are much longer, around 2.0-3.4 kb (Kirk and Blackburn 1995). Although the distal region of micronuclear telomeres is composed of homogeneous 5'-TTGGGG-3' repeats, the inner part contains homogeneous 5'-TTTGGGG-3' repeats, possibly due to a recombination process that can homogenize the inner part of micronuclear telomeres (Kirk and Blackburn 1995).

In *Paramecium tetraurelia*, the mixed synthesis of T_2G_4 and T_3G_3 telomeric repeats is caused by a high frequency of nucleotide misincorporation at a specific template position (McCormick-Graham and Romero 1996; McCormick-Graham *et al.* 1997; Ye and Romero 2002). In human and the plant *Aradidopsis thaliana*, variant telomeric repeat heterogeneity is found in the more internal (centromere proximal) parts of the telomeres, due to accumulation of random mutations independent of telomerase-mediated telomere addition (Brown *et al.* 1990; Richards *et al.* 1992). Template mutations and analysis of a large number of cloned telomeres from *S. pombe* suggest alternative translocation sites and primer slippage may contribute to telomeric repeat heterogeneity in that organism (Leonardi *et al.* 2008). The mechanisms for synthesis of heterogeneous telomeric repeat described above either

require only a single telomerase RNA template or are telomerase-independent and may not be sufficient to explain the described *C. tropicalis* situation.

In laboratory situations, the presence of more than one allele of telomerase may result in heterogeneous telomeric repeats. When a mutant *TER1* gene with an altered template sequence is introduced into species with uniform telomeric repeats, the corresponding mutant telomeric repeats can be incorporated into the ends of telomeres (Yu *et al.* 1990; Underwood *et al.* 2004; Bechard *et al.* 2009; Mravinac *et al.* 2011). If both mutant and wild type templates are present simultaneously, the terminal regions of telomeres will contain mixtures of the two repeat types.

The asexual pathogen *Candida tropicalis*

As mentioned in the earlier section, the third chapter of this dissertation focuses on the telomere and telomerase of *Candida tropicalis*. To better address the topic and to emphasize the importance of the study, a brief introduction of *Candida tropicalis* is given here.

Candida is a genus of budding yeast that contains ~200 species. While many species are commonly found as harmless commensals or endosymbionts in human and other warm-blooded animal hosts, more than 20 *Candida* species are know as human pathogens that can cause yeast infections known as candidiasis (Hazen 1995; Ha *et al.* 2011). Severe systemic infections can occur in the bloodstream and major organs,

especially in immunocompromised patients. Since the 1980s, *Candida* has become the fourth most common cause of bloodstream infections (BSIs) in the United States (Pappas *et al.* 2009). The mortality of invasive candidiasis can be >40% (d'Enfert,C. and Hube, B. (editors) 2007; Pappas *et al.* 2009; Ha *et al.* 2011). Although about half of the *Candida* infections are caused by *C. albicans*, multiple other *Candida* species contribute to the remaining cases of candidiasis, including *C. tropicalis*, *C. glabrata*, *C. parapsilosis*, *C. guilliermondii*, *C. stellatoidea*, *C. lusitaniae etc.* (Hazen 1995), and the prevalence of non-*albicans* infections is increasing (Pfaller and Diekema 2007).

Candida tropicalis is the second most commonly encountered *Candida* pathogen after *C. albicans*. Unlike *C. albicans*, which is often a commensal on human mucous membranes, *C. tropicalis* is more often found with deep fungal infections (Zaugg *et al.* 2001). One patient may be infected by more than one strain of *C. tropicalis* simultaneously or sequentially (Gelfand 1989).

C. tropicalis is closely related to *C. albicans*, and the similarities between them mirror their close phylogenetic relationship. Both *C. albicans* and *C. tropicalis* can undergo colony morphology switching (Suzuki *et al.* 1991). Like many other *Candida* species, *C. tropicalis* exists as either ellipsoidal budding cells or as pseudohyphae. In rare case *C. tropicalis* is capable of forming true hyphae, a feature shared only with *C. albicans* (Ann Chai *et al.* 2010). The CTG codon encodes serine instead of leucine in both *C.*

albicans and *C. tropicalis*, as a consequence of a CTG capture event that occurred 100-200 million years ago in this clade of pathogenic yeasts (Hara *et al.* 2000).

C. tropicalis and *C. albicans* have long been known as asexual diploid yeast. However, two mating type-like (MTL) loci, *MTL***a** and *MTL*α, have been identified in *C. albicans* (Richards *et al.* 1992) and annotated in *C. tropicalis* (Butler *et al.* 2009). In previous studies, a normally heterozygous (*MTL***a**/ α) *C. albicans* strain was manipulated in the laboratory to create **a** and α derivatives, either by deletion of one copy of the entire MTL locus (Hull *et al.* 2000), or by the technique of sorbose selection to homogenize the **a**/α parental version into MTL**a**/**a** and MTLα/α homozygous derivatives (Magee and Magee 2000). Mating was observed in the newly constructed **a** and α strains. The mating products were mononuclear and tetraploid in DNA content since *C. albicans* lacks the meiosis machinery (Tzung *et al*. 2001). When exposed to certain laboratory media, the tetraploid cells were induced to random chromosome loss to form a diploid, or very close to diploid, DNA content (Bennett and Johnson 2003). Moreover, the mating efficiency of *C. albicans* both *in vitro* and *in vivo* was relatively low, in some cases as low as one mating event per ten million potential mating partners (Bennett and Johnson 2005). A comparative genome analysis showed that missing components of meiosis in *C. albicans* are actually missing in all *Candida* species, suggesting that sexual *Candida* species undergo meiosis without them (Butler *et al*. 2009). Thus, it is possible that *C. albicans* and other *Candida* species can undergo meiosis using reduced machinery, or different

machinery, which should have significant variation from the current known meiosis cycles. Mating has never been observed yet in *C. tropicalis*.

Focus of this Dissertation

The studies included in this manuscript focus on two specific topics in telomerase RNA characterization. We first developed a novel approach to predict the core secondary and tertiary structure of yeast telomerase RNAs using a newly developed program called TRFolder. This program is capable of folding candidate TR gene sequences and to predict the common core structural elements found in all known TRs, including the pseudoknot and the triple helix, the template boundary element and the core-closing stem and assemble them automatically, which is beyond the capability of any other existing RNA-structure-profiling techniques. We applied our program to confirm and improve previously studied core structures from *Saccharomyces* and *Kluyveromyces* TRs, and make novel prediction of the core structural elements in *S. pombe*, *C. albicans* and several other yeast species.

We then studied the mechanism of dimorphic telomeric repeat synthesis in *C. tropicalis*. Our results suggest that two versions of telomerase RNA with different templates are responsible for the synthesis of the two types of telomeric repeat in specific *C. tropicalis* strains. We also show evidence that these strains have strikingly high sequence divergence between their allele pairs that appears to be genome-wide. This is an exciting result that could be evidence for mating and recombination in this

supposedly asexual yeast. Both of these topics expand our knowledge on yeast telomerase RNAs, and can help us to have a better understanding of the rapid evolution and of the sequence and structure of both telomerase RNA and telomeric DNA.

References

Ann Chai, L. Y., Denning, D. W. and Warn, P. (2010) *'Candida tropicalis* in human disease', *Crit Rev Microbiol* **36**: 282-298.

Autexier, C. and Greider, C. W. (1995) 'Boundary elements of the *Tetrahymena* telomerase RNA template and alignment domains', *Genes Dev.* **9**: 2227-2239.

Autexier, C. and Triki, I. (1999) '*Tetrahymena* telomerase ribonucleoprotein RNA-protein interactions', *Nucleic Acids Res.* **27**: 2227-2234.

Bachand, F. and Autexier, C. (2001) 'Functional regions of human telomerase reverse transcriptase and human telomerase RNA required for telomerase activity and RNA-protein interactions', *Mol Cell Biol* **21**: 1888-1897.

Bechard, L. H., Butuner, B. D., Peterson, G. J., McRae, W., Topcu, Z. and McEachern, M. J. (2009) 'Mutant telomeric repeats in yeast can disrupt the negative regulation of recombination-mediated telomere maintenance and create an alternative lengthening of telomeres-like phenotype', *Mol Cell Biol* **29**: 626-639.

Beernink, H. T., Miller, K., Deshpande, A., Bucher, P. and Cooper, J. P. (2003) 'Telomere maintenance in fission yeast requires an Est1 ortholog', *Curr Biol* **13**: 575-580.

Bennett, R. J. and Johnson, A. D. (2003) 'Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains', *EMBO J* **22**: 2505-2515.

Bennett, R. J. and Johnson, A. D. (2005) 'Mating in *Candida albicans* and the search for a sexual cycle', *Annu Rev Microbiol* **59**: 233-255.

Bergeron, J. and Drouin, G. (2008) 'The evolution of 5S ribosomal RNA genes linked to the rDNA units of fungal species', *Curr. Genet.* **54**: 123-131.

Berman, J., Tachibana, C. Y. and Tye, B. K. (1986) 'Identification of a telomere-binding activity from yeast', *Proc Natl Acad Sci U S A* **83**: 3713-3717.

Blackburn, E. H. and Gall, J. G. (1978) 'A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena*', *J Mol Biol* **120**: 33-53.

Blasco, M. A., Funk, W., Villeponteau, B. and Greider, C. W. (1995) 'Functional characterization and developmental regulation of mouse telomerase RNA', *Science* **269**: 1267-1270.

Brown, W. R., MacKinnon, P. J., Villasante, A., Spurr, N., Buckle, V. J. and Dobson, M. J. (1990) 'Structure and polymorphism of human telomere-associated DNA', *Cell* **63**: 119-132.

Bryan, T. M., Sperger, J. M., Chapman, K. B. and Cech, T. R. (1998) 'Telomerase reverse transcriptase genes identified in *Tetrahymena thermophila* and *Oxytricha trifallax*', *Proc Natl Acad Sci U S A* **95**: 8479-8484.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., Agrafioti, I., Arnaud, M. B., Bates, S., Brown, A. J., Brunke, S., Costanzo, M. C., Fitzpatrick, D. A., de Groot, P. W., Harris, D., Hoyer, L. L., Hube, B., Klis, F. M., Kodira, C., Lennard, N., Logue, M. E., Martin, R., Neiman, A. M., Nikolaou, E., Quail, M. A., Quinn, J., Santos, M. C., Schmitzberger, F. F., Sherlock, G., Shah, P., Silverstein, K. A., Skrzypek, M. S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M. P., Sudbery, P. E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N. A., Lorenz, M. C., Birren, B. W., Kellis, M. and Cuomo, C. A. (2009) 'Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes', *Nature* **459**: 657-662.

Cech, T. R. (2004) 'Beginning to understand the end of the chromosome', *Cell* **116**: 273-279.

Cesare, A. J., Quinney, N., Willcox, S., Subramanian, D. and Griffith, J. D. (2003) 'Telomere looping in *P. sativum* (common garden pea)', *Plant J* **36**: 271-279.

Chakrabarti, K., Pearson, M., Grate, L., Sterne-Weiler, T., Deans, J., Donohue, J. P. and Ares, M., Jr. (2007) 'Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis', *RNA* **13**: 1923-1939.

Chappell, A. S. and Lundblad, V. (2004) 'Structural elements required for association of the *Saccharomyces cerevisiae* telomerase RNA with the *Est2* reverse transcriptase', *Mol. Cell. Biol.* **24**: 7720-7736.

Chen, J. L., Blasco, M. A. and Greider, C. W. (2000) 'Secondary structure of vertebrate telomerase RNA', *Cell* **100**: 503-514.

Chen, J. L. and Greider, C. W. (2003) 'Template boundary definition in mammalian telomerase', *Genes Dev.* **17**: 2747-2752.

Chen, Z., Koeneman, K. S. and Corey, D. R. (2003) 'Consequences of telomerase inhibition and combination treatments for the proliferation of cancer cells', *Cancer Res* **63**: 5917-5925.

Cohen, S. B., Graham, M. E., Lovrecz, G. O., Bache, N., Robinson, P. J. and Reddel, R. R. (2007) 'Protein composition of catalytically active human telomerase from immortal cells', *Science* **315**: 1850-1853.

Cohn, M. and Blackburn, E. H. (1995) 'Telomerase in yeast', *Science* **269**: 396-400.

Cohn, M., McEachern, M. J. and Blackburn, E. H. (1998) 'Telomeric sequence diversity within the genus *Saccharomyces*', *Curr. Genet.* **33**: 83-91.

Collins, K. and Gandhi, L. (1998) 'The reverse transcriptase component of the *Tetrahymena* telomerase ribonucleoprotein complex', *Proc Natl Acad Sci U S A* **95**: 8485-8490.

Connor, J. M., Gatherer, D., Gray, F. C., Pirrit, L. A. and Affara, N. A. (1986) 'Assignment of the gene for dyskeratosis congenita to Xq28', *Hum Genet* **72**: 348-351.

Conrad, M. N., Wright, J. H., Wolf, A. J. and Zakian, V. A. (1990) 'RAP1 protein interacts with yeast telomeres in vivo: overproduction alters telomere structure and decreases chromosome stability', *Cell* **63**: 739-750.

Counter, C. M., Meyerson, M., Eaton, E. N. and Weinberg, R. A. (1997) 'The catalytic subunit of yeast telomerase', *Proc Natl Acad Sci U S A* **94**: 9202-9207.

d'Adda di Fagagna, F., Reaper, P. M., Clay-Farrace, L., Fiegler, H., Carr, P., Von Zglinicki, T., Saretzki, G., Carter, N. P. and Jackson, S. P. (2003) 'A DNA damage checkpoint response in telomere-initiated senescence', *Nature* **426**: 194-198.

Dandjinou, A. T., Levesque, N., Larose, S., Lucier, J. F., Abou Elela, S. and Wellinger, R. J. (2004) 'A phylogenetically based secondary structure for the yeast telomerase RNA', *Curr. Biol.* **14**: 1148-1158.

de Lange, T., Shiue, L., Myers, R. M., Cox, D. R., Naylor, S. L., Killery, A. M. and Varmus, H. E. (1990) 'Structure and variability of human chromosome ends', *Mol Cell Biol* **10**: 518-527.

d'Enfert,C. and Hube, B. (editors) (2007). '*Candida*: comparative and functional genomics', *Caisier Academic Press*.

Dionne, I. and Wellinger, R. J. (1998) 'Processing of telomeric DNA ends requires the passage of a replication fork', *Nucleic Acids Res* **26**: 5365-5371.

Donate, L. E. and Blasco, M. A. (2011) 'Telomeres in cancer and ageing', *Philos Trans R Soc Lond B Biol Sci* **366**: 76-84.

Feng, J., Funk, W. D., Wang, S. S., Weinrich, S. L., Avilion, A. A., Chiu, C. P., Adams, R. R., Chang, E., Allsopp, R. C., Yu, J. and et al. (1995) 'The RNA component of human telomerase', *Science* **269**: 1236-1241.

Fitzgerald, M. S., McKnight, T. D. and Shippen, D. E. (1996) 'Characterization and developmental patterns of telomerase expression in plants', *Proc Natl Acad Sci U S A* **93**: 14422-14427.

Fitzpatrick, D. A., Logue, M. E., Stajich, J. E. and Butler, G. (2006) 'A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis', *BMC Evol. Biol.* **6**: 99.

Flores, I. and Blasco, M. A. (2010) 'The role of telomeres and telomerase in stem cell aging', *FEBS Lett* **584**: 3826-3830.

Forstemann, K. and Lingner, J. (2001) 'Molecular basis for telomere repeat divergence in budding yeast', *Mol. Cell. Biol.* **21**: 7277-7286.

Forsyth, N. R., Wright, W. E. and Shay, J. W. (2002) 'Telomerase and differentiation in multicellular organisms: turn it off, turn it on, and turn it off again', *Differentiation* **69**: 188-197.

Friedman, K. L. and Cech, T. R. (1999) 'Essential functions of amino-terminal domains in the yeast telomerase catalytic subunit revealed by selection for viable mutants', *Genes Dev* **13**: 2863-2874.

Friedman, K. L., Heit, J. J., Long, D. M. and Cech, T. R. (2003) 'N-terminal domain of yeast telomerase reverse transcriptase: recruitment of Est3p to the telomerase complex', *Mol Biol Cell* **14**: 1-13.

Gelfand, M. S. (1989) '*Candida tropicalis*', *Infect Control Hosp Epidemiol* **10**: 280-283.

Gilley, D. and Blackburn, E. H. (1999) 'The telomerase RNA pseudoknot is critical for the stable assembly of a catalytically active ribonucleoprotein', *Proc. Natl. Acad. Sci. U S A* **96**: 6621-6625.

Granger, M. P., Wright, W. E. and Shay, J. W. (2002) 'Telomerase in cancer and aging', *Crit Rev Oncol Hematol* **41**: 29-40.

Greene, E. C. and Shippen, D. E. (1998) 'Developmentally programmed assembly of higher order telomerase complexes with distinct biochemical and structural properties', *Genes Dev.* **12**: 2921-2931.

Greider, C. W. and Blackburn, E. H. (1985) 'Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts', *Cell* **43**: 405-413.

Greider, C. W. and Blackburn, E. H. (1989) 'A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis', *Nature* **337**: 331-337.

Griffith, J. D., Comeau, L., Rosenfield, S., Stansel, R. M., Bianchi, A., Moss, H. and de Lange, T. (1999) 'Mammalian telomeres end in a large duplex loop', *Cell* **97**: 503-514.

Gunisova, S., Elboher, E., Nosek, J., Gorkovoy, V., Brown, Y., Lucier, J. F., Laterreur, N., Wellinger, R. J., Tzfati, Y. and Tomaska, L. (2009) 'Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements', *RNA* **15**: 546-559.

Guo, L., Zhang, D., Wang, Y., Malmberg, R. L., McEachern, M. and Cai, L. (2011) 'TRFolder: computational prediction of novel telomerase RNA structures in yeast genomes', *Int J Bioinform Res Appl* **7**: 63-81.

Ha, J. F., Italiano, C. M., Heath, C. H., Shih, S., Rea, S. and Wood, F. M. (2011) 'Candidemia and invasive candidiasis: a review of the literature for the burns surgeon', *Burns* **37**: 181-195.

Hanahan, D. and Weinberg, R. A. (2000) 'The hallmarks of cancer', *Cell* **100**: 57-70.

Hara, A., Ueda, M., Misawa, S., Matsui, T., Furuhashi, K. and Tanaka, A. (2000) 'A mutated hygromycin resistance gene is functional in the n-alkane-assimilating yeast *Candida tropicalis*', *Arch. Microbiol.* **173**: 187-192.

Harrington, L., Zhou, W., McPhail, T., Oulton, R., Yeung, D. S., Mar, V., Bass, M. B. and Robinson, M. O. (1997) 'Human telomerase contains evolutionarily conserved catalytic and structural subunits', *Genes Dev* **11**: 3109-3115.

Hassock, S., Vetrie, D. and Giannelli, F. (1999) 'Mapping and characterization of the X-linked dyskeratosis congenita (DKC) gene', *Genomics* **55**: 21-27.

Hazen, K. C. (1995) 'New and emerging yeast pathogens', *Clin Microbiol Rev* **8**: 462-478.

Heiss, N. S., Knight, S. W., Vulliamy, T. J., Klauck, S. M., Wiemann, S., Mason, P. J., Poustka, A. and Dokal, I. (1998) 'X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions', *Nat Genet* **19**: 32-38.

Hemann, M. T., Strong, M. A., Hao, L. Y. and Greider, C. W. (2001) 'The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability', *Cell* **107**: 67-77.

Henderson, E. R. and Blackburn, E. H. (1989) 'An overhanging 3' terminus is a conserved feature of telomeres', *Mol Cell Biol* **9**: 345-348.

Hiraoka, Y., Henderson, E. and Blackburn, E. H. (1998) 'Not so peculiar: fission yeast telomere repeats', *Trends Biochem. Sci.* **23**: 126.

Holt, S. E. and Shay, J. W. (1999) 'Role of telomerase in cellular proliferation and cancer', *J Cell Physiol* **180**: 10-18.

Hsu, M., McEachern, M. J., Dandjinou, A. T., Tzfati, Y., Orr, E., Blackburn, E. H. and Lue, N. F. (2007) 'Telomerase core components protect *Candida* telomeres from aberrant overhang accumulation', *Proc. Natl. Acad. Sci. U S A* **104**: 11682-11687.

Huang, C. and Yu, Y. T. (2010) 'Targeted 2'-O methylation at a nucleotide within the pseudoknot of telomerase RNA reduces telomerase activity *in vivo*', *Mol Cell Biol* **30**: 4368-4378.

Huard, S., Moriarty, T. J. and Autexier, C. (2003) 'The C terminus of the human telomerase reverse transcriptase is a determinant of enzyme processivity', *Nucleic Acids Res* **31**: 4059-4070.

Hughes, T. R., Weilbaecher, R. G., Walterscheid, M. and Lundblad, V. (2000) 'Identification of the single-strand telomeric DNA binding domain of the *Saccharomyces cerevisiae* Cdc13 protein', *Proc Natl Acad Sci U S A* **97**: 6457-6462.

Hull, C. M., Raisner, R. M. and Johnson, A. D. (2000) 'Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host', *Science* **289**: 307-310.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H. T., Rauhut, A.,

Reeb, V., Arnold, A. E., Amtoft, A., Stajich, J. E., Hosaka, K., Sung, G. H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J. M., Slot, J. C., Wang, Z., Wilson, A. W., Schussler, A., Longcore, J. E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P. M., Powell, M. J., Taylor, J. W., White, M. M., Griffith, G. W., Davies, D. R., Humber, R. A., Morton, J. B., Sugiyama, J., Rossman, A. Y., Rogers, J. D., Pfister, D. H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R. A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R. A., Serdani, M., Crous, P. W., Hughes, K. W., Matsuura, K., Langer, E., Langer, G., Untereiner, W. A., Lucking, R., Budel, B., Geiser, D. M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D. S., Lutzoni, F., McLaughlin, D. J., Spatafora, J. W. and Vilgalys, R. (2006) 'Reconstructing the early evolution of Fungi using a six-gene phylogeny', *Nature* **443**: 818-822.

Kachouri-Lafond, R., Dujon, B., Gilson, E., Westhof, E., Fairhead, C. and Teixeira, M. T. (2009) 'Large telomerase RNA, telomere length heterogeneity and escape from senescence in *Candida glabrata*', *FEBS Lett* **583**: 3605-3610.

Kilian, A., Bowtell, D. D., Abud, H. E., Hime, G. R., Venter, D. J., Keese, P. K., Duncan, E. L., Reddel, R. R. and Jefferson, R. A. (1997) 'Isolation of a candidate human telomerase catalytic subunit gene, which reveals complex splicing patterns in different cell types', *Hum Mol Genet* **6**: 2011-2019.

Kim, M. M., Rivera, M. A., Botchkina, I. L., Shalaby, R., Thor, A. D. and Blackburn, E. H. (2001) 'A low threshold level of expression of mutant-template telomerase RNA inhibits human tumor cell proliferation', *Proc Natl Acad Sci U S A* **98**: 7982-7987.

Kirk, K. E. and Blackburn, E. H. (1995) 'An unusual sequence arrangement in the telomeres of the germ-line micronucleus in *Tetrahymena thermophila*', *Genes Dev* **9**: 59-71.

Klobutcher, L. A., Swanton, M. T., Donini, P. and Prescott, D. M. (1981) 'All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus', *Proc Natl Acad Sci U S A* **78**: 3015-3019.

Krauskopf, A. and Blackburn, E. H. (1996) 'Control of telomere growth by interactions of RAP1 with the most distal telomeric repeats', *Nature* **383**: 354-357.

Kuilman, T., Michaloglou, C., Mooi, W. J. and Peeper, D. S. (2010) 'The essence of senescence', *Genes Dev* **24**: 2463-2479.

Kurtzman, C. P. (2003) 'Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*', *FEMS Yeast Res.* **4**: 233-245.

Lai, C. K., Miller, M. C. and Collins, K. (2002) 'Template boundary definition in *Tetrahymena* telomerase', *Genes Dev.* **16**: 415-420.

Lai, C. K., Mitchell, J. R. and Collins, K. (2001) 'RNA binding domain of telomerase reverse transcriptase', *Mol Cell Biol* **21**: 990-1000.

Lansdorp, P. M., Verwoerd, N. P., van de Rijke, F. M., Dragowska, V., Little, M. T., Dirks, R. W., Raap, A. K. and Tanke, H. J. (1996) 'Heterogeneity in telomere length of human chromosomes', *Hum Mol Genet* **5**: 685-691.

Legassie, J. D. and Jarstfer, M. B. (2006) 'The unmasking of telomerase', *Structure* **14**: 1603-1609.

Lendvay, T. S., Morris, D. K., Sah, J., Balasubramanian, B. and Lundblad, V. (1996) 'Senescence mutants of *Saccharomyces cerevisiae* with a defect in telomere replication identify three additional *EST* genes', *Genetics* **144**: 1399-1412.

Leonardi, J., Box, J. A., Bunch, J. T. and Baumann, P. (2008) '*TER1*, the RNA subunit of fission yeast telomerase', *Nat. Struct. Mol. Biol.* **15**: 26-33.

Li, B. and de Lange, T. (2003) 'Rap1 affects the length and heterogeneity of human telomeres', *Mol Biol Cell* **14**: 5060-5068.

Li, B., Oestreich, S. and de Lange, T. (2000) 'Identification of human Rap1: implications for telomere evolution', *Cell* **101**: 471-483.

Lin, J., Ly, H., Hussain, A., Abraham, M., Pearl, S., Tzfati, Y., Parslow, T. G. and Blackburn, E. H. (2004) 'A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein', *Proc. Natl. Acad. Sci. U S A* **101**: 14713-14718.

Lin, J. J. and Zakian, V. A. (1996) 'The *Saccharomyces* CDC13 protein is a single-strand TG1-3 telomeric DNA-binding protein in vitro that affects telomere behavior *in vivo*', *Proc Natl Acad Sci U S A* **93**: 13760-13765.

Lingner, J. and Cech, T. R. (1996) 'Purification of telomerase from *Euplotes aediculatus*: requirement of a primer 3' overhang', *Proc. Natl. Acad. Sci. U S A* **93**: 10712-10717.

Lingner, J., Cooper, J. P. and Cech, T. R. (1995) 'Telomerase and DNA end replication: no longer a lagging strand problem?', *Science* **269**: 1533-1534.

Lingner, J., Hendrick, L. L. and Cech, T. R. (1994) 'Telomerase RNAs of different ciliates have a common secondary structure and a permuted template', *Genes Dev.* **8**: 1984-1998.

Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V. and Cech, T. R. (1997) 'Reverse transcriptase motifs in the catalytic subunit of telomerase', *Science* **276**: 561-567.

Livengood, A. J., Zaug, A. J. and Cech, T. R. (2002) 'Essential regions of *Saccharomyces cerevisiae* telomerase RNA: separate elements for *Est1p* and *Est2p* interaction', *Mol. Cell. Biol.* **22**: 2366-2374.

Lustig, A. J., Kurtz, S. and Shore, D. (1990) 'Involvement of the silencer and UAS binding protein RAP1 in regulation of telomere length', *Science* **250**: 549-553.

Ly, H., Blackburn, E. H. and Parslow, T. G. (2003) 'Comprehensive structure-function analysis of the core domain of human telomerase RNA', *Mol. Cell. Biol.* **23**: 6849-6856.

Magee, B. B. and Magee, P. T. (2000) 'Induction of mating in *Candida albicans* by construction of MTLa and MTLalpha strains', *Science* **289**: 310-313.

Mantell, L. L. and Greider, C. W. (1994) 'Telomerase activity in germline and embryonic cells of *Xenopus*', *EMBO J* **13**: 3211-3217.

Marcand, S., Wotton, D., Gilson, E. and Shore, D. (1997) 'Rap1p and telomere length regulation in yeast', *Ciba Found Symp* **211**: 76-103.

McClintock, B. (1931) 'Cytological observations of deficiencies involving known genes, translocations and an inversion in *Zea mays*', *Mo. Agric. Exp. Res. Stn. Res. Bull.* **163**: 4-30.

McCormick-Graham, M., Haynes, W. J. and Romero, D. P. (1997) 'Variable telomeric repeat synthesis in *Paramecium tetraurelia* is consistent with misincorporation by telomerase', *EMBO J.* **16**: 3233-3242.

McCormick-Graham, M. and Romero, D. P. (1996) 'A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*', *Mol. Cell. Biol.* **16**: 1871-1879.

McEachern, M. J. and Blackburn, E. H. (1994) 'A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts', *Proc. Natl. Acad. Sci. U S A* **91**: 3453-3457.

McEachern, M. J. and Blackburn, E. H. (1995) 'Runaway telomere elongation caused by telomerase RNA gene mutations', *Nature* **376**: 403-409.

McEachern, M. J. and Hicks, J. B. (1993) 'Unusually large telomeric repeats in the yeast *Candida albicans*', *Mol. Cell. Biol.* **13**: 551-560.

Meyerson, M., Counter, C. M., Eaton, E. N., Ellisen, L. W., Steiner, P., Caddle, S. D., Ziaugra, L., Beijersbergen, R. L., Davidoff, M. J., Liu, Q., Bacchetti, S., Haber, D. A. and Weinberg, R. A. (1997) 'hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization', *Cell* **90**: 785-795.

Meyne, J., Ratliff, R. L. and Moyzis, R. K. (1989) 'Conservation of the human telomere sequence (TTAGGG)n among vertebrates', *Proc Natl Acad Sci U S A* **86**: 7049-7053.

Miller, M. C., Liu, J. K. and Collins, K. (2000) 'Template definition by *Tetrahymena* telomerase reverse transcriptase', *EMBO J* **19**: 4412-4422.

Minev, B., Hipp, J., Firat, H., Schmidt, J. D., Langlade-Demoyen, P. and Zanetti, M. (2000) 'Cytotoxic T cell immunity against telomerase reverse transcriptase in humans', *Proc Natl Acad Sci U S A* **97**: 4796-4801.

Mitchell, J. R., Wood, E. and Collins, K. (1999) 'A telomerase component is defective in the human disease dyskeratosis congenita', *Nature* **402**: 551-555.

Morin, G. B. (1989) 'The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats', *Cell* **59**: 521-529.

Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L. and Wu, J. R. (1988) 'A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes', *Proc Natl Acad Sci U S A* **85**: 6622-6626.

Mravinac, B., Mestrovic, N., Cavrak, V. V. and Plohl, M. (2011) 'TCAGG, an alternative telomeric sequence in insects', *Chromosoma* **120**: 367-376.

Muller, H. J. (1938) 'The remaking of chromosomes', *The collecting Net* 182-195.

Munoz-Jordan, J. L., Cross, G. A., de Lange, T. and Griffith, J. D. (2001) 't-loops at trypanosome telomeres', *EMBO J* **20**: 579-588.

Murti, K. G. and Prescott, D. M. (1999) 'Telomeres of polytene chromosomes in a ciliated protozoan terminate in duplex DNA loops', *Proc Natl Acad Sci U S A* **96**: 14436-14439.

Nair, S. K., Heiser, A., Boczkowski, D., Majumdar, A., Naoe, M., Lebkowski, J. S., Vieweg, J. and Gilboa, E. (2000) 'Induction of cytotoxic T cell responses and tumor immunity against unrelated tumors using telomerase reverse transcriptase RNA transfected dendritic cells', *Nat Med* **6**: 1011-1017.

Nakamura, T. M., Morin, G. B., Chapman, K. B., Weinrich, S. L., Andrews, W. H., Lingner, J., Harley, C. B. and Cech, T. R. (1997) 'Telomerase catalytic subunit homologs from fission yeast and human', *Science* **277**: 955-959.

Nugent, C. I., Hughes, T. R., Lue, N. F. and Lundblad, V. (1996) 'Cdc13p: a single-strand telomeric DNA-binding protein with a dual role in yeast telomere maintenance', *Science* **274**: 249-252.

Olovnikov, A. M. (1973) 'A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon', *J. Theor. Biol.* **41**: 181-190.

Ouellette, M. M., Liao, M., Herbert, B. S., Johnson, M., Holt, S. E., Liss, H. S., Shay, J. W. and Wright, W. E. (2000) 'Subsenescent telomere lengths in fibroblasts immortalized by limiting amounts of telomerase', *J Biol Chem* **275**: 10072-10076.

Pappas, P. G., Kauffman, C. A., Andes, D., Benjamin, D. K., Jr., Calandra, T. F., Edwards, J. E., Jr., Filler, S. G., Fisher, J. F., Kullberg, B. J., Ostrosky-Zeichner, L., Reboli, A. C., Rex, J. H., Walsh, T. J. and Sobel, J. D. (2009) 'Clinical practice

guidelines for the management of candidiasis: 2009 update by the Infectious Diseases Society of America', *Clin Infect Dis* **48**: 503-535.

Pfaller, M. A. and Diekema, D. J. (2007) 'Epidemiology of invasive candidiasis: a persistent public health problem', *Clin Microbiol Rev* **20**: 133-163.

Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X. and Chen, J. J. (2008) 'The telomerase database', *Nucleic Acids Res* **36**: 339-343.

Prescott, J. and Blackburn, E. H. (1997) 'Telomerase RNA mutations in *Saccharomyces cerevisiae* alter telomerase action and reveal nonprocessivity *in vivo* and *in vitro*', *Genes Dev.* **11**: 528-540.

Prowse, K. R., Avilion, A. A. and Greider, C. W. (1993) 'Identification of a nonprocessive telomerase activity from mouse cells', *Proc Natl Acad Sci U S A* **90**: 1493-1497.

Qiao, F. and Cech, T. R. (2008) 'Triple-helix structure in telomerase RNA contributes to catalysis', *Nat Struct Mol Biol* **15**: 634-640.

Reichenbach, P., Hoss, M., Azzalin, C. M., Nabholz, M., Bucher, P. and Lingner, J. (2003) 'A human homolog of yeast Est1 associates with telomerase and uncaps chromosome ends when overexpressed', *Curr Biol* **13**: 568-574.

Richards, E. J. and Ausubel, F. M. (1988) 'Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*', *Cell* **53**: 127-136.

Richards, E. J., Chao, S., Vongs, A. and Yang, J. (1992) 'Characterization of *Arabidopsis thaliana* telomeres isolated in yeast', *Nucleic Acids Res* **20**: 4039-4046.

Romero, D. P. and Blackburn, E. H. (1991) 'A conserved secondary structure for telomerase RNA', *Cell* **67**: 343-353.

Schnapp, G., Rodi, H. P., Rettig, W. J., Schnapp, A. and Damm, K. (1998) 'One-step affinity purification protocol for human telomerase', *Nucleic Acids Res.* **26**: 3311-3313.

Seto, A. G., Livengood, A. J., Tzfati, Y., Blackburn, E. H. and Cech, T. R. (2002) 'A bulged stem tethers *Est1p* to telomerase RNA in budding yeast', *Genes Dev.* **16**: 2800-2812.

Seto, A. G., Umansky, K., Tzfati, Y., Zaug, A. J., Blackburn, E. H. and Cech, T. R. (2003) 'A template-proximal RNA paired element contributes to *Saccharomyces cerevisiae* telomerase activity', *RNA* **9**: 1323-1332.

Shampay, J., Szostak, J. W. and Blackburn, E. H. (1984) 'DNA sequences of telomeres maintained in yeast', *Nature* **310**: 154-157.

Shay, J. W. and Bacchetti, S. (1997) 'A survey of telomerase activity in human cancer', *Eur. J. Cancer* **33**: 787-791.

Shay, J. W. and Roninson, I. B. (2004) 'Hallmarks of senescence in carcinogenesis and cancer therapy', *Oncogene* **23**: 2919-2933.

Shay, J. W. and Wright, W. E. (2005) 'Senescence and immortalization: role of telomeres and telomerase', *Carcinogenesis* **26**: 867-874.

Shay, J. W. and Wright, W. E. (2010) 'Telomeres and telomerase in normal and cancer stem cells', *FEBS Lett* **584**: 3819-3825.

Shefer, K., Brown, Y., Gorkovoy, V., Nussbaum, T., Ulyanov, N. B. and Tzfati, Y. (2007) 'A triple helix within a pseudoknot is a conserved and essential element of telomerase RNA', *Mol. Cell. Biol.* **27**: 2130-2143.

Shippen-Lentz, D. and Blackburn, E. H. (1990) 'Functional evidence for an RNA template in telomerase', *Science* **247**: 546-552.

Singer, M. S. and Gottschling, D. E. (1994) '*TLC1*: template RNA component of *Saccharomyces cerevisiae* telomerase', *Science* **266**: 404-409.

Snow, B. E., Erdmann, N., Cruickshank, J., Goldman, H., Gill, R. M., Robinson, M. O. and Harrington, L. (2003) 'Functional conservation of the telomerase protein *Est1p* in humans', *Curr Biol* **13**: 698-704.

Starling, J. A., Maule, J., Hastie, N. D. and Allshire, R. C. (1990) 'Extensive telomere repeat arrays in mouse are hypervariable', *Nucleic Acids Res* **18**: 6881-6888.

Steiner, B. R., Hidaka, K. and Futcher, B. (1996) 'Association of the Est1 protein with telomerase activity in yeast', *Proc Natl Acad Sci U S A* **93**: 2817-2821.

Stewart, J. A., Chaiken, M. F., Wang, F. and Price, C. M. (2011) 'Maintaining the end: Roles of telomere proteins in end-protection, telomere replication and length regulation', *Mutat Res*.

Suzuki, T., Miyamae, Y. and Ishida, I. (1991) 'Variation of colony morphology and chromosomal rearrangement in *Candida tropicalis* pK233', *J Gen Microbiol* **137**: 161-167.

Szostak, J. W. and Blackburn, E. H. (1982) 'Cloning yeast telomeres on linear plasmid vectors', *Cell* **29**: 245-255.

Theimer, C. A., Blois, C. A. and Feigon, J. (2005) 'Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function', *Mol. Cell.* **17**: 671-682.

Tzfati, Y., Fulton, T. B., Roy, J. and Blackburn, E. H. (2000) 'Template boundary in a yeast telomerase specified by RNA structure', *Science* **288**: 863-867.

Tzfati, Y., Knight, Z., Roy, J. and Blackburn, E. H. (2003) 'A novel pseudoknot element is essential for the action of a yeast telomerase', *Genes Dev.* **17**: 1779-1788.

Tzung, K.W., Williams, R.M., Scherer, S., Federspiel, N., Jones, T., Hansen, N., Bivolarevic, V., Huizar, L., Komp, C., Surzycki, R., Tamse, R., Davis, R.W. and Agabian, N. (2001) 'Genomic evidence for a complete sexual cycle in *Candida albicans*', *Proc. Natl. Acad. Sci. U S A* **98**: 3249-3253.

Ulyanov, N. B., Shefer, K., James, T. L. and Tzfati, Y. (2007) 'Pseudoknot structures with conserved base triples in telomerase RNAs of ciliates', *Nucleic Acids Res.* **35**: 6150-6160.

Underwood, D. H., Zinzen, R. P. and McEachern, M. J. (2004) 'Template requirements for telomerase translocation in *Kluyveromyces lactis*', *Mol Cell Biol* **24**: 912-923.

Vonderheide, R. H. (2002) 'Telomerase as a universal tumor-associated antigen for cancer immunotherapy', *Oncogene* **21**: 674-679.

Wang, Z. R., Guo, L., Chen, L. and McEachern, M. J. (2009) 'Evidence for an additional base-pairing element between the telomeric repeat and the telomerase RNA template in *Kluyveromyces lactis* and other yeasts', *Mol Cell Biol* **29**: 5389-5398.

Watson, J. D. (1972) 'Origin of concatemeric T7 DNA', *Nat New Biol* **239**: 197-201.

Webb, C. J. and Zakian, V. A. (2008) 'Identification and characterization of the *Schizosaccharomyces pombe TER1* telomerase RNA', *Nat. Struct. Mol. Biol.* **15**: 34-42.

Weinrich, S. L., Pruzan, R., Ma, L., Ouellette, M., Tesmer, V. M., Holt, S. E., Bodnar, A. G., Lichtsteiner, S., Kim, N. W., Trager, J. B., Taylor, R. D., Carlos, R., Andrews, W. H., Wright, W. E., Shay, J. W., Harley, C. B. and Morin, G. B. (1997) 'Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT', *Nat Genet* **17**: 498-502.

Wellinger, R. J., Wolf, A. J. and Zakian, V. A. (1993) '*Saccharomyces* telomeres acquire single-strand TG1-3 tails late in S phase', *Cell* **72**: 51-60.

White, L. K., Wright, W. E. and Shay, J. W. (2001) 'Telomerase inhibitors', *Trends Biotechnol* **19**: 114-120.

Wright, W. E. and Shay, J. W. (1992) 'The two-stage mechanism controlling cellular senescence and immortalization', *Exp Gerontol* **27**: 383-389.

Xie, M., Mosig, A., Qi, X., Li, Y., Stadler, P. F. and Chen, J. J. (2008) 'Structure and function of the smallest vertebrate telomerase RNA from teleost fish', *J Biol Chem* **283**: 2049-2059.

Ye, A. J. and Romero, D. P. (2002) 'A unique pause pattern during telomere addition by the error-prone telomerase from the ciliate *Paramecium tetraurelia*', *Gene* **294**: 205-213.

Yu, G. L., Bradley, J. D., Attardi, L. D. and Blackburn, E. H. (1990) 'In vivo alteration of telomere sequences and senescence caused by mutated *Tetrahymena* telomerase RNAs', *Nature* **344**: 126-132.

Zappulla, D. C. and Cech, T. R. (2004) 'Yeast telomerase RNA: a flexible scaffold for protein subunits', *Proc. Natl. Acad. Sci. U S A* **101**: 10024-10029.

Zaugg, C., Borg-Von Zepelin, M., Reichard, U., Sanglard, D. and Monod, M. (2001) 'Secreted aspartic proteinase family of *Candida tropicalis*', *Infect Immun* **69**: 405-412.

Zhao, Y., Sfeir, A. J., Zou, Y., Buseman, C. M., Chow, T. T., Shay, J. W. and Wright, W. E. (2009) 'Telomere extension occurs at most chromosome ends and is uncoupled from fill-in in human cancer cells', *Cell* **138**: 463-475.

Zhou, J., Hidaka, K. and Futcher, B. (2000) 'The Est1 subunit of yeast telomerase binds the Tlc1 telomerase RNA', *Mol Cell Biol* **20**: 1947-1955.

Zou, Y., Sfeir, A., Gryaznov, S. M., Shay, J. W. and Wright, W. E. (2004) 'Does a sentinel or a subset of short telomeres determine replicative senescence?', *Mol Biol Cell* **15**: 3709-3718.

Zvereva, M. I., Shcherbakova, D. M. and Dontsova, O. A. (2010) 'Telomerase: structure, functions, and activity regulation', *Biochemistry (Mosc)* **75**: 1563-1583.

Figure 1.1. The leading strand problem of DNA end replication. The leading strand will lose its 3' overhang upon replication since the parental strand does not provide a template. If the 3' overhang is not reestablished, the lagging strand problem will occur in the next round of replication [adapted from (Lingner *et al.* 1995)].

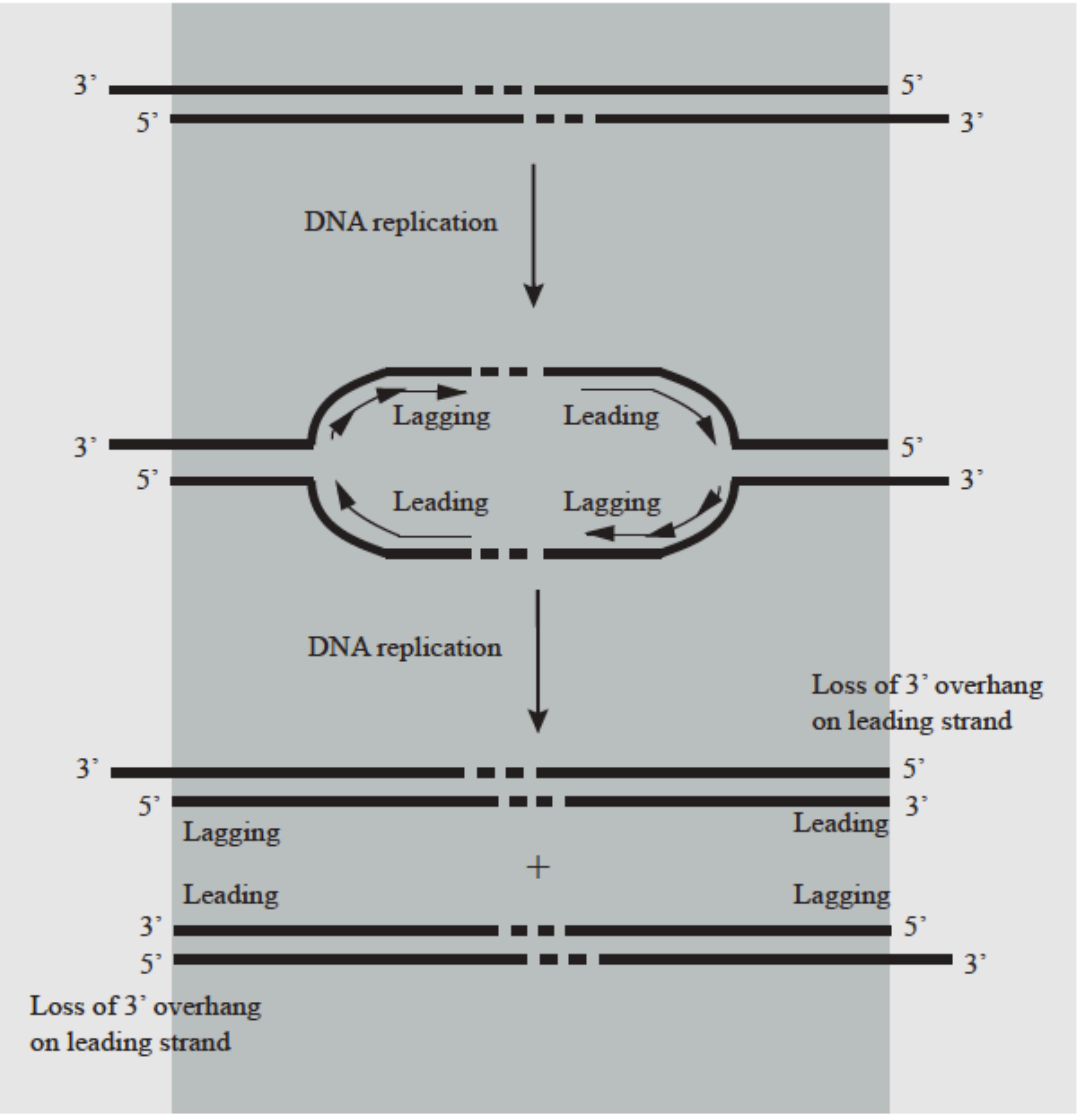

3'
5'
5'
3'
DNA replication
Lagging
Leading
3'
5'
5'
3'
Leading
Lagging
DNA replication
Loss of 3' overhang
on leading strand
3'
5'
5'
3'
Lagging
Leading
+
Leading
Lagging
3'
5'
5'
3'
Loss of 3' overhang
on leading strand

Figure 1.2. Long imperfect direct repeats border the template region of telomerase RNAs from multiple yeast species. Shown are template regions of telomerase RNAs from a variety of yeast species. Terminal repeats of the template are shown underlined. Nonidentical bases within the repeats are indicated by gaps while missing bases (relative to the other repeat) are indicated by carets. *S. cerevisiae* has three possible pairs of direct repeats, as indicated. The second pair in *S. cerevisiae* as well as the pair of direct repeats in *C. guilliermondii* is shown in two independent lines in the figure because of their overlap. The yeast species are shown next to an approximate phylogenetic tree based on previous studies (Kurtzman 2003; Fitzpatrick *et al.* 2006; James *et al.* 2006; Bergeron and Drouin 2008). The templates are aligned according to the region encoding the expected Rap1 binding site of the telomeric repeat (shaded gray). In some species, the Rap1 binding site is encoded by two discontinuous parts of the template, which are copied into consecutive Rap1 binding sites in tandem telomeric repeats.

Wang, Z.R., Guo, L., Chen, L., McEachern, M.J. 2009

Molecular and Cellular Biology. 29(20): 5389-5398.

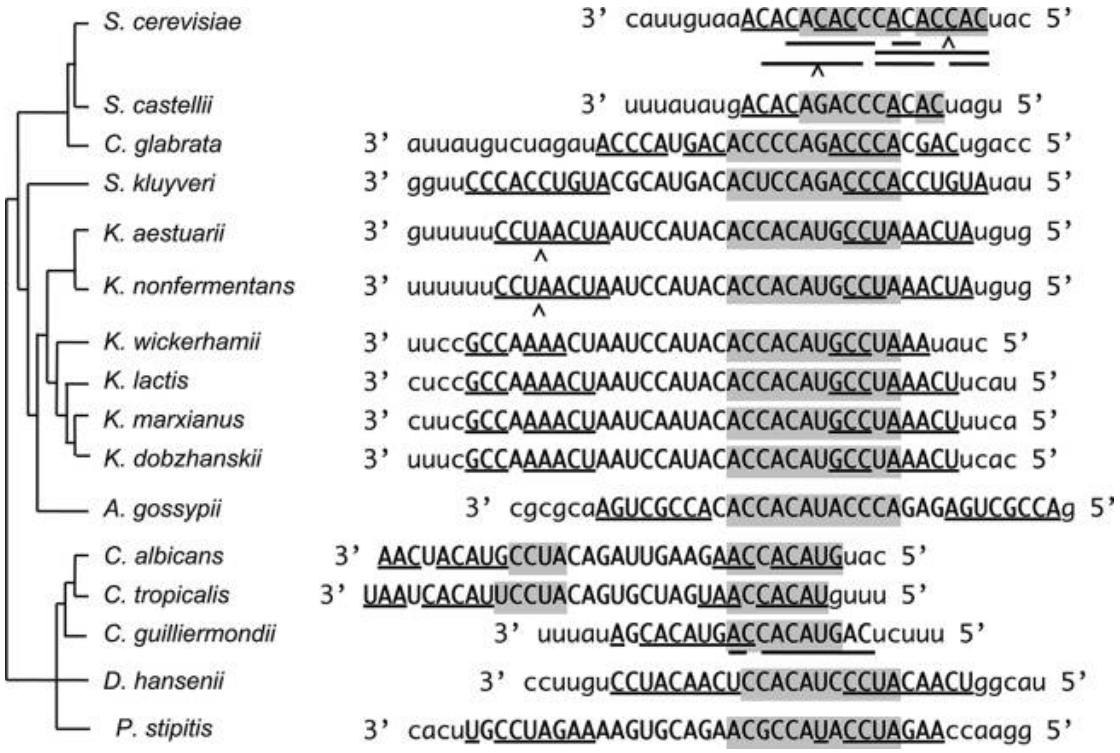

S. cerevisiae
3' cauuguaaACACACACCCACACCACuac 5'
S. castellii
3' uuuauaugACACAGACCCACACuagu 5'
C. glabrata
3' auuaugucuagauACCCAUGACACCCCAGACCCACGACugacc 5'
S. kluyveri
3' gguuCCCACCUGUACGCAUGACACUCCAGACCCACCUGUAuau 5'
K. aestuarii
3' guuuuuCCUAACUAAUCCAUACACCACAUGCCUAAACUAugug 5'
K. nonfermentans
3' uuuuuuCCUAACUAAUCCAUACACCACAUGCCUAAACUAugug 5'
K. wickerhamii
3' uuccGCCAAAACUAAUCCAUACACCACAUGCCUAAAuauc 5'
K. lactis
3' cuccGCCAAAACUAAUCCAUACACCACAUGCCUAAACUucau 5'
K. marxianus
3' cuucGCCAAAACUAAUCAAUACACCACAUGCCUAAACUuuca 5'
K. dobzhanskii
3' uuucGCCAAAACUAAUCCAUACACCACAUGCCUAAACUucac 5'
A. gossypii
3' cgcgcaAGUCGCCACACCACAUACCCAGAGAGUCGCCAg 5'
C. albicans
3' AACUACAUGCCUACAGAUUGAAGAACCACAUGuac 5'
C. tropicalis
3' UAAUCACAUUCCUACAGUGCUAGUAACCACAUguuu 5'
C. guilliermondii
3' uuuauAGCACAUGACCACAUGACucuuu 5'
D. hansenii
3' ccuuguCCUACAACUCCACAUCCCUACAACUggcau 5'
P. stipitis
3' cacuUGCCUAGAAAAGUGCAGAACGCCAUACCUAGAAccaagg 5'

Table 1.1. Telomeric repeat sequences in multiple organisms (adapted from the Telomerase Database [Podlevsky *et al.* 2008)].

Group	Organism	Telomeric repeat sequence (5' to 3')
Vertebrates	Human	TTAGGG
	Mouse	TTAGGG
	Zebrafish	TTAGGG
Nematodes	*Caenorhabditis elegans*	TTAGGC
Fission yeast	*Schizosaccharomyces pombe*	$G_{2–8}$TTAC(A)
Budding yeast	*Saccharomyces cerevisiae*	$T(G)_{2-3}(TG)_{1-6}$
	Kluyveromyces lactis	ACGGATTTGATTAGGTATGTGGTGT
	Candida albicans	ACGGATGTCTAACTTCTTGGTGT
	Candida glabrata	CTGGGTGCTGTGGGGT
	Candida guillermondii	ACTGGTGT
	Candida tropicalis	A[C/A]GGGATGTCACGATCATTGGTTGT
	Debaryomyces hansenii	ATGTTGAGGTGTAGGG
	Ashbya gossypii	GTGTGGTGTATGGGTCTCTCAGCG
	Pichia stipitis	GGATCTTTTCACGTCTTGCGGTA
	Yarrowia lipolytica	GGACGATTG
Plants	*Arabidopsis thaliana*	TTTAGGG
Ciliates	*Tetrahymena thermophila*	TTGGGG
	Paramecium tetraurelia	TT[T/G]GGG
	Euplotes crassus	TTTTGGGG

CHAPTER 2

TRFOLDER: COMPUTATIONAL PREDICTION OF NOVEL TELOMERASE RNA STRUCTURES IN YEAST GENOMES[1]

---

[1]Guo L, Zhang D, Wang Y, Malmberg RL, McEachern MJ and Cai L. 2010. *International Journal of Bioinformatics Research and Applications*. 7(1): 63-81.

Reprinted here with permission of publisher.

Abstract

The identification of Telomerase RNAs (TRs) has been difficult owing to their rapid evolutionary divergence. The common core structure found in all known TRs contains a pseudoknot and a triple helix, which are beyond the capability of existing RNA-structure-profiling techniques. We describe a novel approach to predict the structure of key TR features and to aid the identification of TRs in genomes, using a program we developed, TRFolder. We applied our method to confirm and improve previously studied core structures from *Saccharomyces* and *Kluyveromyces* TRs. We made novel structural predictions of core elements of the TRs from *Schizosaccharomyces pombe*, *Candida albicans*, and several other yeast species.

Introduction

Telomerase RNA (TR) is a vital component of telomerase, the enzyme that functions to ensure complete replication of telomeres (Smogorzewska and de Lange, 2004). TRs contain a short region that is complementary to the telomeric repeat sequence and serves as a template for telomerase to synthesise telomeric DNA repeats (Greider and Blackburn, 1989). Telomere maintenance is critical to cellular immortalisation and there is much interest in telomerase's role in both cancer and aging (Shay and Roninson, 2004; Shay and Wright, 2005). TRs have been identified in ciliates (Greider and Blackburn, 1989; Romero and Blackburn, 1991; Shippen-Lentz and Blackburn, 1990), vertebrates (Blasco *et al*., 1995; Chen *et al*., 2000; Feng *et al*., 1995) and yeast species (Feng *et al*., 1995; Gunisova *et al*., 2009; Hsu *et al*., 2007; Leonardi

et al., 2008; McEachern and Blackburn, 1995; Singer and Gottschling, 1994; Webb and Zakian, 2008). Computational identification of TR sequences by sequence similarity has been difficult owing to the evolutionary divergence of TR sequences.

All TRs examined to date have certain conserved structural features. Fungal, vertebrate and ciliate TRs contain a conserved core structure consisting of four structural elements (Figure 2.1A): the template region, the 5' boundary element upstream of the template, a downstream pseudoknot structure, which binds TERT, and the core-closing stem (a long-range base-pairing element that encloses the template and the pseudoknot) (Chen *et al*., 2000; Chen and Greider, 2003; Comolli *et al*., 2002; Lingner *et al*., 1994; Romero and Blackburn, 1991; Theimer *et al*., 2000, 2003). In addition, a triple helix motif within the pseudoknot has been identified in *Kluyveromyces* (Shefer *et al*., 2007) and human TR (Theimer *et al*., 2005).

The most effective RNA structure prediction method is a structural homologue search, in which the conserved secondary structure of an RNA family is profiled and used to search for genome regions that match the structural profile (Eddy, 2006). Stochastic Context-Free Grammars (SCFGs) are often used to probabilistically model the consensus structures (Eddy and Durbin, 1994; Nawrocki *et al*., 2009; Sakakibara *et al*., 1994). However, existing programs may not perform well on TR structure searches, as the consensus core structure of TRs contains extensive variation in the length of stems and loops. We have used the homology search tools Infernal (Nawrocki *et al*.,

2009) and RNATOPS (Huang *et al*., 2008) to search for the TR of *Kluyveromyces lactis* using the *Saccharomyces* TR consensus structure profile obtained from Rfam (http://rfam.sanger.ac.uk/). Neither program found a hit. Energy-based folding methods (Hofacker, 2003; Zuker, 2003) do not work well for TRs because they are usually longer than 350 nucleotides and contain pseudoknots.

We introduce TRFolder, a utility program that consists of a set of functions for TR-specific structure prediction. Unlike existing general-purpose structure prediction programs, TRFolder is effective in folding sequences with a putative TR template into the best possible TR core structural elements. We test our approach on the well-studied yeast *Saccharomyces* and *Kluyveromyces* TR core structures. We have applied TRFolder to the prediction of TRs in several yeast species for which TR structures were not previously known, including *C. glabrata*, *C. guilliermondii*, *C. tropicalis*, *A. gossypii*, *D. hansenii*, *P. stipitis* and *S. pombe*. In most of these species, the TR genes were recently identified by Gunisova *et al*. (2009), and are confirmed by our independent analysis.

Methods

Our method predicts the TR structure by predicting its core structure components one at a time based on statistical profiles of components of known TR structures. There are three technical steps:

- identify likely TR genes using telomeric homology to the putative template combined with neighbouring gene analysis
- predict the pseudoknot in the vicinity of regions containing the candidate templates and filter out structures that lack a potential triple helix structure
- predict the 5' template boundary element (a stem-loop structure in yeast) and the base-pairing regions closing the structure.

Training data collection

The TR core secondary structure information from the following were used as training data for this work: *S. cerevisiae*, *S. cariocanus*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *K. lactis* (Lin *et al*., 2004), *K. nonfermentans*, *K. aestuarii*, *K. dobzhanskii*, *K. wickerhamii* and *K. marxianus* (Box *et al*., 2008; Dandjinou *et al*., 2004; Shefer *et al*., 2007; Tzfati *et al*., 2003). Each stem and loop length in the four structural elements (pseudoknot, triple helix, boundary element and core-closing stem) is summarised in Table 2.1. The numbering of the stems and loops is shown in Figure 2.1A. The sequences of the *TER1/TLC1* genes were obtained from the GenBank data library: AY639009 (*S. cerevisiae*), AY639010 (*S. cariocanus*), AY639011 (*S. mikatae*), AY639012 (*S. kudriavzevii*), AY639013 (*S. bayanus*), AY639015 (*S. paradoxus*), U31465 (*K. lactis*), AY151277 (*K. aestuarii*), AY151279 (*K. marxianus*), AY151278 (*K. dobzhanskii*), AY151281 (*K. wickerhamii*) and AY151280 (*K. nonfermentans*).

The parameters used in TRFolder were generally defined as the mean size plus/minus two times the standard deviation of the previously determined sizes from the *Saccharomyces* and *Kluyveromyces* core structures (Chen and Greider, 2003; Lin *et al*., 2004; Shefer *et al*., 2007; Tzfati *et al*., 2003).

Structure profiling

For each core structure prediction, we built a 5 × 5 log-odds matrix. These matrices indicate the frequency of association between each two bases in each core structure. We define $P(a, b)$, the base pair probability distribution obtained from rRNAs of Rfam, as the prior frequencies. The probability distribution $F(a, b)$ for base pairs is for any specific structure for a specific family of organisms. The program uses the combined matrix $M = wF + (1 - w) P$, for a chosen number w, $0 \leqslant w \leqslant 1$. The score of base pair between nucleotides a and b is computed as $\log(M(a, b)/q(a)\, q(b))$, where q is the probability for individual nucleotides of the background. There are three steps to obtain $F(a, b)$.

- The probability $q(a)$ of individual nucleotides a is computed from the RNA sequence.
- (I) Counting the number of occurrences for each canonical base pair (i.e., each of AU, UA, GC, CG, GU, UG). (II) Counting the number of occurrences for each gap (i.e., each of A-, -A, -G, G-, C-, -C, U-, -U). (III) To make the predicted stem more stable, we used a different approach to deal with the occurrence of non-canonical base pairs. For each occurrence of non-canonical base pair, we regard it as two gaps A- and

-C, and count both of them. (IV) Adding (II) and (III) for each gap. (V) Summing all counts into a total. (VI) Let c be the pseudocount, add c into the count of each pair and gap. (VII) For each canonical base pair or a gap, compute the frequency as ($count + c$)/($total + 24c$). (VIII) For each non-canonical pair, the frequency = c/($total + 24c$). These become the values of frequency function F.

- The score for a gap (e.g., a-) is computed as $\log(M(a,-)/q(a))$ while the score for a canonical base pair (e.g., between a and b) is computed as $\log (M(a, b)/(q(a) \times q(b))$.

Structure prediction

In general, the prediction of a stem (i.e., a double helix) is accomplished through a pairwise complementary alignment between two regions. The length of the stem needs to be within a statistical range estimated from the training data, the average plus or minus two standard deviations. There are three steps for the pseudoknot and triple helix prediction.

- For a given search window, whose size could be adjusted by users, apply a (semi-global) pairwise complementary alignment to find all meaningful stems for which the score is greater than zero and the length of base pairs is no less than three.
- Each pair of stems that are crossing, but not overlapping will be combined as a pseudoknot; the pseudoknot that has the maximum score will be kept.
- For each pseudoknot candidate, the first loop and the last stem arm are folded to be a triple helix via (local) pairwise complementary alignment.

The sequence segment between the two pairing regions of the stem is not scored but its length needs to be within a statistical range estimated from the training data. Additional conditions, such as the distance from a specific position, may be enforced on a predicted stem by the user.

A third arm is then predicted within the U-rich loop 1 region in the predicted pseudoknot, which can align with the 5' arm of stem 2 in the same direction, and thus form a triple helix with stem 2. After each core structure prediction, there were many structurally similar candidates. TRFolder computes a similarity filter measured in terms of the midpoint of stem arms. For all i, $|mid(Ai)\text{-}mid(Bi)| < (len(Ai) + len(Bi))/4$, where *Ai* and *Bi* are the two arms, *len* is the length function and *mid* is the midpoint position function.

The predicted core structures can be assembled and ranked with user-supplied weights for each of the structural elements. In our experiments, we first set weights for all four elements to be 1, to obtain the top 50 structure predictions by TRFolder in all 12 species. The prediction in each species that is consistent or closest with the previously proposed TR core structure was selected. We then used a 'grid search method' to identify the best weight combination of these structural elements that can rank the selected correct structure within top 3 in all or most of the ten *Saccharomyces* and *Kluyveromyces* species. We explored various possible weights for combining the

scores of the structural features, and found that simply weighting them all equally at 1 worked well.

Selection of templates for TRFolder

We chose six yeasts with fully or partially sequenced genomes and uniform telomeric repeats (*C. glabrata, C. guilliermondii*, *C. tropicalis*, *A. gossypii*, *D. hansenii* and *P. stipitis*) to search for candidate TR template by using telomeric sequences as BLAST queries on the selected genomes. The telomeric repeat sequences were from the Telomerase Database (http://telomerase.asu.edu/) (Podlevsky *et al*., 2008). Genomic sequence downloads were from Genolevures (http://www.genolevures.org/) for *C. glabrata*, *K. lactis*, *D. hasenii* and *A. gossypii* and from the *Candida* Database from the Broad Institute (http://www.broad.mit.edu) for *C. tropicalis* and *C. guilliermondii*. Genomic DNA sequences from *S. cerevisiae* were from *Saccharomyces* Genome Database (http://www.yeastgenome.org/) while those for other *Saccharomyces* (Dandjinou *et al*., 2004) and *Kluyveromyces* species (Seto *et al*., 2002) were from GenBank. Only one candidate with a perfect match of at least 1 bp longer than one copy of the telomeric repeat sequence was found in each of *C. glabrata*, *C. tropicalis* and *A. gossypii*. Two candidates were found in *P. stipitis*; 12 candidates were found in *D. hansenii*.

We next compared the neighbouring genes of the TR candidate regions with the neighboring genes of the identified TR genes in *Saccharomyces* species,

Kluyveromyces species, *C. albicans* and *S. pombe* by checking the annotated genome (Dujon *et al*., 2004; Gattiker *et al*., 2007; Hirschman *et al*., 2006; Jeffries *et al*., 2007; Rossignol *et al*., 2008). The *S. pombe* TR gene (*TER1*) shares a common nearby gene (*DAD1*) with the TR genes in *K. lactis*, *A. gossypii* and *D. hansenii*, even though *S. pombe* is a distant relative of the other species. Neighbouring gene analysis also allowed us to identify the likely TR gene in *C. guilliermondii*, although this species only has a partial genome sequence and a short telomeric repeat (8 bp). A summary of the candidate TR templates is in Table 2.2. Further evidence that the TR candidates were TR genes came from the finding that each of them had the extended base-pairing potential between the 3′ -side of the RNA template and the telomeric DNA, as has been found in *K. lactis* (Wang *et al*., 2009).

Results

TRFolder confirmed the previously proposed TR structures in *Saccharomyces* and *Kluyveromyces*

TRFolder successfully predicted the presence of pseudoknots with an overlapping triple helix and other structural elements in the 4 kb sequence centred on the TR template in the *Saccharomyces* and *Kluyveromyces* species (Table 2.S1). A comparison of *S. cerevisiae* and *K. lactis* TR core structures predicted by TRFolder vs. those of previous studies is shown in Figure 2.2. The comparison in the other species is shown in Table 2.S1. We found that all of the previously proposed pseudoknot structures (Chappell and Lundblad, 2004; Lin *et al*., 2004; Tzfati *et al*., 2003) were in

our predictions, and all but one of them was the top-ranked in Z-score by TRFolder (Table 2.S1). The other predicted structures in our results were highly similar, and, in most species, largely overlapped with the previously proposed structures (data not shown). The only exceptional case is that the predicted structure in the *K. wickerhamii* TR has a completely different stem 1 of the pseudoknot.

The boundary element was predicted as the highest-scoring single stem-loop structure upstream of the TR template beginning 0–3 nt from the TR template (Table 2.1). For all the *Kluyveromyces* and *Saccharomyces* species, we found only one qualified stem-loop structure in each species, and most of them are very similar to those previously proposed. Some of our predictions missed one or a few base pairs.

We defined the position of the core-closing stem based on our prediction of the pseudoknot structure and 5′ boundary element. Also, we set a smaller gap penalty value for the prediction of the core-closing stem to allow a longer stem with more gaps or bulges. The positions of the predicted core-closing stems in *Saccharomyces* species turned out to be highly similar to those predicted in the previous work (Chappell and Lundblad, 2004; Lin *et al*., 2004; Tzfati *et al*., 2003), while the predicted base pairing is slightly different (Lin *et al*., 2004). In *S. kudriavzevii*, *S. paradoxus* and *S. mikatae*, our predictions of the core-closing stems have at least 8 more base pairs compared with those of Lin *et al*. (2004) (Table 2.3).

Prediction of novel core secondary structures

We next predicted the structural elements of TR genes for which limited or no previous structural predictions had been made (*C. glabrata*, *A. gossypii*, *C. albicans*, *C. tropicalis*, *C. guilliermondii*, *P. stipitis*, *D. hansenii* and *S. pombe*). Because the sizes of known yeast TRs are around 1–2 kb, TRFolder filtered out the candidates whose distance between the putative template and the pseudoknot structure is above 2 kb. The candidates were ranked by the summation of scores for the predicted pseudoknot with a triple helix, core-closing stem and boundary element (Tables 2.4 and 2.S2). The TR core structures of *S. pombe*, *P. stipitis*, *A. gossypii* and *D. hansenii*, which were not previously predicted, are shown in Figure 2.3.

In five of the eight species, *A. gossypii*, *C. albicans*, *P. stipitis*, *C. guilliermondii* and *S. pombe*, the top three structures were exactly the same as each other except for the core-closing stems; generally, the top alternatives were strongly overlapping. For the other three yeast species, the three highest-scoring predictions had differences other than just in the core-closing stems. The three highest scoring of the predicted *C. glabrata* TR structures had three completely different pseudoknots-triple helices. The pseudoknot and triple helix of the second highest ranking of these structures is highly similar to those recently proposed by Kachouri-Lafond *et al*. (2009) using comparative sequence analysis. The top two predicted pseudoknots were both surprisingly far (1.4 kb) from the template, implying that the size of the RNA must be at least 2 kb. Northern blot analysis confirmed that the *C. glabrata* TR is unusually

large and contains the region of the predicted pseudoknots (Kachouri-Lafond *et al*., 2009). In the top three predicted structures of *C. tropicalis*, three pseudoknots are located about 400 nt away from the template, and have a nearly identical stem 2. The top and second best structures differ from the third in having a significantly better triple helix. Only one boundary element and one core-closing stem were predicted in the top three structures. The top three predicted secondary structures in *D. hansenii* have exactly the same boundary element. The pseudoknots in the top and third best structures are the same, and both stems highly overlap with those in the second best structure.

In three of the eight yeasts, *C. albicans*, *C. tropicalis* and *C. guilliermondii*, the 5' and 3' ends of the TR have been recently mapped (Gunisova *et al*., 2009; Hsu *et al*., 2007). In each case, our predicted structures fall within the mapped gene. However, the 5' arm of the predicted boundary element and the 5' arm of the core-closing stem of *S. pombe* are outside of the mapped region of the *S. pombe* TR (Leonardi *et al*., 2008) (Table 2.S2). Another study suggested that part of the paired region in the *S. pombe* boundary element overlaps with the template itself (Box *et al*., 2008), a structure not currently permitted by TRFolder. Moreover, the size constraint of the loop applied to the boundary element in our studies was 160–410 nt, which is larger than the loop size (57 nt) proposed by Leonardi and co-workers. When we adjusted the constraint of the loop size in TRFolder to 50–410 nt, and provided a truncated template excluding the overlapping part, the suggested boundary element was in each of the top three

predicted core structures (data not shown). While this new boundary element also led to a different prediction of core-closing stem, exactly the same pseudoknot and triple helix were in each of the top three core structures of this test as in the initial analysis using the full-length template. *S. pombe* is very distantly related to other yeast species, which we used as training data. This is the first proposed pseudoknot structure with a triple helix for the TR of *S. pombe*.

TRs generate significantly higher scores than scrambled sequences folded by TRFolder

We tested our program using random sequences generated in two different ways to serve as a negative control. With the 'random position approach', five 4 kb genomic sequences were excised randomly from the same chromosome where the TR gene was located, to mimic a false positive predicted template. With the 'random shuffle approach', we randomly shuffled the sequence of the chromosome containing TR gene, and took five random 4 kb segments from the shuffled sequence. For species whose genomes have not yet been completely sequenced, the shuffled sequences were generated based on all the available genomic sequences in these species in NCBI. The exact midpoint of the 4 kb sequence was designated as being the position of the template sequence for TRFolder.

The test results on the random sequences are summarised in Table 2.5. The average score of the top-ranked secondary structure on the random sequences from all species

was similar, within the range from 30 to 35, indicating this is the background score of yeast genomes produced by TRFolder. However, the maximum score TRFolder obtained on random sequences in each of the 18 species was usually above 35. In some species, the score of the top-ranked core secondary structure in putative *TER1* was not significantly higher, or was sometimes even lower than the maximum score on random sequence*s* (*D. hansenii* for example). To test whether TR candidates as a group were scoring higher than random sequences, we used a pairwise *t*-test to compare the scores of the top TR candidate and of the random sequences from each species. For each species, the top-scoring predicted structure of a TR candidate was paired with the predicted structure of each of the five random sequences, and the significance of score difference was calculated. The results showed that the highest score of TR candidates was significantly higher than the scores of random sequences ($p < 0.0001$). For the random position approach, there were 40 pairs in total for pairwise *t*-test (10 species, 5 pairs in each species). The results showed that the highest score of TR candidates was also significantly higher than the scores of random sequences ($p < 0.0001$). Since there were six random sequences whose scores were zero, and these might have caused biased results, we re-ran the pairwise *t*-tests leaving out these sequences. The test results were still significant ($p < 0.0001$) in both negative control approaches. These tests show that, as a group, the TR candidates we identified had significantly higher Z-scores than the random sequences.

Discussion

We introduce TRFolder, a program that is capable of TR core structure prediction in distantly related yeast species independent from sequence comparative analysis. TRFolder was applied to multiple yeast species with newly identified TRs, including *C. tropicalis*, *C. guilliermondii*, *C. glabrata*, *A. gossypii*, *D. hansenii, P. stipitis* and *S. pombe*. Once TR gene candidates were identified, the TRFolder program was used to predict candidate structural features. TRFolder is the first program that is specific for TR prediction, and is able to find a pseudoknot structure with a triple helix, together with other key structural features of TRs, over a broad phylogenetic range of yeast species.

Several lines of evidence suggest that the proposed TR core structures identified by TRFolder are correct or close to correct. First, the program correctly identified the core structures previously identified in both the *Saccharomyces* and the *Kluyveromyces* genera. Second, within all the yeast species examined, the best-scoring core structures proposed are typically similar and overlapping to each other. Third, TRFolder's predictions of core secondary structure on most of the novel TRs were within the mapped region of TR transcripts (Gunisova *et al*., 2009; Kachouri-Lafond *et al*., 2009), although the predictions were based on larger 4 kb genomic sequences. Of course, the novel core TR structures predicted here need experimental verification.

This work is the first to propose several structural elements in a few specific species with previously identified TRs, for example, the triple helix in the pseudoknot of *Saccharomyces* TRs; the whole core secondary structure in *C. albicans*; the pseudoknot with a triple helix in *S. pombe*. We have also confirmed the identification and report structural predictions for TRs from *C. glabrata*, *C. guilliermondii*, *C. tropicalis*, *A. gossypii*, *D. hansenii* and *P. stipitis* (Figure 2.3).

There are differences in reliability for the predictions of different structural elements. The pseudoknot structure with a triple helix is the most reliable predictive feature; the boundary element is second; while the core-closing stem is the least reliably predictable structural element.

The predictions on random sequences by TRFolder showed that, in many cases, the known or putative TRs have Z-scores that are much higher than the best scores of the random sequences. The Z-scores of the best-ranked core structures in all known or putative TRs are comparable with or higher than the mean of the Z-score distribution (data not shown). The fact that some random sequences achieved scores as high as some TR candidates indicates that the utility of TRFolder may be limited to situations where the TR gene has already been identified or narrowed down to a small number of candidates.

While TRFolder alone is not currently capable of TR gene identification by scanning a genome, it is capable of narrowing down a list of candidates identified by telomere-homology or other means. For example, in the *D. hansenii* genome, there are 12 sequences containing long enough stretches of telomere homology to be a possible TR template. TRFolder ranked the correct sequence (as judged by neighbouring gene analysis) as one of the four highest Z-scores among the 12 candidates.

The method employed by TRFolder is different from that used in other structure prediction tools such as MFOLD (Mathews *et al*., 1999; Zuker, 2003), which has been used to predict structural elements of yeast TRs (Tzfati *et al*., 2000). For TRs, MFOLD works well only for short sequences and is not capable of pseudoknot prediction. Gunisova *et al*. (2009) mention using a specific computer algorithm, without presenting details, to search for a pseudoknot in *Candida* spp. putative TRs.

The TRFolder utility developed in our study is the first program that is specifically designed for TRs and that can automatically predict and assemble the set of structural elements comprising the TR core. We chose yeast TRs as a model, but the algorithm developed in our study should be useful with other groups as well. TRFolder is freely available for users to download.

Acknowledgements

We thank Yong Wu for testing the original idea of a pseudoknot search algorithm used in TRFolder, and Jing Xu for the statistical analysis in this study. This work was supported by the National Institute of Health [BISTI R01GM072080-01A1, GM 61645].

References

Blasco, M.A., Funk, W., Villeponteau, B. and Greider, C.W. (1995) 'Functional characterization and developmental regulation of mouse telomerase RNA', *Science*, Vol. 269, pp.1267–1270.

Box, J.A., Bunch, J.T., Zappulla, D.C., Glynn, E.F. and Baumann, P. (2008) 'A flexible template boundary element in the RNA subunit of fission yeast telomerase', *J. Biol. Chem.*, Vol. 283, pp.24224–24233.

Brown, Y., Abraham, M., Pearl, S., Kabaha, M.M., Elboher, E. and Tzfati, Y. (2007) 'A critical three-way junction is conserved in budding yeast and vertebrate telomerase RNAs', *Nucleic Acids Res*., Vol. 35, pp.6280–6289.

Chappell, A.S. and Lundblad, V. (2004) 'Structural elements required for association of the *Saccharomyces cerevisiae* telomerase RNA with the *Est2* reverse transcriptase', *Mol. Cell. Biol.*, Vol. 24, pp.7720–7736.

Chen, J.L. and Greider, C.W. (2003) 'Template boundary definition in mammalian telomerase', *Genes Dev.*, Vol. 17, pp.2747–2752.

Chen, J.L., Blasco, M.A. and Greider, C.W. (2000) 'Secondary structure of vertebrate telomerase RNA', *Cell*, Vol. 100, pp.503–514.

Comolli, L.R., Smirnov, I., Xu, L., Blackburn, E.H. and James, T.L. (2002) 'A molecular switch underlies a human telomerase disease', *Proc. Natl. Acad. Sci.*, *USA*, Vol. 99, pp.16998–17003.

Dandjinou, A.T., Levesque, N., Larose, S., Lucier, J.F., Abou Elela, S. and Wellinger, R.J. (2004) 'A phylogenetically based secondary structure for the yeast telomerase RNA', *Curr. Biol.*, Vol. 14, pp.1148–1158.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P. and Souciet, J.L. (2004) 'Genome evolution in yeasts', *Nature*, Vol. 430, pp.35–44.

Eddy, S.R. (2006) 'Computational analysis of RNAs', *Cold Spring Harb. Symp. Quant. Biol.*, Vol. 71, pp.117–128.

Eddy, S.R. and Durbin, R. (1994) 'RNA sequence analysis using covariance models', *Nucleic Acids Res.*, Vol. 22, pp.2079–2088.

Feng, J., Funk, W.D., Wang, S.S., Weinrich, S.L., Avilion, A.A., Chiu, C.P., Adams, R.R., Chang, E., Allsopp, R.C., Yu, J., Le, S., West, M.D., Harley, C.B., Andrews, W.H., Greider, C.W. and Villeponteau, B. (1995) 'The RNA component of human telomerase', *Science*, Vol. 269, pp.1236–1241.

Gattiker, A., Rischatsch, R., Demougin, P., Voegeli, S., Dietrich, F.S., Philippsen, P. and Primig, M. (2007) '*Ashbya* Genome Database 3.0: a cross-species genome and transcriptome browser for yeast biologists', *BMC Genomics*, Vol. 8, p.9.

Greider, C.W. and Blackburn, E.H. (1989) 'A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis', *Nature*, Vol. 337, pp.331–337. Gunisova, S., Elboher, E., Nosek, J., Gorkovoy, V., Brown, Y., Lucier, J.F., Laterreur, N., Wellinger, R.J., Tzfati, Y. and Tomaska, L. (2009) 'Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements', *RNA*, Vol. 15, pp.546–559.

Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C.L., Williams, J., Andrada, R., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Thanawala, M.K., Weng, S., Dolinski, K., Botstein, D. and Cherry, J.M. (2006) 'Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome', *Nucleic Acids Res.*, Vol. 34, pp.D442–D445.

Hofacker, I.L. (2003) 'Vienna RNA secondary structure server', *Nucleic Acids Res.*, Vol. 31, pp.3429–3431.

Hsu, M., McEachern, M.J., Dandjinou, A.T., Tzfati, Y., Orr, E., Blackburn, E.H. and Lue, N.F. (2007) 'Telomerase core components protect *Candida* telomeres from aberrant overhang

accumulation', *Proc. Natl. Acad. Sci., USA*, Vol. 104, pp.11682–11687.

Huang, Z., Wu, Y., Robertson, J., Feng, L., Malmberg, R.L. and Cai, L. (2008) 'Fast and accurate search for non-coding RNA pseudoknot structures in genomes', *Bioinformatics*, Vol. 24, pp.2281–2287.

Jeffries, T.W., Grigoriev, I.V., Grimwood, J., Laplaza, J.M., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., Shapiro, H., Jin, Y.S., Passoth, V. and Richardson, P.M. (2007) 'Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*', *Nat. Biotechnol.*, Vol. 25, pp.319–326.

Kachouri-Lafond, R., Dujon, B., Gilson, E., Westhof, E., Fairhead, C. and Teixeira, M.T. (2009) 'Large telomerase RNA, telomere length heterogeneity and escape from senescence in

Candida glabrata', *FEBS Lett.*, Vol. 583, pp.3605–3610.

Leonardi, J., Box, J.A., Bunch, J.T. and Baumann, P. (2008) '*TER1*, the RNA subunit of fission yeast telomerase', *Nat. Struct. Mol. Biol.*, Vol. 15, pp.26–33.

Lin, J., Ly, H., Hussain, A., Abraham, M., Pearl, S., Tzfati, Y., Parslow, T.G. and Blackburn, E.H. (2004) 'A universal telomerase RNA core structure includes

structured motifs required for binding the telomerase reverse transcriptase protein', *Proc. Natl. Acad. Sci*., *USA*, Vol. 101, pp.14713–14718.

Lingner, J., Hendrick, L.L. and Cech, T.R. (1994) 'Telomerase RNAs of different ciliates have a common secondary structure and a permuted template', *Genes Dev.*, Vol. 8, pp.1984–1998.

Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) 'Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure', *J. Mol. Biol.*, Vol. 288, pp.911–940.

McEachern, M.J. and Blackburn, E.H. (1995) 'Runaway telomere elongation caused by telomerase RNA mutations', *Nature*, Vol. 376, pp.403–409.

Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) 'Infernal 1.0: inference of RNA alignments', *Bioinformatics*, Vol. 25, pp.1335–1337.

Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X. and Chen, J.J. (2008) 'The telomerase database', *Nucleic Acids Res.*, Vol. 36, pp.D339–D343.

Romero, D.P. and Blackburn, E.H. (1991) 'A conserved secondary structure for telomerase RNA', *Cell*, Vol. 67, pp.343–353.

Rossignol, T., Lechat, P., Cuomo, C., Zeng, Q., Moszer, I. and d'Enfert, C. (2008) 'CandidaDB: a multi-genome database for *Candida* species and related *Saccharomycotina*', *Nucleic Acids Res.*, Vol. 36, pp.D557–D561.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (1994) 'Stoc Vol. hastic context-free grammars for tRNA modeling', *Nucleic Acids Res.*, Vol. 22, pp.5112–5120.

Seto, A.G., Livengood, A.J., Tzfati, Y., Blackburn, E.H. and Cech, T.R. (2002) 'A bulged stem tethers *Est1p* to telomerase RNA in budding yeast', *Genes Dev.*, Vol. 16, pp.2800–2812.

Shay, J.W. and Roninson, I.B. (2004) 'Hallmarks of senescence in carcinogenesis and cancer therapy', *Oncogene*, Vol. 23, pp.2919–2933.

Shay, J.W. and Wright, W.E. (2005) 'Senescence and immortalization: role of telomeres and telomerase', *Carcinogenesis*, Vol. 26, pp.867–874.

Shefer, K., Brown, Y., Gorkovoy, V., Nussbaum, T., Ulyanov, N.B. and Tzfati, Y. (2007) 'A triple helix within a pseudoknot is a conserved and essential element of telomerase RNA', *Mol. Cell. Biol.*, Vol. 27, pp.2130–2143.

Shippen-Lentz, D. and Blackburn, E.H. (1990) 'Functional evidence for an RNA template in telomerase', *Science*, Vol. 247, pp.546–552.

Singer, M.S. and Gottschling, D.E. (1994) '*TLC1*: template RNA component of *Saccharomyces cerevisiae* telomerase', *Science*, Vol. 266, pp.404–409.

Smogorzewska, A. and de Lange, T. (2004) 'Regulation of telomerase by telomeric proteins', *Annu. Rev. Biochem.*, Vol. 73, pp.177–208.

Theimer, C.A., Blois, C.A. and Feigon, J. (2005) 'Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function', *Mol. Cell.*, Vol. 17, pp.671–682.

Theimer, C.A., Finger, L.D., Trantirek, L. and Feigon, J. (2003) 'Mutations linked to dyskeratosis congenita cause changes in the structural equilibrium in telomerase RNA', *Proc. Natl. Acad. Sci*., *USA*, Vol. 100, pp.449–454.

Tzfati, Y., Fulton, T.B., Roy, J. and Blackburn, E.H. (2000) 'Template boundary in a yeast telomerase specified by RNA structure', *Science*, Vol. 288, pp.863–867.

Tzfati, Y., Knight, Z., Roy, J. and Blackburn, E.H. (2003) 'A novel pseudoknot element is essential for the action of a yeast telomerase', *Genes Dev.*, Vol. 17, pp.1779–1788.

Wang, Z.R., Guo, L., Chen, L. and McEachern, M.J. (2009) 'Evidence for an additional basepairing element between the telomeric repeat and the telomerase RNA template in *Kluyveromyces lactis* and other yeasts', *Mol Cell Biol*, Vol. 29, pp.5389–5398.

Webb, C.J. and Zakian, V.A. (2008) 'Identification and characterization of the *Schizosaccharomyces pombe TER1* telomerase RNA', *Nat. Struct. Mol. Biol.*, Vol. 15, pp.34–42.

Zuker, M. (2003) 'Mfold web server for nucleic acid folding and hybridization prediction', *Nucleic Acids Res.*, Vol. 31, pp.3406–3415.

Figure 2.1. Inferring telomerase RNA structure by TRFolder. (A) the conserved core secondary structure of telomerase RNAs. The names of the structural elements used in our paper are shown in italic. The numbering of stems and loops as used in this work are indicated as shown. Previously used names for these elements (Lin *et al*., 2004) are shown in parentheses and (B) flow chart of TRFolder

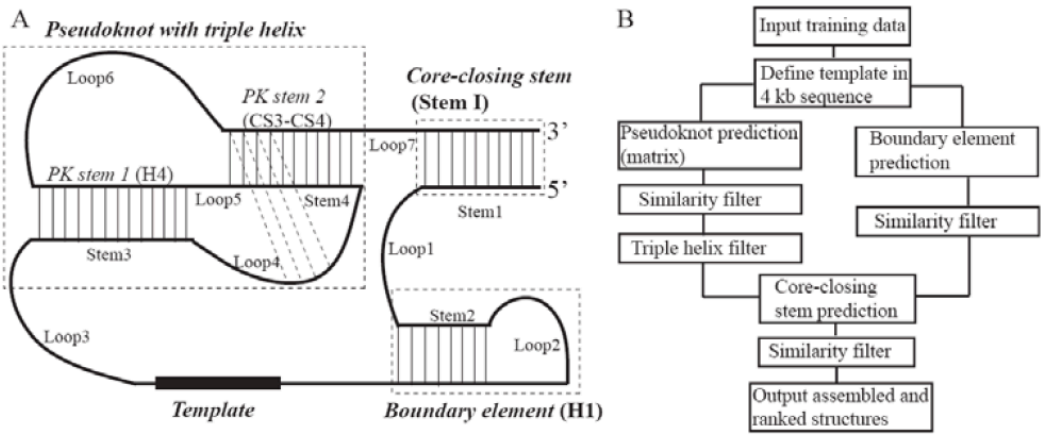
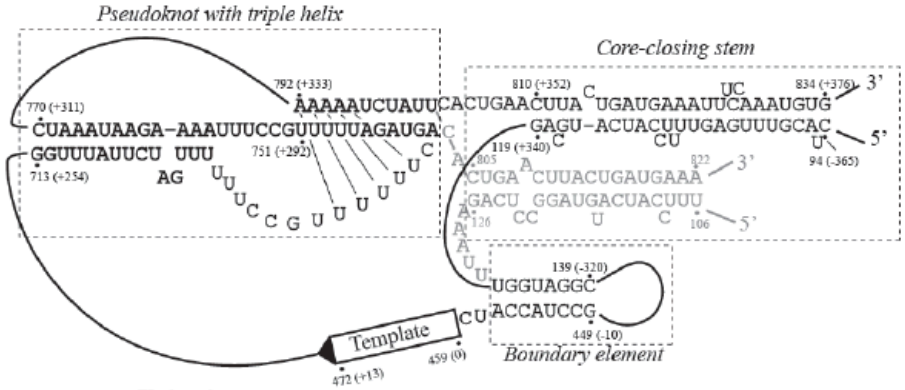

A
Pseudoknot with triple helix
Loop6
PK stem 2
(CS3-CS4)
Core-closing stem
(Stem I)
3'
5'
Loop7
PK stem 1 (H4)
Loop5
Stem4
Stem1
Stem3
Loop4
Loop1
Stem2
Loop2
Loop3
Template
Boundary element (H1)
B
Input training data
Define template in 4 kb sequence
Pseudoknot prediction (matrix)
Boundary element prediction
Similarity filter
Similarity filter
Triple helix filter
Core-closing stem prediction
Similarity filter
Output assembled and ranked structures

Figure 2.2. Differences between our predictions of the TR core structure of *S. cerevisiae* and *K. lactis*, and the predictions of previous studies. The previously predicted structures (Lin *et al*., 2004; Shefer *et al*., 2007; Tzfati *et al*., 2000, 2003) are shown in grey if different from our predictions. Two systems for numbering TRs are shown. Numbers not in parentheses use the 5' end of the telomerase RNA molecules as position 1 (Brown *et al*., 2007; Dandjinou *et al*., 2004; Lin *et al*., 2004; Shefer *et al*., 2007). The numbers in parentheses show the numbering system used in this work, which are counting from the first nucleotide at the 5' end of the template. The numbers at the ends of the template indicates the first and last nucleotide of the template region.

S. cerevisiae



K. lactis

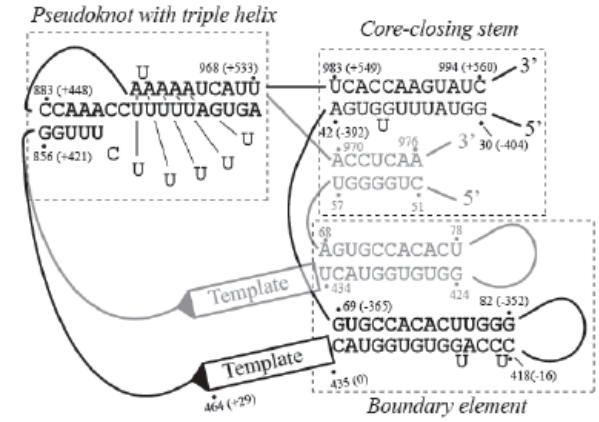
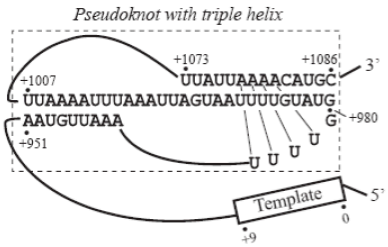
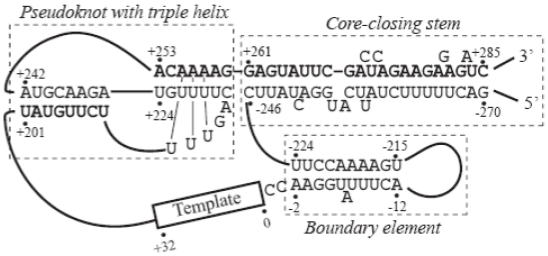


Figure 2.3. The core secondary structure of telomerase RNAs predicted by TRFolder in *S. pombe*, *P. stipitis*, *A. gossypii* and *D. hansenii*. The numbering system used count position 1 as the first nucleotide at the 5' end of the template. The numbers at both ends of the template indicate the first and last nucleotide of the template region. In *S. pombe*, the predicted boundary element and the 5'-arm of the core-closing stem are outside of the mapped TR region (Leonardi et al., 2008) so only the template region and the pseudoknot with triple helix are shown here. The complete prediction of the core secondary structure of *S. pombe* is listed in Table 2.S2.

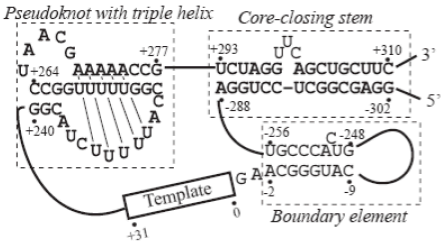
S. pombe*



P. stipitis



A. gossypii



D. hansenii

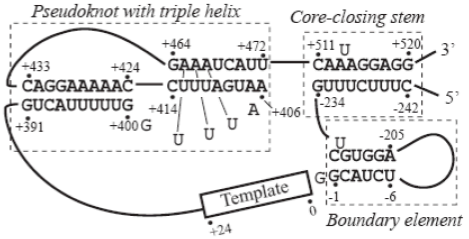


Table 2.1. Summary of core structural elements of TR genes from *Saccharomyces* and *Kluyveromyces* species used as training data in TRFolder. The positions of the stems and loops are shown in Figure 2.1.

Species	Stem and loop lengths (nt)												
	Stem1	Loop1	Stem2	Loop2	Stem3	Loop3	Stem4	Loop4	Loop5	Loop6	Loop7	Template to Stem 2	Triple helix
S. cerevisiae	28	5	8	322	13	256	11	13	6	21	2	1	/
S. cariocanus	28	5	8	278	13	257	10	13	7	22	3	2	/
S. paradoxus	29	5	8	277	13	257	10	13	7	21	2	2	/
S. mikatae	28	5	9	303	13	258	11	21	7	19	2	2	/
S. kudriavzevii	28	5	8	316	13	256	10	13	7	24	3	2	/
S. bayanus	27	5	8	331	17	313	11	9	3	22	2	2	/
K .lactis	7	10	11	345	5	421	9	6	3	76	1	0	5
K. nonfermentans	/	/	12	185	5	349	11	6	1	43	/	/	5
K. aestuarii	/	/	16	179	4	341	11	7	2	42	/	3	5
K. dobzhanskii	/	/	11	339	5	484	9	6	3	64	/	0	5
K. wickerhamii	/	/	11	223	5	374	9	6	3	45	/	2	5
K. marxianus	/	/	11	327	5	435	10	6	2	80	/	1	5
Parameters in TRFolder	8-40	3-48	6-11	160-410	1-21	0-2000	6-14	1-25	0-15	5-107	0-50	0-3	3-6

‘/’ indicates that the information is not currently available.

Table 2.2. Telomerase RNA template candidates in six yeast species. Sequences shown were identified by BLAST searches using tandem telomeric repeats and/or known neighboring genes of TR genes as queries as described in text. Each template candidate has > 1 full repeat of perfect homology to telomeric sequences from the species in question. The underlined parts of the predicted template sequence indicate the direct repeats on the 5' and 3' sides of the template.

Species	Telomeric Repeat Size (bp)	Template Length (bp)	Template Sequence
C. glabrata	16	19	3' ACCCAUGACACCCCAGACCCACGAC 5'
C. tropicalis	23	28	3' UAAUCACAUUCCUACAGUGCUAGUAACCACAU 5'
A. gossypii	24	32	3' AGUCGCCACACCACAUACCCAGAGAGUCGCCA 5'
P. stipitis	24	31	3' UGCCUAGAAAAGUGCAGAACGCCAUACCUAGAA 5'
D. hansenii	16	25	3' CCUACAACUCCACAUCCCUACAACU 5'
C. guilliermondii	8	16	3' AGCACAUGACCACAUGAC 5'

Table 2.3. Differences between the highest-scoring predicted structures and previous predictions for TRs of five *Kluyveromyces* and *Saccharomyces* species. Stem lengths include non-paired nucleotides, so we list the lengths of both arms.

Species		*K. lactis*	*K. dobzhanskii*	*K. nonfermentans*	*K. wickerhamii*	*K. aestuarii*
Pseudoknot with Triple Helix	Distance to template (nt)	391	451	331	323	309
	PK Stem 1 length (nt)	5/5	8/8	9/9	10/10	4/4
	PK Stem 2 length (nt)	9/9	7/7	15/14	10/9	9/9
	Triple Helix length (nt)	4/4	4/4	3/3	3/3	5/5
	Comparison	Stem 2 missing 1 bp after A bulge	Stem 1 has extra 3 bp; stem 2 missing 1 bp after U-U gap	Different pseudoknot	Different stem 1; stem 2 missing 1 bp after U bulge	Stem 2 missing 1 bp after U-U gap
Boundary Element	Distance to template (nt)	0	0	1	0	3
	Stem length (nt)	14/14	14/16	12/12	12/11	11/10
	Loop length (nt)	335	330	185	318	189
	Comparison	Extra 4 bps after U bulge	Extra 4 bps with U bulge	N/A	Different stem-loop	Missing 4 bps after U bulge
Core-closing Stem	Stem length (nt)	13/12	39/40	22/19	17/12	24/24
	Distance to Pseudoknot (nt)	15	2	6	4	6
	Distance to Boundary Element (nt)	26	9	25	8	17
	Comparison	/	/	/	/	/
Total size of TR core structure (nt)		964	1031	761	816	697

Table 2.3. Differences between the highest-scoring predicted structures and previous predictions for TRs of five *Kluyveromyces* and *Saccharomyces* species. Stem lengths include non-paired nucleotides, so we list the lengths of both arms (continued).

Species		*S. cerevisiae*	*S. kudriavzevii*	*S. cariocanus*	*S. paradoxus*	*S. mikatae*
Pseudoknot with Triple Helix	Distance to template (nt)	240	239	241	238	242
	PK Stem 1 length (nt)	15/13	9/9	9/9	17/18	10/10
	PK Stem 2 length (nt)	11/11	12/12	10/10	10/10	17/14
	Triple Helix length (nt)	5/5	5/5	4/4	4/4	5/5
	Comparison	Stem 1 has extra 3 A-U pair	Stem 1 missing 3 bp after A-C gap	Stem 1 missing 2 bp with AA-AG gap	Stem 2 missing 1 bp after A bulge	Stem2 has extra A-U
Boundary Element	Distance to template (nt)	1	1	1	1	1
	Stem length (nt)	8/8	8/8	10/12	10/12	9/9
	Loop length (nt)	309	316	272	271	303
	Comparison	Exactly the same	Missing U-G pair	Same position; pairing slightly different	Same position; pairing slightly different	Exactly the same
Core-closing Stem	Stem length (nt)	26/25	43/46	26/30	42/43	41/39
	Distance to Pseudoknot (nt)	8	3	10	3	4
	Distance to Boundary Element (nt)	12	5	15	5	5
	Comparison	Same position; pairing slightly different	Same position; pairing different; ours is longer	Similar position; pairing different; ours has less bulges	Same position and pairing; ours is 8 bp longer	Same position and pairing; ours is 8 bp longer
Total size of TR core structure (nt)		740	776	720	729	761

Table 2.4. Summary of the predictions of the core structural elements of TR genes from eight yeast species. Stem lengths include non-paired nucleotides, so the lengths of both arms are listed for each stem, as they are sometimes different. The eight yeast species include *C. albicans* (*C. alb*), *C. tropicalis* (*C. tro*), *A. gossypii* (*A. gos*), *C. glabrata* (*C. gla*), *C. guilliermondii* (*C. gui*), *D. hansenii* (*D. han*) and *S. pombe*. The last two columns show the ranges in size of each structural element predicted in this study in the TRs from the *Saccharomyces* and *Kluyveromyces* species. The predicted structures are listed in Table 2.S1 and 2.S2.

Species			*C. alb*	*C. tro*	*A. gos*	*C. gla*	*C. gui*	*P. sti*	*D. han*	*S. pom*	*Saccharomyces*	*Kluyveromyces*
Pseudo-knot with triple helix	Distance to template (nt)		326		208	1348	686	168	366	941	238-242	310-452
	PK stem1 length (nt)	5’	3	4	3	18	20	8	10	9	9-18	4-10
		3’	3	4	3	18	20	8	10	9		
	PK stem2 length (nt)	5’	14	10	8	6	11	7	9	14	10-17	7-9
		3’	14	10	8	7	10	7	9	14		
	Triple helix (nt)		3	5	5	3	3	3	3	4	4-5	4-5
Boundary element	Distance to template (nt)		0	3	1	1	2	1	0	3	1	1-3
	Stem length (nt)	5’	9	7	9	9	11	10	7	13	8-12	10-15
		3’	9	8	8	8	8	11	6	12		
	Loop length (nt)		199	187	238	332	340	202	198	300	271-316	186-337
Core-closing stem	Stem length (nt)	5’	8	24	15	17	18	25	9	20	25-43	11-37
		3’	7	25	18	14	19	25	10	23		
	Distance to pseudoknot (nt)		0	17	15	31	8	1	38	17	3-10	3-38
	Distance to boundary element (nt)		14	13	31	26	29	21	22	6	5-15	3-27
Total size of TR core structure (nt)			672	718	612	1941	1277	555	762	1480	720-776	697-1031

Table 2.5. Summary of negative control tests. The highest Z-scores of the predicted secondary structure in known/putative TRs, as well as the average of top Z-scores and standard deviation in the tested random sequences are listed in the table. The average scores and standard deviation are calculated based on non-zero Z-scores of random sequences. The genome sequences in *Saccharomyces* and *Kluyveromyces* species except *S. cerevisiae* and *K. lactis* are not yet available, so no results are listed in the random position testing for these species.

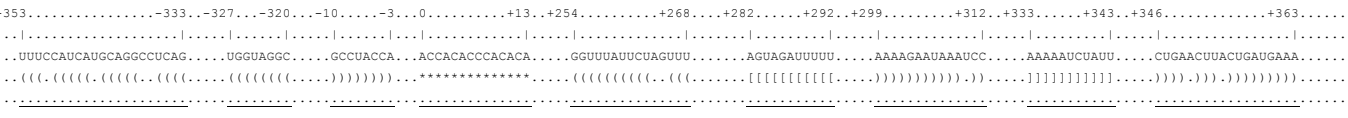
Species	Known/Putative TR	Shuffled Sequence				Random Position			
		MAX	MIN	AVG	SD	MAX	MIN	AVG	SD
K. lactis	43.73	38.15	29.1	33.20	3.51	40.77	33.34	36.38	3.89
K. aestuarii	46.16	37.3	31.25	33.67	2.38	/	/	/	/
K. dobzhanskii	50.63	37.4	28.61	33.68	3.33	/	/	/	/
K .nonfermentans	50.37	38.92	30.51	34.90	3.45	/	/	/	/
K. wickerhamii	36.68	37.22	32.76	33.91	2.87	/	/	/	/
S. cerevisiae	55.15	40.56	32.46	35.22	3.26	34.78	30.44	32.50	1.85
S. cariocanus	47.09	36.4	32.4	34.26	1.46	/	/	/	/
S. kudriavzevii	55.53	37.02	33.21	34.75	1.89	/	/	/	/
S. mikatae	54.68	39.06	33.74	36.13	2.44	/	/	/	/
S. paradoxus	53.15	37.95	29.83	34.41	3.36	/	/	/	/
A. gossypii	37.05	35.06	30.26	32.64	2.01	38.33	32.22	34.61	3.27
D. hansenii	35.97	37.66	29.96	33.94	2.84	41.26	30.72	35.81	3.56
C. albicans	40.21	0	0	0	0	37.33	30.7	34.36	3.12
C. glabrata	38.19	35.88	30.71	33.58	2.08	37.26	29.42	34.82	3.65
C. guilliermondii	37.82	37.39	31.76	34.73	2.07	32.1	27.91	30.48	2.25
C. tropicalis	40.79	33.07	30.86	32.05	0.95	37.32	30.22	33.68	2.72
P. stipitis	39.57	36.48	30.43	34.33	2.76	41.26	30.72	34.14	4.19
S. pombe	40.79	38.54	33.57	35.52	2.24	38.97	30.59	34.51	3.20

Table 2.S1. Predicted core structures for the three highest scoring TR candidates of five *Saccharomyces* and five *Kluyveromyces* species. The data are based on the outputs from program TRFolder, formatted and scaled to fit into the text width. The candidates are ranked according to their total scores. The users can set different weights to each structural element based on their interest. Default weight was set to be 1 in our test. There are five lines in the structure annotation. The first two lines make up the ruler; the third is the sequence; line four is the predicted secondary structure including the pseudoknot; and the last line indicates a predicted stem that forms a triple helix with a base-pairing region in the predicted pseudoknot indicated in line four. Asterisks underscore the predicted template. The previously predicted structures were also listed in each species for comparison.

<u>The predicted core structures of top 3 TR candidates of *S. cerevisiae*</u>:

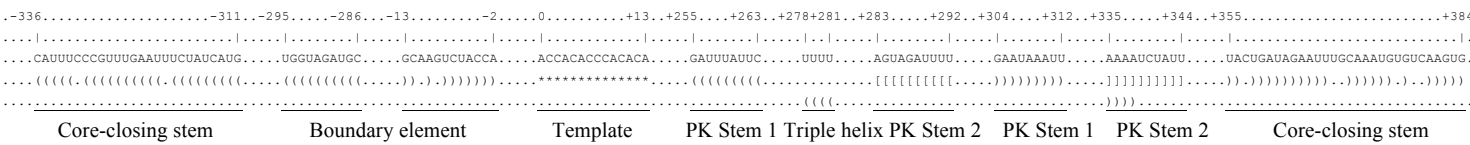
```
Total:55.15 =pk Score:1*24.39 +Triple Score:1.0*11.29 +boundaryEle Score:1*12.26 +Core Closing Score:1.0*7.21
..-365.....................-340..-327...-320...-10.....-3.....0..........+13..+254.........+268..+276+280..+282......+292..+299........+311..+333......+343..+352.....................+376
.....|.........................|.....|......|.....|......|.....|.............|.....|.............|.....|...|.....|.........|.....|............|.....|.........|.....|.......................|.....
.....CUACGUUUGAGUUUUCCAUCAUGCAG.....UGGUAGGC.....GCCUACCA.....ACCACACCCACACA.....GGUUUAUUCUAGUUU.....UUUUU.....AGUAGAUUUUU.....AAAAGAAUAAAAUC.....AAAAAUCUAUU.....CUUACUGAUGAAAAUUUCCAAAUGUG.....
.....(.(((((((((((.(.(((((((.((.....(((((((((.....)))))))))).....**************.....((((((((((..(((.................[[[[[[[[[[.....))))))))))))))).....]]]]]]]]]].....)))).)))))))))).).)))))))
.....______________________________.....________.....________.....________________.....________________.....(((((.....____________.....____________.....)))))......____.....____________________________.....
      Core-closing stem          Boundary element       Template          PK Stem 1    Triple helix    PK Stem 2     PK Stem 1      PK Stem 2       Core-closing stem

Total:52.65 =pk Score:1*24.39 +Triple Score:1.0*11.29 +boundaryEle Score:1*12.26 +Core Closing Score:1.0*4.71
..-362....-354..-327...-320...-10.....-3.....0..........+13..+254.........+268..+276+280..+282......+292..+299........+311..+333......+343..+383.....+392
.....|.......|.....|......|.....|......|.....|.............|.....|.............|.....|...|.....|.........|.....|............|.....|.........|.....|.........|.....
.....CGUUUGAGU.....UGGUAGGC.....GCCUACCA.....ACCACACCCACACA.....GGUUUAUUCUAGUUU.....UUUUU.....AGUAGAUUUUU.....AAAAGAAUAAAAUC.....AAAAAUCUAUU.....ACAUCGAACG.....
.....(((((((((.....(((((((((.....)))))))))).....**************.....((((((((((..(((.................[[[[[[[[[[.....))))))))))))))).....]]]]]]]]]].....)).)))))))
.....................................................................................(((((.............................................)))))..............

Total:50.74 =pk Score:1*24.39 +Triple Score:1.0*11.29 +boundaryEle Score:1*12.26 +Core Closing Score:1.0*2.8
..-375...-368..-327...-320...-10.....-3.....0..........+13..+254.........+268..+276+280..+282......+292..+299........+311..+333......+343..+357..+363
.....|......|.....|......|.....|......|.....|.............|.....|.............|.....|...|.....|.........|.....|............|.....|.........|.....|.....|.....
.....UUUGUACA.....UGGUAGGC.....GCCUACCA.....ACCACACCCACACA.....GGUUUAUUCUAGUUU.....UUUUU.....AGUAGAUUUUU.....AAAAGAAUAAAAUC.....AAAAAUCUAUU.....UGAUGAA.....
.....(((((.((.....(((((((((.....)))))))))).....**************.....((((((((((..(((.................[[[[[[[[[[.....))))))))))))))).....]]]]]]]]]].....)))))))
.....................................................................................(((((.............................................)))))..............
```

Comparison: previously published structure

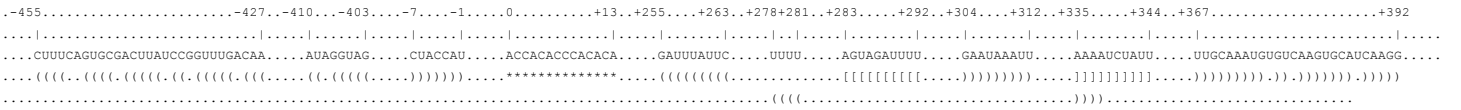


The predicted core structures of top 3 TR candidates of *S. cariocanus*:

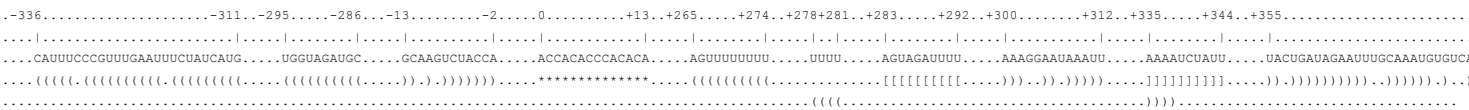
Total:47.09 =pk Score:1*20.11 +Triple Score:1.0*9.03 +boundaryEle Score:1*9.92 +Core Closing Score:1.0*8.03



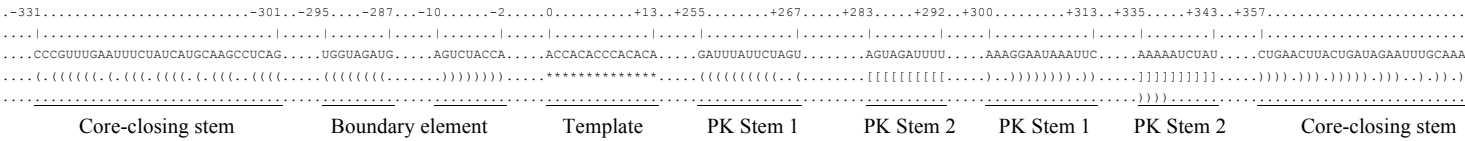
Total:44.57 =pk Score:1*20.11 +Triple Score:1.0*9.03 +boundaryEle Score:1*7.83 +Core Closing Score:1.0*7.6



Total:44.21 =pk Score:1*17.23 +Triple Score:1.0*9.03 +boundaryEle Score:1*9.92 +Core Closing Score:1.0*8.03

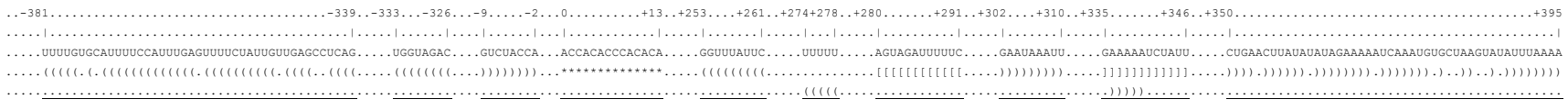


Comparison: previously published structure



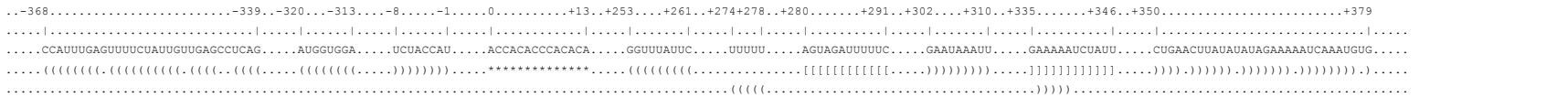
The predicted core structures of top 3 TR candidates of *S. kudriavzevii*:

```
Total:55.53 =pk Score:1*22.24 +Triple Score:1.0*11.29 +boundaryEle Score:1*10.95 +Core Closing Score:1.0*11.05
```

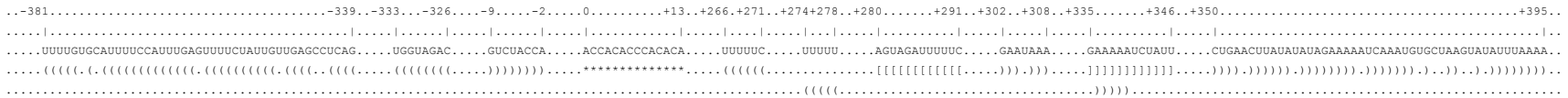


Core-closing stem | Boundary element | Template | PK Stem 1 | Triple helix | PK Stem 2 | PK Stem 1 | PK Stem 2 | Core-closing stem

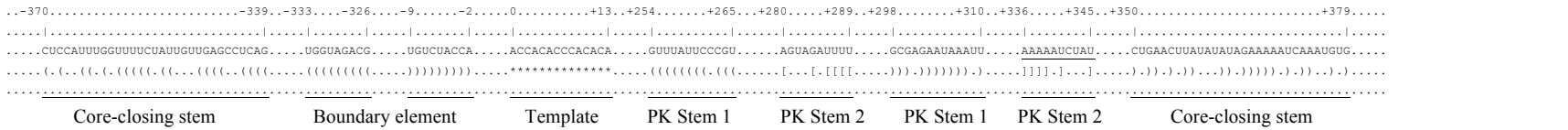
```
Total:53.17 =pk Score:1*22.24 +Triple Score:1.0*11.29 +boundaryEle Score:1*9.64 +Core Closing Score:1.0*10.0
```



```
Total:52.56 =pk Score:1*19.27 +Triple Score:1.0*11.29 +boundaryEle Score:1*10.95 +Core Closing Score:1.0*11.05
```



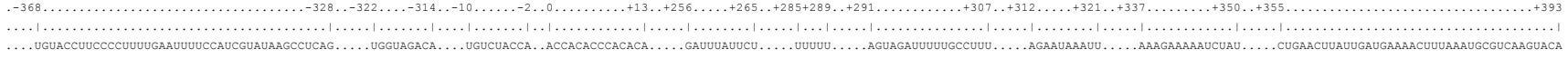
```
Comparison: previously published structure
```

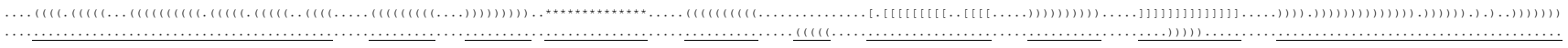


Core-closing stem | Boundary element | Template | PK Stem 1 | PK Stem 2 | PK Stem 1 | PK Stem 2 | Core-closing stem

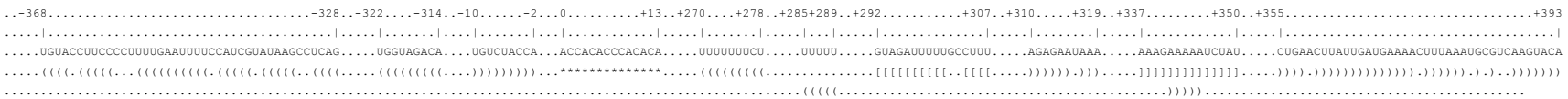
The predicted core structures of top 3 TR candidates of *S. mikatae*:

```
Total:54.68 =pk Score:1*22.27 +Triple Score:1.0*11.29 +boundaryEle Score:1*11.82 +Core Closing Score:1.0*9.3
```

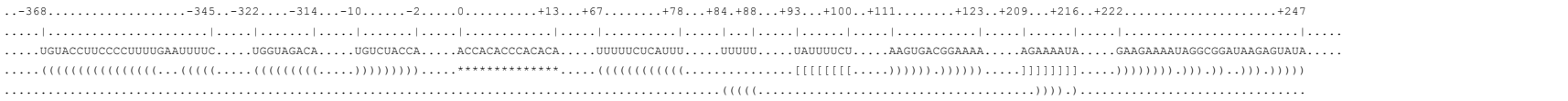




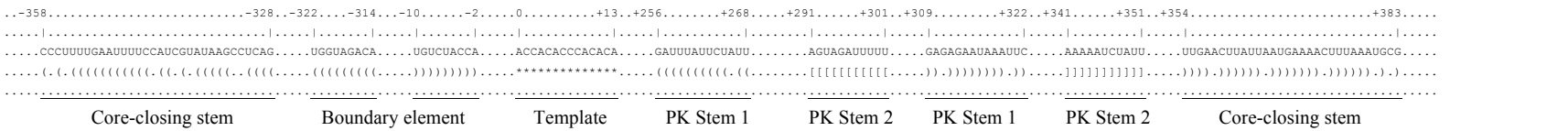
Total:53.92 =pk Score:1*21.51 +Triple Score:1.0*11.29 +boundaryEle Score:1*11.82 +Core Closing Score:1.0*9.3



Total:49.83 =pk Score:1*19.63 +Triple Score:1.0*9.55 +boundaryEle Score:1*11.82 +Core Closing Score:1.0*8.83

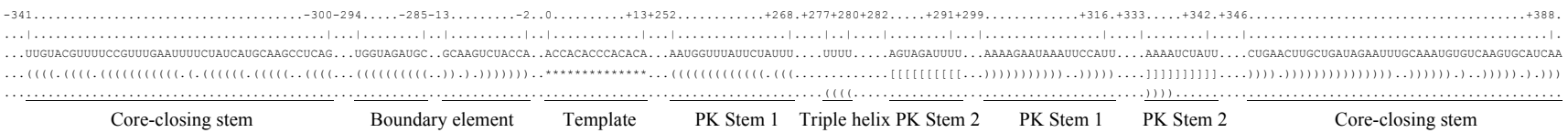


Comparison: previously published structure

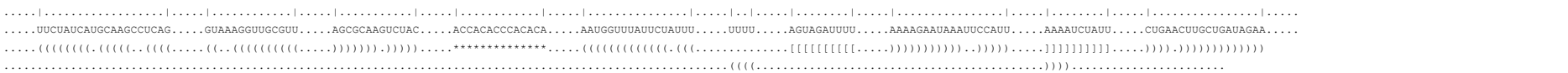


The predicted core structures of top 3 TR candidates of *S. paradoxus*:

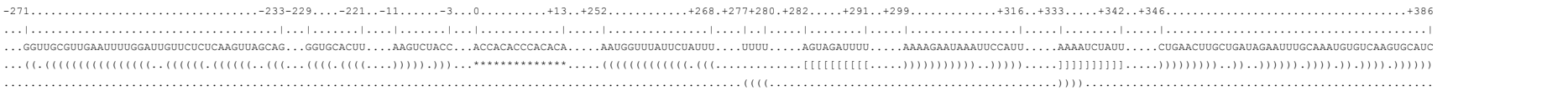
Total:53.15 =pk Score:1*24.42 +Triple Score:1.0*9.03 +boundaryEle Score:1*9.92 +Core Closing Score:1.0*9.78



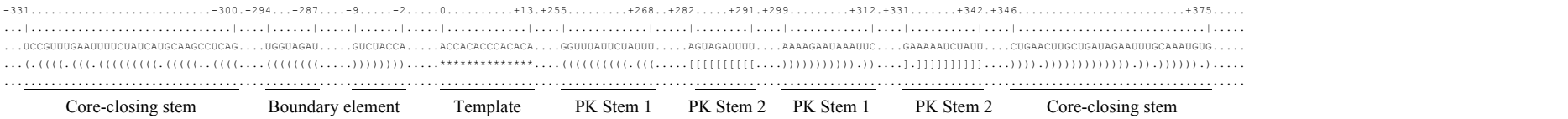
Total:48.47 =pk Score:1*24.42 +Triple Score:1.0*9.03 +boundaryEle Score:1*7.45 +Core Closing Score:1.0*7.57



Total:48.28 =pk Score:1*24.42 +Triple Score:1.0*9.03 +boundaryEle Score:1*6.88 +Core Closing Score:1.0*7.95

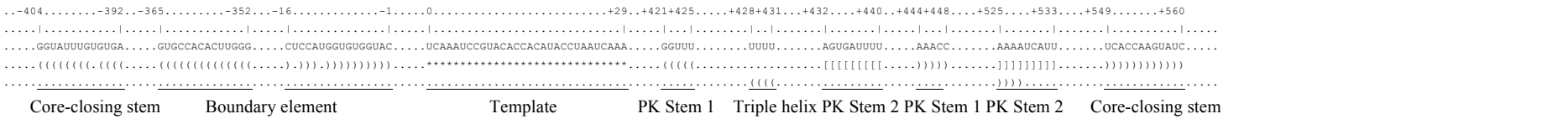


Comparison: previously published structure

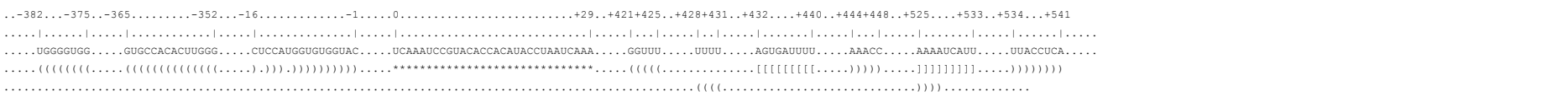


The predicted core structures of top 3 TR candidates of *K. lactis*:

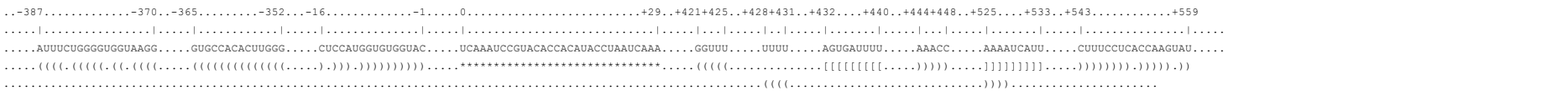
Total:43.73 =pk Score:1*17.14 +Triple Score:1.0*9.12 +boundaryEle Score:1*11.87 +Core Closing Score:1.0*5.6



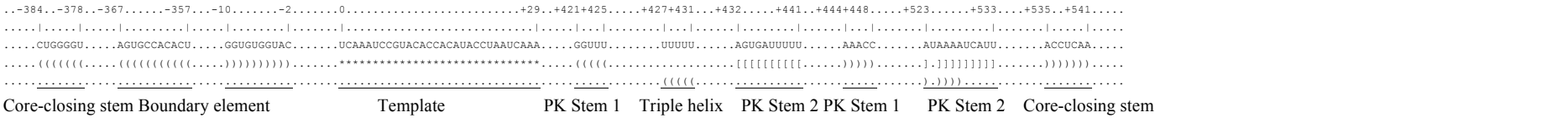
Total:42.03 =pk Score:1*17.14 +Triple Score:1.0*9.12 +boundaryEle Score:1*11.87 +Core Closing Score:1.0*3.9



Total:41.43 =pk Score:1*17.14 +Triple Score:1.0*9.12 +boundaryEle Score:1*11.87 +Core Closing Score:1.0*3.3



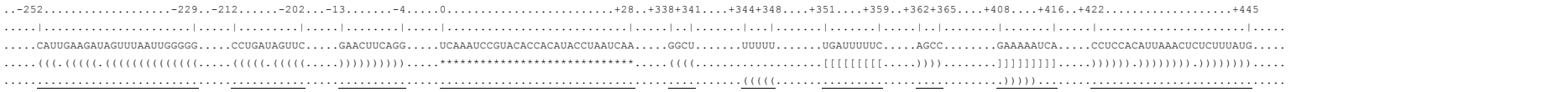
Comparison: previously published structure



Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

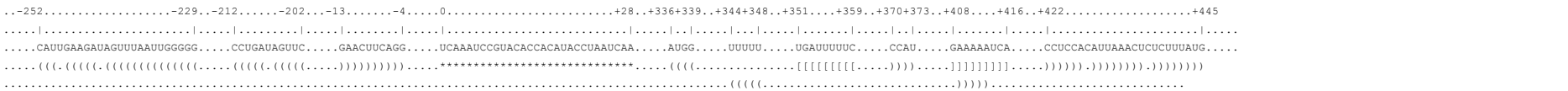
The predicted core structures of top 3 TR candidates of *K. aestuarii*:

Total:46.16 =pk Score:1*16.62 +Triple Score:1.0*11.22 +boundaryEle Score:1*9.89 +Core Closing Score:1.0*8.43

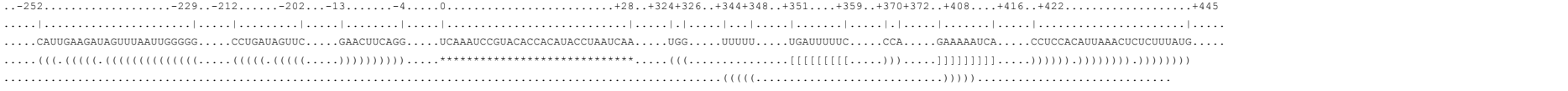


Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

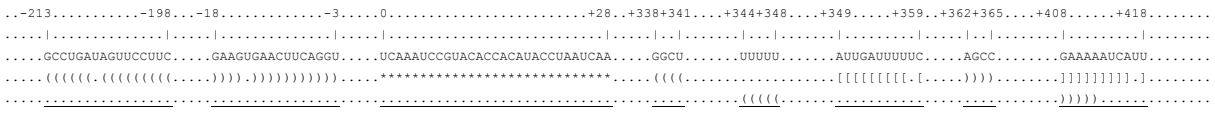
Total:45.73 =pk Score:1*16.19 +Triple Score:1.0*11.22 +boundaryEle Score:1*9.89 +Core Closing Score:1.0*8.43



Total:45.24 =pk Score:1*15.7 +Triple Score:1.0*11.22 +boundaryEle Score:1*9.89 +Core Closing Score:1.0*8.43



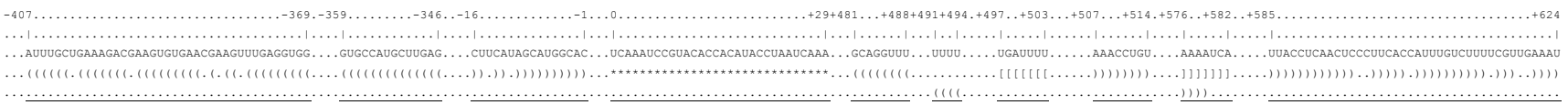
Comparison: previously published structure



Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2

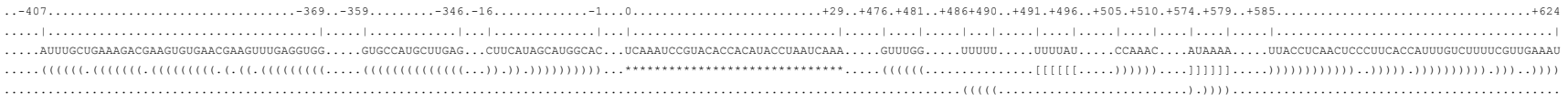
The predicted core structures of top 3 TR candidates of *K. dobzhanskii*:

Total:50.63 =pk Score:1*18.0 +Triple Score:1.0*9.12 +boundaryEle Score:1*14.97 +Core Closing Score:1.0*8.54

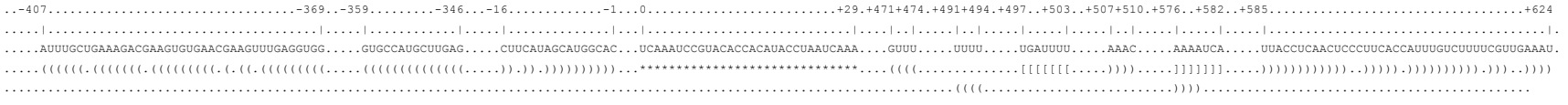


Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

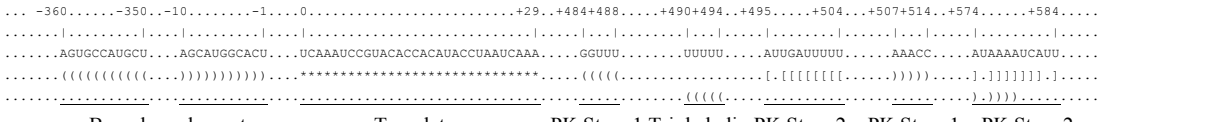
Total:49.6 =pk Score:1*16.41 +Triple Score:1.0*9.68 +boundaryEle Score:1*14.97 +Core Closing Score:1.0*8.54



Total:47.8 =pk Score:1*15.17 +Triple Score:1.0*9.12 +boundaryEle Score:1*14.97 +Core Closing Score:1.0*8.54



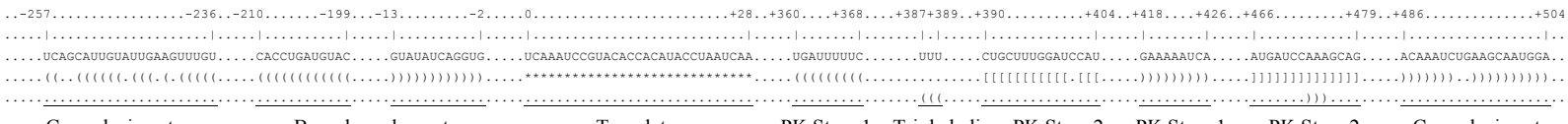
Comparison: previously published structure



Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2

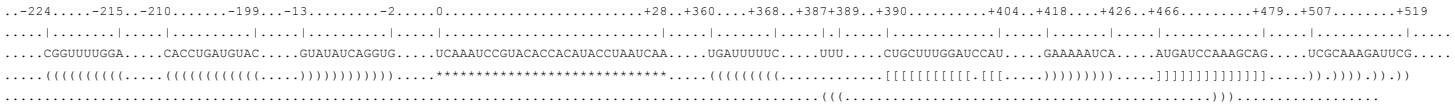
The predicted core structures of 3 top TR candidates of *K. nonfermentans*:

```
Total:50.37 =pk Score:1*26.37 +Triple Score:1.0*6.83 +boundaryEle Score:1*11.34 +Core Closing Score:1.0*5.83
```

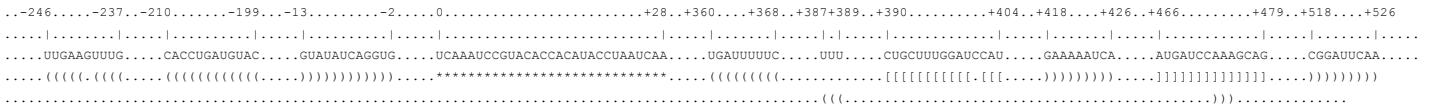


Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

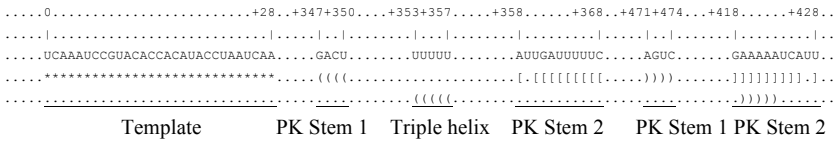
```
Total:48.44 =pk Score:1*26.37 +Triple Score:1.0*6.83 +boundaryEle Score:1*11.34 +Core Closing Score:1.0*3.9
```



```
Total:47.97 =pk Score:1*26.37 +Triple Score:1.0*6.83 +boundaryEle Score:1*11.34 +Core Closing Score:1.0*3.43
```



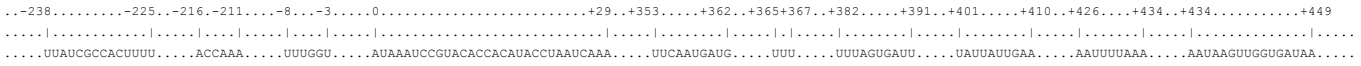
```
Comparison: previously published structure
```

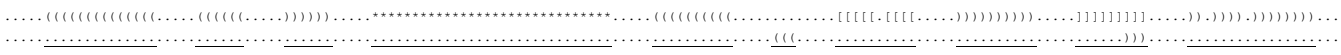


Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2

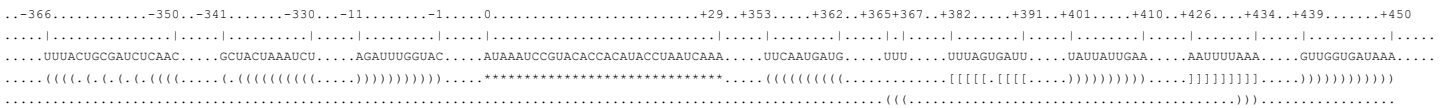
The predicted core structures of 3 top TR candidates of *K. wickerhamii*:

```
Total:36.68 =pk Score:1*16.02 +Triple Score:1.0*6.84 +boundaryEle Score:1*5.72 +Core Closing Score:1.0*8.1
```

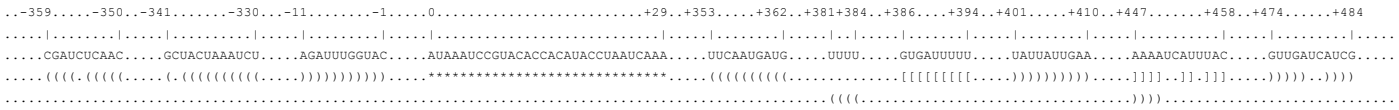




Total:34.7 =pk Score:1*16.02 +Triple Score:1.0*6.84 +boundaryEle Score:1*8.09 +Core Closing Score:1.0*3.75



Total:34.52 =pk Score:1*14.04 +Triple Score:1.0*9.12 +boundaryEle Score:1*8.09 +Core Closing Score:1.0*3.27



Comparison: previously published structure

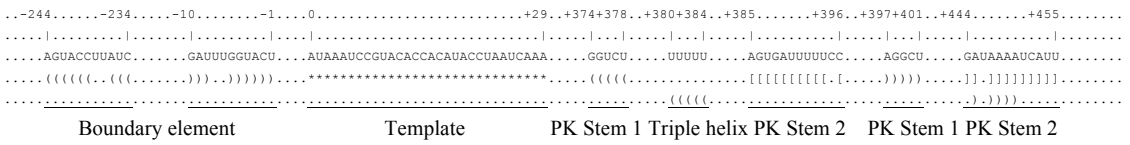


Table 2.S2. Predicted core structures for the three top scoring TR candidates of six yeast species: *A. gossypii, C. albicans, C. glabrata, C. guilliermondii, C. tropicalis, D. hansenii, P. stipites* and *S. pombe*, based on outputs from program TRFolder, formatted and scaled to fit into the text width. The candidates are ranked according to their scores. There are five lines in the structure annotation. The first two lines are the ruler; the third is the sequence; line four is the predicted secondary structure including the pseudoknot; and the last line indicates a predicted stem that forms a triple helix with a base-pairing region in the predicted pseudoknot indicated in line four. Asterisks underscore the predicted template. In *S. pombe*, pounds in the third line indicates the mapped 5' and 3' end of the telomerase RNA, which are at position -226 and +1185. Therefore, our predicted boundary element and the 5'-arm of the core-closing stem are outside of the mapped region (Leonardi, *et al.*, 2008).

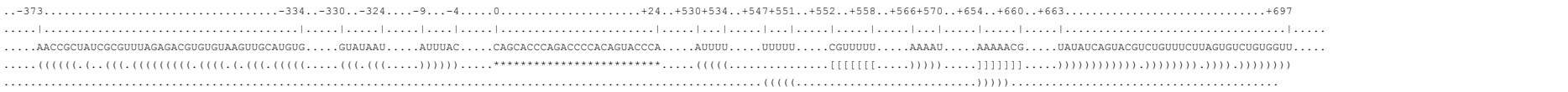
<u>The predicted core structures of top 3 TR candidates for *C. glabrata*:</u>

```
Total:38.19 =pk Score:1*18.21 +Triple Score:1*6.79 +boundaryEle Score:1*6.26 +Core Closing Score:1*6.93
..-393...........-377..-350....-342....-9.....-2.....0....................+24.+1373...........+1390.+1393+1395....+1398+1403.+1405...........+1422.+1497.+1503...+1535........+1548
.....|...............|.....|.......|.....|......|.....|.......................|.....|................|.....|.|.......|....|.....|................|.....|.....|.......|............|.....
.....UUUAAGUAGGGUUUUGU.....UGUGUAAGU.....AUUUACCA.....CAGCACCCAGACCCCACAGUACCCA.....CAGGAGGAUAGAAUAUUU.....UUU.......UUUUUG.....AAAUGUUAUAUCCUUCUG.....CAAGAAA.......ACAAAAUCUACAAA.....
.....(((..(((((.((((((.....((.((((((.....)))))))).....*************************.....((((((((((((((((((...............[[[[[[.....)))))))))))))))))).....]]].]]].......)))))))))))))).....
.........................................................................................................(((.........................................)))...........................
```

Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

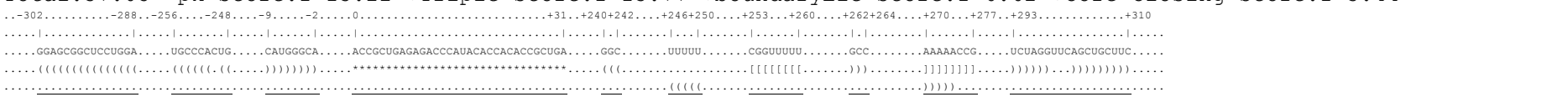
```
Total:37.93 =pk Score:1*17.95 +Triple Score:1*6.79 +boundaryEle Score:1*6.26 +Core Closing Score:1*6.93
..-393...........-377..-350....-342....-9.....-2.....0....................+24.+1430.+1436.+1440+1442.+1445..+1452.+1454.+1460.+1501..+1508.+1535........+1548
.....|...............|.....|.......|.....|......|.....|.......................|.....|.....|.....|.|.....|......|.....|.....|.....|......|.....|............|.....
.....UUUAAGUAGGGUUUUGU.....UGUGUAAGU.....AUUUACCA.....CAGCACCCAGACCCCACAGUACCCA.....AUCGUAU.....UUU.....UGGGUUUU.....AUACGAU.....AAAACUCA.....ACAAAAUCUACAAA.....
.....(((..(((((.((((((.....((.((((((.....)))))))).....*************************.....(((((((.............[[[[[[[[.....))))))).....]]]]]]]].....))))))))))))))
.....................................................................................................(((...............................)))...................
```

Total:37.7 =pk Score:1*16.19 +Triple Score:1*11.32 +boundaryEle Score:1*4.37 +Core Closing Score:1*5.82



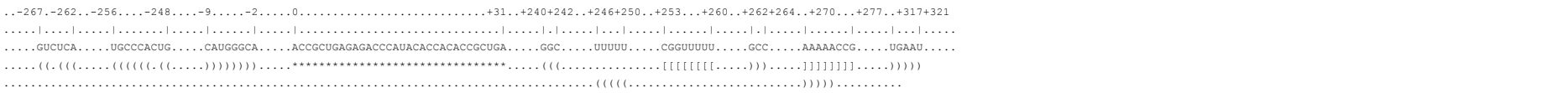
The predicted structures of top 3 candidates for *A. gossypii*:

Total:37.05 =pk Score:1*13.22 +Triple Score:1*13.77 +boundaryEle Score:1*6.62 +Core Closing Score:1*3.44

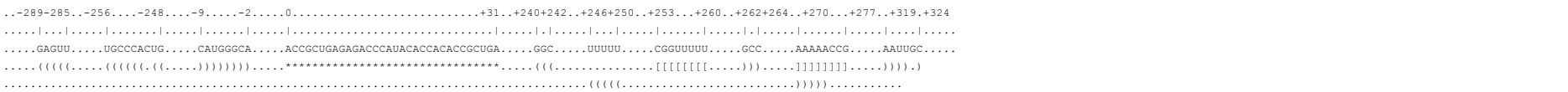


Core-closing stem Boundary element Template PK Stem 1 Triple helix PK Stem 2 PK Stem 1 PK Stem 2 Core-closing stem

Total:36.47 =pk Score:1*13.22 +Triple Score:1*13.77 +boundaryEle Score:1*6.62 +Core Closing Score:1*2.86

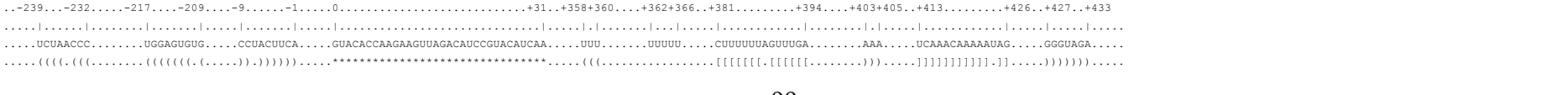


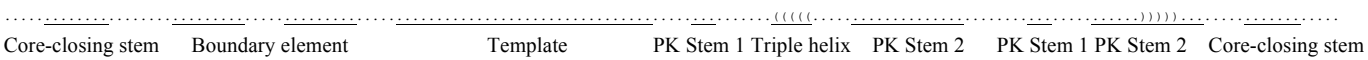
Total:36.24 =pk Score:1*13.22 +Triple Score:1*13.77 +boundaryEle Score:1*6.62 +Core Closing Score:1*2.63



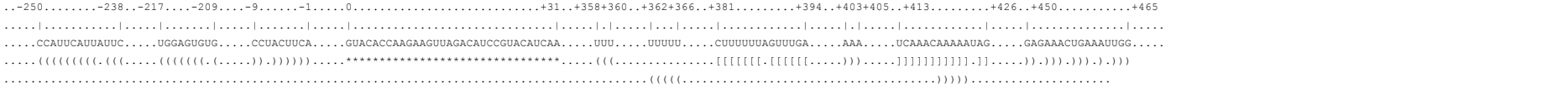
The predicted core structures of top 3 TR candidates of *C. albicans*:

Total:40.21 =pk Score:1*16.77 +Triple Score:1*10.53 +boundaryEle Score:1*7.63 +Core Closing Score:1*5.28

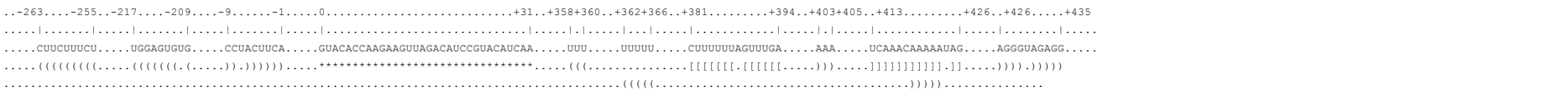




Total:39.55 =pk Score:1*16.77 +Triple Score:1*10.53 +boundaryEle Score:1*7.63 +Core Closing Score:1*4.62

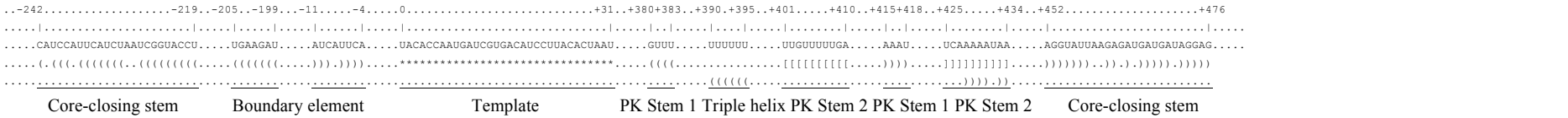


Total:39.54 =pk Score:1*16.77 +Triple Score:1*10.53 +boundaryEle Score:1*7.63 +Core Closing Score:1*4.61

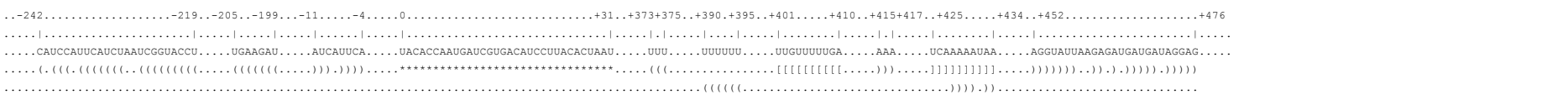


The predicted core structures of top 3 TR candidates of *C. tropicalis*:

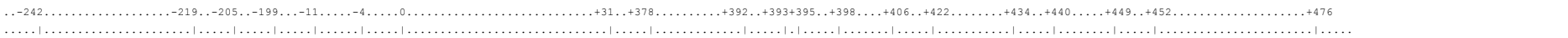
Total:40.79 =pk Score:1*15.86 +Triple Score:1*10.76 +boundaryEle Score:1*6.3 +Core Closing Score:1*7.87

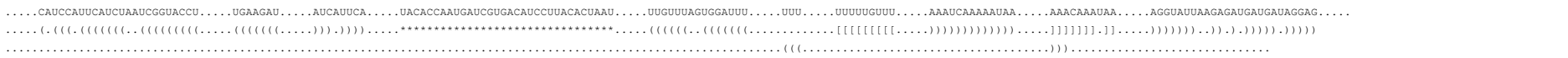


Total:40.5 =pk Score:1*15.57 +Triple Score:1*10.76 +boundaryEle Score:1*6.3 +Core Closing Score:1*7.87



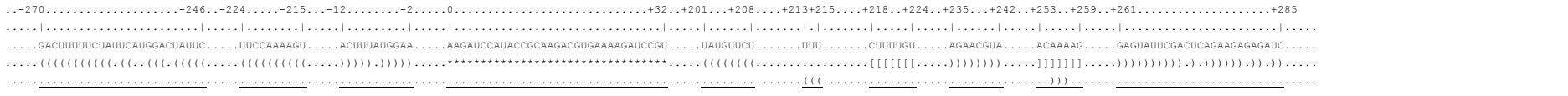
Total:38.76 =pk Score:1*18.3 +Triple Score:1*6.29 +boundaryEle Score:1*6.3 +Core Closing Score:1*7.87





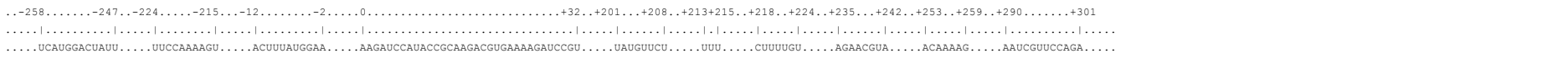
The predicted core structures of top 3 TR candidates of *P. stipitis*:

Total:39.57 =pk Score:1*17.9 +Triple Score:1*7.06 +boundaryEle Score:1*8.49 +Core Closing Score:1*6.12

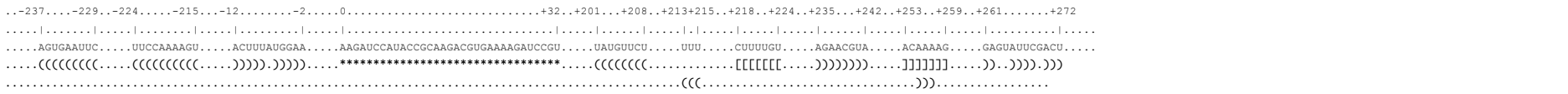


Core-closing stem | Boundary element | Template | PK Stem 1 | Triple helix | PK Stem 2 | PK Stem 1 | PK Stem 2 | Core-closing stem

Total:36.99 =pk Score:1*17.9 +Triple Score:1*7.06 +boundaryEle Score:1*8.49 +Core Closing Score:1*3.54

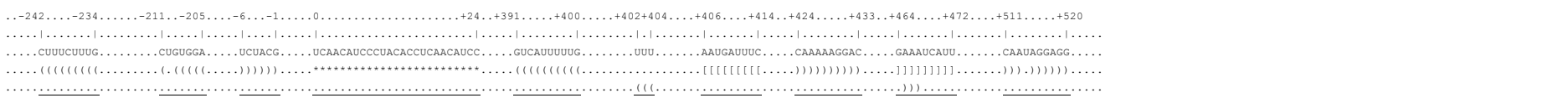


Total:36.5 =pk Score:1*17.9 +Triple Score:1*7.06 +boundaryEle Score:1*8.49 +Core Closing Score:1*3.05



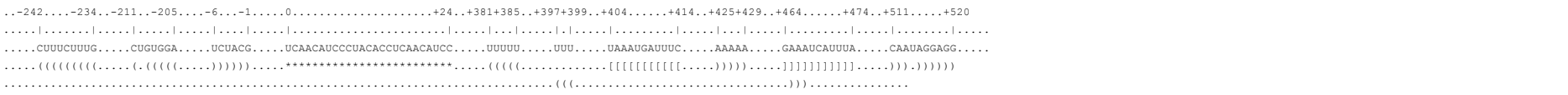
The predicted core structures of top 3 TR candidates of *D. hansenii*:

Total:35.97 =pk Score:1*19.52 +Triple Score:1*6.8 +boundaryEle Score:1*4.85 +Core Closing Score:1*4.8

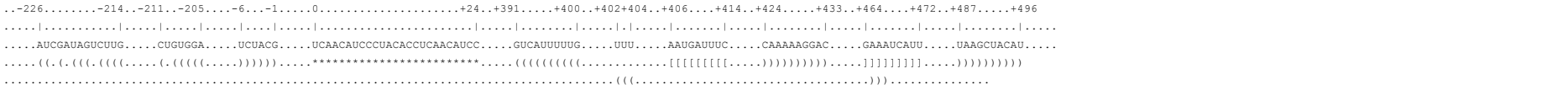


Core-closing stem | Boundary element | Template | PK Stem 1 | Triple helix | PK Stem 2 | PK Stem 1 | PK Stem 2 | Core-closing stem

Total:35.72 =pk Score:1*19.27 +Triple Score:1*6.8 +boundaryEle Score:1*4.85 +Core Closing Score:1*4.8

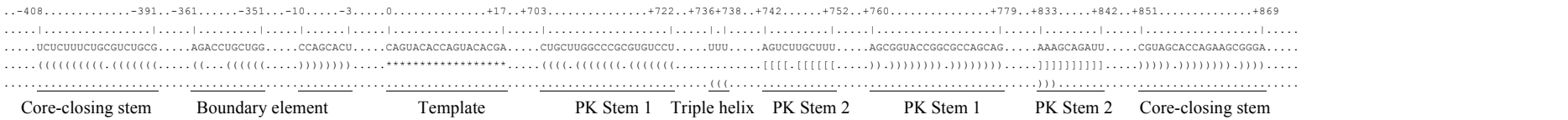


Total:34.71 =pk Score:1*19.52 +Triple Score:1*6.8 +boundaryEle Score:1*4.85 +Core Closing Score:1*3.54

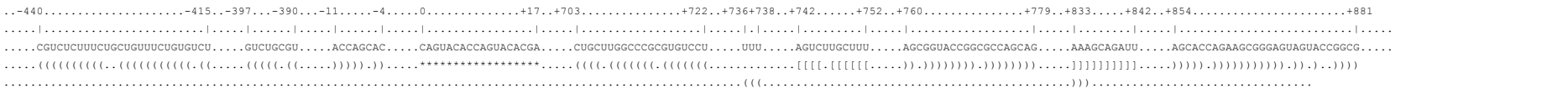


The predicted core structures of top 3 TR candidates of *C. guilliermondii*

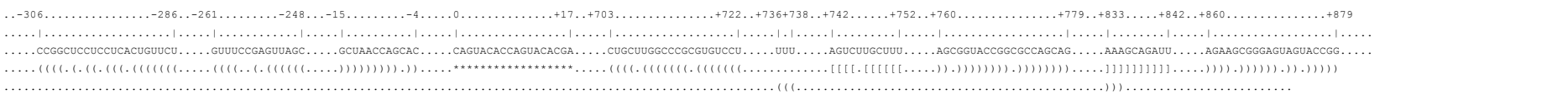
Total:37.82 =pk Score:1*17.08 +Triple Score:1*7.33 +boundaryEle Score:1*5.49 +Core Closing Score:1*7.92



Total:37.36 =pk Score:1*17.08 +Triple Score:1*7.33 +boundaryEle Score:1*5.51 +Core Closing Score:1*7.44

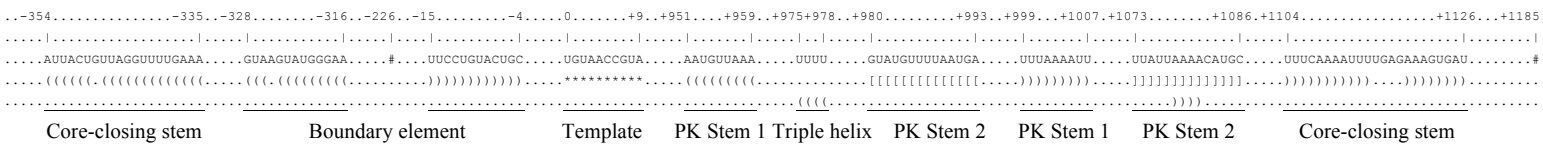


Total:35.54 =pk Score:1*17.08 +Triple Score:1*7.33 +boundaryEle Score:1*5.2 +Core Closing Score:1*5.93

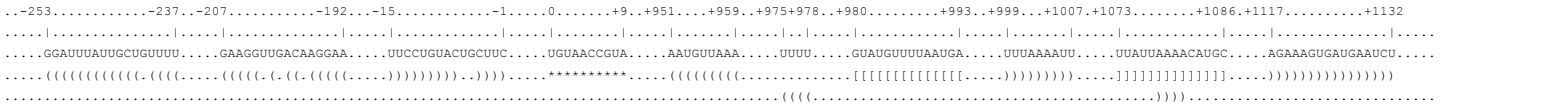


The predicted core structures of top 3 TR candidates of *S. pombe*

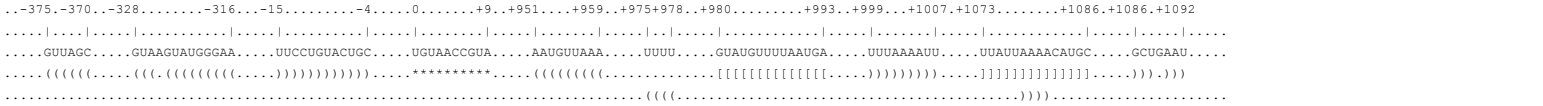
Total:40.79 =pk Score:1*16.75 +Triple Score:1*8.75 +boundaryEle Score:1*10.54 +Core Closing Score:1*4.75



Total:40.25 =pk Score:1*16.75 +Triple Score:1*8.75 +boundaryEle Score:1*8.19 +Core Closing Score:1*6.56



Total:38.81 =pk Score:1*16.75 +Triple Score:1*8.75 +boundaryEle Score:1*10.54 +Core Closing Score:1*2.77



CHAPTER 3

CANDIDA TROPICALIS STRAINS CONTAINING DIMORPHIC TELOMERIC REPEATS ARE ASSOCIATED WITH HIGH ALLELIC DIVERGENCE IN THE TELOMERASE RNA GENE AND ELSEWHERE IN THE GENOME[2]

[2]Guo, L, Dunwoody, R. and McEachern, M. J. To be submitted to *Eukaryotic Cell*.

Abstract

It has been previously reported that certain *C. tropicalis* strains contain two forms of telomeric repeat that differ by a single base pair. In this study, we found that one of these strains has two versions of the telomerase RNA gene (*TER1*), one that essentially matches the allele from the sequenced MYA-3404 strain and the other with 7% sequence divergence from it. The latter allele includes a template mutation specifically for the atypical type of telomeric repeats. To our knowledge, this is the first identification of a natural organism that has dimorphic telomeric repeats as a result of having two different *TER1* alleles. A second *C. tropicalis* strain with both types of telomeric repeat contained only the divergent type of *TER1*. It is likely that this strain previously had two different *TER1* alleles but that mitotic recombination led to loss of heterozyzosity. Our study also showed that the two strains with dimorphic telomeric repeats showed unusually high allelic divergence for most of multiple genes examined across the genome. Although there has been accumulating evidence of sexual reproduction in the *Candida* clade, mating has never been observed in *C. tropicalis*. The presence of highly divergent alleles in these strains supports the possibility that genetic recombination can take place in this "asexual diploid" yeast.

Introduction

Candida is a genus of yeast, of which many species are endosymbionts of animal hosts including humans. *Candida tropicalis* is an asexual diploid organism. Certain strains of *C. tropicalis* are well-known pathogens of humans and other organisms, making it the second most common pathogenic Candida species after *Candida albicans* (Ann Chai *et*

al., 2010; Kothavade *et al.*, 2010). The diploid genome is approximately 30 Mb, containing 5-6 pairs of chromosomes. Although the haploid states of *C. albicans* and *C. tropicalis* have not yet been observed, their genomes both contain a mating type-like locus (MTL) whose genes complement *Sacharomyces cerevisiae* homologous genes (Hull and Johnson, 1999). When the MTL locus is made haploid, *C. albicans* is able to undergo mating but stays at tetraploid possibly because it lacks some components required for meiosis (Hull *et al.*, 2000; Magee and Magee, 2000). Though there is not currently any evidence of mating in *C. tropicalis*, similar to *C. albicans*, the genome contains the MTL locus but lacks some components required for meiosis, suggesting the possibility of mating in *C. tropicalis* using a similar mechanism as *C. albicans*.

A previous study has shown that some strains of *C. tropicalis* have telomeres containing two forms of telomeric repeats differing by a single nucleotide (A-repeat and C-repeat) that are present all the way to the chromosomal terminus (McEachern and Blackburn, 1994) (Figure 3.1A). Telomeres are typically composed of tandem repeats that serve as a 'cap' on chromosome ends, preventing them from eliciting DNA damage responses like end-to-end fusion or exonucleolytic degradation (Paeschke *et al.*, 2010; Donate and Blasco, 2011). Most of the known species have uniform telomeric repeats. However, several organisms including some protozoa, fungi, slime molds and plants, have irregular telomere repeat sequences (Podlevsky *et al.*, 2008). This is most pronounced in *Saccharomyces cerevisia*e and *Schizosaccharomyces pombe* with the telomere consensus sequences $(TG)_{1\text{-}4}G_{2\text{-}3}$ (Cohn *et al.*, 1998) and $GGTTACA(G)_{1\text{-}4}$ (Hiraoka *et al.*, 1998), respectively. Even in human telomeres with invariant 5'-GGTTAG-3' repeats, random

mutations can accumulate and increase the repeat heterogeneity in the more internal (centromere proximal) parts of the telomeres (Brown *et al.* 1990; Richards *et al.* 1992).

Telomerase is a ribonucleoprotein polymerase whose RNA component (telomerase RNA) contains a short region that is complementary to the telomeric repeat sequence and serve as a template for telomerase to synthesize telomeric DNA repeats (Greider and Blackburn, 1989). In all telomerase RNAs, the template is defined by terminal repeats of ~3-10 nucleotides (nt) at both ends (Yu *et al*., 1990, Wang *et al*., 2009). These repeats are thought to be involved in the alignment with the last telomeric repeat before a new round of telomeric repeat synthesis. The sequence synthesized based on the 5' TR of the template can align perfectly or with a mismatch with 3' TR of the template after translocation, providing 3-10 nt of complementarity between the telomere and the template. This appears to account for the uniformity in telomeric repeat sequences within most species.

Different models have been put forward to explain the synthesis of variable telomeric repeats. In *S. cerevisiae*, two possible mechanisms may account for the heterogeneous telomeric repeat synthesis: redundant alignment possibilities within the template region; and abortive reverse transcription (Forstemann and Lingner, 2001). In *Tetrahymena thermophila*, slippage of the template in a stretch of four G/C base pairs was proposed to account for the synthesis of poly(dG) observed *in vitro* (Harrington and Greider, 1991). In *Paramecium tetraurelia*, the mixed synthesis of T_2G_4 and T_3G_3 telomeric repeats is caused by a high frequency of nucleotide misincorporation at a specific template position

(McCormick-Graham and Romero, 1996; McCormick-Graham *et al*., 1997; Ye and Romero, 2002). Template mutations and analysis of a large number of cloned telomeres from *S. pombe* suggest alternative translocation sites and primer slippage may contribute to telomeric repeat heterogeneity (Leonardi *et al*., 2008). In species with uniform telomeric repeats, when *TER1* with mutated template was introduced, the corresponding telomeric repeats can be incorporated into the ends of telomeres (Yu *et al*., 1990; Underwood *et al*., 2004; Bechard *et al*., 2009; Mravinac *et al*., 2011).

In this study we report that *C. tropicalis* has two versions of telomerase RNA gene (*TER1*), each contains a template specifically for the synthesis of one type of telomeric repeats. Our study also showed that, besides the *TER1* region, the whole chromosome set they are located on, have around 5% sequence divergence. This striking high sequence divergence indicates a possible mating or recombination event occurred in this "asexual" diploid yeast. This is the first identification of more than one version of functional telomerase RNA genes in one species, and it can help us to have a better understanding of the rapid evolution and function maintenance of telomerase RNA structure and template sequence.

Methods and Materials

Strains

The *Candida tropicalis* strains B-4438, B-4439 and B-4443 used in this study are National Institute of Health strains generously provided by B. Wickes. The *C. tropicalis* strain 1739-82 was provided by I. Polacheck (The Hebrew University-Hadassah Medical

Center, Jerusalem). The *Candida tropicalis* strain used in FACS analysis is American Type Culture Collection (ATCC) 750. *S. cerevisiae* haploid strains arg-**a** and arg-α, diploid strain arg-**a**/α, *K. lactis* haploid strain SAY516 (Kegel *et al*. 2006) and diploid strain GG1958*SAY516 were used as controls in the FACS analysis. GG1958 was a gift from Zonneveld BJ. Diploid strain GG1958*SAY516 was from mating between two haploid strains SAY516 and GG1958 on malt plates and isolating diploids on *K. lactis* selective medium.

Blotting and Hybridization

DNA was blotted from 0.8% agarose gels onto Nytran membranes (GE healthcare, Amersham Hybond-XL) according to procedure recommended by manufacturer. Hybridizations were carried out according to the procedure of McEachern and Blackburn (1994) in Na_2HPO_4 and 7% (wt/vol) SDS. The hybridization probes used in RFLP analysis were ^{32}P-labeled DNA fragment generated from PCR of *C. tropicalis* B-4443 genomic DNA (Table 1). Hybridization was carried out at 55°C, with washes at the same temperature using 100 mM Na_2HPO_4 and 2% SDS.

Flow cytometry analysis

Cells for flow cytometry analysis are prepared using a protocol adapted from Nicoletti *et al*. (1991). Cells were collected at exponential status ($5X10^6$ cells/ml), washed with 50 mM Tris (pH 7.5), fixed with 70% ethanol, treated with 0.5 mg/ml RNaseA and 20 mg/ml Proteinase K and stained with propidium iodide (PI). Samples were sonicated for

10 seconds immediately before analysis. Flow cytometry was carried out using a FACSCalibur flow cytometer (Becton Dickinson).

Oligonucleotides for PCR and sequencing

The oligonucleotides used for PCR and sequencing of *TER1* and probes for RFLP analysis used in this study are listed in Table 1.

Results

Confirming the presence of two types of telomeric repeats in certain *C. tropicalis* strains.

It was previously reported that certain *C. tropicalis* strains contain two forms of telomeric repeats that differ by a single base pair (McEachern and Blackburn, 1994) (Figure 3.1A). We will refer to these two forms as the A-repeat and the C-repeat. As a first step to study the mechanism of how *C. tropicalis* produces these two types of repeats, we repeated hybridizations to detect their presence. Genomic DNA from several *C. tropicalis* strains, including B-4414, B-4438, B-4439, B-4443 and 1739-82 was hybridized to oligonucleotide probes specific for either A-repeat or C-repeat (Figure 3.1B). As expected, all strains hybridized well to the probe specific to the A-repeat, consistent with A-repeats being present in telomeres of all the strains. In contrast, only strains B-4414 and 1739-82 hybridized with the probe specific to the C-repeat. These results were consistent with the previously reported data and clearly indicated that both B-4414 and 1739-82 contained at least two forms of telomeric repeats, while the B-4438, B-4439 and B-4443 strains likely contained only the A-repeat. We also observed that the ratio of A-repeats to C-repeats was different between B-4414 and 1739-82. The ratio of C-probe

signal strength to the A-probe signal strength was 30% higher in the 1739-82 strain compared to the B-4414 strain. We will hereafter refer to *C. tropicalis* strains with only A-repeats as A strains and strains with both A-repeats and C-repeats as A/C strains.

At least two hypotheses could explain the existence of two types of telomeric repeats in *C. tropicalis* strains. The first is that there are two different alleles of the telomerase RNA gene (*TER1*), each with a different template sequence responsible for the synthesis of one type of telomeric repeat (Figure 3.2A). The reported *TER1* template sequence has a sequence consistent with the synthesis of the A-repeat [(Podlevsky *et al*., 2008; Gunisova *et al.*, 2009 and Guo *et al*., 2011) and Figure 3.2]. An alternative hypothesis is that only a single template sequence is sufficient for the synthesis of both A-repeats and C-repeats. As shown in Figure 3.2B, if the synthesis of one telomeric repeat does not stop at the expected last nucleotide of the template (32U), but reads through to the next nucleotide (33G) instead, a cytosine would be incorporated onto the telomeric end. If this end were then able to translocate back to the start of the template and be extended by telomerase in spite of a mismatched terminal base, a C-repeat would be the result synthesized.

<u>There is significant sequence divergence between *TER1* genes in the A-strains and the A/C-strains.</u>

In order to distinguish between two models for dimorphic telomeric repeats, we first attempted to PCR amplify and sequence the *TER1* region from both A strains and A/C strains. This analysis showed that the template sequence in the A-strain B-4443 perfectly matched the sequence from strain MYA-3404 from which the *C. tropicalis* genome was

sequenced (http://www.broadinstitute.org/). However, despite multiple attempts of sequencing using the same primers, we failed in getting reliable sequencing results in A/C-strain 1739-82. In addition, PCR and sequencing results suggested that A/C-strain B-4414 had the same template as in A-strain B-4443 and MYA-3404.

We then used RFLP (Restriction Fragment Length Polymorphism) analysis to test the template sequence in both A/C strains and in a control A-strain. The two-template model for formation of dimorphic telomeric repeats predicted that there should be an additional *Rsa*I site caused by the single nucleotide polymorphism (Figure 3.2, Figure 3.3A). However, instead of the an *Rsa*I fragment becoming 23 bp shorter than the 228 bp fragment of the A-strain control, as expected from this model, we observed one band that was ~228 bp and another band of ~1.2 kb in the A/C-strain B-4414 strain and only an ~1.2 kb band in the A/C-strain 1739-82 (Figure 3.3B). This result suggested that the A/C-strain 1739-82 probably contained two *TER1* alleles that was different from those in the A-strain and that the A/C-strain B-4414 may have one of each type of allele. The unexpectedly large *Rsa*I fragment sizes of the unusual alleles can potentially be explained as one of the expected *Rsa*I sites being missing. However, whether the unusual alleles identified contain the expected template mutation matching the C-repeat could not be determined from this analysis (Figure 3.3A).

Other restriction enzymes were then used for RFLP analysis around the *TER1* template region in the A/C- and A-strains. Strikingly, six of ten enzymes showed RFLPs between A-strain and A/C-strains (Figure 3.3B). In all ten digestions, the sizes of hybridizing

fragments in the A-strain were consistent with the expected sizes from the genome database. In the *BsmA*I, *Dra*I and *Alu*I digestions, similar to the *Rsa*I digestion, the A/C-strain 1739-82 had a single band of a different fragment size than seen in the A-strain and the A/C-strain B-4414 had two bands, one matching each of the single fragment sizes of the A/C-strain 1739-82 and A-strain B-4443. With the *Hae*III digestion, both A/C-strains had a single band of the same size, which was different from the band size in A-strain. With *Apo*I, the A-strain had two fragments, one of which matched the expected size, while the A/C-strain 1739-82 had a single band that was slightly smaller than the expected size. The A/C-strain B-4414 in contrast had two bands, one matched the size of the lower band in B-4443 and the other was a unique size. This complicated band pattern was presumably caused by multiple RFLPs between different alleles in these strains. Assuming each RFLP was due to a single base change, our results indicated that 5% of the bases in the restriction sites sampled by our digests were different in the unusual *TER1* alleles compared to the alleles present in the A-strain. These results suggested that the unusual alleles were surprisingly divergent from the normal allele. Such a level of divergence would greatly exceeded the 0.17% average sequence divergence between alleles over the genome and was considerably higher than the 1.6 % divergence seen in the most divergent 5000 bp region of the sequenced genome.

Allele specific PCR and sequencing of the *TER1* templates in A/C-strains

The strikingly high sequence divergence of *TER1* between A/C-strains and A-strains was possibly the reason why it was not possible to achieve PCR of this gene in the A/C-strain 1739-82, since the PCR primers were designed based on sequence from the A-strain

MYA-3404. In the A/C-strain B-4414, the PCR primers might favor the *TER1*^{A} allele if there was any sequence divergence in the primers.

The RFLP analysis suggested that A/C-strain 1739-82 had two copies of an unusual *TER1* allele and that the A/C-strain B-4414 had one standard allele and one unusual one. In order to amplify these unusual *TER1* alleles, we attempted PCR amplification using multiple primer pairs designed in the A-strain MYA-3404, and were able to get the complete sequence of the *TER1* gene. Primers used for PCR and sequencing the A/C-strain 1739-82 and A-strain are listed in Table 1. The *TER1* allele in A/C-strain 1739-82 strain was found to contain a telomerase RNA template whose sequence was fully complementary to the C-repeat. SNP (Single Nucleotide Polymorphism) map of complete *TER1* gene and surrounding regions in CBS94 (EU477490.1), MYA-3404 (http://www.broadinstitute.org/), A-strain B-4443 and A/C-strain 1739-82 was shown in Figure 3.4. Complete alignment of *TER1* gene and surrounding regions in these strains is available in Figure 3.S1 of Supplementary Materials. In Figure 3.4, the sequence of the *TER1*^{C} allele and surrounding regions in the A/C-strain 1739-82 strain was found to have 6.92% sequence divergence with the *TER1*^{A} allele and surrounding regions from the A-strain B-4443 strain. This degree of sequence polymorphism is fully consistent with degree of divergence inferred from the RFLP analysis.

We then designed allele-specific PCR and sequencing primers based on the *TER1*^{A} and *TER1*^{C} sequences in which the first nucleotide in the 3' end of both primers was a SNP. The allele-specific primers used are listed in Table 2. With these primers, we were able to

separately amplify the two *TER1* alleles in the A/C-strain B-4414. The PCR products obtained were tested by RFLP analysis to make sure that only one *TER1* allele was amplified (data not shown). The sequencing results from these PCR products in A/C-strain B-4414 showed that, as expected, the divergent allele had a template specific to the C-type telomeric repeat while the other allele had a template matching A-type repeats (Figure 3.4). The ~300 bp of *TER1* (C repeat type) sequence obtained from the A/C-strain B-4414 is 100% the same as the *TER1* in A/C-strain 1739-82. The ~500 bp sequence of *TER1* (A repeat type) from A/C-strain B-4414 has two SNPs relative to *TER1* in A-strain B-4443. One of these SNPs in CBS94 matched the *TER1* (A repeat type) in the A/C-strain B-4414.

Comparison of the predicted core secondary structure of the telomerase RNAs from the divergent *TER1* alleles

The core secondary structure of the *C. tropicalis* (CBS94, A-strain) telomerase RNA was previously predicted using the TR-Folder program (Guo *et al*., 2011). The total size of the predicted core structure is 718 nt, which contains a pseudoknot with a triple helix that is ~400 nt upstream of the template region, a template boundary element that begins four nt downstream of the template, and a long pairing core-closing stem that brings the 3' and 5' parts of the core structure together (Figure 3.5). We used TR-Folder to predict the core structure of *TER1* from both the A-strain B-4443 and the C-repeat specific *TER1* from the A/C-strain 1739-82. The predicted core structures from both RNAs were 719 nt. This region of the two alleles contained 31 SNPs (4.3% sequence divergence) (Figure 3.4). Among these 31 SNPs, one was located in the template region, accounting for the

dimorphic telomeric repeats and only two were located in bases involved in pairing regions of the predicted structural elements. The remaining 28 were located in either the loop between pseudoknot and the template or the loop of the template boundary element (Figure 3.5). The pseudoknot and its triple helix as well as the pairing stem of boundary element were exactly the same for both alleles. The core-closing stem of C-type *TER1* contained two C to U transitions that each change G-C pairs in A-type *TER1* to G-U pairs in $TER1^{C}$.

A/C-strains contain divergent alleles in genes throughout the genome

We next asked whether dramatic sequence divergence was specific to *TER1* alleles or was present elsewhere in the genome of A- and A/C-strains. To address this, we performed RFLP analysis using multiple probe sequences from different regions of the genome. The probes used in this RFLP analysis were designed based on sequence available from the A-strain MYA-3404. The PCR primers used for generating the RFLP probes are listed in Table 1. To minimize complications of interpreting results, the probes were designed to be small (222-620 bp) and not cut by the enzymes used to digest genomic DNA. The same six enzymes that detected RFLPs around *TER1* region (*Alu*I, *Apo*I, *BsmA*I, *Dra*I, *Hae*III and *Rsa*I) were used to look for RFLPs in the A-strain B-4443 and the B-4414 and 1739-82 A/C-strains. We first examined eight regions at specified distances (from 2 kb to 100 kb) on either side of *TER1* on the same supercontig (Supercontig 1). We also examined six randomly chosen genome sequences located on six different supercontigs covering the majority of the assembled supercontigs above 0.5

Mb as well as two genes, EST1 (on Supercontig 11) and EST2 (on Supercontig 5), that are other components of telomerase.

Table 3.2 summarizes the RFLP data from the three categories of sequence probes examined: those from *TER1*, those from the *TER1* supercontig and those from other non-*TER1* supercontigs. For our calculations, all strains were presumed to be diploid. In the *TER1* region, ten enzymes were used for digestions. Two alleles per strain times 10 digests yield 20 allelic tests per strain. For the RFLP analysis of both the non-*TER1* sequences on the *TER1* supercontig ("same supercontig" in Table 3.2) and the other probes ('different supercontigs' in Table 3.2), 96 allelic tests per strain (8 probes × 6 digestions × 2 alleles) were performed. Our results from this analysis showed that the ratio of alleles showing RFLPs in the A-strain B-4443 were 4-5% in each of the three groups of probes (Table 3.2). However, the ratio of alleles showing RFLPs in the A/C-strain B-4414 was much higher at 15-25%. The A/C-strain 1739-82 had an even higher ratio of alleles showing RFLPs in the different supercontigs category (29%) but showed none (0%) with the *TER1* sequence and showed only 3% in sequences from the *TER1* Supercontig 1. The data with probes from the same supercontig as *TER1* were consistent with that of *TER1* itself, in which the A-strain B-4443 and the A/C-strain 1739-82 were generally homozygous while the A/C-strain B-4414 often had highly divergent alleles. However, the results with probes from the other supercontigs suggested that both A/C-strains were commonly heterozygous and had highly divergent alleles.

In order to see relationships between strains, we then estimated the number of alleles in one strain that were different from alleles present in other strains. This comparison, shown in Table 3.3, was done in pair-wise combinations of strains. In this test, any strain compared to itself showed no allelic differences while comparisons between divergent strains would be expected to show increasingly greater differences. The allele difference between the A-strain B-4443 and the sequenced A-strain MYA-3404 was relatively low and comparable in all three categories of sequence probes: 5% with *TER1*, 4% with same supercontig and 4% with different supercontigs.

However, allelic comparisons between all other pairs of strains showed much greater differences (Table 3.3). The A/C-strain 1739-82 compared to the A-strain B-4443 showed the most divergence; 55% with *TER1*, 60% with same supercontig and 39% with different supercontigs. Notably, the two A/C strains showed large allelic differences with one another: 30% with *TER1*, 49% with same supercontig and 25% with different supercontigs. This indicates that despite the fact that they contain virtually the same highly unusual *TER1* allele, they are not, for the most part, closely related. As would be expected, the allelic difference levels between A-strain MYA-3404 and the A/C-strains B-4414 and 1739-82 are very similar to the levels between A-strain B-4443 and the A/C-strains (data not shown). We conclude from our RFLP analysis that the high sequence divergence of the C-type *TER1* alleles is not peculiar to *TER1* but rather is widespread throughout the genome in the A/C-strains.

Flow cytometry suggests that DNA content differs between *C. tropicalis* strains.

We next used flow cytometry analysis to examine the DNA content of A- and A/C-strains of *C. tropicalis*. Control experiments with *Saccharomyces cerevisiae* and *Kluyveromyces lactis* cells confirmed that the protocol in our hands was able to readily detect separate G1 and G2 peaks in each species (Figure 3.6A). Diploid cells of both species showed the expected rightward shifts in both peaks consistent with a doubled DNA content. In contrast, among four *C. tropicalis* strains examined (A-strains ATCC750 and B-4443, and A/C strains B-4414 and 1739-82), none showed two well-separated peaks (Figure 3.6B). This result was consistent with previous studies with *Candida tropicalis* strain ATCC750 (Dvorak *et al*. 1987). This was suggested to be due to a particularly large proportion of *C. tropicalis* cells being in the G2-M phase during exponential growth (Dvorak *et al*. 1987). Interestingly, multiple flow cytometric analyses reproducibly indicated that different *C. tropicalis* strains did not have the same DNA content. The A/C-strain B-4414 had the highest DNA content, A-strain B-4443 had the lowest, while A-type strain ATCC750 and A/C-strain 1739-82 had intermediate DNA contents that were similar to each other (Figure 3.6B). These results are inconsistent with the strains examined all being true diploids with similar DNA contents.

We quantified the relative signal of the various bands produced in our RFLP analysis in order to look for evidence that alleles of any tested probe being present in a strain in other than a 1:1 ratio or whether any strain might carry a greater or lesser number of alleles of some sequences. Our data suggested that the different *TER1* alleles present in the A/C strain B-4414 were present in a 1:1 ratio (data not shown). While other evidence hinted

that a few other polymorophic sequences might be present in ratios other than 1:1, our limited data set and the variability of quantification of signal from Southern blots made firm conclusions impossible.

Discussion

In this study, we have shown that the presence of two types of telomeric repeats (A-repeat and C-repeat) in *C. tropicalis* strain B-4414 is associated with the presence of two different alleles of the telomerase RNA gene *TER1*. These alleles contained template sequences differing by a single nucleotide at a position consistent with each being specific for the synthesis of one of the two types of detected telomeric repeats present in the strain. Although formal proof awaits biochemical or genetic tests, we conclude that the simultaneous presence of A-type and C-type telomeric repeats in B-4414 is due to the presence of two different *TER1* template alleles. To our knowledge, this is the first identification of a natural organism that has dimorphic telomeric repeats as a result of having two different telomerase RNA gene alleles.

Although previous hybridization data (McEachern and Blackburn 1994) indicated that *C. tropicalis* strain 1739-82 had telomeres containing both A-type and C-type telomeric repeats, our data here showed that only C-type *TER1* template alleles were present in this strain. This result addresses why the 1739-82 strain contains a higher percentage of C-type repeats than does the B-4414 strain but raises the question of why 1739-82 has A-type repeats at all. A simple possibility is that the strain may have earlier been heterozygous at *TER1* and contained both forms of the template but that an evolutionarily

recent mitotic recombination event led to loss of that heterozygosity. Work in the yeast *Kluyveromyces lactis* has shown that wild type telomeric repeat can linger at the bases of telomeres for at least 2500 cell divisions after the wild type *TER1* template was experimentally replaced with a template mutant specifying the synthesis of an altered form of telomeric repeat (McEachern *et al*., 2002). The inner areas of telomeres are the most resistant to sequence turnover because they are generally copied by standard DNA replication rather than being generated *de novo* by telomerase. Whether the A-type repeats present in the 1739-82 strain are located in the innermost parts of telomeres has not been determined. Although we have no direct evidence that the 1739-82 strain was recently heterozygous at *TER1*, we note that our RFLP analysis of this strain indicates that it has a much lower degree of heterozygosity on the *TER1* supercontig than it does elsewhere in its genome. Extended regions of homozygosity have been found in *C. tropicalis* A-strain MYA-3404 as well as in *C. albicans* SC5314 and WO-1 (Butler *et al*. 2009). Chromosome regions devoid of SNPs can be as large as 1 Mb in *C. tropicalis* MYA-3404.

In addition to specifying the synthesis of C-type telomeric repeats, the U to G mutation at the +10 position of the template in the C-type telomerase RNA (Figure 3.2) formally leads to a shift in the position of the whole template. While the A-type telomerase RNA template can be described as 32 nt long with 9 nt terminal repeats containing one mismatch, the C-type template can be described as 33 nt long with 10 nt terminal repeats containing one mismatch (Figure 3.2C). This occurs because the G at the +10 position creates a match with the G at position +33, thereby lengthening the size of the terminal

repeats bordering the template from 9 nt to 10 nt. We have no reason to believe that the *C. tropicalis* telomerase can copy telomeric sequence further along the C-type template than it can using an A-type template. Nonetheless, the one nt expansion in the potential template size with the C-type template presents the evolutionary capacity for that to eventually arise. A comparison of telomerase RNA template sequences from related yeasts suggests that this sort of shift in the size and position of template terminal repeats is a regular feature of yeast telomerase RNA evolution (Guo *et al*. 2009).

Our results also provide some insights into the functioning of the *C. tropicalis* telomerase RNA in regions outside the template. In the comparison of secondary structures in the 719 nt core region between the A-type and C-type telomerase RNAs, 28 of 31 SNPs were located in the loop region between base pairing elements. Even for the 2 SNPs in the base pairing region, the C to U transitions still maintains the capability of forming base pairs. The overall lack of SNPs within the predicted core structural features is consistent with the hypothesis that most of the sequence changes are likely to be selectively neutral. This provides additional evidence for the accuracy of the structure predicted by the TR-Folder program (Guo *et al.*, 2011).

Our RFLP analysis suggests that the randomly chosen A-strain (B-4443) in our study has a relatively low degree of heterozygosity and is closely related to the sequenced *C. tropicalis* strain MYA-3404. Our studies also show that the A/C strains B-4414 and 1739-82 have much higher levels of heterozygosity and have genes throughout their genomes with alleles that are surprisingly divergent from characterized *C. tropicalis*

genes. Despite having *TER1* alleles that are highly similar if not identical, our results with sequences elsewhere in the genome indicate that B-4414 and 1739-82 are, overall, not closely related to one another. These results are seemingly at odds with the idea that *C. tropicalis* is an obligately asexual organism. They are instead more consistent with the possibility that the two A/C strains shared a common ancestor that was significantly divergent from more standard *C. tropicalis* strains but provided the source of a common C-type *TER1* allele.

In fact, there is accumulating evidence of sexual reproduction in the "asexual" *Candida* clade in the past ten years. Mating has been observed in *C. albicans* between laboratory constructed homozygous MTL**a** and MTLα strains (Magee and Magee 2000). In addition to traditional heterothallic mating, homothallic mating between cells of the same sex has recently been described (Alby *et al*., 2009). The tetraploid mating products can go back to diploid, or very close to diploid DNA content by gradual chromosome loss instead of traditional meiosis (Bennett and Johnson 2003). Thus, progeny formed by this parasexual cycle may not be true diploids but rather aneuploids instead (Bennett and Johnson 2003; Forche *et al*. 2008). *Candida dubliniensis*, the species most closely related to *C. albicans*, exhibits strong similarities with *C. albicans* in its heterothallic mating mechanism (Pujol *et al*. 2004). Another diploid species in the *Candida* clade, *Lodderomyces elongisporus*, has been shown to undergo mating through a homothallic mating program (Lockhart *et al*. 2008), and yet all MTL loci are missing in this species (Butler *et al*. 2009). Two haploid *Candida* species, *Candida lusitaniae* and *Candida guilliermondii*, have defined heterothallic sexual cycles (Butler *et al.*, 2009; Reedy *et al*., 2009).

Although mating has never been reported in *C. tropicalis*, some other evidence suggests that this species is more complicated than simply being an asexual diploid yeast. First, the *C. tropicalis* annotated genome (of MYA-3404) contains two mating type-like (MTL) loci, *MTL***a** and *MTL*α, which are orthologs of the *MTL* loci in *C. albicans*. Moreover, *C. tropicalis* contains a similar set of meiotic genes as *C. albicans* (Butler *et al.* 2009). *C. tropicalis* is the most closely related species to *C. albicans* and *C. dubliniensis*, and is therefore likely to share the heterothallic mating program and parasexual cycle with these two diploid species. Second, multilocus sequence typing (MLST) studies in *C. tropicalis* and *C. albicans* isolates showed that both species exhibit high levels of recombination (Tavanti *et al*. 2005; Jacobsen *et al*. 2008a, b; Odds and Jacobsen 2008). Third, the flow cytometry analysis in this study suggests that different *C. tropicalis* strains have different DNA contents and that some may not be euploid. It is possible that an increase in ploidy or DNA content will enhance the pathogenesis in these species, perhaps because of the expansion of gene families related to virulence. It should be noted that aneuploidy is a common method for *C. albicans* to adapt to different environmental stresses (Alby *et al*., 2010).

Our work predicts that it may be possible to find divergent *C. tropicalis*-like strains that were the original source of the C-type *TER1* alleles in the A/C strains in our study. Such strains might be expected to have relatively low levels of heterozygosity but have 5-10% sequence divergence from standard *C. tropicalis* strains. Our results also suggest that PCR-based methods of characterizing *C. tropicalis* genetic diversity (Tavanti *et al*. 2005;

Jacobsen *et al*. 2008b; Odds and Jacobsen 2008) may sometimes fail to detect the unexpected highly allelic divergence such as that seen in the A/C strains in this study. Much future work will be needed in order to have a good understanding of the genetic structure of this pathogen.

References

Ann Chai, L. Y., Denning, D. W. and Warn, P. (2010) '*Candida tropicalis* in human disease', *Crit Rev Microbiol* **36**: 282-298.

Bechard, L. H., Butuner, B. D., Peterson, G. J., McRae, W., Topcu, Z. and McEachern, M. J. (2009) 'Mutant telomeric repeats in yeast can disrupt the negative regulation of recombination-mediated telomere maintenance and create an alternative lengthening of telomeres-like phenotype', *Mol Cell Biol* **29**: 626-639.

Brown, W. R., MacKinnon, P. J., Villasante, A., Spurr, N., Buckle, V. J. and Dobson, M. J. (1990) 'Structure and polymorphism of human telomere-associated DNA', *Cell* **63**: 119-132.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., Agrafioti, I., Arnaud, M. B., Bates, S., Brown, A. J., Brunke, S., Costanzo, M. C., Fitzpatrick, D. A., de Groot, P. W., Harris, D., Hoyer, L. L., Hube, B., Klis, F. M., Kodira, C., Lennard, N., Logue, M. E., Martin, R., Neiman, A. M., Nikolaou, E., Quail, M. A., Quinn, J., Santos, M. C., Schmitzberger, F. F., Sherlock, G., Shah, P., Silverstein, K. A., Skrzypek, M. S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M. P., Sudbery, P. E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N. A., Lorenz, M. C., Birren, B. W., Kellis, M. and Cuomo, C. A.

(2009) 'Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes', *Nature* **459**: 657-662.

Cohn, M., McEachern, M. J. and Blackburn, E. H. (1998) 'Telomeric sequence diversity within the genus *Saccharomyces*', *Curr. Genet.* **33**: 83-91.

Donate, L. E. and Blasco, M. A. (2011) 'Telomeres in cancer and ageing', *Philos Trans R Soc Lond B Biol Sci* **366**: 76-84.

Dvorak, J. A., Whelan, W. L., McDaniel, J. P., Gibson, C. C. and Kwon-Chung, K. J. (1987) 'Flow cytometric analysis of the DNA synthetic cycle of *Candida* species', *Infect Immun* **55**: 1490-1497.

Forstemann, K. and Lingner, J. (2001) 'Molecular basis for telomere repeat divergence in budding yeast', *Mol. Cell. Biol.* **21**: 7277-7286.

Greider, C. W. and Blackburn, E. H. (1989) 'A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis', *Nature* **337**: 331-337.

Gunisova, S., Elboher, E., Nosek, J., Gorkovoy, V., Brown, Y., Lucier, J. F., Laterreur, N., Wellinger, R. J., Tzfati, Y. and Tomaska, L. (2009) 'Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements', *RNA* **15**: 546-559.

Guo, L., Zhang, D., Wang, Y., Malmberg, R. L., McEachern, M. and Cai, L. (2011) 'TRFolder: computational prediction of novel telomerase RNA structures in yeast genomes', *Int J Bioinform Res Appl* **7**: 63-81.

Harrington, L. A. and Greider, C. W. (1991) 'Telomerase primer specificity and chromosome healing', *Nature* **353**: 451-454.

Hiraoka, Y., Henderson, E. and Blackburn, E. H. (1998) 'Not so peculiar: fission yeast telomere repeats', *Trends Biochem. Sci.* **23**: 126.

Hull, C. M. and Johnson, A. D. (1999) 'Identification of a mating type-like locus in the asexual pathogenic yeast *Candida albicans*', *Science* **285**: 1271-1275.

Hull, C. M., Raisner, R. M. and Johnson, A. D. (2000) 'Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host', *Science* **289**: 307-310.

Kegel, A., Martinez, P., Carter, S. D. and Astrom, S. U. (2006) 'Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*', *Nucleic Acids Res* **34**: 1633-1645.

Kothavade, R. J., Kura, M. M., Valand, A. G. and Panthaki, M. H. (2010) '*Candida tropicalis*: its prevalence, pathogenicity and increasing resistance to fluconazole', *J Med Microbiol* **59**: 873-880.

Leonardi, J., Box, J. A., Bunch, J. T. and Baumann, P. (2008) '*TER1*, the RNA subunit of fission yeast telomerase', *Nat. Struct. Mol. Biol.* **15**: 26-33.

Magee, B. B. and Magee, P. T. (2000) 'Induction of mating in *Candida albicans* by construction of MTL**a** and MTLalpha strains', *Science* **289**: 310-313.

McCormick-Graham, M., Haynes, W. J. and Romero, D. P. (1997) 'Variable telomeric repeat synthesis in *Paramecium tetraurelia* is consistent with misincorporation by telomerase', *EMBO J.* **16**: 3233-3242.

McCormick-Graham, M. and Romero, D. P. (1996) 'A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*', *Mol. Cell. Biol.* **16**: 1871-1879.

McEachern, M. J. and Blackburn, E. H. (1994) 'A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts', *Proc. Natl. Acad. Sci. U S A* **91**: 3453-3457.

McEachern, M. J., Underwood, D. H. and Blackburn, E. H. (2002) 'Dynamics of telomeric DNA turnover in yeast', *Genetics* **160**: 63-73.

Mravinac, B., Mestrovic, N., Cavrak, V. V. and Plohl, M. (2011) 'TCAGG, an alternative telomeric sequence in insects', *Chromosoma* **120**: 367-376.

Paeschke, K., McDonald, K. R. and Zakian, V. A. (2010) 'Telomeres: structures in need of unwinding', *FEBS Lett* **584**: 3760-3772.

Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X. and Chen, J. J. (2008) 'The telomerase database', *Nucleic Acids Res* **36**: D339-343.

Richards, E. J., Chao, S., Vongs, A. and Yang, J. (1992) 'Characterization of *Arabidopsis thaliana* telomeres isolated in yeast', *Nucleic Acids Res* **20**: 4039-4046.

Underwood, D. H., Zinzen, R. P. and McEachern, M. J. (2004) 'Template requirements for telomerase translocation in *Kluyveromyces lactis*', *Mol Cell Biol* **24**: 912-923.

Wang, Z. R., Guo, L., Chen, L. and McEachern, M. J. (2009) 'Evidence for an additional base-pairing element between the telomeric repeat and the telomerase RNA template in *Kluyveromyces lactis* and other yeasts', *Mol Cell Biol* **29**: 5389-5398.

Ye, A. J. and Romero, D. P. (2002) 'A unique pause pattern during telomere addition by the error-prone telomerase from the ciliate *Paramecium tetraurelia*', *Gene* **294**: 205-213.

Yu, G. L., Bradley, J. D., Attardi, L. D. and Blackburn, E. H. (1990) 'In vivo alteration of telomere sequences and senescence caused by mutated *Tetrahymena* telomerase RNAs', *Nature* **344**: 126-132.

Figure 3.1. Presence of two types of telomeric repeats in certain *C. tropicalis* strains. A) Alignment of telomeric repeat sequences in *C. tropicalis* and *C. albicans*. The sequences used as probe specific to A-repeat and C-repeat are underlined. The SNP between A-repeat and C-repeat is shown in gray. B) Two types of telomeric repeats in A/C-strain B-4414 and 1739-82. Genomic DNAs from five representative strains of *C. tropicalis* were digested with *Cla*I, electrophoresed in an 0.8% agarose gel, and probed with oligonucleotides specific to either A-type telomeric repeat (left) or C-type telomeric repeat (right).

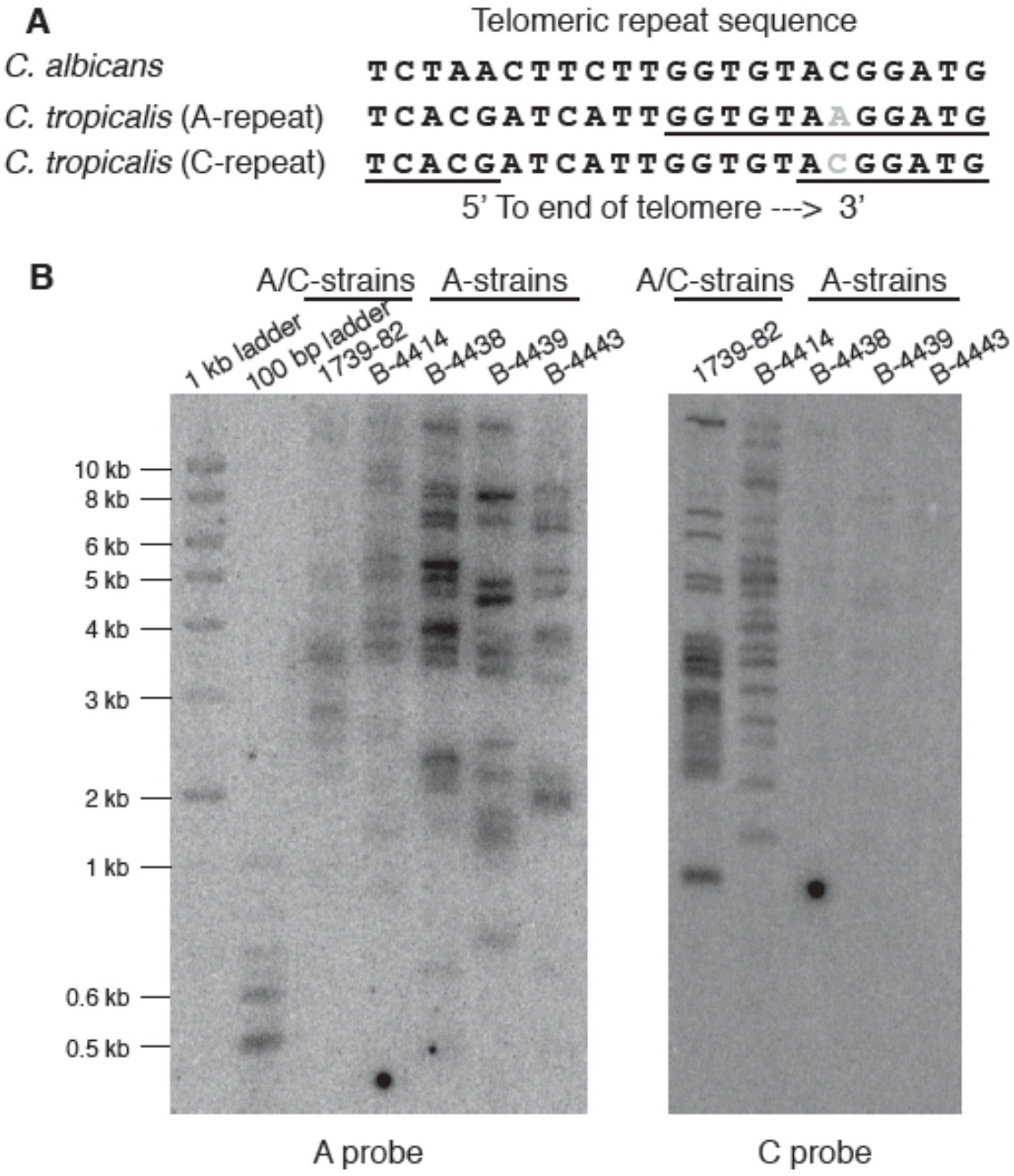
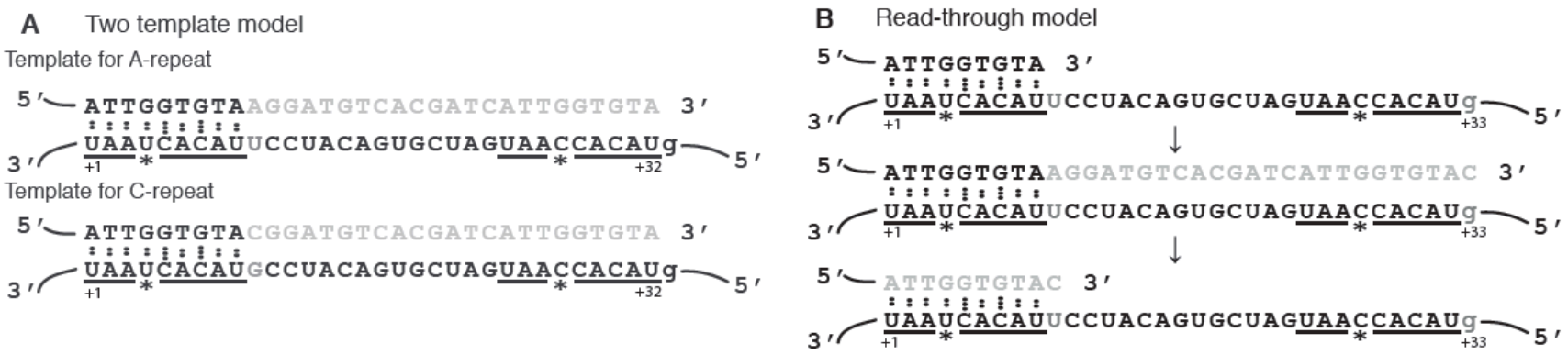

A
Telomeric repeat sequence
C. albicans
TCTAACTTCTTGGTGTACGGATG
C. tropicalis (A-repeat)
TCACGATCATTGGTGTAAGGATG
C. tropicalis (C-repeat)
TCACGATCATTGGTGTACGGATG
5' To end of telomere ---> 3'
B
A/C-strains
A-strains
A/C-strains
A-strains
1 kb ladder
100 bp ladder
1739-82
B-4414
B-4438
B-4439
B-4443
1739-82
B-4414
B-4438
B-4439
B-4443
10 kb
8 kb
6 kb
5 kb
4 kb
3 kb
2 kb
1 kb
0.6 kb
0.5 kb
A probe
C probe

Figure 3.2. Two hypotheses to explain the existence of two types of telomeric repeats in *C. tropicalis* strains. A) Diagram of two-template model. A single-stranded end of telomeric DNA is base paired with the region immediately 3' to the TR template. The top panel shows the synthesis of an A-repeat from the A-template. The bottom panel shows the synthesis of a C-repeat from the C-template. Dots between bases indicate hydrogen bonds of base pairs. Numbers indicate nomenclature used for base positions in this work. The single nucleotide difference between these templates, and the newly incorporated nucleotides at the telomere end are shown in gray. The imperfect direct repeats bordering the template are underlined, and the mismatch is marked with asterisk. B) Diagram of one-template read-through model. The top panel shows the synthesis of an A-repeat from the A-template. The middle panel shows the incorporation of a cytosine when the synthesis of a new telomeric repeat reads through to 33G. The bottom panel shows the same sequences after translocation of the newly synthesized DNA and the synthesis of a new telomeric repeat. C) Alignment of the two versions of *C. tropicalis* TR templates in the two-template model with *K. lactis* and *C. albicans* TR templates. The direct repeats bordering the template are underlined and the mismatch is marked with asterisks. The template regions are shown in capital bold letters. The conserved telomeric Rap1 binding site encoded by the templates are shaded in gray. Note that the split sites in the *Candida* templates will be one contiguous site in the telomere.



C

Rap1 binding site

K. lactis TER1 3′ cuccGCCAAAACUAAUCCAUACACCACAUGCCUAAACUucau 5′

C. albicans TER1 3′ AACUACAUGCCUACAGAUUGAAGAACCACAUGuac 5′

C. tropicalis TER1 with C-template 3′ UAAUCACAUGCCUACAGUGCUAGUAACCACAUGuuu 5′

C. tropicalis TER1 with A-template 3′ UAAUCACAUUCCUACAGUGCUAGUAACCACAUguuu 5′

Figure 3.3. RFLP analysis shows significant sequence divergence between *TER1* genes in A- and A/C-strains. A) Lower; Diagram of *Rsa*I enzyme cutting map in A-template and C-template. The template region is shown in with a thick black line. The probe used is indicated by the thick gray line. Upper; The template sequence is shown as both DNA sequence and RNA sequence. The direct repeats are underlined, and the mismatch is marked with asterisk. B) RFLPs between A-strain B-4443 and A/C-strains B-4414 and 1739-82 are shown in six digestions.

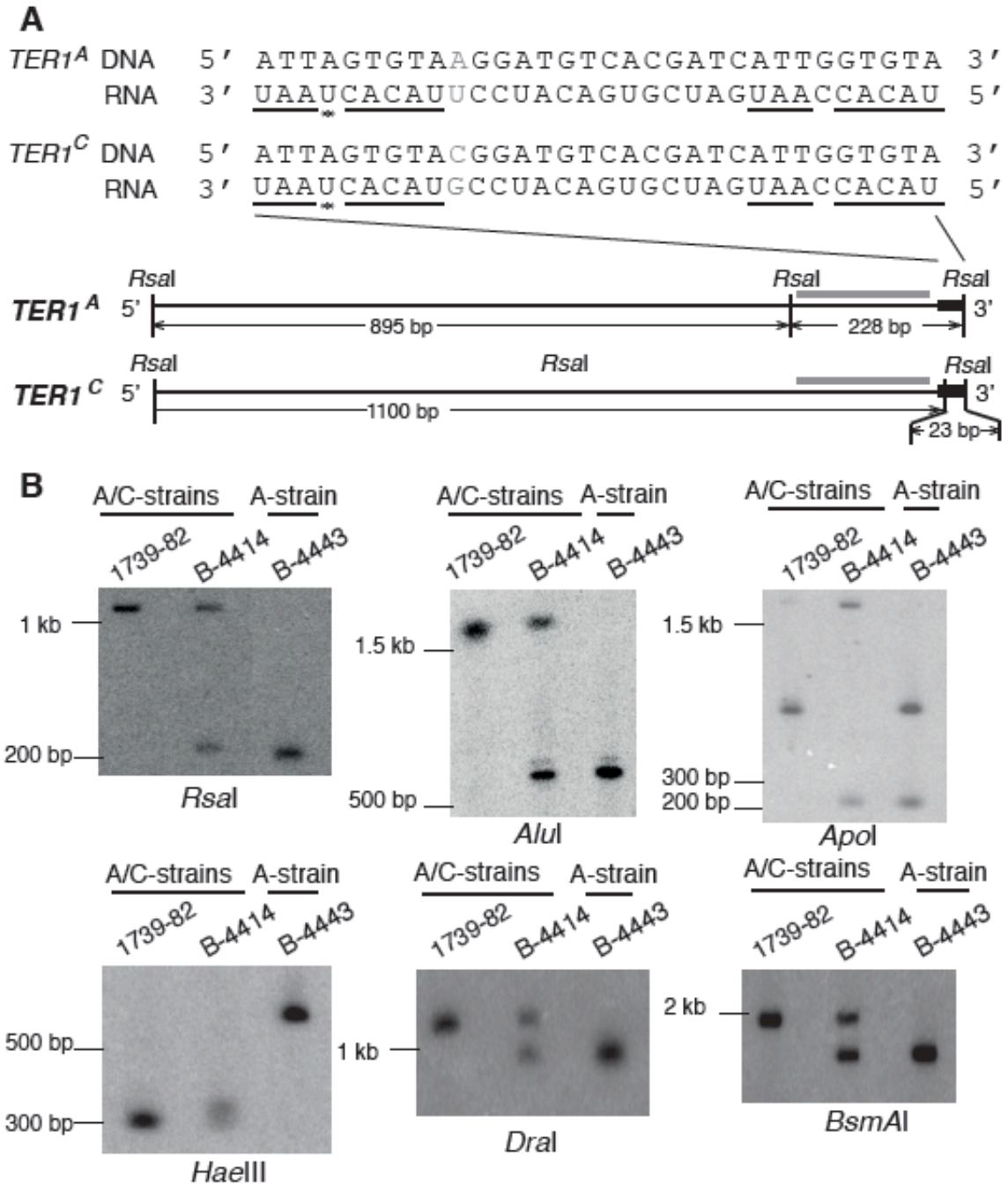

A
TER1A DNA 5′ ATTAGTGTAAGGATGTCACGATCATTGGTGTA 3′
RNA 3′ UAAUCACAUUCCUACAGUGCUAGUAACCACAU 5′
TER1C DNA 5′ ATTAGTGTACGGATGTCACGATCATTGGTGTA 3′
RNA 3′ UAAUCACAUGCCUACAGUGCUAGUAACCACAU 5′
RsaI
RsaI
RsaI
TER1A 5'
3'
895 bp
228 bp
RsaI
RsaI
RsaI
TER1C 5'
3'
1100 bp
23 bp
B
A/C-strains
A-strain
1739-82
B-4414
B-4443
1 kb
200 bp
RsaI
1.5 kb
500 bp
AluI
1.5 kb
300 bp
200 bp
ApoI
500 bp
300 bp
HaeIII
1 kb
DraI
2 kb
BsmAI

Figure 3.4. SNP map of *TER1* and surrounding regions in multiple A- and A/C-strains. Complete *TER1* and surrounding regions are shown for CBS94 (EU477490.1), MYA-3404 (http://www.broadinstitute.org/), A-strain B-4443 and A/C-strain 1739-82. Partial *TER1* sequences surrounding the telomerase RNA template region in both alleles are shown for the A/C-strain B-4414. Dots indicate positions of SNPs. Alleles specifying a template predicted to make C-type telomeric repeats are indicated as $TER1^{C}$. Alleles specifying a template predicted to make A-type telomeric repeats are indicated as $TER1^{A}$. The alignment is performed using T-COFFEE webserver (Di Tommaso *et al*. 2011).

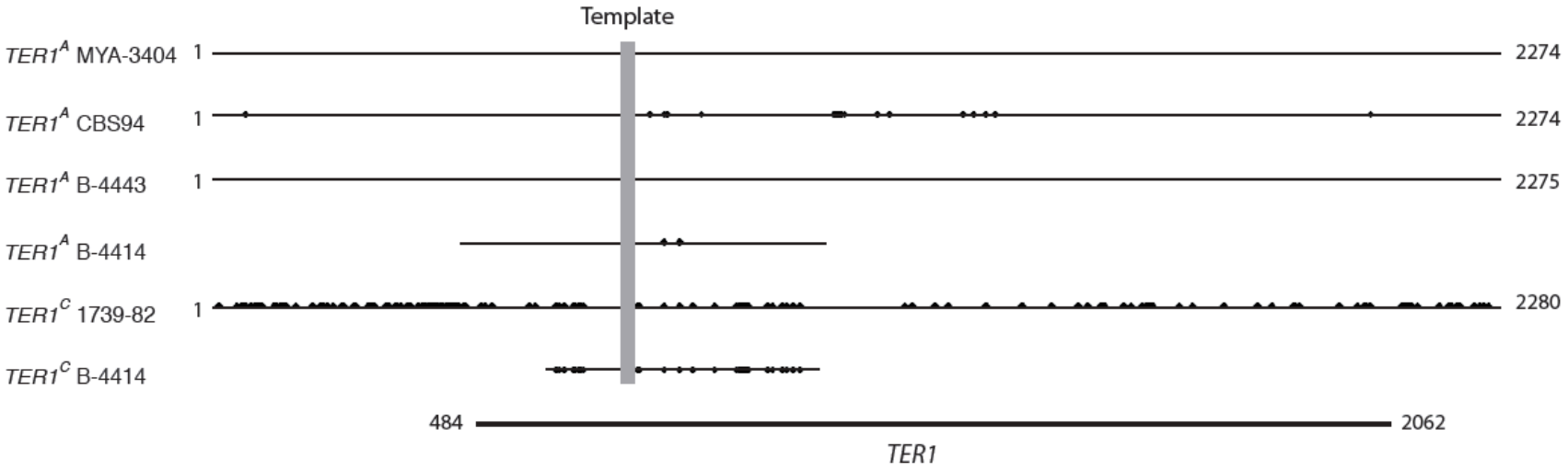
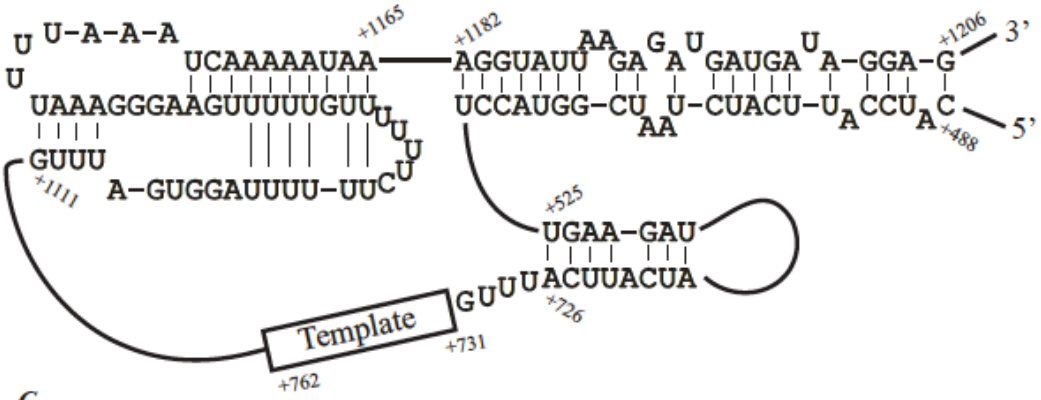

Template
TER1A MYA-3404
1
2274
TER1A CBS94
1
2274
TER1A B-4443
1
2275
TER1A B-4414
TER1C 1739-82
1
2280
TER1C B-4414
484
2062
TER1

Figure 3.5. Comparison of the predicted core secondary structure in the telomerase RNAs with A-template and C-template. The core structure of the telomerase RNA with A-template from A-strain B-4443 and the telomerase RNA with C-template from A/C-strain 1739-82 were both predicted using the TR-Folder program (Guo *et al*. 2011). The predicted core structures from both RNAs are 719 nt, with 31 SNPs in total. 28 SNPs located in the loop regions are shown by red asterisks. 2 SNPs in the base pairing regions are highlighted in red. One of these is in the template region, accounting for the dimorphic telomeric repeats. Numbers indicate nomenclature used for base positions in Figure 3.S1.

TER1 ^{A}


U U-A-A-A
UCAAAAAUAA
UAAAGGGAAGUUUUUGUU
GUUU
A-GUGGAUUUU-UUC
AGGUAUU GA A GAUGA A-GGA-G
UCCAUGG-CU U-CUACU-U CCU C
AA
UGAA-GAU
ACUUACUA
GUU
Template
+1165
+1182
+1206
3'
5'
+488
+1111
+525
+726
+731
+762

TER1 ^{C}

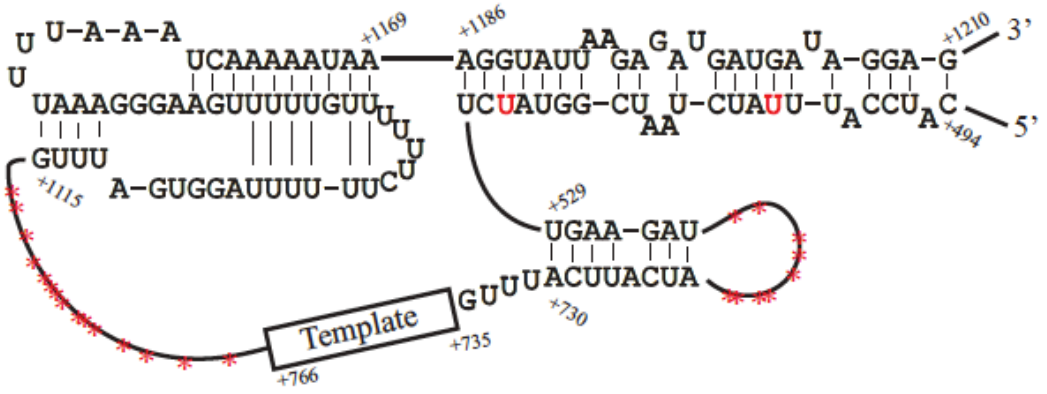

U U-A-A-A
UCAAAAAUAA
UAAAGGGAAGUUUUUGUU
GUUU
A-GUGGAUUUU-UUC
AGGUAUU GA A GAUGA A-GGA-G
UCUAUGG-CU U-CUAU U-U CCU C
AA
UGAA-GAU
ACUUACUA
GUU
Template
+1169
+1186
+1210
3'
5'
+494
+1115
+529
+730
+735
+766

Figure 3.6. DNA content of *C. tropicalis* A- and A/C-strains by flow cytometry analysis. Haploid and diploid controls are shown in A) *S. cerevisiae* strains and B) *K. lactis* strains. Results with *C. tropicalis* strains were shown in C). The X axes in the experiments are in arbitrary units that are comparable for strains within the same species but not comparable for strains between species.

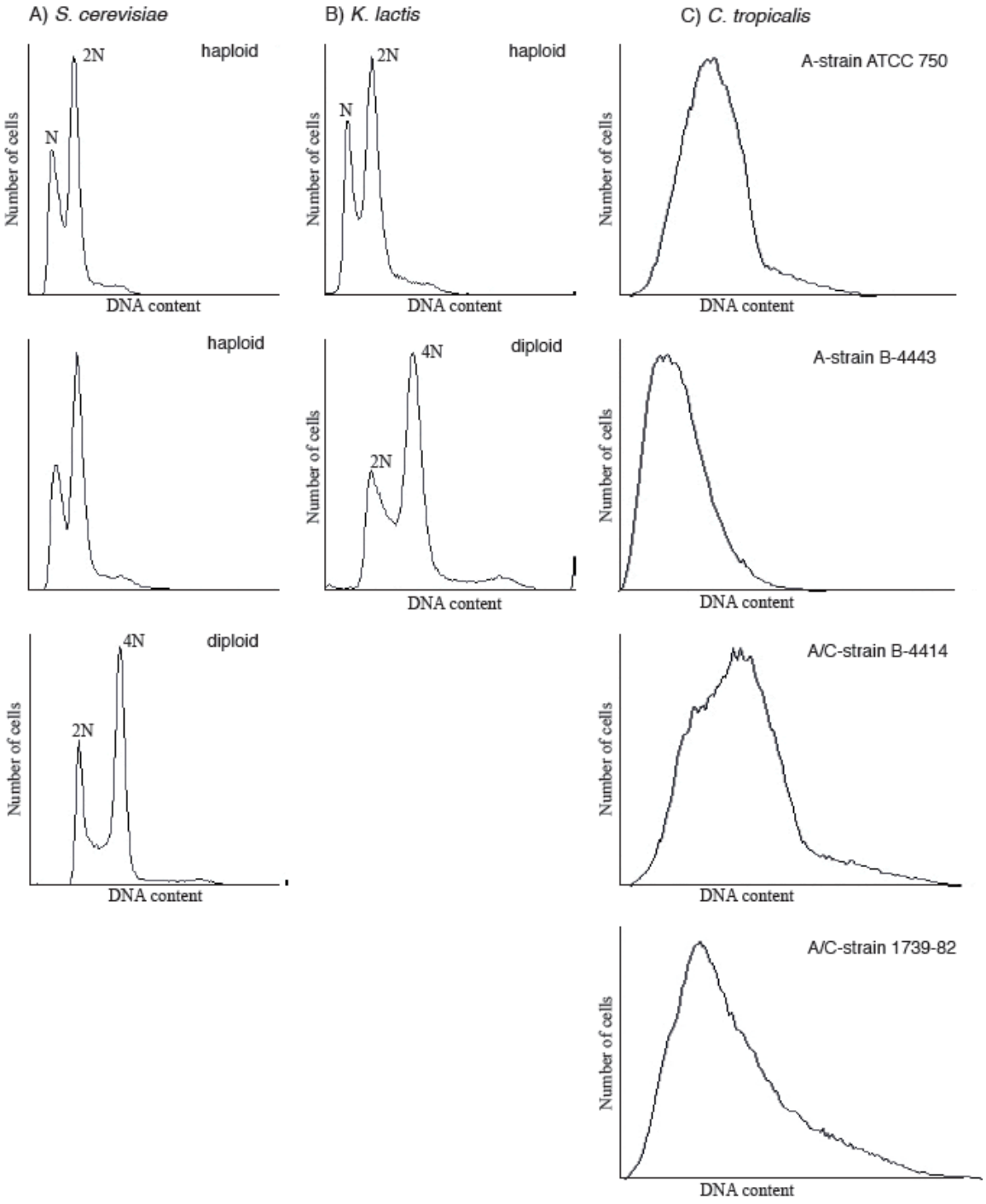

A) S. cerevisiae
B) K. lactis
C) C. tropicalis
haploid
2N
N
Number of cells
DNA content
haploid
2N
N
Number of cells
DNA content
A-strain ATCC 750
Number of cells
DNA content
haploid
Number of cells
DNA content
diploid
4N
2N
Number of cells
DNA content
A-strain B-4443
Number of cells
DNA content
diploid
4N
2N
Number of cells
DNA content
A/C-strain B-4414
Number of cells
DNA content
A/C-strain 1739-82
Number of cells
DNA content

Figure 3.S1. Complete alignment of *TER1* and surrounding regions in multiple A- and A/C-strains. Complete sequences of *TER1* and surrounding regions are shown for CBS94 (EU477490.1), MYA-3404 (http://www.broadinstitute.org/), A-strain B-4443 and A/C-strain 1739-82. Partial *TER1* sequences surrounding the telomerase RNA template region in both alleles are shown for A/C-strain B-4414. SNPs are shaded in gray. The base pairing regions of the predicted core secondary structures as well as the telomerase RNA template region are underlined, with the names indicated beneath. Whether the sequence is predicted to generate A-type telomeric repeats or C-type telomeric repeats is indicated at the left.

```
A MYA-3404    1 AAGTGTTGTCTGGAATATTATTTTATTTGAAACCTAATTTTGGTTGTAAAACGAAGCAACATCAATAGTTGGAGATTTGGTTTCAGGACTTCTTTTT--T 98
A CBS94       1 AAGTGTTGTCTGGAATATTATTTTATTTGAAACCTAATTTTGGTTGTAAAACGAAGCAACATCAATAGTTGGAGAyTTGGTTTCAGGACTTCTTTTT--T 98
A B-4443      1 AAGTGTTGTCTGGAATATTATTTTATTTGAAACCTAATTTTGGTTGTAAAACGAAGCAACATCAATAGTTGGAGATTTGGTTTCAGGACTTCTTTTT--T 98
C 1739-82     1 AAATGTTGTCTGGAATATTATTTTATTTTAAACCTAATTTTGGTTGTAAAACGAAGCAAAATCAATAGCTGGTTGTTTGGTTTCAGAACTTTTTTTTTCT 100

A MYA-3404   99 TCTTTTGTTTCAGAAAATGTAAACATCAAAGATCGTTTAAAAATGCAATGAAAAATGTCGCAAATTAAAGAGTGAGAAAGCAAATGCAACATACAGTTAA 198
A CBS94      99 TCTTTTGTTTCAGAAAATGTAAACATCAAAGATCGTTTAAAAATGCAATGAAAAATGTCGCAAATTAAAGAGTGAGAAAGCAAATGCAACATACAGTTAA 198
A B-4443     99 TCTTTTGTTTCAGAAAATGTAAACATCAAAGATCGTTTAAAAATGCAATGAAAAATGTCGCAAATTAAAGAGTGAGAAAGCAAATGCAACATACAGTTAA 198
C 1739-82   101 TCTCCTGTTTCAGAAAATGTAAACATAAAGAATCGTTAGCAAACGCAAAGAAAAATGTCGCAAACTAAAGAGTGAGAAAGCAAATGCAACATACTGTTAA 200

A MYA-3404  199 TTAGCACCATCATGCTAGTGATAAATATACCTTATTAGTTGCTACTTTGCCAACACGTTTAATTTTAAGTATCAATGGTGTGATTGGTGAATATATAGCA 298
A CBS94    199 TTAGCACCATCATGCTAGTGATAAATATACCTTATTAGTTGCTACTTTGCCAACACGTTTAATTTTAAGTATCAATGGTGTGATTGGTGAATATATAGCA 298
A B-4443   199 TTAGCACCATCATGCTAGTGATAAATATACCTTATTAGTTGCTACTTTGCCAACACGTTTAATTTTAAGTATCAATGGTGTGATTGGTGAATATATAGCA 298
C 1739-82  201 TTAGCAACATCTTACTAGTAATAAAAATACTTTATTAGTTGCTATTTGGCAACACGTTTAATTTTTAGCAGCAATGGTGTGATTGGTGAAGATGTAGGA 300

A MYA-3404  299 GCTGAATATTCATTGTTACATTGCTTCTTTGAATCTATTTTTTCTCAAATTTTAATGTCGTGCACCTAAATGTCGTGGTACTAAATTCACAAAGGGAG-- 396
A CBS94    299 GCTGAATATTCATTGTTACATTGCTTCTTTGAATCTATTTTTTCTCAAATTTTAATGTCGTGCACCTAAATGTCGTGGTACTAAATTCACAAAGGGAG-- 396
A B-4443   299 GCTGAATATTCATTGTTACATTGCTTCTTTGAATCTATTTTTTCTCAAATTTTAATGTCGTGCACCTAAATGTCGTGGTACTAAATTCACAAAGGGAG-- 396
C 1739-82  301 GCAGAATATTCATTGTTACATTGATTCCCTGAATCTATTTTCTTTCAAATATTAATGTCGTTTACCCAAATGTCGTGGTATTAAATTTACAAGGGGAGAG 400

A MYA-3404  397 --GGGGGATTGACCTTTCGCTTCGTCTTTCACAAGTTTACTTTACTTCTTTTTTTTTTTTTTACATACTTGTTCTTTTTCCTGTTAAGTCAGACATCCATT 494
A CBS94    397 --GGGGGATTGACCTTTCGCTTCGTCTTTCACAAGTTTACTTTACTTCTTTTTTTTTTTTTTACATACTTGTTCTTTTTCCTGTTAAGTCAGACATCCATT 494
A B-4443   397 --GGGGGATTGACCTTTCGCTTCGTCTTTCACAAGTTTACTTTACTTCTTTTTTTTTTTTTTACATACTTGTTCTTTTTCCTGTTAAGTCAGACATCCATT 494
C 1739-82  401 AGGGAGAATTTACTTTTCGTCTTTCTTTTATGAGCTCTATTT--TTTTTTTGTTTGTTTTTAACACACTTGTTCTTTTTCCTGTTAAGTCAGGCATCCATT 498
A B-4414                                                  TTTTTTTTTTTTTACATACTTGTTCTTTTTCCTGTTAAGTCAGACATCCATT
Helices                                                                                    Core-closing Stem

A MYA-3404  495 CATCTAATCGGTACCTTAATCATTCATGGTGAAGATATGCTTAATTTCCCACTATTTTCAGAGATATGTTTATCCATTCAGTTATCCGTCAACAGAACAA 594
A CBS94    495 CATCTAATCGGTACCTTAATCATTCATGGTGAAGATATGCTTAATTTCCCACTATTTTCAGAGATATGTTTATCCATTCAGTTATCCGTCAACAGAACAA 594
A B-4443   495 CATCTAATCGGTACCTTAATCATTCATGGTGAAGATATGCTTAATTTCCCACTATTTTCAGAGATATGTTTATCCATTCAGTTATCCGTCAACAGAACAA 594
C 1739-82  499 TATCTAATCGGTATCTTAATCATTCATGGTGAAGATATGCTTAATTTCCCACTATTTTCAGAGATATGTTTATCCATTCTGTTATCCGTCTACAGAACAA 598
A B-4414      CATCTAATCGGTACCTTAATCATTCATGGTGAAGATATGCTTAATTTCCCACTATTTTCAGAGATATGTTTATCCATTCAGTTATCCGTCAACAGAACAA
Helices       Core-closing Stem           TBE

A MYA-3404  595 CTAAAGACTGACGAAGTTTCCATCCTACATGGCACGTCTAAAATATAATGAATAACTCAGTCCTCATGACATTGTGAGCCATTTGAATCCATTTACCTTA 694
A CBS94    595 CTAAAGACTGACGAAGTTTCCATCCTACATGGCACGTCTAAAATATAATGAATAACTCAGTCCTCATGACATTGTGAGCCATTTGAATCCATTTACCTTA 694
A B-4443   595 CTAAAGACTGACGAAGTTTCCATCCTACATGGCACGTCTAAAATATAATGAATAACTCAGTCCTCATGACATTGTGAGCCATTTGAATCCATTTACCTTA 694
C 1739-82  599 CTAAAGACTGACGAAGTTTCCATCCAACATGACACGTCTAGAATATAATGAATAACTTAATCCTCGTTACATAGTGAGCCATTTGAATCCATTTACCTTA 698
A B-4414      CTAAAGACTGACGAAGTTTCCATCCTACATGGCACGTCTAAAATATAATGAATAACTCAGTCCTCATGACATTGTGAGCCATTTGAATCCATTTACCTTA
C B-4414            GACGAAGTTTCCATCCAACATGACACGTCTAGAATATAATGAATAACTTAATCCTCGTTACATAGTGAGCCATTTGAATCCATTTACCTTA
```

```
A MYA-3404  695 GTTGGCCCTTCTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCTTACACTAATAAAAACGATGTTTTATTTACATTGTATGCTGAA 794
A CBS94    695 GTTGGCCCTTCTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCTTACACTAATAAAAACGATGTTTTATTTACATTGTATGyTGAA 794
A B-4443   695 GTTGGCCCTTCTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCTTACACTAATAAAAACGATGTTTTATTTACATTGTATGCTGAA 794
C 1739-82  699 GTTGGCCCTTCTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCGTACACTAATAAAGGCGATGTTTTATTTACATTGTATGCTGAA 798
A B-4414      GTTGGCCCTTCTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCTTACACTAATAAAAACGATGTTTTATTTACATTGTATGCTGAA
C B-4414      GTTGGCCCTTTTGATAGTTTGAGATCATTCATTTGTACACCAATGATCGTGACATCCGTACACTAATAAAGGCGATGTTTTATTTACATTGTATGCTGAA
Helices                          TBE              Template

A MYA-3404  795 TATTTTTCAGGTTTGAAAGGCAACAAAATCTTCTATTTGCACTAACCTTCGTGAAAACCACGTAAAAAACTTTACCAGGATTTCGACGATTTAATTTGCT 894
A CBS94    795 TATTTTTCAGGTTTGAAAGGsAACAAATTCTTCTATTTGCACTAACCTTCGTGAAAACCACGTAAAAAACTTTACCAGGATTTCGAsGATTTAATTTGCT 894
A B-4443   795 TATTTTTCAGGTTTGAAAGGCAACAAAATCTTCTATTTGCACTAACCTTCGTGAAAACCACGTAAAAAACTTTACCAGGATTTCGACGATTTAATTTGCT 894
C 1739-82  799 TATCTTTCAGGTTTGAAAGGTAACAAAATCTTCTATTTGCACTAATCTTCGTGAAAACCACGTAAAAAGCTTTACCAGGATTTCGACGATTTAATTTGCT 898
A B-4414      TATTTTTCAGGTTTGAAAGGGAACAAATTCTTCTATTTGCACTAACCTTCGTGAAAACCACGTAAAAAACTTTACCAGGATTTCGACGATTTAATTTGCT
C B-4414      TATCTTTCAGGTTTGAAAGGTAACAAAATCTTGTATTTGCACTAATCTTCGTGAAAACCACGTAAAAAGCTTTACCAGGATTTTGACGATTTAATTTGCT

A MYA-3404  895 GGGTCATATCGAAGAATGTTGATTATTTGCTTTACACAAGGACTCTGCTTTCTAGTTTTCGAGTACTCTATGCAATAATCATGTCATTCCGGCAGATGGG 994
A CBS94    895 GGGTCATATCGAAGAATGTTGATTATTTGCTTTACACAAGGACTCTGCTTTCTAGTTTTCGAGTACTCTATGCAATAATCATGTCATTCCGGCAGATGGG 994
A B-4443   895 GGGTCATATCGAAGAATGTTGATTATTTGCTTTACACAAGGACTCTGCTTTCTAGTTTTCGAGTACTCTATGCAATAATCATGTCATTCCGGCAGATGGG 994
C 1739-82  899 GGGTCCTATCGAAGAATGTTGATTATTTGCTTTACACAAGGACTATGCTTTATGCTTTTCAGGTATACTATGCAATAATCATGTCATTCCGGCAGATGGA 998
A B-4414      GGGTCATATCGAAGAATGTTGATTATTTGCTTTACAC
C B-4414      GGGTCCTATCGAAGAATGTTGATTATT

A MYA-3404  995 CTCATCCTAAGATCAATACGAAGTTAGCTAGCAACACTTGGAAGACCCTTCTTTCTCATAAAGTTATTATTCTTTAGCCCGGGGTATTCAGCAAAATTTC 1094
A CBS94    995 CTCATCCTAAGATCAATACGAAGTTAGCTAGCAACACTTGGAAGACCTTTCTTTCTCATAAAGTTATTATTCTTTAGCCCGGGGTATTCAGCAAAATTTC 1094
A B-4443   995 CTCATCCTAAGATCAATACGAAGTTAGCTAGCAACACTTGGAAGACCCTTCTTTCTCATAAAGTTATTATTCTTTAGCCCGGGGTATTCAGCAAAATTTC 1094
C 1739-82  999 CTCATCCTAGGATCAATACGAAGTTGGCTAGCAATACTTGGAAGGCCTTTCTTTCTCATAAAGTTATTATTCTTTAGCCCGGGGTATTCAGCAAAATTTC 1098

A MYA-3404 1095 CCATACTTTTTTTTTGTTTAGTGGATTTTTT-CTTTTTT-GTTTTTGAAGGGAAATTTTAAATCAAAAATAATTATCAAACAAATAACAGGTATTAAGAG 1192
A CBS94   1095 CCATACTTTTTTTTTGTTTAGkGrATT-TTTyC-TTTTTkGTTTTTGAAGGGAAATTTTAAATCAAAAATAATTATCAAACAAATAACAGGTATTAArAG 1192
A B-4443  1095 CCATACTTTTTTTTTGTTTAGTGGATTTTTT-CTTTTTT-GTTTTTGAAGGGAAATTTTAAATCAAAAATAATTATCAAACAAATAACAGGTATTAAGAG 1192
C 1739-82 1099 CCATACTTTTTTTTTGTTTAGTGGATTTTTT-CTTTTTT-GTTTTTGAAGGGAAATTTTAAATCAAAAATAATTATCAAACAAATAACAGGTATTAAGAG 1196
Helices                pk Stem1 triple helix    pk Stem2    pk stem1     pk stem 2          Core-closing Stem

A MYA-3404 1193 ATGATGATAGGAGACCAAGCCTGGTTTTTTTTTGGTTTCACTCATTTTGCGTGGGAGTATGCACTTTGGGATGATCTATACTGGGGAGTTTGATTTACCT 1292
A CBS94   1193 ATGATGATAGGAGACCAAGCCkGGTTTTTTTTTGGTTTCACTCATTTTGCGTGGGAGTATGCACTTTGGGATGATCTATACTGGGGAGTTTGATTTACCT 1292
A B-4443  1193 ATGATGATAGGAGACCAAGCCTGGTTTTTTTTTGGTTTCACTCATTTTGCGTGGGAGTATGCACTTTGGGATGATCTATACTGGGGAGTTTGATTTACCT 1292
C 1739-82 1197 ATGATGATAGGAGACCAAGCCTGGTTTTTTTTTGGTTTCACTCTTTTTGCGTGGGATTATGCACTTTGGGATGATCTATACTGGGGAGTTTGATTTGCCT 1296
Helices       Core-closing stem

A MYA-3404 1293 ACTGTTTTTGTTTTGATTTTCTTGGGGTGCCATGTTCTTCGGATTGTGGTGG-CCGTTCCAGGATTCAATCAGAAACTTTTTTGGGTAAATTCTCCTTTC 1391
A CBS94   1293 ACTGTTTTTGTTTTGATTTTCyTGGGGTGCCATGTTCTTCGGATTGTGGT-GrCCGTTCCAGGATTCrATCAGAAACTTTTTTGGGTrAATTCTCCTTTC 1391
```

```
A B-4443   1293 ACTGTTTTTGTTTTGATTTTCTTGGGGTGCCATGTTCTTCGGATTGTGGTGG-CCGTTCCAGGATTCAATCAGAAACTTTTTTGGGTAAATTCTCCTTTC 1391
C 1739-82  1297 ACTGTTTTTGTTTTGATTTTCTTGGGGTGCCATGTTCTTCGGATTGTGGTGG-CCGTTCCAGGATTCAATCAGAAACTTTTTTGGGTGGATTCTCCTTTC 1395

A MYA-3404 1392 TTGTCTCCATTTATGGAGTCAGATAAAAAGTTGGTTGTCTTTGCTAATTCTCACTAGAGTGTTTCTGGCATTGCCGAAATCAGGTATTCCTTTTCATTGC 1491
A CBS94    1392 TTGTCyCCATTTATGGAGTCAGATAAAAAGTTGGTTGTCTTTGCTAATTCTCACTAGAGTGTTTCTGGCATTGCCGAAATCAGGTATTCCTTTTCATTGC 1491
A B-4443   1392 TTGTCTCCATTTATGGAGTCAGATAAAAAGTTGGTTGTCTTTGCTAATTCTCACTAGAGTGTTTCTGGCATTGCCGAAATCAGGTATTCCTTTTCATTGC 1491
C 1739-82  1396 TTGTCTCCATTTATGGAGTCAGATAAAAAGTTGGTTGTCTTTGCTAATTCCCACCAGAGTGTTTCTGGCATTGCCGAAATCAGGTATTCCTTTTCATTGC 1495

A MYA-3404 1492 ACAATTTCTTATCCTGCAGTGAGAAACCTAGAGATCTAACTAAAAGAAATATAGCTGTATCTTTGGTTT-TTGGTGTGAATATGGGATCACTGTAAGGAC 1590
A CBS94    1492 ACAATTTCTTATCCTGCAGTGAGAAACCTAGAGATCTAACTAAAAGAAATATAGCTGTATCTTTGGTTT-TTGGTGTGAATATGGGATCACTGTAAGGAC 1590
A B-4443   1492 ACAATTTCTTATCCTGCAGTGAGAAACCTAGAGATCTAACTAAAAGAAATATAGCTGTATCTTTGGTTT-TTGGTGTGAATATGGGATCACTGTAAGGAC 1590
C 1739-82  1495 ACAATCTCTTATCCTGCAGTGAGAAACCTAGAGATCTAACTAAAAGAAATATGGCTGTATCTTTGGTTTCTTGGTGTGAATATGGGATCACTATATGAAC 1595

A MYA-3404 1591 TAATGTCAGATTTAGGATGTTATTCAGAAAATCATGGTCTTTTCCTAACTAATTATCTCTTGTTTTTCTATTTCATAGAAGACTGGAGGGAAGAGTGGTT 1690
A CBS94    1591 TAATGTCAGATTTAGGATGTTATTCAGAAAATCATGGTCTTTTCCTAACTAATTATCTCTTGTTTTTCTATTTCATAGAAGACTGGAGGGAAGAGTGGTT 1690
A B-4443   1591 TAATGTCAGATTTAGGATGTTATTCAGAAAATCATGGTCTTTTCCTAACTAATTATCTCTTGTTTTTCTATTTCATAGAAGACTGGAGGGAAGAGTGGTT 1690
C 1739-82  1595 TAATGTCAGATTTAGGATGTTATTCAGAAAATCGTGGTCTTTTCCCAACTAATTATCTCTTGTTCTTCTATTTTGTGGAGGCTAGAGGGAAGAGTGGTT 1695

A MYA-3404 1691 GCGAGGGATTAAAATTACTCGTTGGAGTTATTCCAGTTGAGTTTTCTTGTTTACATTTAAAGTTCATTCCAGTTTCAGATCTAAGGACTTCCTAATTAGT 1790
A CBS94    1691 GCGAGGGATTAAAATTACTCGTTGGAGTTATTCCAGTTGAGTTTTCTTGTTTACATTTAAAGTTCATTCCAGTTTCAGATCTAAGGACTTCCTAATTAGT 1790
A B-4443   1691 GCGAGGGATTAAAATTACTCGTTGGAGTTATTCCAGTTGAGTTTTCTTGTTTACATTTAAAGTTCATTCCAGTTTCAGATCTAAGGACTTCCTAATTAGT 1790
C 1739-82  1696 GCGAGGGATTAAAATTACTCGTTGGAGTTCTTCCAGTTGAGTTTTCTTGTTTTCATTTAAAGTTCATTCCAGTTTCAGATCTAAGGACTTCCTAATTAGT 1795

A MYA-3404 1791 GGTTTGTTACCAATGTTCATGTCGGGCATGCTAGACTCTTGAACCGTGTTTGAAGTTTTATGTACAAAAATGGAGTAATTGGCGATCTTACTGATTCCGT 1890
A CBS94    1791 GGTTTGTTACCAATGTTCATGTCGGGCATGCTAGACTCTTGAACCGTGTTTGAAGTTTTATGTACAAAAATGGAGTAATTGGCGATCTTACTGATTCCGT 1890
A B-4443   1791 GGTTTGTTACCAATGTTCATGTCGGGCATGCTAGACTCTTGAACCGTGTTTGAAGTTTTATGTACAAAAATGGAGTAATTGGCGATCTTACTGATTCCGT 1890
C 1739-82  1796 GGTTTGTTTCCAATGTTCATGTCGGGCATGCTAGACTCTTGAACCGTGTTTGAAGTTTTATGTACAAGAATGGAGTAATTGGCGATCTTACTGGTTCCGT 1895

A MYA-3404 1891 ATTAGTTTGGGACTGAAGACAATTTGGAGTGTCCAGGATGGGGAGGACGATGAGTCGTTCTTGGAGAGGGATTTACATAGCACGAATGTGATATTGGAAT 1990
A CBS94    1891 ATTAGTTTGGGACTGAAGACAATTTGGAGTGTCCAGGATGGGGAGGACGATGAGTCGTTCTTGGAGAGGGATTTACATAGCACGAATGTGATATTGGAAT 1990
A B-4443   1891 ATTAGTTTGGGACTGAAGACAATTTGGAGTGTCCAGGATGGGGAGGACGATGAGTCGTTCTTGGAGAGGGATTTACATAGCACGAATGTGATATTGGAAT 1990
C 1739-82  1896 ATTAGTTTGGGACTGAAGACAATTTGGAGTGGCTAGGATGGGAAGGACGATGAGTCGTTCTTGGAGAGGGATTTACATAGCACGAATGTGATATTGGAAT 1995

A MYA-3404 1991 GATATGAAAGTGGGATTTTAAGATAGATTAGATTTGGGTGTCTCCTTCCTTTATTACAAACCTTTTT-TTTTTGTATGTATTAATTTTCATGTCCTAGTT 2089
A CBS94    1991 GATATGAAAGTGGGATTTTAAGATAGATTAGATTTGGGTGTCTCCTTCCTTTATTACAAACCTTTTTTTTTTTGTATGTATTAATTTTCATGTCCTAGTT 2090
A B-4443   1991 GATATGAAAGTGGGATTTTAAGATAGATTAGATTTGGGTGTCTCCTTCCTTTATTACAAACCTTTTT-TTTTTGTATGTATTAATTTTCATGTCCTAGTT 2089
C 1739-82  1996 GATATGAAAGTGAGATTTTAAGATAGATTAGATTTGGGTGTCTCCCTCCTTTATTATTCACCTTTT---TTTTGTATGTATTAATTTTCATGTCCTAGTT 2092

A MYA-3404 2090 GTTGCAAAATAAACTAGCATTACTAACTACTAATTAATTTC---TTTTACTATGTCATCTATTAATTACTATTAAAGCATAAAAATTTCTATGTGTATGA 2186
A CBS94    2091 GTTGCAAAATAAACTAGCATTACTAACTACTAATTAATTTC---TTTTACTATGTCATCTATTAATTACTATTAAAGCATAAAAATTTCTATGTGTATGA 2187
A B-4443   2090 GTTGCAAAATAAACTAGCATTACTAACTACTAATTAATTTC---TTTTACTATGTCATCTATTAATTACTATTAAAGCATAAAAATTTCTATGTGTATGA 2186
```

C 1739-82 2092 GTTGCAAAATAAACTAGCATTACTAAC**A**ACT**T**ATTA**T**TTT**ATTT**TTTTACTATGT**T**ATCTATTAATTACTATTAAAGCATAAAAATTTCTAT**A**TGTATGA 2192

A MYA-3404 2187 AAGAAAAATAGTCAATACATAACTAACACCTTTCTTTTCAGTTTGTAGAATATGGTTGATTCTTCGAGGTTATTCTGTAACTATTAAT 2274
A CBS94   2188 AAGAAAAATAGTCAATACATAACTAACACCTTTCTTTTCAGTTTGTAGAATATGGTTGATTCTTCGAGGTTATTCTGTAACTATTAAT 2275
A B-4443   2187 AAGAAAAATAGTCAATACATAACTAACACCTTTCTTTTCAGTTTGTAGAATATGGTTGATTCTTCGAGGTTATTCTGTAACTATTAAT 2274
C 1739-82 2193 AAGAAAAA**C**AGT**T**AATACATAAC**C**A**GTT**CCTTTCTTTTCAGTTTGTAGAATAT**AT**TTGATTCTTC**T**AGGT**A**ATTCTGT**G**ACTATTAAT 2280

Table 3.1. List of oligonucleotides used in this study. A) A probe and C probe are the oligonucleotides specific for the A-type repeat and C-type telomeric repeats, respectively. *TER1*L and *TER*1R are PCR primers for probe used in RFLP analysis for *TER1* region. B) and C) list all the oligonucleotides used as PCR and sequencing primers for *TER1* region in A-strain B-4443 and A/C-strain 1739-82. D) lists all the oligonucleotides used as allele-specific PCR and sequencing primers for *TER1* template and surrounding regions in A/C-strain B-4414. E) lists all the oligonucleotides used as PCR primers for probes used in RFLP analysis for *TER1* surrounding regions on Supercontig 1 and other Supercontigs. 01L to 08R are primers for RFLP analysis in *TER1* surrounding regions on Supercontig 1. S2L to S8R are primers for RFLP analysis in other supercontigs. S stands for Supercontig, and the following number represents the supercontig number. *EST2*_L and *EST2*_R are primers for RFLP analysis in annotated *EST2* gene on Supercontig 5. *EST1*_L and *EST1*_R are primers for RFLP analysis in annotated *EST1* gene on Supercontig 11.

Name		Sequence (5'—3')	Name		Sequence (5'—3')
A) Southern	A probe	GTGTAAGGATG	E) PCR primers for RFLP probes	01L	TTGAGGAGAAGAATTGAACGTG
	C probe	ACGGATGTCACG		01R	TCAAGTTACAGCTGCAAGGAAA
	*TER1*L	TCAGGTTTGAAAGGCAACAA		02L	CAACAATTCGGTTGCAAAACT
	*TER1*R	TCGAAAACTAGAAAGCAGAGTCC		02R	ATTTGTCATTACGTCGCTGGT
B) PCR and sequencing primers for A-strain B-4443 and A/C-strain 1739-82	L01	ATTCTTCGATATGACCCAGCA		03L	GAAATCAAAAGTGTGTATACAGCAAA
	L02	AAAACCAGGCTTGGTCTCCTA		03R	GTGGTGTGCAGGAGGAGAAG
	L03	TTAAAATTTCCCTTCAAAAACAAAA		04L	TGGTTGAAAGGTGATAGTCAAGG
	L04	TTCTGATTGAATCCTGGAACG		04R	CGTTGTCTTGGATTTGATGG
	L05	ATGGGAAATTTTGCTGAATACC		05L	ATCCACACACTCGGACTCCT
	L06	ATTTTAATCCCTCGCAACCAC		05R	GGAAGAAGTTTTCCAGCCTTT
	L07	CAAAACCACCGAATCCAAAC		06L	TGTAGATGAAGATGAAGATGAAGACA
	R01	TCGCTTCGTCTTTCACAAGTT		06R	AAGCAACAAAAACTACAGCAAAGA
	R02	TTCACTGGTATCCATGGAAGTTT		07L	TTGAACTCAAATCCATCGGTAA
	R03	GCAATGAAAAATGTCGCAAAT		07R	TTGGATGTTGAAGAATGTCCTG
C) Additional PCR and sequencing primers for	L08	CTTCCTTCTTTTCAGCAGG		08L	CAATCATGGATGACGCAGAA
	R04	TTTGTGCTCTCAGGAGCTGT		08R	CCTGAGCCCTGTCTGGTCTA
	R05	AATCGTGGTCTTTTCCCAACT		S2L	TTGCTCTGTTCTGTTCTGGTG
	R06	GCATTGCCGAAATCAGGTAT		S2R	GCTGCTTGTTTATCCGCAAT
	R07	AAGGACTATGCTTTATGCTTTTCAG		S3L	TGGAAATCCATACTATCACTAACATTT
	R08	GAATATTCATTGTTACATTGATTCCC		S3R	ACGTCCTGCTCAACGTTTT
D) Allele specific PCR and sequencing primers for A/C-strain B-4414	*TER1*A0	AACATCCTAAATCTGACATTAGTCCT		S4L	AAGCATGGTGTGGTTCTTCC
	*TER1*A0	GATTATTGCATAGAGTACTCGAAAAC		S4R	CCTGCTGGACCACAAGAAGT
	*TER1*A0	GTAAAGTAAACTTGTGAAAGACGAA		S5L	TTTGCTTCTGGAGAAATGAAGA
	*TER1*A0	TGTGAATATGGGATCACTGTAAGG		S5R	ATCAAGAACGCCACCAATTC
	*TER1*A0	GGACTCTGCTTTCTAGTTTTCGA		S6L	GCCATTTGAGTGCTCCAGTT
	*TER1*A0	CATACAGTTAATTAGCACCATCATG		S6R	TGCAAAATACGAAGAACACGA
	*TER1*C0	AACATCCTAAATCTGACATTAGTTCA		S8L	CCAAACCCATCTCCGATTTA
	*TER1*C0	TTGCATAGTATACCTGAAAAGCAT		S8R	TGTTGGAAGAAGAAGACAATGG
	*TER1*C0	AAAGACGAAAAGTAAATTCTCCCTC		*EST2*	TCACTGGCAAACCTCTTCAA
	*TER1*C0	GGTGTGAATATGGGATCACTATATGA		*EST2*	CAAGGTTTCCCTTAGCATATTTTG
	*TER1*C0	AAGGACTATGCTTTATGCTTTTCAG		*EST1*	CAAAAAGATTCCATTTTCCAAGA
	*TER1*C0	GAATATTCATTGTTACATTGATTCCC		*EST1*	TCACTGGCAAACCTCTTCAA

Table 3.2. Heterozygosity detected as RFLPs in three *C. tropicalis* strains.

Denominator represents the total number of alleles tested in the RFLP analysis, assuming each strain is diploid. For the *TER1* region, 10 digestions in total were performed, so 20 allelic tests in total were performed for each strain. Numerator represents the number of digested DNA regions showing RFLPs between alleles in the same strain. For the eight DNA sequences on the same supercontig as *TER1* (not counting *TER1* itself) and for the eight sequences on different supercontigs, 6 digestions were performed for each strain for RFLP analysis.

	(A) B-4443	**(A/C) B-4414**	**(A/C) 1739-82**
TER1	**1/20 (5%)**	**5/20 (25%)**	**0/20 (0%)**
Same Supercontig (total)	**4/96 (4%)**	**15/96 (16%)**	**3/96 (3%)**
99 kb upstream	1/12	0/12	0/12
18 kb upstream	0/12	2/12	1/12
7 kb upstream	0/12	3/12	1/12
2 kb upstream	0/12	2/12	0/12
3 kb downstream	0/12	4/12	1/12
8 kb downstream	0/12	1/12	0/12
18 kb downstream	2/12	2/12	0/12
98 kb downstream	1/12	1/12	0/12
Different Supercontigs (total)	**5/96 (5%)**	**22/96 (23%)**	**28/96 (29%)**
Supercontig 2	1/12	3/12	5/12
Supercontig 3	1/12	3/12	3/12
Supercontig 4	0/12	3/12	5/12
Supercontig 5	1/12	2/12	2/12
Supercontig 6	0/12	4/12	4/12
Supercontig 8	1/12	1/12	3/12
EST1 (Supercontig 11)	0/12	4/12	2/12
EST2 (Supercontig 5)	1/12	2/12	4/12

Table 3.3. Allelic RFLP comparisons between *C. tropicalis* A- and A/C-strains. Sequence of A-strain MYA-3404 is from *Candida* database (http://www.broadinstitute.org/). Denominator represents the total number of alleles tested in RFLP analysis. Numerator represents the number of alleles that show RFLPs between two strains. All strains were presumed to be diploid for this analysis.

	(A) B-4443 to (A) MYA-3404	(A) B-4443 to (A) B-4443	(A/C) B-4414 to (A) B-4443	(A/C) B-4414 to (A/C) 1739-82	(A/C) 1739-82 to (A) B-4443
TER1	1/20 (5%)	0/20 (0%)	7/20 (35%)	6/20 (30%)	11/20 (55%)
Same Supercontig	4/96 (4%)	0/96 (0%)	15/96 (16%)	47/96 (49%)	58/96 (60%)
Other Supercontigs	5/96 (4%)	0/96 (0%)	27/96 (28%)	24/96 (25%)	37/96 (39%)

CHAPTER 4

CONCLUSIONS AND PERSPECTIVES

Identification and characterization of telomerase RNAs from organisms has been one of the most important topics ever since the establishment of the telomere and telomerase field in the 1970s. The new method and results reported in this thesis has expanded our knowledge on yeast telomerase RNAs.

In this dissertation, we designed TRFolder, a program that is capable of telomerase RNA core structure prediction in distantly related species independent of sequence comparative analysis. For prediction and study of large non-coding RNAs with complicated secondary structure like telomerase RNAs, structural profiling is much more efficient than sequence alignment. Although the current TRFolder program has certain limitations, it is the first program that is specific for telomerase RNA prediction, and is able to find a pseudoknot structure with a triple helix, together with other key structural features of TRs over a broad phylogenetic range of yeast species. Previous studies were mainly based on homology comparison analysis and manual alignment of specific protein binding regions, both of which require a very close relationship between the TRs being studied. Besides the universally conserved core structural elements, each group of telomerase RNAs have their own specific structural or sequence elements. For example, the 3' half of vertebrate

TR contains the CR4-CR5 domain and Box H-ACA conserved motifs, which are essential for telomerase activity and stability in vivo (Mitchell *et al.* 1999; Tesmer *et al.* 1999). More specifically, a highly conserved stem loop structure in the CR4-CR5 domain named P6.1 is critical for telomerase activity (Chen *et al.* 2002). A conserved three-way junction (TWJ), a structure similar to the P6.1 stem loop in vertebrate TRs, has been identified in *Saccharomyces* and *Kluvyeromyces* TRs (Brown *et al.* 2007; Gunisova *et al.* 2009). In ciliate TRs, a conserved sequence motif flanking the template boundary element exhibits high affinity to TERT (Miller *et al.* 2000; Lai *et al.* 2002). In the future, these features can be added as filters to TRFolder if more detailed structural predictions are needed in specific species. The specific filters together with a customized weighting system can surely help to broaden the usage of TRFolder into TRs other than yeast, such as vertebrates and ciliates. Additionally, further structure studies, such as NMR, or structure disruption by mutation can help us to verify the proposed structures and to improve the accuracy and efficiency of our TRFolder program. The algorithm used in TRFolder can also be applied to other large non-coding RNAs with complicated structures, pseudoknot and triple helix for example. TRFolder and its algorithm will contribute in identification and study of telomerase RNA structure as well as other large non-coding RNAs.

Also in this study, we showed that the presence of two types of telomeric repeats (A-repeat and C-repeat) in *C. tropicalis* strain B-4414 is associated with the presence of two different alleles of the telomerase RNA gene *TER1*. At least two future experiments

could be done to fully demonstrate that C-repeats are synthesized from the telomerase RNA with the C-template and A-repeats are synthesized from the telomerase RNA with the A-template. First, a *TER1* with a C-template could be cloned and introduced into a *C. tropicalis* A-strain such as B-4443. Then it could be checked if the C-repeats are incorporated into the newly synthesized telomere ends by either hybridization with a probe specific to C-repeats, or by cloning and sequencing of telomeres in the newly constructed A-strain. Both integrative (Haas *et al.* 1990; Hara *et al.* 2001) and autonomously replicating (Hara *et al.* 2000) plasmids are available in *C. tropicalis*. A mutated version of hygromycin B resistance gene (HYG^r), with all nine CTG codons in the coding sequence of HYG changed to another leucine codon (CTC) by site-directed mutagenesis, can be successfully expressed in *C. tropicalis* as a dominant selective marker (Hara *et al.* 2000). Second, primer extension assays using both versions of telomerase from appropriate strains could be performed to answer if C-repeats can be added by telomeric synthesis only in the presence of a C-template *in vitro*. Active telomerase fractions from *C. tropicalis* can be prepared as previously described for *C. albicans* (Singh *et al.* 2002), with modifications, if necessary.

We have shown that the telomerase RNA with a C-template effectively has a 33 bp template with 10 nt terminal repeats, compared to a 32 bp template with 9 nt terminal repeats in the telomerase RNA encoded by the A-template *TER1*. Previous studies showed that although the telomerase RNA templates in yeast species are highly divergent both in sequence and length, they can be discerned to be able to encode a conserved Rap1

binding sequence (Cohn *et al.* 1998; Wang *et al.* 2009). In Figure 3.2C, both versions of TR template in *C. tropicalis* are aligned with the template in *C. albicans* and *K. lactis*. Both TR templates in *C. tropicalis* maintain the conserved sequence of a Rap1 binding site. Additionally, in closely related species, a small number of nucleotide differences between the TR templates can result in an extension of the template and/or a shift in the permutation of the template. A comparison of telomerase RNA template sequences from related yeasts suggests that this sort of shift in the size and position of template terminal repeats is a regular feature of telomerase RNA evolution (Wang *et al.* 2009). For example, *K. wickerhamii*, a *Kluvyeromyces* species closely related to *K. lactis*, has a 32 nt template that is perfectly aligned with the 34 nt template in *K. lactis*, except for the last 2 nucleotides at the 5' end direct repeat. The two nucleotide differences disrupt their alignment potential with the 3' direct repeat of the template in *K. wickerhamii*, which results in a shortened template bordering by 7 nt direct repeats with a mismatch, instead of 9 nt direct repeats with a mismatch as in *K. lactis* (Figure 1.2). In the *C. tropicalis* C-template, 33G provides an additional base pair between the telomeric repeat and telomerase RNA template and can now considered to be the new 5' boundary of the template. The 10G in *C. tropicalis* C-template aligns with 9G in *C. albicans* template (Figure 1.2). The 3' end of the *C. tropicalis* C-template matches the *C. tropicalis* A-template and its 5' end matching the *C. albicans* template. A detailed study of template sequences in other *Candida* species and *C. tropicalis* strains should give us a better idea about the history of template evolution.

Although there is no reason to believe that the *C. tropicalis* telomerase can copy telomeric sequence further along the C-template than it can using an A-template, the identification of C-template in *C. tropicalis* might capture an intermediate step of template extension and a shift in template permutation. The expansion of potential template size is probably the first step of template boundary redefinition and/or template shift of permutation. While the direct repeats bordering the TR template are involved in the translocation of telomeric repeat synthesis, the template boundary element of the RNA functions as a barrier for reverse transcription beyond the end of the template. Mutations in this element can lead to the synthesis of a telomeric repeat that is lengthened by sequence copied from beyond the normal template (Autexier and Greider 1995; Prescott and Blackburn 1997; Miller *et al.* 2000; Tzfati *et al.* 2000). If the expansion of the potential template region is followed a shift of the template boundary element position, the evolutionary shift of template permutation will eventually arise. Two ways for a template boundary element to shift over a nucleotide would be an insertion of a base between the template and the boundary element or by a base change that disrupted the pairing of the most basal base pair in the stem of the boundary element.

Our results showed that the presence of highly divergent alleles and different DNA content in specific *C. tropicalis* strains supports the possibility that genetic recombination can take place in this "asexual diploid" yeast. While originally designated as a cluster of asexual yeast species, the *Candida* clade is now known to have species with a sexual or a parasexual cycle. Many remaining *Candida* spp may have sexual or parasexual cycles

that have yet to be discovered. Although mating in *C. tropicalis* hasn't been directly observed, our study strongly suggest that A/C-strain B-4414 and 1739-82 could be products of mating between two highly divergent *C. tropicalis* strains. The previous evidence of mating in diploid yeast were in laboratory constructed *C. albicans* and *C. dubliniensis* species. It is striking since our results suggest that *C. tropicalis* can mate in nature. In fact, haploid species in the *Candida* clade, e.g. *C. lusitaniae* and *C. guilliermondii*, are infrequent human pathogens while the diploid *Candida* species (*C. albicans* and *C. tropicalis* in particular) are the major causes of *Candidiasis* (Alby *et al*. 2010). It is possible that the increase of ploidy will enhance the pathogenesis in these species, perhaps because of the expansion of gene families related to virulence. In addition, aneuploidy is a common method for *C. albicans* to adapt to different environmental stresses (Alby and Bennett 2010). *Candida* pathogens including *C. tropicalis* and *C. albicans* may use the parasexual cycle to enlarge ploidy and generate multiple aneuploidy combinations, in order to increase their fitness and develop pathogenesis. Given the medical importance of *Candida tropicalis*, our study will facilitate the expansion of the current understanding of sexual reproduction and pathogenesis in *C. tropicalis* and other pathogenic *Candida* species.

References

Alby, K. and Bennett, R. J. (2010) 'Sexual reproduction in the *Candida* clade: cryptic cycles, diverse mechanisms, and alternative functions', *Cell Mol Life Sci* **67**: 3275-3285.

Autexier, C. and Greider, C. W. (1995) 'Boundary elements of the *Tetrahymena* telomerase RNA template and alignment domains', *Genes Dev.* **9**: 2227-2239.

Brown, Y., Abraham, M., Pearl, S., Kabaha, M. M., Elboher, E. and Tzfati, Y. (2007) 'A critical three-way junction is conserved in budding yeast and vertebrate telomerase RNAs', *Nucleic Acids Res.* **35**: 6280-6289.

Chen, J. L., Opperman, K. K. and Greider, C. W. (2002) 'A critical stem-loop structure in the CR4-CR5 domain of mammalian telomerase RNA', *Nucleic Acids Res* **30**: 592-597.

Cohn, M., McEachern, M. J. and Blackburn, E. H. (1998) 'Telomeric sequence diversity within the genus *Saccharomyces*', *Curr. Genet.* **33**: 83-91.

Gunisova, S., Elboher, E., Nosek, J., Gorkovoy, V., Brown, Y., Lucier, J. F., Laterreur, N., Wellinger, R. J., Tzfati, Y. and Tomaska, L. (2009) 'Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements', *RNA* **15**: 546-559.

Haas, L. O., Cregg, J. M. and Gleeson, M. A. (1990) 'Development of an integrative DNA transformation system for the yeast *Candida tropicalis*', *J Bacteriol* **172**: 4571-4577.

Hara, A., Arie, M., Kanai, T., Matsui, T., Matsuda, H., Furuhashi, K., Ueda, M. and Tanaka, A. (2001) 'Novel and convenient methods for *Candida* tropicalis gene disruption using a mutated hygromycin B resistance gene', *Arch Microbiol* **176**: 364-369.

Hara, A., Ueda, M., Misawa, S., Matsui, T., Furuhashi, K. and Tanaka, A. (2000) 'A mutated hygromycin resistance gene is functional in the n-alkane-assimilating yeast *Candida tropicalis*', *Arch. Microbiol.* **173**: 187-192.

Lai, C. K., Miller, M. C. and Collins, K. (2002) 'Template boundary definition in *Tetrahymena* telomerase', *Genes Dev* **16**: 415-420.

Miller, M. C., Liu, J. K. and Collins, K. (2000) 'Template definition by *Tetrahymena* telomerase reverse transcriptase', *EMBO J* **19**: 4412-4422.

Mitchell, J. R., Cheng, J. and Collins, K. (1999) 'A box H/ACA small nucleolar RNA-like domain at the human telomerase RNA 3' end', *Mol Cell Biol* **19**: 567-576.

Prescott, J. and Blackburn, E. H. (1997) 'Telomerase RNA mutations in *Saccharomyces cerevisiae* alter telomerase action and reveal nonprocessivity *in vivo* and *in vitro*', *Genes Dev.* **11**: 528-540.

Singh, S. M., Steinberg-Neifach, O., Mian, I. S. and Lue, N. F. (2002) 'Analysis of telomerase in *Candida albicans*: potential role in telomere end protection', *Eukaryot. Cell.* **1**: 967-977.

Tesmer, V. M., Ford, L. P., Holt, S. E., Frank, B. C., Yi, X., Aisner, D. L., Ouellette, M., Shay, J. W. and Wright, W. E. (1999) 'Two inactive fragments of the integral RNA cooperate to assemble active telomerase with the human protein catalytic subunit (hTERT) in vitro', *Mol Cell Biol* **19**: 6207-6216.

Tzfati, Y., Fulton, T. B., Roy, J. and Blackburn, E. H. (2000) 'Template boundary in a yeast telomerase specified by RNA structure', *Science* **288**: 863-867.

Wang, Z. R., Guo, L., Chen, L. and McEachern, M. J. (2009) 'Evidence for an additional base-pairing element between the telomeric repeat and the telomerase RNA template in *Kluyveromyces lactis* and other yeasts', *Mol Cell Biol* **29**: 5389-5398.