The Advantage of Instability in Maintaining Function Within the Repetitive CTD of RNA Polymerase II

A thesis submitted by

Alexandra E. Exner

In partial fulfillment of the requirements for the degree of
Master of Science
in
Biology

TUFTS UNIVERSITY
August 2016
Advisor: Dr. Stephen Fuchs

Abstract

The carboxyl-terminal domain (CTD) of Rpb1, the largest subunit of RNA polymerase II, is composed of a tandemly repeating 7 amino acid sequence that is conserved amongst eukaryotes. Previous studies in *S.cerevisiae* have shown that mutations that shorten the CTD to eight repeats significantly impair growth, and spontaneous suppressor mutations arise in which the CTD coding region expands to restore growth. Here, we introduced stop codons after the first eight repeats to measure the ability of yeast to catalyze contraction of the CTD coding region in order to delete the stop codon containing repeats. We propose that the *RPB1* gene has evolved with more than the necessary number of repeats to be able to contract to delete mutations while still maintaining functions essential for transcription. We determined that contractions are likely occurring through a Rad5p mediated pathway that leads to replication fork regression and misalignment due to regions of microhomology.

Acknowledgements

Firstly, I would like to thank my advisor, Dr. Stephen Fuchs for welcoming me into his lab. Thank you for your guidance and support over the last few years. I really appreciate how you have always been there to answer questions and have encouraged me to be a better scientist. I would also like to thank my lab members, past and present. Thank you Adam for all your help solving problems in the lab and taking the time to explain concepts whenever I had questions. Thank you Mike, Ish, Kristina, Summer, Brad, and Mohammed for your insights and making the lab a great place to work. Thank you to my committee members Stephen Fuchs, Mitch McVey, and Juliet Fuhrman for your support and encouragement. Thank you Mitch for teaching me how to be a good teacher in Pedagogy, and inspiring my interest in DNA repair in your Molecular Biology course. Thank you Juliet for letting me be your TA for Experiments in Molecular Biology, and always being so helpful.

Thank you to my friends and family for being there when I need you. Special thanks to my Mom for believing in me and encouraging my decisions. Thank you to Tom Dixon for being so reassuring and bringing me coffee while I wrote this thesis. Thank you Emma for being my best friend for the last 18 years.

Thank you to the entire Tufts Biology Department and Tufts University for inviting me to be a part of your wonderful community and encouraging and inspiring me throughout my time here.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Repetitive sequences and their propensity for instability

The C-terminal domain of RNA Polymerase II

RNA polymerase II is an essential protein complex with a unique unstructured repetitive C-terminal domain (CTD) tail attached to its largest subunit Rpb1p. The CTD tail is composed of tandem repeats of the amino acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser ($Y_1S_2P_3T_4S_5P_6S_7$). The repetitive sequence is highly conserved amongst Eukaryotes with 26 repeats found in yeast and 52 in humans (Allison, Moyle, Shales, & James Ingles, 1985; Yang & Stiller, 2014). In general, the number of repeats is greater in more complex organisms; however, this is not always the case (Yang & Stiller, 2014). The CTD is an integral part of the transcription machinery and its repetitive nature may play a role in ensuring its essential functions are never compromised.

Variation in the number of CTD repeats

There is a general correlation between CTD length and complexity, however, the simple organism the *Hydra* has a greater number of repeats than humans (60 and 52 repeats respectively). There is a theory that some of the "extra" repeats seen in the *Hydra* were only amplified recently and will soon be lost, resulting in a CTD length that parallels the *Hydra's* complexity (Yang & Stiller, 2014). This presents the CTD as a dynamic structure that maybe functions best in a given organism with a number of repeats within a certain range rather than a very specific number. In our previous paper, 36 strains of *Saccharmoyces*

cerevisiae were sequenced and aligned to examine natural variation (Morrill, Exner, Babokhov, Reinfeld, & Fuchs, 2016). These strains had between 24 and 26 repeats, and several SNPs, insertions, and deletions of repeats were seen. Though the amino acid sequence is generally the same in each repeat, the nucleotide sequence varies making it possible to identify specific repeat insertions/deletions. These findings prompt a number of questions about the CTD such as: how many repeats are necessary? And do different repeats have different functions? Previous research showed that *S. cerevisiae* is viable if the CTD tail has at least 10 repeats (Nonet, Sweetser, & Young, 1987). CTD constructs with 10-12 repeats were conditionally viable showing sensitivity to different conditions and those with at least 13 repeats showed wild type growth. This seems to indicate that there are extra repeats. Or possibly a longer tail is simply better able to interact with the rest of the protein complex during interactions necessary for transcription. Potentially some of the repeats are actually "extra" and serve as templates to be used in DNA repair. It is possible that having extra repeats provides a template for homologous recombination in case a mutation occurs within one of the repeats. Then the mutation could be excised and the gap filled in using the sequence from another repeat as a template. The extra repeats may also be there to prevent loss of function. If some repeats are not fully functional as a result of mutations or errors in DNA replication, repeats could simply be deleted by the cell to ensure essential functions are never lost.

Functions of the CTD tail

RNA polymerase II (RNAPII) is responsible for the transcription of DNA into RNA (mRNA as well as small noncoding RNAs) (Nonet et al., 1987). The CTD tail extends from the Rpb1p subunit close to the RNA exit channel, and functions as a scaffold for factors involved in transcription and RNA processing (Phatnani & Greenleaf, 2004). These factors are recruited to the site of transcription in response to patterns of post translation modifications on the CTD (Fuchs, Laribee, & Strahl, 2009; Phatnani & Greenleaf, 2006). Initially when RNA polymerase II is at the promoter of a gene it is generally unphosphorylated and is interacting with the mediator complex and other factors necessary for preinitiation (Komarnitsky, Cho, & Buratowski, 2000; Plaschka et al., 2015). The mediator complex interacts directly with RNAPII near the CTD and is thought to aid in phosphorylation of Ser5 of the CTD (Plaschka et al., 2015). Phosphorylation at Ser5 peaks at promoter regions and is associated with the initiation of transcription (Komarnitsky et al., 2000). Serine 5 phosphorylation is decreased during elongation and elongation of the transcript is promoted by increased Ser2 phosphorylation and binding of elongation factors (Carty & Greenleaf, 2002; Komarnitsky et al., 2000). At the termination of transcription, Ser2 phosphorylation predominates and 3'- end processing factors associate with the CTD (Ahn, Kim, & Buratowski, 2004). The phosphate groups are removed and RNAPII is ready to begin transcription again (Fuchs et al., 2009). Though patterns of phosphorylation are known, it is not exactly known how many repeats are phosphorylated at different time points, or which repeats are modified

(Phatnani & Greenleaf, 2006). This raises interesting questions about whether the repeats are all equally important or if some have different functions. The prolines in the repeats are also an additional regulatory control and can be found in either *cis* or *trans* conformation. Changing conformation allows the CTD to be a fluid structure and to bring transcription factors into contact with the transcript as it is being processed. Proline conformations may also be important for the recruitment of specific factors. For example, the phosphatase Ssu72 that removes the phosphates from Ser5, only recognizes the substrate with Pro6 in the *cis* conformation (Werner-Allen et al., 2010). There are likely other regulatory functions of the proline conformations as well. The specific composition of amino acids within the repeats of the CTD is integral to its function. While the repetitive nature of the CTD is important for its functions in transcription, we consider the evolution of the number of repeats that comprise the CTD.

The CTD is essential for the process of transcription and its distinctive repetitive amino acid structure is important for recruiting transcription factors and bringing them to the transcription site. Since we have seen that many of the repeats can be deleted while still maintaining normal growth, there must be an advantage to having so many seemingly extra repeats. We propose that the CTD has evolved with more than the necessary number of repeats to be able to cope with mutations, polymorphisms, and rearrangements. Having plasticity allows the CTD to expand and contract without sacrificing viability. Here we investigate this concept and how such events may occur.

Repetitive Amino Acid Sequences

The presence of fluctuations in the CTD repeat number is not particularly surprising since repetitive sequences tend to be unstable and prone to expansions and contractions (Pearson, Edamura, & Cleary, 2005). There are different types of repetitive sequences, repetitive amino acid sequences and repetitive nucleotide sequences. Both of these types of repeats have been shown to change in repeat number through expansion and contraction events, though this instability results in diverse consequences.

Repetitive amino acid sequences make up approximately 15-20% of the human genome and are often associated with structural fibers such membrane proteins, fibrous muscle proteins, neurofilaments, or spider silks (Beckwitt, Arcidiacono, & Stote, 1998; Mularoni, Ledda, Toll-Riera, & Alba, 2010). In yeast, the ability of cells to adhere to one another in "flocs" (flocculation) has been linked to repetitive sequences within the *FLO* genes (Verstrepen, Jansen, Lewitter, & Fink, 2005; Verstrepen & Klis, 2006). This cell-cell adhesion allows cells to form relatively large mats consisting of thousands of cells, which can help keep them protected in stressful conditions, or adapt to new environments. Cells' adhesion capabilities are the result of "adhesin" proteins on the cell surface, which have a large middle domain composed of serine- and threonine-rich repeats. In yeast, the genes *FLO5, FLO9, FLO10*, and *FLO11* encode either cell-cell adhesion proteins or proteins with the ability to adhere to other substrates. Longer stretches of repeats in the adhesin genes result in superior adhesive abilities (Verstrepen et al., 2005). Significant variation in the length of these

repeats is seen in *S. cerevisiae* and this variation is thought to be the result of instability within this region. Expansions and contractions often occur through strand slippage or recombination events during replication of this repetitive sequence (Verstrepen et al., 2005; Verstrepen & Klis, 2006). In the pathogenic strain *C.albicans* frequent changes in repeat number help to create new variability in the cell surface leading to protection against host immune response (Hoyer, Green, Oh, & Zhao, 2008). The adhesion genes are also activated as a stress response to help protect cells in the middle of a floc, or allow cells to float in aqueous environments (Verstrepen & Klis, 2006). The repetitive nature of these adhesin genes results in their frequent expansion and contraction resulting in altered adhesive capabilities and the ability to adapt to new environments.

Stretches of repetitive amino acids sequences can also play a role in physical morphology. An interesting study showed that variation in tandem amino acid repeats has been linked to phenotypic differences between dog breeds (Fondon & Garner, 2004). Repeat contraction events in the coding region of the *Alx-4* gene was linked to bilateral polydactyly (additional rear claw) unique only to the Great Pyrenees and absent in 88 other breeds. During the same study, researchers discovered another phenotype that was attributed to variation in number of amino acid repeats within a gene. *Runx-2*, a regulator of osteoblast formation, encodes 18-20 glutamines followed by 12-17 alanines (23Q and 17A in humans). Variations in the Q/A ratio were linked to craniofacial differences seen across dog breeds (Fondon & Garner, 2004). Using 3D computer modeling and sequencing, a lower ratio of Q/A was linked to a shorter skull, mid-face

hypoplasia, and a low nasal bridge. This low Q/A ratio has also been seen in humans with cleidocranial dysplasia (CDD) a disease that results in skeletal and cranial irregularities (Fondon & Garner, 2004; Mundlos, Otto, & Mundlos, 1997). In the cases of these genes, expansions or contractions within regions of amino acid repeats can alter morphology or cause physical defects. Though there may be evolutionary advantages of such repetitive sequences, they can also be regions of detrimental instability.

Repetitive DNA sequences

Repetitive regions of DNA are often associated with unstable regions of the genome and are frequently connected to disease. These regions of DNA are prone to expansion events during replication leading to adverse consequences. Instability within regions of DNA repeats has been linked to at least 40 human diseases including Huntington's disease, Friedrich's ataxia, and myotonic dystrophy (Shah & Mirkin, 2015). While trinucleotide repeats (TNRs) make up the largest category of DNA repeats linked to disease, there are also tetranucleotide, pentanucleotide, minisatellite and megasatelite DNA repeats that causes various human diseases (Mirkin, 2006; Pearson et al., 2005). Pentanuclotide expansions for example, have been linked to spinocerebellar ataxia, while ministaellite expansions are associated with epilepsy (Pearson et al., 2005). Specific trinucleotides are associated with certain diseases; Friedrich's ataxia for example is caused by expansions of $(GAA)_n$ repeats within the first intron of the frataxin gene, and it has been shown that the greater the number of

GAA repeats, the higher the likelihood of expansion (Mirkin, 2006). This is a general principle of instability that has been shown in regions of repetitive DNA; the longer the stretch of repeats, the more likely expansions are to occur (Pearson et al., 2005).

The role of secondary structures in expansions and contractions within repetitive regions

Expansion and contraction events in repetitive elements tend to occur more frequently than point mutations since these sequences are likely to cause strand-slippage or other obstacles during replication (Fondon & Garner, 2004). If there is slippage of the nascent strand this will result in an expansion in the following round of replication (Mirkin, 2006). Repetitive elements can cause secondary structures, which can lead to increased strand-slippage and expansions or contractions during replication. For example, repeats of CGG, CTG and CAG have all been shown to form hairpin-like structures (Mirkin, 2006). The GAA repeats associated with Friedrich's ataxia have been seen to form an intermolecular triplex, which interferes with transcription elongation leading to expansions. Inverted repeats are capable of forming hairpin or cruciform structures which can lead to double strand breaks which often cause expansions/contractions in the repair process (Shah & Mirkin, 2015). Secondary structures can also block polymerases during replication and cause replication fork stalling or fork reversal. Fork reversal can lead to a structure similar to a

Holliday junction, which when resolved can result in a double strand break (Shah & Mirkin, 2015; Wang & Vasquez, 2014).

The composition of bases within repeats also affects secondary structure formation and stability. G-quadruplexes (G_4-DNA) can form when repeats are guanine rich. When a single strand has at least four stretches of three or more guanine residues it can form an intramolecular G_4-DNA structure, and when multiple strands associate they form intermolecular G_4-DNA. The G-quadruplex is created when three G-quartets are stacked (Byrd & Raney, 2015; Lopes, Kriegsman, Foiani, & Nicolas, 2011). Work done in *S. cerevisiae* has shown that the helicase Pif1p binds to and unfolds G_4 structures. The CTD tail of RNAPII has a greater quantity of cytosine bases on its coding strand resulting in a guanine rich noncoding strand. Previously we have seen by circular dichroism that regions of the CTD coding region are capable of forming structures similar to what has been reported for complex mixed parallel/antiparallel G_4-DNA (Morrill et al., 2016). Potentially G_4-DNA or other secondary structures are involved in the stability of the CTD.

These secondary structures are likely to form on single stranded DNA during replication when the repetitive strands unwind and separate at the replication fork (Mirkin, 2006). Depending on which strand the structures form on, expansions or contractions are more likely to occur. If a structure forms on the newly forming lagging strand, an expansion would occur, while if the structure forms on the lagging strand template, a contraction would transpire.

DNA repair involvement in expansions and contractions

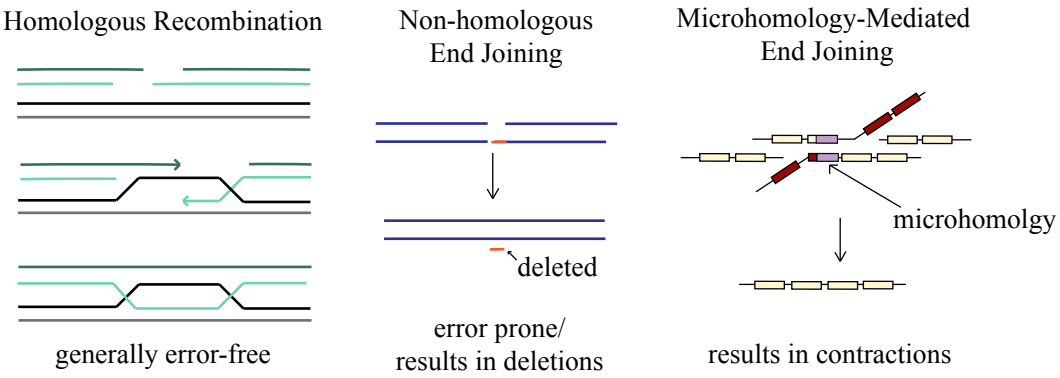


Figure 1. **DNA double strand break repair mechanisms that may be involved in expansions and contraction events.** Homologous recombination is a generally error free pathway that utilizes a sister chromosome/homologous template to repair a double strand break. Non-homologous end joining simply re-ligates broken ends and often results in deletions when overhangs are resected. Microhomology-mediated end joining uses short regions of 5-25 bases of homology to re-anneal strands and repair double strand breaks. This mechanism often leads to contractions as the DNA between the regions of homology is deleted.

Double strand breaks that occur during replication of repetitive elements can also lead to contraction or expansion events. There are primarily two mechanisms of repairing double strand breaks (DSBs) in eukaryotic cells, homologous recombination (HR) and non-homologous end joining (NHEJ) (Shrivastav, De Haro, & Nickoloff, 2008) (figure 1). Generally HR is thought of as "error-free", while NHEJ is considered "error prone". A repair pathway is though to be error-prone when there is a greater frequency of the polymerase incorporating incorrect bases or causing misalignment (Goodman, 2002). HR uses a homologous template such as a homologous chromosome or a repeated sequence within the same chromosome, to repair a DSB (Goodman, 2002;

Shrivastav et al., 2008). The components of repair pathways are generally conserved among eukaryotes though more complicated in humans. In yeast, both HR and NHEJ are started with end-processing by the MRX complex (Mre11/Rad50/Xrs2). In HR ssDNA ends are coated with RPA, which is then displaced by Rad52p. This in turn promotes the binding of Rad51p as well as Rad55p and Rad 57p (Symington 2002, Shrivastav 2008). The strand searches for and binds a homologous donor sequence and is then extended and re-annealed with the non-invading strand. If both strands invaded then this could lead to a double-Holliday junction (Shrivastav et al., 2008). NHEJ involves the Ku70-Ku80 heterodimer, which binds the broken ends at the site of the DSB. Other repair factors are recruited and the broken ends are ligated together. Depending on the break, some bases can be filled in but NHEJ can often lead to small deletions (Helleday, Lo, Van Gent, & Engelward, 2007; Tseng & Tomkinson, 2004).

Another repair pathway of interest that has been associated with deletions is microhomology mediated end joining (MMEJ). MMEJ relies on 5-25 bases of homology during the repair of a double strand break and has been shown to result in deletions and instability (figure 1)(Lee & Lee, 2007). Rad52p (as well as Rad51p and Rad59p) and Ku70p have been identified as non-essential for MMEJ. Many other repair proteins are thought to be involved in MMEJ, such as Pol4p (essential for gap-filling in NHEJ) and Srs2p (interacts with Rad51p during HR) (Lee & Lee, 2007; McVey & Lee, 2008)

Expansions and contractions are common in regions of repetitive DNA or repetitive amino acid sequences. The method by which these events occur can be

influenced by many things, including secondary structure formation, base composition or length of the repetitive region.

Hypothesis on the advantage of plasticity within the CTD

My work attempts to explore the advantages of having a long stretch of amino acid repeats within the coding region of an essential gene. I examine the plasticity of the CTD of *S. cerevisiae* and its ability to contract, and attempt to determine how these contractions may be occurring to gain further insight into the contraction mechanism and evolution of the repetitive CTD. I hypothesize that the CTD is repetitive and contains more than the necessary number of repeats to be able to expand and contract if necessary. I tested the idea that there are extra repeats so if repeats become non-functional due to DNA damage or other mutations, repeats can be deleted so RNAPII can maintain its essential functions.

Chapter 2: Exploring the ability of the CTD to contract to delete mutations

NOTE: A significant portion of the work in this chapter was published in The Journal of Biological Chemistry. Morrill, Summer A., Exner, Alexandra E., Babokhov, Michael, Reinfeld, Bradley I., Fuchs, Stephen M. DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. *Journal of Biological Chemistry.* 2016; *291*(22), 11540–11550. © The American Society for Biochemistry and Molecular Biology
The full text of this article for reference is in Appendix 1. The western blot in this thesis (figure 7) was performed by Michael Babokhov. Expansion experiments referenced in this text were performed by Summer Morrill. All contraction experiments were performed by Alexandra Exner.

INTRODUCTION:

The CTD of the RNA polymerase II subunit Rpb1p is composed of 26 heptapeptide repeats in *S. cerevisiae*, yet previous work has shown that only 8 repeats are necessary for growth (though severely diminished growth) (Morrill et al., 2016; Nonet et al., 1987). The C-terminal domain (CTD) of the RNAPII, is an essential part of the protein complex and an integral part of the transcription machinery (see chapter 1). When the tail is reduced to between 8-12 repeats growth is diminished or cells become extra sensitive to stress, yet when the tail has only 13 repeats, cells exhibit wild type growth (Nonet et al., 1987). This raises interesting questions about why there are twice as many repeats as are necessary for growth and if different repeats have different functions. Work done in the lab by Summer Morrill showed that a CTD construct with only eight repeats could expand to a stable number of repeats, restoring wild type growth. However, it was difficult to determine how these expansions were happening, in part because her CTD_8 construct was composed of synthetic repeats, making each repeat identical. Though the amino acid sequence of each repeat is generally the same, the DNA

sequence has natural variation, allowing us to distinguish one repeat from another. Other ongoing experiments in the lab are trying to ascertain if certain repeats were more important than others, or if different repeats had different functions. The thought was that even though the amino acid sequence of the repeats was the same, maybe the location of the repeat within the 26 influenced its function. These experiments (started by Mohammed Mosaheb) were conducted by mutating all serines to alanines in blocks of eight repeats in the first, middle, or last third of the CTD. These mutations interfered with phosphorylation, consequently preventing other proteins from binding and interfering with CTD functions. During the course of this work, suppressor mutants arose that had contracted and deleted some of the repeats. This inadvertently showed that the CTD could not only expand, but it could also contract to delete mutated repeats.

This led to the development of a new construct to better explore contractions within the CTD. The construct was created using site-directed mutagenesis to insert four stop codons after the first eight repeats (the minimum number needed for growth) while leaving the rest of the CTD tail intact. The stop codons would cause only the first eight repeats to be translated into a shortened protein product since the rest of the repeats would be non-coding. The impaired growth caused by the four stop codons could theoretically be remediated by either expansion within the first eight repeats, or contraction and deletion of the repeats containing stop codons. We refer to this construct as the 4Stop construct.

Since our constructs are on plasmids and transformed into S. *cerevisiae* rather than integrated into the genome, we have two copies of our gene of interest

present. Therefore we use a Tet-off promoter system adapted from the design of the Strathern lab to repress transcription of the genomic copy of *RPB1* (Figure 2)(Malagon et al., 2006)). The genomic copy of *RPB1* is under the control of a tetracycline-responsive promoter. Addition of Doxycycline to the media prevents the Tet transactivator (tTA) from binding the promoter, resulting in repression of transcription. When the genomic copy of *RPB1* is repressed, the cells are forced to use our CTD variant on the plasmid.

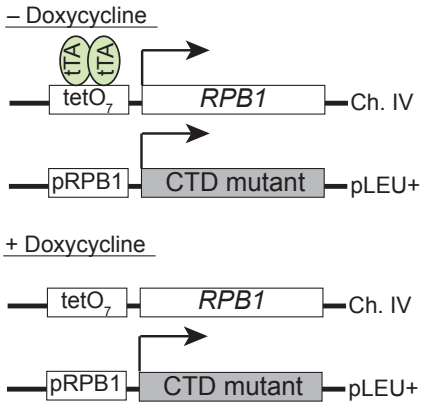


Figure 2. **Doxycycline is used to repress transcription of the genomic copy of *RPB1* so yeast cell growth is dependent on our mutant CTD plasmid.** When doxycycline is present in the media it prevents the tetracycline transactivator (tTA) from binding to the genomic promoter of *RPB1*. This prevents the genomic copy of the gene from being transcribed, and the cells are forced to rely on the mutant copy of the gene that is present on a plasmid.

The fact that the CTD has many more repeats than are necessary to maintain growth, combined with it's ability to expand and contract led to our theory of advantageous plasticity. The CTD is essential for transcription and cell viability. Perhaps there are extra repeats to prevent loss of function. If a mutation occurs in a repeat of the CTD rendering it non-functional, being able to delete the repeat while maintaining full function would clearly be beneficial. Potentially

there is redundancy within the repeats so if one is deleted, another repeat can fulfill the same functions. Our main focus is how the CTD adapts to mutations. Possibly *S. cerevisiae* can either contract to delete whole repeats /parts of repeats, or repair mutations using another repeat as a homologous template (see homologous recombination in chapter 1).

To distinguish what repair mechanism might be leading to contraction events, we tested our 4Stop construct in several repair protein deletion strains (*rad52, pol32, ku70, rad5*) as well as one with a helicase mutation. If deletion of any of these repair proteins leads to a significant change in the occurrence of contraction events it will help to determine how contractions are occurring. The RNA PolII CTD gene has the potential to form DNA secondary structures due to the inherently repetitive nature of its DNA. Secondary structure formation has previously been linked to strand slipping and expansion/contraction events (see chapter 1). Specifically, the non-coding strand of the CTD is very guanine-rich which leads to the possibility of G-quadruplex structures forming. In our previous paper it was shown through circular dichroism that stable secondary structures can form within this region that are consistent with parallel/antiparallel G_4-DNA structures (Morrill et al., 2016). We tested our construct in a strain with a *Pif1-m2* mutation, which disrupts Pif1p activity in the nucleus. This should interfere with the unwinding of G4 DNA thus affecting the rate of contractions, and allowing us to see if these structures are involved in the contraction process. The other protein candidates have functions in various DNA repair pathways and their roles have been examined in regards to expansions of the CTD. Since contractions may be

the result of double strand breaks, we tested the effects of proteins involved in homologous recombination and non-homologous end joining, the classic double strand break repair pathways. Rad52p is good candidate considering there are three types of HR (break-induced replication, single-strand annealing, and gene conversion) all of which require Rad52p (Helleday et al., 2007). Ku70p is part of the Ku70/Ku80 heterodimer, which is recruited to sites of double strand breaks and is essential for non-homologous end joining (Shrivastav et al., 2008). Pol32p has multiple roles in replication and repair. Pol32p is a subunit of DNA polymerase δ, and it shown to be important for its processivity (Johansson, Garg, & Burgers, 2004). Pol32p interacts with Polα (which initiates replication) and Polδ is then involved in the elongation of okazaki fragments (Johnson, Klassen, Prakash, & Prakash, 2015). Pol32p is also involved in the mutagenic translesion synthesis pathway (TLS) (Huang, Rio, Galibert, & Galibert, 2002). Pol32p has been seen to form a four subunit complex with Pol31p, Rev3p, and Rev7p referred to as Polζ_4. The PCNA binding domain of Pol32p is thought to contribute to the functional interaction between Polζ_4 and PCNA in TLS (Makarova, Stodola, & Burgers, 2012). Deletion of *pol32* has been seen to increase mutations and increase deletions of sequences flanked by exact repeats. These contractions seen in *pol32Δ* mutants are also shown to be Rad52p and Rad5p independent (Huang et al., 2002). If we see an increase in mutations or a decrease in contractions, Pol32p may be involved in contractions within the CTD. Rad5p is involved in post replication repair and polyubiquitination of PCNA, leading to template switching (Blastyák et al., 2007). Rad5p is involved in an error-free

repair pathway that involves replication fork reversal. Stalling/reversing of replication forks has been seen to lead to deletions and expansions of replicative sequences (see chapter 1).

Here I present the system we developed to study contractions within the CTD of RNAPII and explain my results regarding the involvement of repair proteins in the contraction process. Based on my findings, I have composed a model of how contractions may be occurring and propose that the repetitive nature of the CTD has evolved to cope with mutations and prevent loss of essential functions.

METHODS:

Strains and Plasmids

Strains used were derived from GRY3019 (*MATa his3*Δ *leu2*Δ *lys2*Δ *met15*Δ *trp1*Δ: : *hisG URA*: : *CMV-tTA kanRPtetO7-TATA- RPB1*). The *rad52Δ, rad5Δ, pol32Δ, ku70Δ, and pif1-m2* strains were previously created by other members of the lab by heterologous gene replacement and confirmed by PCR (Janke et al., 2004).

Site-directed mutagenesis was used to construct the 4Stop plasmid. A Quikchange Lightening mutagenesis kit was used, however the parameters were adapted to the following. Initial denaturation for 2 minutes, followed by 18 cycles of denaturation at 95°C for 20 seconds, annealing at 53°C for 20 seconds and elongation at 68°C for 8 minutes, followed by a final 20 minute elongation at 68°C. Primers were designed with homology to repeat number 9 with the intent of

replacing Tyrosine1 with a stop codon and Serine2 with a Tryptophan in repeats 9 and 10 (table 1). Due to the inherently repetitive nature of the CTD sequence, we produced a construct with four instead of two mutated repeats. The resulting plasmid was digested with *Dpn1* for 15 minutes instead of the recommended 5 to digest the parental DNA. 3µl of the plasmid was then transformed into the XL10-GOLD Ultracompetent cells using a heat shock method according to protocol. Cells were plated on LB-AMP plates.

Plasmids with serine to alanine mutations were created previously through the use of synthetic building blocks and recursive directional ligation by plasmid reconstruction using plasmid JMD2.

Table 1. **Primers used in this work**

Sequence 5' to 3'	Description
CACCAACGTCACCATCATAATGGCCAACGTCACCATCATAT	QuikChange stop codon forward
ATATGATGGTGACGTTGGCCATTATGATGGTGACGTTGGTG	QuikChange stop codon reverse
GATCGATGAGGAGTCACTGG	RPB1 forward PCR primer
TTTACTAGCGCCGTTGGTTT	RPB1 reverse PCR primer
GGCATAGCAATGACAAATTCAAAAGAAGACGCCGACATAGAG GAGAAGCATATGTACAATCTCCAAAGTTAACCATTTGT	Primer to integrate sequences in the genome forward
CAACATTCCAAAATTTGTCCCAAAAAGTCTTTGGTTCATGAT CTTCCCATACAATTGCCTCCTTTCCCAATGCTCTGCGGAT	Primer to integrate sequences in the genome reverse
GATCGATGAGGAGTCACTGG	Plasmid specific primer forward
ATATGCGTCAGGCGACCTCT	Plasmid specific primer reverse

Fluctuation assays

The 4Stop plasmid was transformed in GRY3019 or a mutant strain using the *Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method* according to protocol and plated on SC-LEU media (Gietz & Schiestl, 2007). Transformations were re-streaked after three days to confirm the

transformation was successful. Transformations were then re-streaked for single colonies and allowed to grow 4-5 days until individual colonies were very large. Twelve colonies (or fewer if there were not 12 of sufficient size) were re-suspended in 250µl of water in a 96 well plate. Four serial dilutions were made from each of these suspensions (25µl into 225µl four times) across the plate. 200µl of the original suspension was plated on SC-LEU+DOX plates while 200µl of the final dilution was plated on YPD plates (figure 3). All plates were incubated at 30°C. The YPD plates were removed from the incubator after three days while the SC-LEU+DOX plates were allowed to grow for four days.

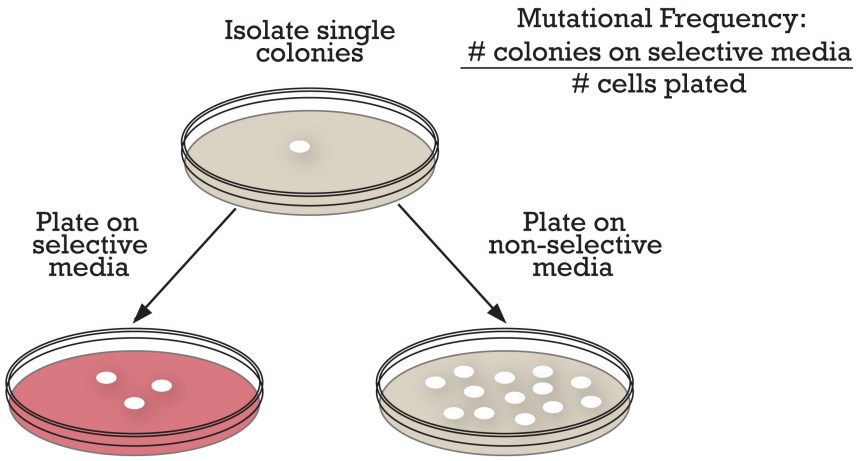


Figure 3. **Fluctuation assays are multi-plate assays performed to ascertain mutation rates and generate suppressor mutants.** After the strain being tested has been transformed with our 4Stop plasmid, it is streaked out for single colonies on an SC-LEU plate, and allowed to grow for 4-5 days until colonies are sufficiently large. Each colony is re-suspended and plated on an SC-LEU+DOX plate while a dilution of the suspension is plated on a YPD plate. This allows us to estimate how many cells were plated on the selection plate and then calculate a mutation rate based on how many suppressor colonies grew from those cells.

Colonies were counted on all plates and used to calculate mutation rates using the fluctuation analysis calculator, which uses the MMS-Maximum Likelihood Estimator method (MMS-MLE)(Hall, Ma, Liang, & Singh, 2009).

Yeast Colony PCR

Four colonies (two large, two small) from each selection plate were selected, restreaked, and allowed to grow over night. Yeast colony PCR was then performed on these colonies using the following method: dissolve colonies in Lyticase to lyse the cells and incubate at 30°C for 15 minutes, add 50µl sterile water to each digest and heat to 95°C in a thermocycler for 5 minutes, spin at 2500rpm for 5-10 minutes until a pellet forms, use 1.5µl of the supernatant for PCR. PCR was performed with primers flanking the CTD repeats (*RPB1* forward and reverse, Table 1). PCR products were run on 1.5% agarose gels.

Spotting assays

Yeast strains containing the plasmids of interest were grown in 5ml overnight cultures of SC-LEU liquid media. Those overnight cultures were used to start fresh cultures (also in in SC-LEU) the next morning at an A_{600} of 0.2. These were allowed to grow 4-5 hours until they reached an A_{600} of 1.0. After reaching this density, 250µl of each culture was transferred to the wells of a 96 well plate. These were then serially diluted 5-fold across the plate. Cells were spotted on to three SC-LEU+DOX plates and three YPD plates using a 48-pin replicating tool. Plates were incubated at 30°C and pictures were taken after two

and three days. Images of spotting assays are representative examples of three trials.

Smash and grab transformation and Sanger sequencing

Sequencing was done initially by amplifying around the CTD with plasmid specific primers (Table 1) so the genomic copy of the CTD would not be amplified. Later we switched to a method of purifying plasmids from culture and sequencing those. Yeast containing the plasmid of interest was grown in 3ml overnight cultures of SC-LEU media. DNA was then extracted using the "Smash and Grab" protocol and re-suspended in 25µl of sterile water (ttp://web.mit.edu/biology/guarente/protocols/quickprep.html). 3µl of each smash and grab product was then transformed into DH5α E. *coli* through electroporation and plated on LB-AMP to maintain the plasmid. Plates were grown overnight at 37°C. Resulting colonies were grown overnight in 3ml of LB-AMP media and the following morning, plasmids were extracted using a Quiagen miniprep kit. Plasmids were sent elsewhere for Sanger sequencing and alignments were done with alignment tools found at the European Bioinformatics Institute website. Alignments were completed by hand as the online tools do not align the contraction-suppressor mutants correctly to the original 4Stop sequence due to their inherently repetitive nature.

RESULTS:

Preliminary assays with serine to alanine mutant constructs

I ran preliminary assays using the constructs containing serine to alanine mutations in repeats 2-9 or 10-17 of the CTD as those produced the most diminished growth phenotypes and would potentially produce suppressor mutations. I transformed these plasmids into our WT strain and ran small-scale fluctuation assays for both. These constructs with mutations in repeats 2-9, or 10-17 will be denoted $(S\text{-}A)_{2\text{-}9}$ or $(S\text{-}A)_{10\text{-}17}$. The aim of these experiments was to examine the types/sizes of contraction mutations that would arise. From both strains I isolated and sequenced suppressor plasmid DNA. I hypothesized that examining the types of deletions might tell us something about how these events were occurring. In the suppressors arising from the $(S\text{-}A)_{2\text{-}9}$ construct I saw two mutants that had deleted six of the mutated repeats leaving 20 remaining repeats. However, it was difficult to tell which repeats had been deleted since these constructs were built with identical synthetic repeats (Figure 4).

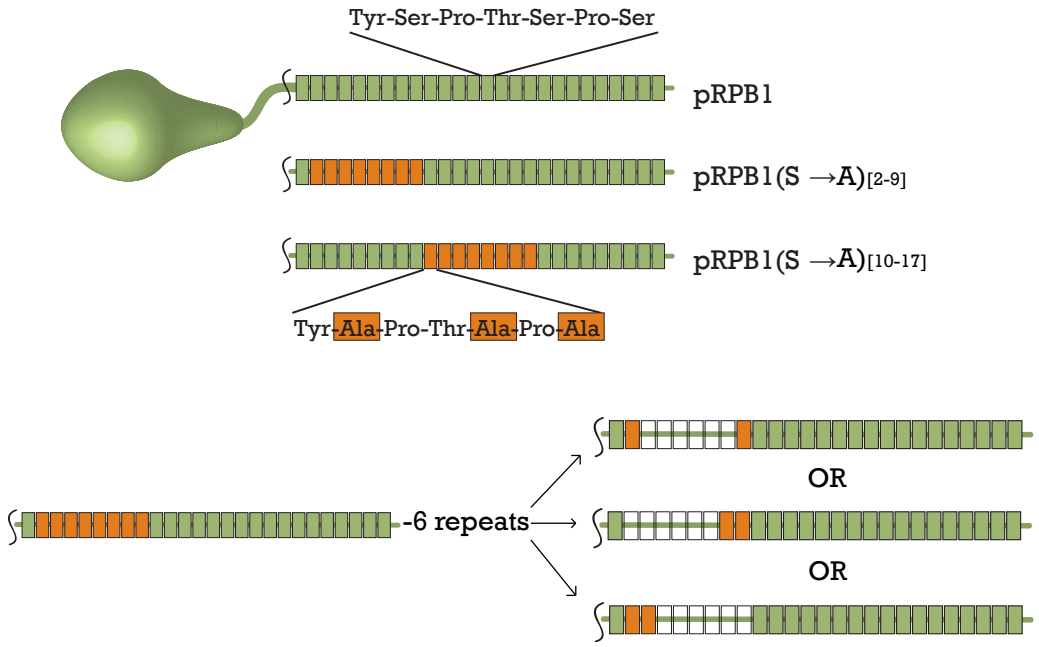


Figure 4. **Preliminary assays involving serine to alanine mutants showed the CTD can contract, though the synthetic constructs made it difficult to understand contractions.** When serine to alanine mutations are made in the first or middle third region of the CTD, growth is significantly diminished. Suppressors arise that delete some or all of the repeats with alanines. Since these constructs were made using synthetic building blocks, all the repeats are identical on the DNA level. This can make it difficult to tell which repeats were deleted. In this example six of the eight repeats containing alanines were deleted, though we cannot tell exactly which ones.

Colony PCR was performed on a subset of the suppressors seen to see how often contraction events were occurring. It was seen that 18 out of 26 samples showed contraction events, indicating that contractions were happening fairly frequently and it would be possible to find more suppressors. From the $(S\text{-}A)_{10\text{-}17}$ construct we saw two different deletions, one where 8 repeats had been deleted and one where 12 repeats had been deleted, leaving 16 and 14 repeats remaining. Aside from the repeats that had contained, alanines, it was again difficult to determine which repeats had been deleted due to their identical nature.

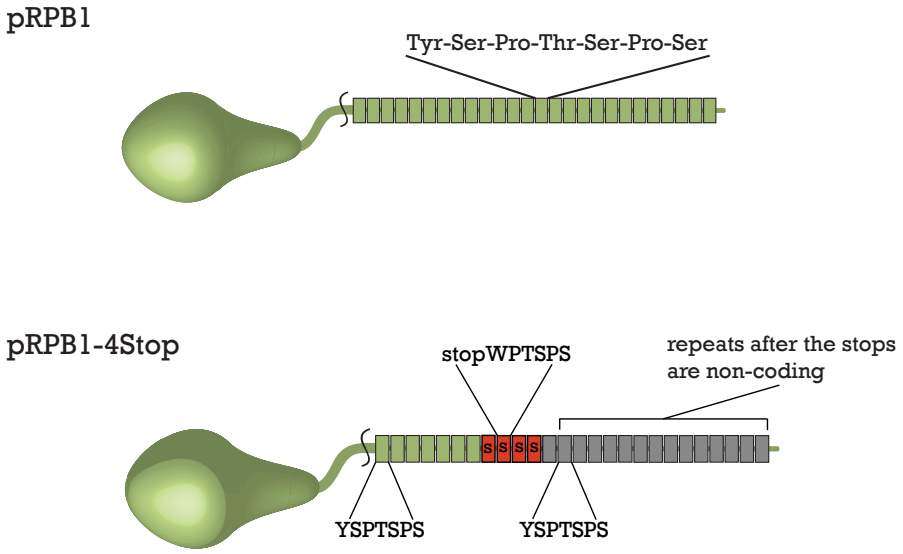


Figure 5. **RNA Polymerase II and the 4Stop construct.** RNA Pol II has a unique repetitive C-terminal tail attached to its largest subunit Rpb1p. The tail is made up of the amino acid sequence YSPTSPS repeated 26 times in *Saccharomyces cerevisiae* and 52 times in humans. Our 4Stop construct was based on the fact that 8 repeats is the minimum number needed for growth in *S. cerevisiae* (but very sick). A stop codon replaces the tyrosine at the beginning of repeats 8-11. The original design only contained two stop codons so we also replaced Serine2 with tryptophan to add an additional challenge for the cells to overcome. The remaining repeats are non-coding since the stop codons lead to a shortened protein product.

We developed the 4Stop construct as an improved way to study contractions, and to be able to identify exactly which repeats were being deleted. The idea was to insert stop codons after the first eight repeats, as that is the minimum number for survival. The construct would produce a shortened protein product, yet the rest of the DNA sequence would remain there to potentially be used as a template that could be used in DNA repair/recombination. The construct was created using site directed mutagenesis of the wild type CTD rather than constructed from synthetic, identical building blocks. Using the wild type CTD would better allow us to determine which repeats are deleted and how these repair

events were occurring. Potentially we would also be able to examine both contractions and expansions with this construct as it could in theory contract to delete the repeats containing the stop codons or it could expand within the first eight repeats. The original design had two stop codons replacing the tyrosines at the beginning of repeats 9 and 10, followed by a tryptophan replacing Serine2 of these repeats. The tryptophan would serve as an additional challenge for DNA repair (since serine2 is important in transcription). Due to the inherently repetitive nature of the CTD, site directed mutagenesis of our WT CTD plasmid resulted in four instead of two stop-codon repeats. Multiple trials resulted in a stop codon (and tryptophan) at the beginning of repeats 8-11 (figure 5). However, this 4Stop construct has been tested and works well as a model to study contractions within the CTD.

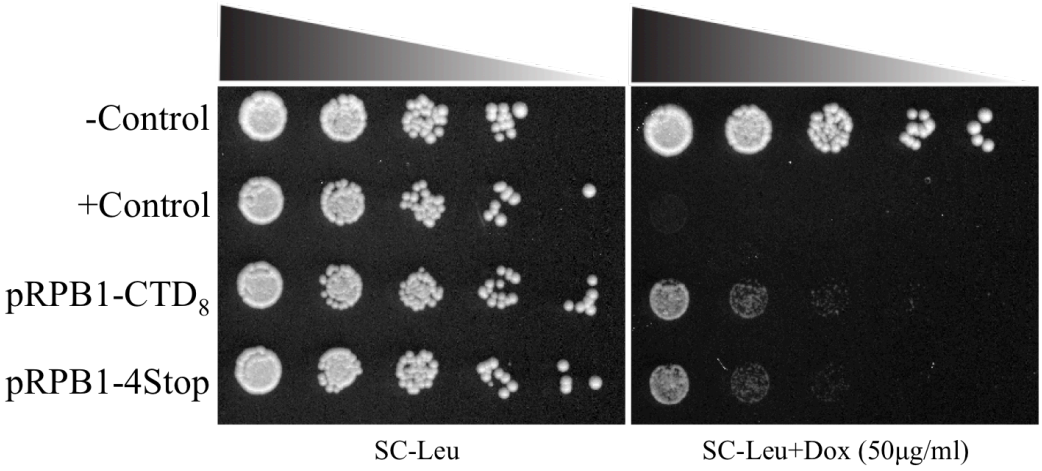
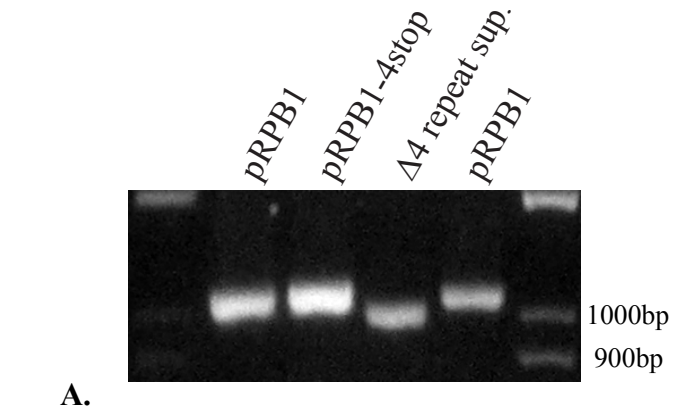
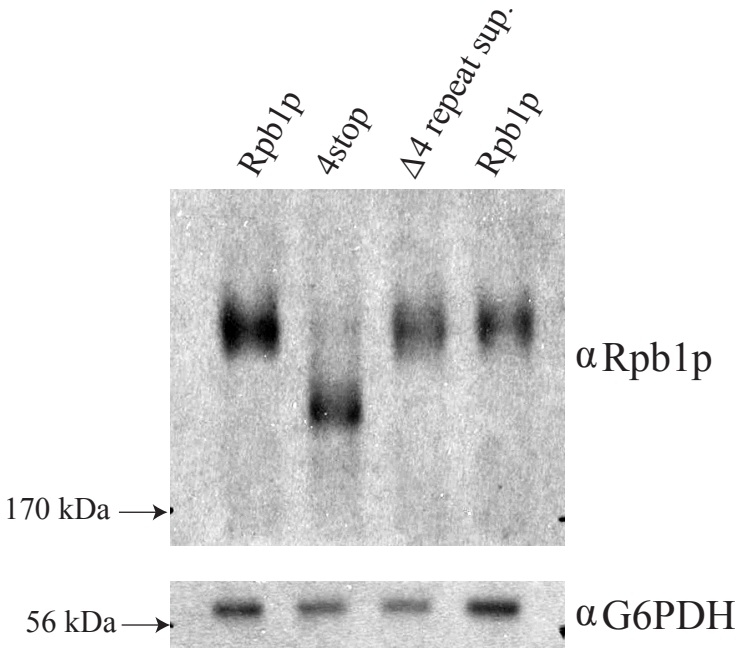


Figure 6. **Spotting assay comparing the health of the wild type GRY3019 strain when growth is dependent on either an 8-repeat CTD plasmid or the 4Stop plasmid.** Growth of GRY3019 transformed with the 4Stop plasmid is comparable to growth seen when the strain is transformed with the 8-repeat CTD plasmid. This confirms that the 4Stop plasmid is conferring the diminished growth that we intended when developing the construct. The positive control is a plasmid containing a full length WT CTD, while the negative control is a plasmid without a copy of the CTD. Spotting assays are a representative example of at least three independent trials.

The 4Stop construct negatively affects growth in a way comparable to the CTD_8 construct as we intended. I performed a spotting assay to evaluate the growth phenotype of the construct (Figure 6). Our positive control is the yeast strain (GRY3019) transformed with a plasmid containing the WT full length *RPB1* sequence while the negative control is transformed with an empty plasmid that does not contain a copy of *RPB1*. From left to right are dilutions of our WT strain transformed with the plasmid of interest. When DOX is added to the media, the genomic copy of Rpb1 is turned off and growth is then dependent only on the plasmid present. When cell growth is dependent on our 4Stop construct, we see severely diminished growth comparable to the growth caused by a CTD plasmid with only eight repeats (pCTD8). We also confirmed by western blot that the 4Stop construct is producing a protein much shorter than the wild type Rpb1p (figure 7B). When suppressor mutations arise that delete the four repeats with the stop codons (Δ4 repeats), a protein comparable to full size is produced, even though the DNA sequence is shorter than the original construct with the stop codons (Figure 7A,B).



A.



B.

Figure 7. **DNA gel and western blot comparing the size of the DNA and the protein produced by the 4Stop plasmid, to the DNA and the protein produced by a 4Stop-suppressor mutant.** A. The DNA of the 4stop suppressor mutant is shorter than the DNA of the 4Stop construct, as can be seen on a gel. B. The 4Stop plasmid does produce a shortened protein product as was intended when building the construct. The 4Stop suppressor mutant shows that deletion of the four stop codon-containing repeats restores the full length protein product comparable to the WT Rpb1p protein. Though the DNA sequence is shorter, the protein is larger when the stop codons are deleted and all repeats are once again coding for the protein. Western blot performed by Michael Babokhov.

Effects of repair protein deletions on contractions

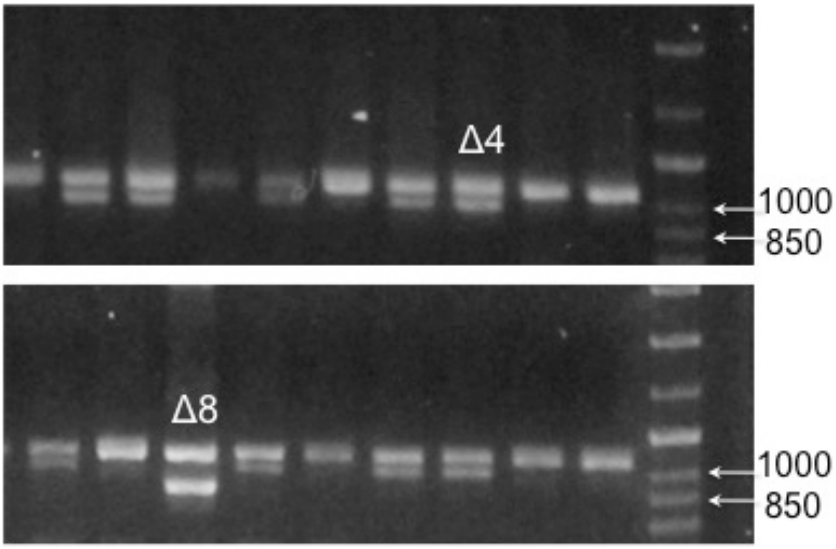


Figure 8. **Yeast colony PCR serves as a simple way to screen for contraction suppressors.** Colony PCR uses primers that flank the CTD sequence both in the genome and on the plasmid. Both copies of the CTD start out the same length but when a contraction event occurs in the plasmid copy, the two bands become easily distinguishable. A suppressor that shows a deletion of the four repeats containing stop codons is only a contraction of 84 bases while a suppressor that deleted eight repeats results in a loss of 168 bases.

Deletion of repair proteins changes the percentage of contraction events we see (Table 2). Yeast colony PCR was performed on four suppressor colonies from each of the 12 plates from each fluctuation assay and contraction events were observed in approximately 51% of these colonies (in WT strain). The PCR was analyzed by gel electrophoresis and the percentage of contraction events is simply the total number of contraction events in relation to the total number of colonies analyzed. These events are easily detected on a gel as the PCR primers amplify both the genomic and the plasmid copy of the CTD tail. When a contraction event occurs, the plasmid CTD is shortened and two bands are seen

(figure 8). Deletions of the four repeats contain stop codons or larger contractions of eight repeats are easily observable this way.

Table 2. **Mutation rates and contraction frequencies of wild type and mutant strains transformed with the 4Stop plasmid**. Mutation rates were calculated by comparing the number of colonies that grew on selective SC-Leu+Dox plates and a dilution plated on non-selective YPD plates. The rates (ranges in parenthesis) were generated using the MSS-Maximum Likelihood Estimator/FALCOR fluctuation analysis calculator. Percent of contractions refers to the fraction of contraction events seen after screening respective strains through yeast colony PCR. Compared to the wild type strain, the *pol32Δ, rad52Δ,* and *rad5Δ* strains show significantly different rates of contractions. The category "other" indicates that a mutation/event occurred elsewhere in the genome that is allowing the colony to grow, though it is not the result of a contraction.

Rates of contraction observed in pRPB1-4Stop in WT and mutant strains

	Mutation Rate (per 10^6/ cell/generation)	Contractions	Expansions	Other	Total	% Contractions
WT	26.39 (10.9-37.9)	59	1	55	115	51
pif1-m2	5.84 (4.3-8.0)	47	0	35	82	57
pol32Δ	21.06 (7.3-43.0)	77	0	10	87	89
ku70Δ	5.75 (2.35-8.56)	76	0	50	126	60
rad52Δ	3.6 (2.5-4.7)	51	0	1	52	98
rad5Δ	4.72 (3.9-5.2)	1	0	105	106	0.9

When the 4Stop plasmid is transformed into our WT strain, approximately 51% (59/115) of the suppressor colonies analyzed showed contraction events (Table 2). When the plasmid is transformed into the *pif1-m2,* or *ku70Δ* strains, there is no significant difference in the percentage of contraction events seen with 57% and 60% respectively (Table 2, Figure 9). The percentage of contraction events observed increases significantly from 51% to 89% when *pol32* is deleted and from 51% to 98% when *rad52* is deleted. The decrease in the percentage of contraction events seen when rad5 is deleted is extremely significant; the

proportion of contraction events goes from 51% to 0.9% with only 1 contraction seen out of 106 suppressors screened (Table 2, Figure 9). Proportions of contraction events were compared to WT rates by chi-square analysis. The difference in the proportion of contraction events observed in WT compared to the *pol32Δ, rad5Δ,* and *rad52Δ* strains was considered to be extremely statistically significant with P-value of less than 0.0001 (Figure 9).

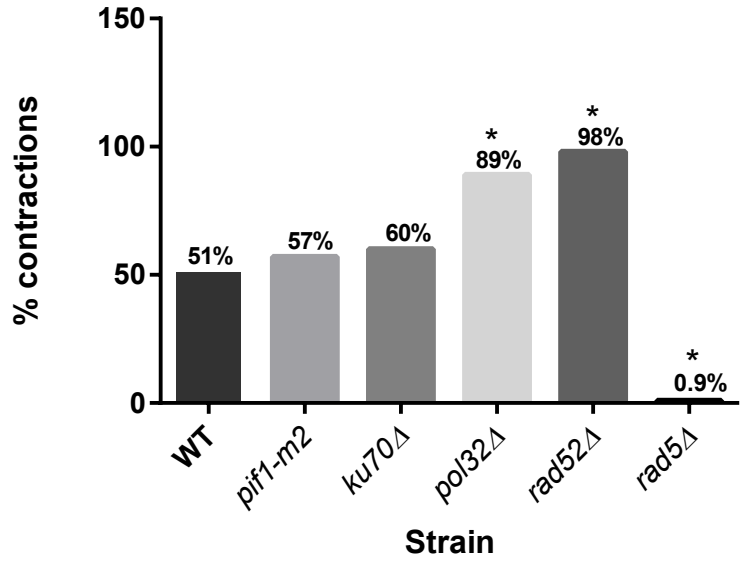


Figure 9. **Percentage of contraction events observed is changed in repair-protein delete strains.** Percentage of contractions was calculated by taking the total number of contraction events seen out of the total number of colonies screened through yeast colony PCR. The exact values are displayed in Table 2, this is simply a more visual representation of the data. Chi-square analysis was performed comparing fractions of contractions seen in each strain to the wild type. The difference in rate of contractions seen in the wild type compared to the *pol32Δ, rad5Δ,* and *rad52Δ* strains is considered to be extremely statistically significant with a P-value of less than 0.0001 (indicated by asterisks).

Deleting specific repair proteins did not have the same effect on contractions that it did on expansions. Deletion of *rad52* had the opposite effect on the proportion of expansion events observed compared to the proportion of contractions. While the percentage of contraction events rose from 51% to 98%,

the percentage of expansions decreased from 38% to 5% (Table 3). It is a little challenging to compare the results of the contractions assay to the results of the expansion assay performed in the *pol32Δ* strain. The frequency of expansions seen in the *pol32Δ* strain was only based on two trials of very different results. One trial resulted in an expansion frequency of 24% while the other showed a frequency of only 2.7%. However, the percentage of contraction events occurring increased substantially in the *pol32Δ* strain, yet the percentage of expansions seemed to decrease. A similar but more drastic affect was seen as a result of deleting *rad52*, the proportion of contraction events increased substantially while the percentage of expansions decreased. In the *rad5*Δ strain there was a decrease in both expansions and contractions but it was much more significant in regard to contractions. Expansions decreased from 38% to 14%, however, contractions became almost non-existent decreasing from 51% to 0.9%.

Table 3. **Deletion of repair proteins does not have the same effect on expansions as it does on contractions.** A change in the percentage of contraction events is seen with the deletion or mutation of certain proteins. These proteins do not seem to have the same effect on expansion events, indicating that expansions and contractions are occurring through different mechanisms. While the percentage of contraction events observed is significantly higher in *pol32Δ* and *rad52Δ* strains, the percentage of expansion events is lower in both *pol32Δ* and *rad52Δ* strains. Deletion of *rad5* decreases both contractions and expansions but has a more significant effect on contractions. Asterisks indicate percentages that are significantly different from the WT as determined by chi-square analysis. Note: expansion percentages all provided by Summer Morrill and *pol32Δ* results are based only on two trials of varying results.

Strain	WT	*pif1-m2*	*pol32Δ*	*ku70Δ*	*rad52Δ*	*rad5Δ*
4Stop % contraction events	51	57	89*	60	98*	0.9*
CTD_8 % expansion events	38	56	13.35 (n=2)	34	5*	14*

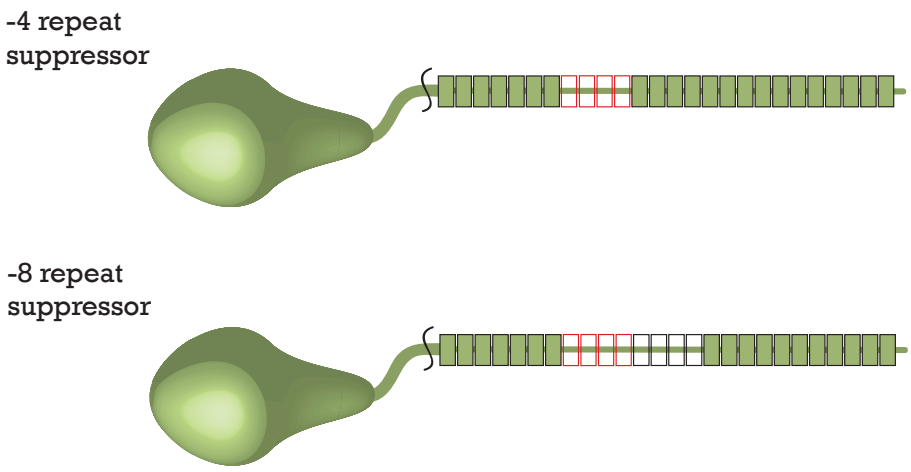


Figure 10. **Two different contractions seen in 4Stop suppressors that restore WT growth.** Most often we see a contraction event resulting in deletion of repeats 8-11 that contained the stop codons. Less frequently we see deletion of those four repeats and the following four repeats. The reason for this is most likely that there is greater homology before and after the four repeats than exists before and after the eight repeats.

When we sequenced the suppressor mutants, we saw several different types of contraction events. Sequencing for the WT strain was done simply by using plasmid specific primers designed for this purpose and amplifying around the CTD. Sequencing for the *ku70* and *pol32* strains was done by extracting DNA from the yeast, then transforming a fraction of the DNA into *E.coli*, purifying the plasmid, and sequencing that. We wanted to see if different deletion strains would show different ratios of particular contraction events or types contractions (error-prone or not). Of the 13 contraction events we sequenced in our WT strain, 12 of those showed a deletion of just the four repeats that contained stop codons, while one suppressor mutant showed a deletion of eight repeats (figures 10 and figure 11). The *pol32Δ* suppressors showed the most variability in size of contractions. Of the 16 suppressors, 12/16 deleted the four stop codon repeats, 2/16 deleted five repeats (the four stop repeats plus one repeat before), 1/16 deleted six repeats (the

four stop repeats plus two repeats before), and 1/16 deleted eight (the four stop repeats plus four repeats before). The Ku70 suppressor sequences showed 10/11 suppressors deleted the four stop repeats while 1/11 deleted five repeats. All deletions of 5, 6, or 8 repeats seen were deletions of the same repeats (ex: all deletions of 5 repeats were a deletion of the four stop codon-containing repeats and one repeat before).

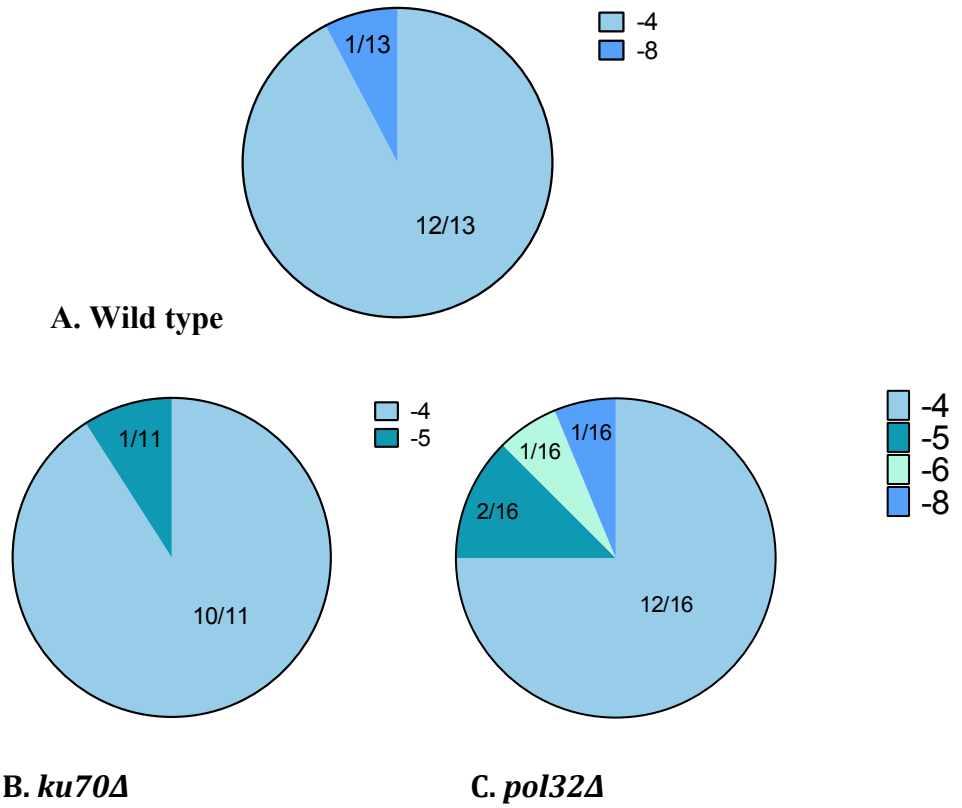


Figure 11. **Ratios of various contraction events seen in WT, *ku70Δ*, and *pol32Δ* strains.** Ratios of contraction events are based on Sanger sequencing results. **A.** Most contraction events (12/13) seen were deletions of just the four stop codon-containing repeats, however one deletion of eight repeats was seen. **B.** When *ku70* was deleted, most of the deletions (10/11) were also seen to be deletions of the four stop-codon repeats, but there was also one suppressor showing a deletion of five repeats which was previously unseen in the WT strain. **C.** When *pol32* was deleted, the most variety was seen in suppressor mutations.

Most showed deletion of 4 repeats though there were also deletions of 5, 6, and 8 repeats.

When examining the alignments of the different contraction events seen, I noticed regions of homology before and after all the deletions. The eight repeat deletion has 11 bases of homology in front of and after the deleted region, while the four repeat deletion has 14 bases of homology in front of and following the deletion (Figure 12 A,B). The suppressor mutants with the five repeat deletions also had homologous regions of 14 bases before and after the contraction (not shown). The deletion of six repeats had homologous regions of 11 bases before and after the contraction (not shown).

A.

```
                 Y  S  P  T  S  P  S   Y  S  P  T  S  P  S   Y  S  P  T  S  P  S
4Stop            TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG
Δ8-repeats       TTTTCTCCAACTTCCCCAACA TACTCTCTTACCTCTCCAGCG TACTCACCAACATCACCATCG
                       Repeat 1              Repeat 2              Repeat 3

4Stop            TACTCACCAACATCACCATCG TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG
Δ8-repeats       -------------------- -------------------- --------------------
                       Repeat 4              Repeat 5              Repeat 6

                                 Stop                  Stop
4Stop            TATTCACCAACGTCACCATCA TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCA
Δ8-repeats       -------------------- -------------------- --------------------
                       Repeat 7              Repeat 8              Repeat 9

                 Stop                  Stop
4Stop            TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC
Δ8-repeats       -------------------- -------------------- TATTCGCCAACGTCGCCTTCC
                       Repeat 10             Repeat 11             Repeat 12

4Stop            TACTCTCCCACGTCGCCAAGC
Δ8-repeats       TACTCTCCCACGTCGCCAAGC
                       Repeat 13
```

B.

```
                 Y  S  P  T  S  P  S   Y  S  P  T  S  P  S  Stop S  P  T  S  P  S
4Stop            TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TAATGGCCAACGTCACCATCA
Δ4-repeats       TACTCACCTACATCACCATCG TATTCACCAACGTCACCATC- --------------------
                       Repeat 6              Repeat 7              Repeat 8

                Stop                  Stop                  Stop
4Stop            TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCG
Δ4-repeats       -------------------- -------------------- --------------------G
                       Repeat 9              Repeat 10             Repeat 11

4Stop            TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC
Δ4-repeats       TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC
                      Repeat 12             Repeat 13
```

Figure 12. **Microhomology exists around deletions in the 4Stop suppressors.** Regions of homology are highlighted **A**. Eleven bases of homology is seen before and after a deletion of eight repeats. **B.** Fourteen bases of homology is seen around a deletion of four repeats. Greater homology around the deletion of four repeats could explain why these deletions are more common.

DISCUSSION

The repetitive CTD of RNAPII is essential and highly conserved amongst eukaryotes. Though the CTD found in *S.* cerevisiae is composed of 26 repeats only 13 are necessary for wild type growth and less than that is sufficient for diminished growth (Morrill et al., 2016; Nonet et al., 1987). The fact that so many repeats can be deleted while maintaining growth, combined with the fact that the CTD has shown the ability to expand and contract inspired our interest. We hypothesize that there are more repeats than necessary to be able to cope with mutations and still maintain normal functions. Repeats can be deleted if they acquire mutations without compromising viability or essential transcription functions.

We showed that the CTD could contract to delete four repeats containing premature stop codons to restore normal growth. Through deletion of various DNA repair proteins we have constructed a theory and model of how this is occurring. Deletion of certain genes coding for DNA repair proteins had a significant effect on contractions and has allowed us to access the involvement of multiple repair pathways.

In addition to the percentage of contractions changing with gene deletions, mutation rate also changes (Table 2). A change in mutation rate is likely due to effects that these deletions/mutations are having elsewhere in the genome.

Mutation rate is a reflection of how many colonies grow in relation to how many cells are plated. For example deletion of *rad52* causes a notable decrease in the mutation rate. Since *rad52* is essential for homologous recombination, deletion of this gene interferes with homologous recombination throughout the genome and as a result, fewer cells are surviving.

Deletion of rad52 (important for HR) did not decrease the percentage of contraction events compared to the WT. Therefore, we conclude that contractions are occurring through a recombination independent mechanism. We also saw a similar increase in the occurrence of contractions in the *pol32Δ* strain (89% in *pol32Δ* v. 95% in *rad52Δ*), therefore Pol32p is therefore likely not involved in contraction events either. Rad5p however is a credible candidate involved in the contraction process, since the deletion of this protein essentially eliminated all contraction events.

Rad5p was initially seen to possess a specialized helicase activity utilized in the regression of replication forks (Blastyák et al., 2007). Rad5 has since been shown to be involved in the post replication repair pathways of template switching leading to either fork reversal or a recombination-mediated repair pathway (Ghosal & Chen, 2013). Both of these are error-free pathways. Since Rad52p is involved in homologous recombination and deletion of *rad52* did not decrease contractions, contractions are likely not occurring through the recombination mediated pathway. It is also worth noting that when looking at sequencing alignments of suppressor sequences, we can tell that recombination with the genome has not occurred. There is a Sac1 restriction site closely

following the 4Stop sequence as well as the CTD_8 sequence. When the CTD_8 sequence would recombine with the yeast genome, the restriction site was lost. We have never seen a loss of the restriction site in any of the contraction suppressors sequenced regardless of strain.

Since deletion of *pol32* significantly increases the occurrence of contraction events we are not pursuing Pol32p as a candidate involved in the process of contractions. Previous studies have seen that deletion of *pol32* led to an increase in mutations and a notable increase in deletions of sequences flanked by identical repeats (Huang et al., 2002). The deletions ranged from 8 to 237 bp in length and were always between exact homologous repeats of 3 to 10 bp. The frequency of these contraction events was also seen to be unaffected by deletions of *rad52* or *rad5*. We have also not seen any instances of mismatches or other errors in the suppressors that we have sequenced. A potential explanation regarding repair errors is that only error-free contraction mutants are viable, which is why we never see errors.

A critical step in the initiation of post replication repair is the mono- or poly- ubiquitination of the sliding clamp PCNA (proliferating cell nuclear antigen) in response to DNA damage (Hoege, Pfander, Moldovan, Pyrowolakis, & Jentsch, 2002). Depending on whether PCNA is mono- or poly- ubiquitinated, repair is directed down the translesion synthesis pathway or the template switching pathway respectively (Ghosal & Chen, 2013).While Rad6p and Rad18p are involved in monoubiquitination, Rad5 is necessary for recruiting the

ubiquitin-conjugating enzyme and adding on additional ubiquitin for polyubiquitination of PCNA (Blastyák et al., 2007).

Since the deletion of *RAD5* almost eliminates contractions, conceivably the deletion is forcing repair down an alternative repair pathway that does not lead to contractions. One possibility is that since polyubiquitination is being prevented in this deletion, only monoubiquitination is occurring, leading repair down a translesion synthesis pathway instead. Possibly, it is more difficult for contractions to occur this way. It could also be forcing repair down other pathways that don't lead to contractions such as the homologous recombination pathway.

Evidence of the Involvement of Fork Stalling in the Contraction Process

The deletion of *rad5* in our 4Stop mutant strain essentially stops contractions from occurring (only 1/115). Since Rad5p has been shown to be important in the pathway of fork reversal in post replication repair, this is potentially the pathway leading to the contraction events we see (Blastyák et al., 2007; Ghosal & Chen, 2013). Replication forks may slip forward or backward during replication of regions of genomic complexity, such as when they encounter DNA structures, DNA-protein complexes, or defective DNA polymerases (Branzei & Foiani, 2007). Fork stalling can lead to fork collapse, double strand breaks and template switching. This often leads to rearrangements such as contractions and duplications. When the replication fork encounters a obstacle, the newly synthesized strands will separate from the fork, the parental strands will

re-anneal as the fork regresses, and then the newly synthesized strands will begin to anneal to each other creating a four way junction (also known as a chicken-foot structure) (Neelsen & Lopes, 2015). When the fork regresses and single stranded DNA is exposed, misalignment and incorrect re-annealing could occur (Branzei & Foiani, 2007; Neelsen & Lopes, 2015). This could be how our contractions are occurring. Since there are regions of homology within the CTD, potentially the DNA is re-annealing in the wrong location and deleting the region in between these homologous regions.

Structures occurring on the lagging strand template strand have been shown to lead to deletions when the DNA is skipped by an okazaki fragment (Mirkin, 2006). This principle could apply to what we are seeing. If an okazki fragment is skipping over a structure or if similar misalignment is occurring with the lagging strand and okazaki fragments, this could produce a contraction (figure 13).

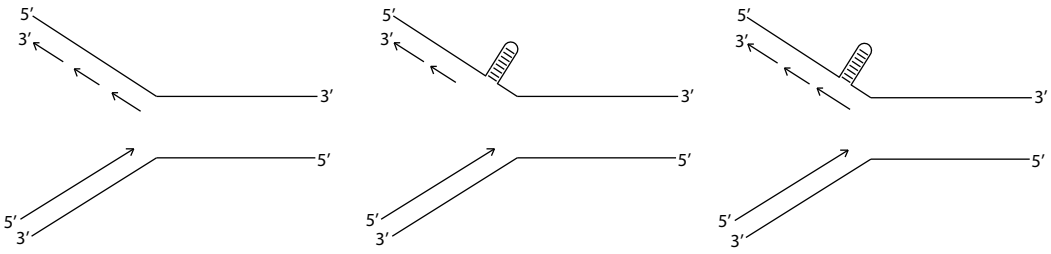


Figure 13. **Fork reversal/fork collapse may be involved in the contraction events seen with the 4Stop construct.** If secondary structures form on the lagging strand template, it can result in okazaki fragments skipping the structure and leading to contractions. Alternatively, when strands of the fork are separated, the okazaki fragments may align to the wrong repeat in a similar manner, leading to contractions.

A way to potentially examine the effect of fork collapse further in our 4Stop construct would be to increase the frequency of fork reversal. This can be achieved by using anticancer agents or topoisomerase inhibitors. Using low enough concentrations of such agents has been shown to induce fork reversal while avoiding double strand breaks (Ray Chaudhuri et al., 2012). If we used a Topoisomerase 1 inhibitor in our system we would predict that an increase in the frequency of fork reversal could lead to an increase in the percentage of contraction events that we see. This could help confirm our hypothesis.

Evidence of the Involvement of Microhomology Mediated End Joining in the Contraction Process

The majority of the contraction events we see are deletion of four repeats. A question that arose was, why we see so few deletions of eight repeats or other deletions. When examining sequencing alignments we discovered regions of homology in front of and after the deletions. There are eleven bases of homology before and after deletions of eight repeats and fourteen bases of homology around deletions of four repeats. This greater homology could be reason we see a larger number of contraction suppressors with a deletion of four repeats. When we sequenced suppressors from the *pol32Δ* strain and the *ku70Δ* strain, we saw deletions of five and in one case six repeats (figure 11). Sequencing alignments also showed 14 bases of homology before and after the deletion of five repeats and only 11 bases of homology in the deletion of six. These results still help support the idea that the greater the region of homology, the more likely we are to

see that deletion. However it is interesting that we see more variety in the contractions in the *pol32* and *ku70* delete strains (especially the *pol32Δ* strain). It is unclear if this has something to do with the deletion of *pol32* or if we simply have not sequenced enough suppressors from the WT strain to make a fair comparison. Pol32p has been shown to play an important role in processivity of Polδ (Johansson et al., 2004). If processivity is impaired, the polymerase could be behaving differently and falling off at different points leading to the variety of contraction events we see. Potentially it is also because we only have a sample size of 13 WT suppressors. The *pol32Δ* suppressor mutants that were sequenced were taken from four fluctuation assays while most of the WT suppressors were from two of the three fluctuation assays done. Potentially that affects the variety we saw. There was also a much greater number of contraction events seen in the *pol32Δ* assay (89% compared to 51%) while the mutation rate was fairly similar (table 2). Perhaps since we saw a much greater number of contractions in the *pol32Δ* assay, it is also more likely that we see more variety.

The regions of homology seen around the deletions point to potential involvement of microhomology mediated end joining (MMEJ) in the contraction process. MMEJ is a more recently studied pathway that is becoming well characterized. MMEJ is generally distinguished by a double strand break that resects and re-anneals using 5-25 bases of homology (McVey &Lee 2008). It is usually associated with deletions and rearrangements. Potentially contractions are the result of a misalignment due to multiple regions of homology within the CTD (Figure 14).

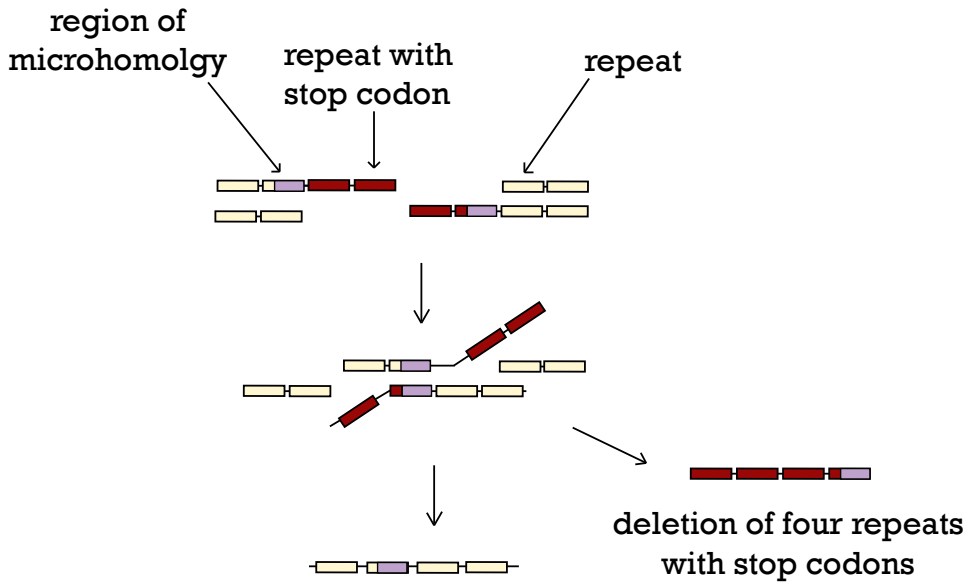


Figure 14. **Contractions may occur through misalignment due to multiple regions of homology within the CTD.** Following a break, strands re-anneal using regions of homology surrounding the four repeats containing stop codons and leading to a contraction. Repeats containing stop codons shown in red, regions of homology shown in purple.

MMEJ has been shown to be a separate pathway from NHEJ, and deletion of NHEJ essential factors does not suppress MMEJ (Lee & Lee, 2007). Our Ku70 deletion seemed to have very little effect on the occurrence of contractions, which is consistent with the involvement of MMEJ. MMEJ is also Rad52p independent, which agrees with our results showing that contractions seem to being occurring through a Rad52p independent pathway. A previous study showed that deletion of *pol32* led to a decrease the frequency of MMEJ, yet we saw an increase in contractions when we deleted it in our model (Lee & Lee, 2007). However, there are other polymerases shown to be involved in MMEJ such as those associated with Pol4p, Rad30p, and Rev3p. In the study they determined that MMEJ was

occurring if >5bp was used for re-annealing. Potentially the particular polymerase employed is determined by how many bases of homology are used for base pairing. Possibly certain polymerases are used when regions of homology are smaller and others are used when the homologous regions are larger. We cannot reject the involvement of MMEJ based on the increase in contractions seen in *pol32Δ* strain. One way to help access the involvement of MMEJ in contractions would be to test our construct in strains with other MMEJ associated proteins knocked out to see if the frequency of contractions increases or decreases.. Other NHEJ and SSA proteins are also shown to be involved in MMEJ such as the MRX complex, Nej1, and the Rad1/10 complex (Lee & Lee, 2007).

Since our *ku70* deletion did not have a significant effect on the percentage of contractions, and the *rad52* and *pol32* deletion results showed an increase in contractions, we do not believe contractions are occurring through HR or NHEJ. Therefore, proteins involved in MMEJ and HR or NHEJ may be good candidates to delete. Deletion of the NHEJ polymerase Pol4p was shown to have a greater effect on the survival of cells that relied on MMEJ than deletion of *pol32, so* that could be a good target for deletion. Other good candidates that are essential for MMEJ are Mre11p, which is involved in HR and NHEJ, or the NHEJ protein Nej1p. Deletion of any of these three genes should lead to a very significant decrease in contractions if they are occurring through MMEJ.

Conclusions

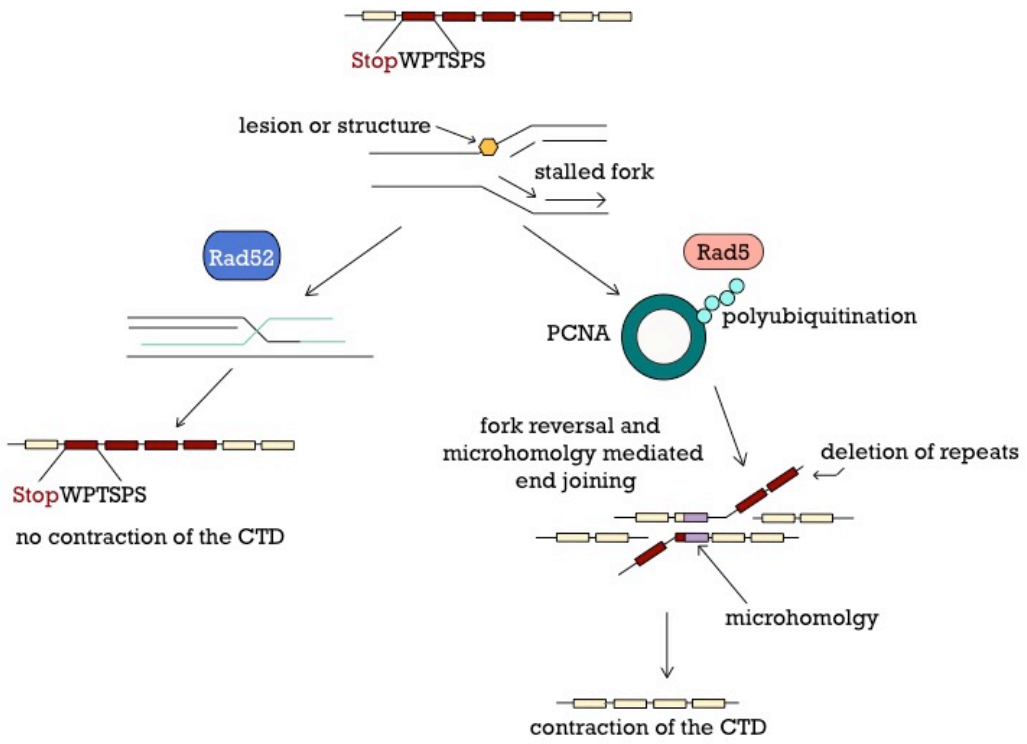


Figure 15. **Proposed model for contraction events involving Rad5p, template switching, and possibly microhomology mediated end joining.** The mechanism of contraction within the CTD seems to be independent of Rad52p, Pol32p, or Ku70p. This implies that the contractions are occurring through a non-HR or NHEJ pathway. The deletion of *rad5* had a very significant effect and basically eliminated contractions. Therefore we propose that the contractions are happening through fork stalling. When the replication fork encounters an obstacle, it stalls and reverses or collapses. In response, PCNA is polyubiquitinated by Rad5p. The contraction potentially happens when the single strands of the replication fork misalign. MMEJ may be involved since this mechanism occurs independently of both Rad52p and Ku70p. Also, there are regions of homology (either 11 or 14 bases long) seen in front of and after the contractions similar to what is seen in MMEJ.

Our current model of how these CTD contractions are occurring is that during replication, the fork encounters an obstacle, possibly a secondary structure forming within the CTD or a DNA-protein complex, that leads to recruitment of Rad5p and fork reversal (figure 15). Since there are regions of homology

surrounding all 4Stop contractions, this is likely an important feature involved in the contraction process. We propose that during fork reversal misalignment is occurring when single stranded DNA is exposed and this is leading to contractions (figure 16). When the single stranded lagging strand DNA is re-annealing with its complementary strand it misaligns and anneals at the wrong region of homology deleting the sequence in between. MMEJ may be involved, as contractions are occurring independently of Rad52p and Ku70p, and flanked by regions of microhomology. Further experiments will help clarify this theory.

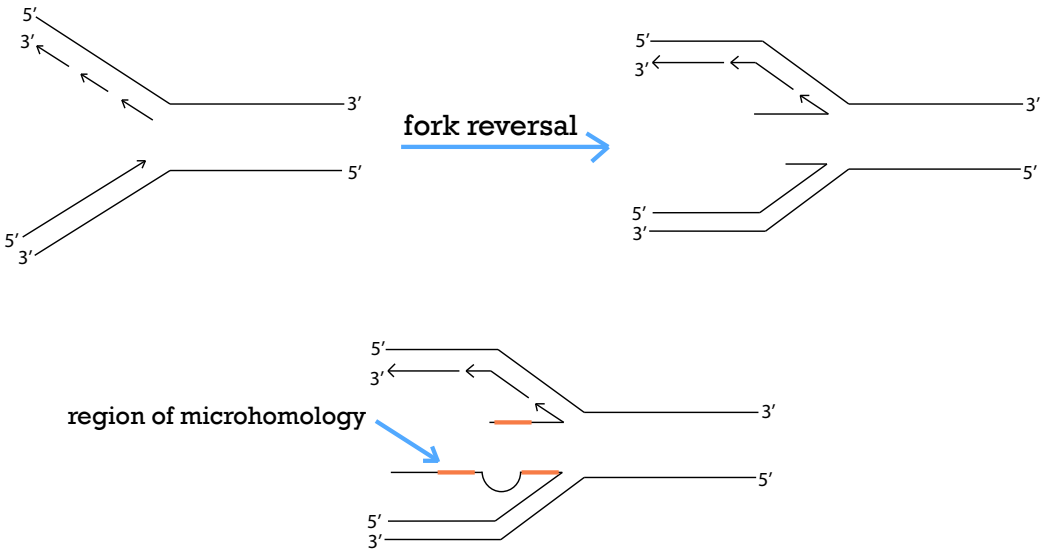


Figure 16. **Regions of homology within the CTD are likely leading to misalignment during fork reversal.** When the replication fork regresses, single stranded DNA is exposed and can misalign if there are multiple regions of homology. Potentially misalignment is occurring using the regions of homology surrounding the repeats containing stop codons and this is leading to the contraction events we see. Regions of homology are indicated in orange.

Our results also help support our hypothesis that since the CTD is essential, it contains more repeats than are necessary for wild type growth as a fail-safe to maintain viability. If repeats are mutated or altered in some way, then

they can be deleted without the cells becoming impaired or inviable. Our construct showed that contractions happen at a higher rate than previously seen in expansions (26 compared to 4.5 per 10^{-6}/cell/generation). I propose that this is because contraction events are sometimes necessary to maintain normal growth and avoid the negative affects of mutations. Our study helps explain how these contractions may be occurring and how having a long repetitive region in the CTD may actually help confer plasticity. It seems that this repetitive structure may be adaptive and advantageous and it would be interesting to find other essential genes that have evolved with similar mechanisms to maintain their function. If similar DNA repair mechanisms are utilized in other repetitive peptide situations such as flocculation (see chapter 1) or other binding mechanisms, knowledge of the proteins involved in expansion/contraction may be useful to develop an application for such proteins. Flocculation for example, is utilized by brewers to aggregate the yeast after fermentation.

This work helps to gain some insight into an interesting mechanism that has evolved in an essential gene and potentially will help understand other repetitive peptide sequences.

Chapter 3: Future Directions

The C-terminal domain of the RNAPII subunit Rpb1p is a dynamic repetitive heptapeptide sequence that is highly conserved amongst eukaryotes (26 repeats in *S. cerevisiae* and 52 repeats in humans) (Allison et al., 1985; Yang & Stiller, 2014). The CTD is essential for all steps of transcription and for cell viability. Though yeast has evolved with 26 repeats, it has been shown that only 13 repeats are needed to maintain normal growth, and when the tail is abbreviated to 8-12 repeats, growth is diminished but cells are still conditionally viable (Morrill et al., 2016; Nonet et al., 1987). This brings up noteworthy questions about why the CTD has evolved with so many seemingly-extra repeats. When the CTD tail is shortened to a length that causes diminished growth, suppressor mutations can arise that restore wild type growth (Morrill et al., 2016). We have also seen the CTD contract to delete whole repeats/multiple repeats containing mutations. This plasticity of the CTD seems to be an evolutionary strategy to cope with mutations without sacrificing essential functions.

My work has focused on understanding the abilities of the CTD to adapt and delete non-functional/mutated repeats to restore optimal growth. We suspect contractions are occurring through a Rad5p mediated pathway that involves replication fork stalling (figure 15). When the strands of the replication fork separate, misalignment likely occurs due to the repetitive nature of the CTD DNA sequence. If misalignment occurs with a newly formed okazaki fragment, a deletion may occur (figure 16).

Expansion and contraction in other repetitive amino acid sequences

Approximately 15-20% of the human genome contains repetitive amino acid sequences and potentially some of what we have learned from this study can be applied to other sequences (Mularoni et al., 2010). It would be interesting to examine other repetitive regions of proteins to determine if this is a general mechanism used for contraction of repetitive amino acid sequences. If this mechanism is broadly used then it could be helpful to understand and manage other systems. For example, the ability of *S. cerevisiae* cells to flocculate (adhere to other cells in groups) is exploited by brewers (see chapter 1). Flocculation is conferred by a large repetitive domain within the adhesin-encoding *FLO* genes. The longer the repetitive region, the better the adhesive abilities of the yeast (Verstrepen & Klis, 2006). After fermentation is conducted yeast can be removed easily since the flocculating yeast aggregate and float at the top of the tank (Verstrepen & Klis, 2006). Brewers have tried to increase the flocculation abilities of yeast to make this process easier. If contraction could be prevented and expansion increased in this system it would optimize the flocculation abilities of the yeast and simplify the process. Our proposed contraction mechanism could be easily tested and similar experiments run to better understand the expansion and contraction mechanisms of the *FLO* genes.

Pathogenic yeast also utilize the ability to adhere to other substances through similar repetitive amino acid sequences. The *ALS* genes (Agglutinin-Like Sequence) allows *Candida albicans* to adhere to host tissues (Hoyer et al., 2008). The middle region of the *ALS* genes is similar to that of the *FLO* genes, consisting

of 108-bp long, Ser/Thr-rich amino acid repeats. This repetitive region of the *ALS* genes leads to frequent recombination of cell surface proteins allowing the yeast to avoid recognition and thus avoid attack from the host organism's immune system (Verstrepen et al., 2005). This strategy is not unique to *C. albicans* or the *ALS* gene repeat region. This mechanism of frequent recombination within the repetitive region of an adhesin gene allows many other pathogenic strains to have similar abilities (Verstrepen et al., 2005). Understanding the mechanisms of amino acid repeat expansion and contraction in these systems may be beneficial in dealing with such pathogens in the future. It would be worthwhile to examine our contraction model in other systems and conduct similar assays to examine the instability/plasticity in other such repetitive amino acid sequences.

Experiments specific to our model of contractions within the CTD

Our hypothesis on the contraction mechanisms of the CTD of RNAPII could be further examined to help refine our model. We suspect that fork stalling and MMEJ are involved in the deletion of repeats, and other experiments will help to confirm or reject this. A way of possibly confirming the involvement of fork stalling in the contraction process would be to use a Topoisomerase 1 inhibitor to increase fork reversal. Misalignment may be occurring during fork stalling or fork reversal. Since there are multiple homologous regions of DNA within the CTD, it is plausible that the exposed single stranded DNA could misalign resulting in a deletion between the homologous regions. Chaudhuri et al. found that concentrations of 50 μM of the Top1p inhibitor CPT was sufficient to delay the

progression of 79% of replication forks in *S. cerevisiae* without inducing DSBs (Ray Chaudhuri et al., 2012). Using human cells they were then able to confirm this by visualizing the four-way junction created during fork reversal as a result of the Top1 inhibition. If fork stalling or fork reversal is involved in the contraction mechanism, we should see an increase in the percentage of contraction events helping us to confirm or reject this theory.

It would also be worthwhile to test deletions of proteins essential for MMEJ to help determine if contractions are occurring through this pathway. There are a number of proteins shown to be involved in MMEJ such as Pol4p, Rad30p, Rev3p, the Mre11/Rad50/Xrs2 complex, Nej1p, and the Rad1/Rad10 complex (Lee & Lee, 2007). Pol4 is a non-processive DNA polymerase seen to be involved in gap filling in NHEJ (Tseng & Tomkinson, 2004). Pol4p interacts directly with the NHEJ Dnl4/Lif1ligase complex, deletion of which leads to a partial decrease in MMEJ (Lee & Lee, 2007; Ma, Kim, Haber, & Lee, 2003). Rad30p (Polη) is one of the translesion synthesis (TLS) polymerases that are recruited in response to monoubiquitination of PCNA (Ghosal & Chen, 2013). Rev3p is the catalytic subunit of DNA Polζ and is also involved in lesion bypass. Polη is responsible for inserting a base opposite the lesion following which Polζ might extend the insertion. Polζ is then replaced by a replicative polymerase that finishes the repair (Ghosal & Chen, 2013).

The MRX complex, Mre11/Rad50/Xrs2, is necessary for MMEJ, however, this complex is involved in HR and NHEJ as well. The MRX complex is also involved in cell cycle check point regulation and telomere maintenance (Ma et al.,

2003). The Rad1/Rad10 endonuclease complex has also been shown to play a role in MMEJ. The Rad1/Rad10 heterodimer complex is important for removing 3' flaps in HR and is presumed to have a similar function in MMEJ (Lee & Lee, 2007). Rad1p has also been shown to be involved in nucleotide excision repair (NER) and single strand annealing (SSA). Various SSA and NHEJ factors are also involved in MMEJ. Sae2 promotes the nuclease activity of the Mre11/Rad50/Xrs2 and is involved in holding the DNA ends in the SSA pathway (Lee & Lee, 2007). Deletion of Sae2 leads to a decrease in MMEJ and an increase in NHEJ. Deletion of Nej1 also leads to a substantial decrease in MMEJ. Nej1 is another NHEJ protein responsible for localizing the essential Lif1 (ligase interfacing factor 1) to the nucleus and the site of the double strand break (Valencia et al., 2001).

Since contractions do not seem to be occurring through a HR or NHEJ mechanism, proteins that are involved in either of these repair pathways as well as MMEJ could be good targets for future deletion experiments. Deletion of Pol4, Mre11 and Nej1 all had very significant effects on the survival of cells that relied on MMEJ for repair (Lee & Lee, 2007). Deletion of Rad10 also caused a substantial decrease in MMEJ. These genes could be good candidates for deletion to further examine the involvement of MMEJ.

Stress response in connection with instability in repetitive regions

We are currently investigating whether stress has an affect on mutation rate and/or the types of mutations that occur in *S. cerevisiae* by using the CTD of

RNAPII as a model. By growing our strain under stressful conditions and performing fluctuation assays, we hope to better understand the effects of stress on expansions and contractions. Mutagenic DNA repair in response to stress has been seen in bacteria, but is not well characterized in yeast (Shor, Fox, & Broach, 2013a). Mammalian cells have been seen to actually increase mutagenic DNA repair in response to stress (Chatterjee, Lin, Santillan, Yotnda, & Wilson, 2015). Recently it was shown that certain forms of stress in human cells increased the frequency of mutations within trinucleotide repeats (Chatterjee et al., 2015). For 24 hours, cells were exposed to hypoxic or oxidative stress or temperature stress (30°C or 44°C) then allowed to recover and screened for contractions. The frequency of contractions increased three to sevenfold after two or three days. Tests were repeated in strains with certain stress response factors knocked out to confirm that the increase in contractions was actually due to the exposure to stress. These knockouts negated all or part of the increase in contraction frequency.

S. cerevisiae has been shown to sense and respond to environmental stresses such as changes in temperature, osmolarity, or oxidative stress, in what is termed the environmental stress response (ESR)(Gasch et al., 2000). Connections have recently been made linking the ESR in yeast to error-prone DNA repair and DNA damage bypass pathways in response to proteotoxic (Shor, Fox, & Broach, 2013b). However, the DNA repair mechanisms employed when yeast is under stress have not been well characterized and the influence of the ESR in mutagenic repair seem to be dependent on the type of stress. The ESR does not necessarily

provide protection from the stress, but rather affects the expression of certain genes. The ESR in yeast is characterized by the repression of about 600 genes and the induction of about 300 genes (totaling approximately 14% of the yeast genome)(Gasch et al., 2000). Repressed genes include genes involved in growth-related events such as initiation of transcription, RNA processing, and nucleotide production, as well as genes coding for ribosomal proteins. More than half of the induced genes are not characterized, but some of those that are characterized are involved in the synthesis, storage, or metabolism of carbohydrates such as trehalose or glycogen. This makes sense since cells under stress would want to preserve their carbohydrate and sugar stores to maintain cellular functions as long as possible. Adhesion genes in yeast are also shown to be activated in response to stress (Smukalla et al., 2008; Verstrepen & Klis, 2006). Flocculation helps cells under stress by protecting cells in the middle of a "floc" or by creating a biofilm to prevent being washed away in aqueous environments.

Currently we are performing preliminary fluctuation assays to see if growing *S. cerevisiae* in a stressful environment affects the rate or type of contraction events we see. These assays are being done with the 4Stop construct as well as the CTD8 construct. The plasmid of interest is transformed into GRY3019, allowed to grow normally on normal SC-LEU for three days, and re-streaked to confirm the transformation is correct. The strain is then re-streaked for large colonies on a stress plate, or re-streaked on a normal SC-LEU plate and put in a stressful environment. Stress plates include SC-LEU+1M Sorbitol (a less desirable sugar source), SC-LEU+1M NaCl, SC-LEU+1M KCL, and stress

conditions include growing cells on normal media at 37°C, 15°C, and room temperature. Under normal conditions colonies grow for 4-5 days until they are large enough to use for a fluctuation assay. In all stress conditions colonies grew very slowly. Colonies did not grow large enough to analyze at all when plated on media containing NaCl. Growth in the other conditions took about a 7-8 days. The 4Stop transformation did not grow a significant amount at 15°C, however the CTD8 transformation grew enough to be analyzed after almost a month.

So far we only have single trials from which to calculate mutation rates. When yeast transformed with the 4Stop plasmid is grown at 37°C instead of 30°C the mutation rate is 36.5 per 10^{-6}, which is a small increase when compared to the WT rate of 26.39 per 10^{-6}. When cells containing the 4Stop construct are grown on a SC-LEU+1M Sorbitol plate the mutation rate is only 2.32 per 10^{-6}. This rate is substantially lower than the WT mutation rate and the experiment needs to be repeated to confirm the significance of the results. When yeast transformed with the CTD8 construct was grown at 15°C the mutation rate was found to be 19.02 per 10^{-6}. This is notably higher than the wild type rate of 4.5 per 10^{-6}. This experiment should also be redone to confirm these results. Our limited preliminary data does seem to indicate that mutation rates may be different under stress conditions and these experiments should be replicated and explored further. Currently, we are determining if the frequency and types of contraction events we see is different in yeast exposed to stress. Since exposure to stress is seen to significantly increase the frequency of contractions in mammalian cells, we will potentially see a similar increase in contractions within the CTD.

Integration of the 4Stop construct into the genome of *S. cerevisiae*

Another experiment in progress is integrating the 4Stop sequence into the genome rather than having it on a plasmid. There has been debate concerning whether or not having the 4Stop-CTD on a plasmid is having an effect on the results that we have observed. Potentially we are transforming multiple plasmids into GRY3019 and if multiple plasmids are being transformed, the number of plasmids is not necessarily the same each time we do a transformation. This could lead to the plasmids using each other as temples and recombining. However, we do not believe that the contractions are occurring thorough a homologous recombination pathway as deletion of *rad52* did not decrease the frequency of contractions (see chapter 2). Since the CTD is essential and we cannot delete it, we are currently attempting to integrate the 4Stop sequence into the *CAN1* locus on chromosome 5 (*RPB1* is located on chromosome 4). We have built primers to amplify the 4Stop CTD sequence and its promoter, and produce 60 base extensions on either end that are homologous to the *CAN1* locus. Using a high efficiency yeast transformation protocol these linear PCR products can be transformed with GRY3019 and should become integrated at the *CAN1* locus. Transformations that have integrated into the genome will be able to grow on SC-ARG+CAN plates. The *CAN1* gene encodes a plasma membrane arginine permease and canavanine is a toxic structural analogue of arginine. When the *CAN1* gene is interrupted the strain should be resistant to canavanine. The strain can then be used on plates containing Doxycycline, which will turn off the wild type CTD so only our mutant CTD is being transcribed. We will then be able to

run some preliminary fluctuation assays and determine if there is a change in mutation rate or the frequency of contraction events. This will eliminate any concern that having the mutant sequence on a plasmid is interfering with the accuracy of our results.

Examining the role of DNA sequence in regions of instability

DNA composition affects the stability of a sequence and its ability to form secondary structures. Currently we are designing experiments to explore the effect that the DNA sequence of the CTD has on its stability. We are examining this by building constructs that code for the CTD amino acid sequence, yet have alternative DNA sequences (Table 4). The amino acid sequence of the CTD is repetitive but the DNA sequence varies slightly from repeat to repeat and can be altered slightly to test different hypotheses. For example, building a CTD with DNA more likely to form secondary structures such as G-quadruplexes could be interesting in examining the roles and effects of structure on the CTD. There is also an imbalance of Gs and Cs in the CTD with the coding strand having many cytosines while the non-coding strand has many guanines (referred to as G-C skew) (Morrill et al., 2016). To further examine this we could build CTDs with increased and decreased G-C skew. Potentially this will just increase the formation of G-quadruplexes, or maybe there are other effects of this skew. Another possible construct is one that has as many or as few Gs and Cs as possible. Since G-C pairs are joined by more bonds and are therefore more stable, increasing or decreasing this stability may have effects on frequency of

contraction/expansion events. Building alternative CTD constructs could help to understand if there is something inherently important about the DNA composition of the CTD. The insight gained from such experiments can also be applied on a larger scale. The concepts being examined here, such as structure formation and instability or G-C skew, are not unique only to the CTD, they are seen throughout the genome. The CTD may be a good model system to examine these concepts as we already have some insight into the mechanisms and rates of instability in this region. The system that we have developed would easily allow us to examine other DNA sequence themes and the principles learned may be applicable on a broader scale.

Table 4. **Alternative DNA sequences to build CTD constructs.** The DNA sequence can be changed while still maintaining the consensus amino acid sequence. Creating constructs with alternative DNA sequences can help understand the effects of features such as G4 structures and G-C skew within the CTD. We can also examine the effect that %GC content may have on contractions.

Construct	Y	S	P	T	S	P	S	%GC	G-C
WT	TAC	TCA	CCA	ACA	TCA	CCA	TCG	48	-8
High imbalance	TAC	TCC	CCC	ACC	TCC	CCC	TCC	72	-15
Low imbalance	TAC	AGT	CCG	ACG	AGT	CCG	AGT	52	0
G4 optimized	TAC	TCC	CCA	ACA	AGC	CCC	AGC	56	-10

The CTD of the RNAPII subunit Rpb1p is an interesting repetitive structure on its own and what we learn from its examination may help to understand other repetitive sequences. The sequence is highly conserved amongst

eukaryotes with a general trend towards an increased number of repeats in more complex organisms. The CTD has evolved with more than the necessary number of repeats needed for normal growth in what we propose is an evolutionary strategy to maintain essential functions. When repeats acquire mutations the CTD can contract to delete these repeats without compromising viability. Our work here has helped to understand the ability of the CTD to contract and the mechanism by which these contractions occur. The mechanisms that contribute to the plasticity of the CTD may help us understand similar repetitive regions and repetitive amino acid sequences as whole.

References

Ahn, S. H., Kim, M., & Buratowski, S. (2004). Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3' End Processing. *Molecular Cell*, *13*, 67–76.

Allison, L. A., Moyle, M., Shales, M., & James Ingles, C. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell*, *42*(2), 599–610. http://doi.org/10.1016/0092-8674(85)90117-5

Beckwitt, R., Arcidiacono, S., & Stote, R. (1998). Evolution of repetitive proteins: Spider silks from Nephila clavipes (Tetragnathidae) and Araneus bicentenarius (Araneidae). *Insect Biochemistry and Molecular Biology*, *28*(3), 121–130. http://doi.org/10.1016/S0965-1748(97)00083-0

Blastyák, A., Pintér, L., Unk, I., Prakash, L., Prakash, S., & Haracska, L. (2007). Yeast Rad5 Protein Required for Postreplication Repair Has a DNA Helicase Activity Specific for Replication Fork Regression. *Molecular Cell*, *28*(1), 167–175. http://doi.org/10.1016/j.molcel.2007.07.030

Branzei, D., & Foiani, M. (2007). Template Switching: From Replication Fork Repair to Genome Rearrangements. *Cell*. http://doi.org/10.1016/j.cell.2007.12.007

Byrd, A. K., & Raney, K. D. (2015). A parallel quadruplex DNA is bound tightly but unfolded slowly by Pif1 helicase. *Journal of Biological Chemistry*, *290*(10), 6482–6494. http://doi.org/10.1074/jbc.M114.630749

Carty, S. M., & Greenleaf, A. L. (2002). Hyperphosphorylated C-terminal Repeat Domain-associating Proteins in the Nuclear Proteome Link Transcription to

DNA/Chromatin Modification and RNA Processing. *Molecular & Cellular Proteomics*, *1*(8), 598–610. http://doi.org/10.1074/mcp.M200029-MCP200

Chatterjee, N., Lin, Y., Santillan, B. A., Yotnda, P., & Wilson, J. H. (2015). Environmental stress induces trinucleotide repeat mutagenesis in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3764–9. http://doi.org/10.1073/pnas.1421917112

Fondon, J. W., & Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(52), 18058–18063. http://doi.org/10.1073/pnas.0408118101

Fuchs, S., Laribee, N., & Strahl, B. (2009). Protein Modifications in Transcription Elongation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1789*(1), 26–36. http://doi.org/10.1016/j.bbagrm.2008.07.008.Protein

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., … Brown, P. O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes □ D. *Molecular Biology of the Cell*, *11*, 4241–4257.

Ghosal, G., & Chen, J. (2013). DNA damage tolerance: a double-edged sword gurading the genome. *Transl Cancer Res*, *2*(3), 107–129. http://doi.org/10.3978/j.issn.2218-676X.2013.04.01.DNA

Gietz, R. D., & Schiestl, R. H. (2007). Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols*, *2*(1), 35–37.

http://doi.org/10.1038/nprot.2007.14

Goodman, M. F. (2002). Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual Review of Biochemistry*, *71*(1), 17–50. http://doi.org/10.1146/annurev.biochem.71.083101.124707

Hall, B. M., Ma, C.-X., Liang, P., & Singh, K. K. (2009). Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics (Oxford, England)*, *25*(12), 1564–1565. http://doi.org/10.1093/bioinformatics/btp253

Helleday, T., Lo, J., Van Gent, D. C., & Engelward, B. P. (2007). DNA double-strand break repair: From mechanistic understanding to cancer treatment, *6*(2). http://doi.org/10.1016/j.dnarep.2007.02.006

Hoege, C., Pfander, B., Moldovan, G., Pyrowolakis, G., & Jentsch, S. (2002). RAD6 -dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO, *419*(September), 135–141.

Hoyer, L. L., Green, C. B., Oh, S.-H., & Zhao, X. (2008). Discovering the secrets of the Candida albicans agglutinin-like sequence (ALS) gene family--a sticky pursuit. *Medical Mycology*, *46*(1), 1–15. http://doi.org/10.1080/13693780701435317

Huang, M.-E., Rio, A.-G., Galibert, M.-D., & Galibert, F. (2002). Pol32, a Subunit of Saccharomyces cerevisiae DNA Polymerase □, Suppresses Genomic Deletions and Is Involved in the Mutagenic Bypass Pathway. *Genetics*, *160*(4), 1409–1422.

Janke, C., Magiera, M. M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., …

Knop, M. (2004). A versatile toolbox for PCR-based tagging of yeast genes: New fluorescent proteins, more markers and promoter substitution cassettes. *Yeast*, *21*(11), 947–962. http://doi.org/10.1002/yea.1142

Johansson, E., Garg, P., & Burgers, P. M. J. (2004). The Pol32 Subunit of DNA Polymerase Contains Separable Domains for Processive Replication and Proliferating Cell Nuclear Antigen (PCNA) Binding. *Journal of Biological Chemistry*, *279*(3), 1907–1915. http://doi.org/10.1074/jbc.M310362200

Johnson, R. E., Klassen, R., Prakash, L., & Prakash, S. (2015). A major role of DNA polymerase δ in replication of both the leading and lagging DNA strands. *Molecular Cell*, *59*(2), 163–175. http://doi.org/10.1016/j.molcel.2015.05.038

Komarnitsky, P., Cho, E. J., & Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes and Development*, *14*(19), 2452–2460. http://doi.org/10.1101/gad.824700

Lee, K., & Lee, S. E. (2007). Saccharomyces cerevisiae Sae2- and Tel1-dependent single-strand DNA formation at DNA break promotes microhomology-mediated end joining. *Genetics*, *176*(4), 2003–14. http://doi.org/10.1534/genetics.107.076539

Lopes, J., Kriegsman, B., Foiani, M., & Nicolas, A. (2011). G-quadruplex-induced instability during leading- strand replication. *The EMBO Journal*, *30*(19), 4033–4046. http://doi.org/10.1038/emboj.2011.316

Ma, J.-L., Kim, E. M., Haber, J. E., & Lee, S. E. (2003). Yeast Mre11 and Rad1

Proteins Define a Ku-Independent Mechanism To Repair Double-Strand Breaks Lacking Overlapping End Sequences. *Molecular and Cellular Biology*, *23*(23), 8820–8828. http://doi.org/10.1128/MCB.23.23.8820-8828.2003

Makarova, A. V, Stodola, J. L., & Burgers, P. M. (2012). A four-subunit DNA polymerase ζ complex containing Pol δ accessory subunits is essential for PCNA-mediated mutagenesis. *Nucleic Acids Research*, *40*(22), 11618–26. http://doi.org/10.1093/nar/gks948

Malagon, F., Kireeva, M. L., Shafer, B. K., Lubkowska, L., Kashlev, M., & Strathern, J. N. (2006). Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. *Genetics*, *172*(4), 2201–9. http://doi.org/10.1534/genetics.105.052415

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics*. Elsevier. http://doi.org/10.1016/j.tig.2008.08.007

Mirkin, S. M. (2006). DNA structures, repeat expansions and human hereditary disorders This review comes from a themed issue on Nucleic acids Edited. *Current Opinion in Structural Biology*, *16*, 351–358. http://doi.org/10.1016/j.sbi.2006.05.004

Morrill, S. A., Exner, A. E., Babokhov, M., Reinfeld, B. I., & Fuchs, S. M. (2016). DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. *Journal of Biological Chemistry*, *291*(22), 11540–11550. http://doi.org/10.1074/jbc.M115.696252

Mularoni, L., Ledda, A., Toll-Riera, M., & Alba, M. M. (2010). Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, *20*(6), 745–754. http://doi.org/10.1101/gr.101261.109

Mundlos, S., Otto, F., & Mundlos, C. (1997). Mutations Involving the Transcription Factor CBFA1 Cause Cleidocranial Dysplasia. *Cell*, *89*, 773–779.

Neelsen, K. J., & Lopes, M. (2015). Replication fork reversal in eukaryotes: from dead end to dynamic response. *Nature Reviews. Molecular Cell Biology*, *16*(4), 207–220. http://doi.org/10.1038/nrm3935

Nonet, M., Sweetser, D., & Young, R. A. (1987). Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell*, *50*(6), 909–915. http://doi.org/10.1016/0092-8674(87)90517-4

Pearson, C. E., Edamura, K. N., & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, *6*(10), 729–742. http://doi.org/10.1038/nrg1689

Phatnani, H. P., & Greenleaf, A. L. (2004). Identifying PhosphoCTD-Associating Proteins. In *mRNA Processing and Metabolism* (Vol. 257, pp. 17–28).

Phatnani, H. P., & Greenleaf, A. L. (2006, November 1). Phosphorylation and functions of the RNA polymerase II CTD. *Genes and Development*. Cold Spring Harbor Lab. http://doi.org/10.1101/gad.1477006

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., … Cramer, P. (2015). Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature*, *518*(7539), 376–380.

http://doi.org/10.1038/nature14229

Ray Chaudhuri, A., Hashimoto, Y., Herrador, R., Neelsen, K. J., Fachinetti, D., Bermejo, R., … Lopes, M. (2012). Topoisomerase I: poisoning results in PARP-mediated replication fork reversal. *Nature Publishing Group*, *19*(4), 417–424. http://doi.org/10.1038/nsmb.2258

Shah, K. A., & Mirkin, S. M. (2015). The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair*, *32*, 106–112. http://doi.org/10.1016/j.dnarep.2015.04.020

Shor, E., Fox, C. A., & Broach, J. R. (2013a). The Yeast Environmental Stress Response Regulates Mutagenesis Induced by Proteotoxic Stress. *PLoS Genetics*, *9*(8), e1003680. http://doi.org/10.1371/journal.pgen.1003680

Shor, E., Fox, C. A., & Broach, J. R. (2013b). The Yeast Environmental Stress Response Regulates Mutagenesis Induced by Proteotoxic Stress. *PLoS Genetics*, *9*(8), e1003680. http://doi.org/10.1371/journal.pgen.1003680

Shrivastav, M., De Haro, L. P., & Nickoloff, J. A. (2008). Regulation of DNA double-strand break repair pathway choice. *Cell Research*, *18*(1), 134–147. http://doi.org/10.1038/cr.2007.111

Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., … Verstrepen, K. J. (2008). FLO1 Is a Variable Green Beard Gene that Drives Biofilm-like Cooperation in Budding Yeast. *Cell*, *135*(4), 726–737. http://doi.org/10.1016/j.cell.2008.09.037

Tseng, H. M., & Tomkinson, A. E. (2004). Processing and joining of DNA ends coordinated by interactions among Dnl4/Lif1, Pol4, and FEN-1. *Journal of*

Biological Chemistry, *279*(46), 47580–47588. http://doi.org/10.1074/jbc.M404492200

Valencia, M., Bentele, M., Vaze, M. B., Herrmann, G., Kraus, E., Lee, S. E., … Haber, J. E. (2001). NEJ1 controls non-homologous end joining in Saccharomyces cerevisiae. *Nature*, *414*(6864), 666–669. http://doi.org/10.1038/414666a

Verstrepen, K. J., Jansen, A., Lewitter, F., & Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nature Genetics*, *37*(9), 986–990. http://doi.org/10.1038/ng1618

Verstrepen, K. J., & Klis, F. M. (2006, April). Flocculation, adhesion and biofilm formation in yeasts. *Molecular Microbiology*. Blackwell Publishing Ltd. http://doi.org/10.1111/j.1365-2958.2006.05072.x

Wang, G., & Vasquez, K. M. (2014). Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair*, *19*, 143–51. http://doi.org/10.1016/j.dnarep.2014.03.017

Werner-Allen, J. W., Lee, C.-J., Liu, P., Nicely, N. I., Wang, S., Greenleaf, A. L., & Zhou, P. (2010). cis-Proline-mediated Ser(P) 5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72 * □ S. http://doi.org/10.1074/jbc.M110.197129

Yang, C., & Stiller, J. W. (2014, April 22). Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. *Proceedings of the National Academy of Sciences of the United States of America*. National Acad Sciences. http://doi.org/10.1073/pnas.1323616111

Appendix 1: Previously Published Work

This research was originally published in The Journal of Biological Chemistry. Morrill, Summer A., Exner, Alexandra E., Babokhov, Michael, Reinfeld, Bradley I., Fuchs, Stephen M. DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. *Journal of Biological Chemistry.* 2016; *291*(22), 11540–11550.

,

DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain

Summer A. Morrill, Alexandra E. Exner, Michael Babokhov, Bradley I. Reinfeld, and Stephen M. Fuchs*

From the Department of Biology, Tufts University, Medford, MA, 02155, USA

Running title: *CTD instability*

To whom correspondence should be addressed: Prof. Stephen M. Fuchs, Department of Biology, Tufts University, 200 Boston Ave Suite 4700, Medford, MA, USA 02155 Tel: 1-617-627-0310; Fax: 1-617-627-0309; Email: stephen.fuchs@tufts.edu

Keywords: RNA polymerase, DNA instability, G-quadruplex, tandem repeat

ABSTRACT

The C-terminal domain (CTD) of RNA polymerase II in eukaryotes is comprised of tandemly repeating units of a conserved seven amino acid sequence. The number of repeats is, however, quite variable across different organisms. Furthermore, previous studies have identified evidence of rearrangements within the CTD coding region suggesting DNA instability may play a role in regulating or maintaining CTD repeat number. The work described here establishes a clear connection between DNA instability and CTD repeat number in *S. cerevisiae*. First, analysis of 36 diverse *S. cerevisiae* isolates revealed evidence of numerous past rearrangements within the DNA sequence that encodes the CTD. Interestingly, the total number of CTD repeats was relatively static (24-26 repeats in all strains), suggesting a balancing act between repeat expansion and contraction. In an effort to explore the genetic plasticity within this region we measured the rates of repeat expansion and contraction using novel reporters and a doxycycline-regulated expression system for *RPB1*. In efforts to determine the mechanisms leading to CTD repeat variability we identified the presence of DNA secondary structures, specifically G-quadruplex-like DNA, within the CTD coding region. Furthermore, we demonstrated that mutating *PIF1*, a G-quadruplex specific helicase, results in increased CTD repeat length polymorphisms. We also determined that *RAD52* is necessary for CTD repeat expansion but not contraction, identifying a role for recombination in repeat expansion. Results from these DNA rearrangements may help explain the CTD copy number variation seen across eukaryotes, as well as support a model of CTD expansion and contraction to maintain CTD integrity and overall length.

RNA polymerase II (RNAPII) is an essential protein complex responsible for the transcription of messenger RNA (mRNA) in eukaryotes. It shares many components with other RNA polymerases but its most distinctive

feature is a C-terminal domain (CTD) in the largest and catalytic subunit Rpb1p, comprised of repeating units of a seven amino acid sequence. The CTD is essential for proper transcription and acts as a binding domain for a variety of proteins involved in transcription elongation, mRNA processing, chromatin remodelling, and DNA repair (1,2).

In all eukaryotes the CTD is comprised of nearly identical repeats of a seven amino acid sequence, YSPTSPS. However, the number of CTD repeats is quite different across organisms. For example, mammals can have as many as 52 repeats. In contrast, budding yeast have only 26 repeats. Surprisingly, previous studies to understand the role of CTD repeats revealed that as few as 8 repeats will support growth (3). It is still unclear why so many CTD repeats are conserved if they are not essential for function.

Early mutational studies noted that cells could undergo spontaneous mutations within the CTD coding region to repair premature stop codons (within the repeat) that compromised function (4). Although never fully explored, these investigators observed that the DNA encoding the CTD underwent either expansion or contraction, resulting in a new sequence that produced a longer protein product. This led us to question what aspect of the CTD coding sequence was promoting DNA instability and to hypothesize that the non-essential CTD repeats primarily act as a template for DNA rearrangement.

Repetitive elements in the genome are well known to be unstable. For example, trinucleotide repeat sequences such as CAG repeats are known to form stable hairpin secondary structures which interfere with DNA replication, resulting in DNA breakage. To overcome DNA secondary structures cells either deploy certain helicases to unwind these structures or implement DNA repair pathways to repair DNA damage associated with these non-canonical DNA conformations (5,6). This repair process leads to changes in repeat length which in the case of many trinucleotide repeats can lead to a number of severe neurodegenerative diseases (7).

The repetitive CTD region differs from most other repetitive regions in two important ways: 1) The coding sequence is comprised of a degenerate 21 bp repeat, and 2) it is a conserved sequence in an essential protein. It is therefore not intuitive as to what aspect of the CTD sequence might make it prone to rearrangements. Furthermore, we were interested in what potential evolutionary advantage there would be in having an unstable region within an essential protein.

Here we describe the development of a genetic system to examine DNA rearrangements in the repetitive CTD region of *RPB1*. We measure genetic expansion and contraction frequencies and identify cellular factors that regulate this process, including the recombination factor Rad52p and the G-quadruplex DNA (G_4-DNA) associating helicase, Pif1p (8). Lastly, we have uncovered an

unusual signature within the DNA of the CTD coding region that forms secondary structure *in vitro*. Based on these findings we propose a model for CTD repeat instability and maintenance in budding yeast and speculate as to how CTD length has evolved in different organisms.

EXPERIMENTAL PROCEDURES

Yeast strains and plasmids – Strains used in this study were derived from GRY3019 (*MATa his3 Δ leu2 Δ lys2 Δ met15 Δ trp1 Δ::hisG URA::CMV-tTA kanRPtetO7-TATA-RPB1*) (9). The *rad52*Δ strain was constructed using heterologous gene replacement (10) and verified by PCR using primers described in **Table 1**. The *pif1-m2* mutation was made following the procedure described previously (11). Yeast were grown on synthetic complete (SC) dropout medium or YPD as indicated. Doxycycline (+DOX, 50 μg/ml) was added to plates to control the expression of genomic *RPB1*. When the selection drugs G418 or ClonNat were added to SC plates, ammonium sulfate was replaced by monosodium glutamate (1g/L) as a nitrogen source (12).

In order to create repetitive CTD constructs of precise length and sequence, we used the recursive directional ligation by plasmid reconstruction (PRe-RDL) technique and plasmid JMD2 described by the Chilkoti lab (13). CTD segments were iteratively ligated together to synthesize CTD regions that were verified by sequencing. This region was amplified using T7 forward and T7 reverse primers complementary to JMD2, digested with XmaI and SacI, and subcloned into a variant of pL-RPB1 (9), which was altered by Quikchange site-directed mutagenesis to introduce unique XmaI and SacI restriction sites upstream and downstream of the CTD coding region within *RPB1*. We intended to design a CTD mutant where repeat 9 was interrupted by a stop codon that would result in a protein that could not support growth. The CTD-4stop plasmid was created by Quikchange mutagenesis using primers with homology to repeat 9 of the CTD. Due to the repetitiveness of the CTD we were never able to isolate a mutant with a single insertion of this sequence and therefore used the isolated pRPB1-4stop plasmid for our experiments. Plasmids used in this study and relevant sequence information are listed in **Table 2**. CTD-containing plasmids and control vectors (14) were transformed into yeast and selected on SC-LEU media (15).

Spotting assays – For phenotypic growth assays, yeast were grown overnight in synthetic complete medium lacking leucine (SC-LEU) to ensure retention of the plasmid copy of *RPB1* (pRPB1). Saturated cultures were used to start fresh cultures in the same medium at an A_{600} of 0.2. Cells were allowed to double at least two times, harvested and resuspended to an A_{600} of 1.0 in sterile water in a 96-well plate. Cells were 5-fold serially diluted and then spotted onto plates containing necessary selections using a 48-pin replicating tool. Plates were incubated at indicated temperatures and imaged daily.

Determining mutational frequency of the truncated CTD – Fluctuation analysis was modified from Aksenova et al. (2013) to fit the doxycycline (DOX) selection system (16). Briefly, strains were streaked for isolation onto SC-LEU to maintain pRPB1, and incubated at 30°C for 3–5 days (approximately when colonies ~10^6 cells, which varies by strain). For each initial colony, cells were suspended in liquid and plated for selection on +DOX media; the same cell suspension was diluted 10^4-fold and plated on YPD as a cell count control. For each experiment, cells derived from 12 individual colonies were plated. Suppressors were identified by colony formation on SC-LEU + DOX media after ~4 days. Colony counts from both the DOX selection and the matched control plates were used as input for the Fluctuation Analysis Calculator (FALCOR), a means of calculating the mutational frequency for each strain using the MSS-Maximum Likelihood Estimator method (MSS-MLE) (17). For mutation-type analysis, a maximum of 4 colonies from any plate were analyzed by colony PCR. Primers flanking the CTD were used to amplify the CTD coding region using a modified PCR protocol (18). To prepare a crude DNA template from the yeast colonies, cells were lysed in 50 µl 0.4% SDS at 90°C for 4 minutes and briefly spun; 1µl of the supernatant was used as template for a 50ul PCR reaction. To modify the PCR reaction mix, we added Triton X-100 to a final concentration of 1%. Unless otherwise noted, gel electrophoresis was done using 2% agarose gels, 1X TBE, SYBR Safe DNA gel stain and run with a 100bp ladder.

Western bloting – Yeast cultures were grown in synthetic defined medium with or without DOX (50 µg/mL) from a starting A_{600} of between 0.1 and 0.2 and grown to mid-log phase (A_{600} 0.6–1.0). In all cases, five optical units of cells were harvested by centrifugation, and extracts were prepared as previously described (19). For Western analysis of Rpb1p, proteins were separated on 8% SDS-PAGE gel made with 19:1 acrylamide:bis-acrylamide. Gels were transferred to PVDF for 90 min at 45 mA and dried in methanol to block according to the manufacturer's instructions (Millipore-Immobilon-P). Dried membranes were rehydrated briefly in methanol and incubated with primary antibodies (Y-80 Santa Cruz) overnight at 4 °C in PBST (phosphate-buffered saline including 0.05% Tween 20) containing 4% milk. Western blots were visualized using HRP-conjugated secondary antibodies and ECL Plus chemiluminescence (GE Healthcare).

Circular dichroism – Circular dichroism measurements were taken using a Jasco 810 instrument at 25˚C in 2 mM sodium phosphate buffer pH 7.0 with or without 100 mM KCl. Samples were heated to 95˚C and slowly cooled to room temperature where they equilibrated for a minimum of 48 hours prior to analysis. Measurements were taken at 0.5 nm intervals from 200–320 nm. Traces

shown are the average of three measurements for each sample.

RESULTS

Variability within the RNAPII CTD sequence of natural and laboratory isolates – Many groups have investigated the evolution of the unusual repetitive sequence that comprises the CTD of RNAPII across organisms (20-22). We wanted to look more closely for evidence of recent expansion or contractions within this coding region, and took advantage of strains available from the Saccharomyces Genome Resequencing Project (SGRP) (23). Due to overlapping reads implemented in Next generation sequencing, repetitive domains are often difficult to accurately characterize. Therefore, we resequenced the CTD region of 36 of these strains using Sanger sequencing. Careful multiple sequence alignment revealed that the 36 strains could be arranged into 14 different sequence groups based on sequence similarity (**Figure 1**). Half of these groups were defined by the addition or subtraction of a 21 base pair repeat unit at different locations within the sequence, shown in **Figure 1** as gaps in the sequence. These gaps are evidence of a rearrangement with the CTD repeat coding sequence. There were 11 instances of single nucleotide polymorphisms (SNPs) across all 36 strains (shown in orange in **Figure 1**). The full alignments can be found in **Figure 2**.

Yeast correct mutations within the CTD primarily through contraction – Sequencing of strains allowed us to uncover past instances of CTD rearrangement, but this does not tell us anything about the relative frequency at which these arrangements occur. In order to directly observe rearrangements occurring within the CTD coding sequence, we synthesized a variant of the CTD (pRPB1–4stop) in which repeats 8–11 each contained a stop codon Tyr1→stop (**Figure 3B**) and a non-coding Ser2→Trp mutation. Previous studies had determined that 8 or more CTD repeats were required for efficient growth and we predicted this mutant RPB1-4stop gene would not be capable of supporting growth. To test this, we devised a Tet-Off expression system based off the work of Strathern and others (9). Briefly, we introduced CTD variants on CEN/ARS-containing plasmids into yeast under the control of the native *RPB1* promoter. When appropriate, doxycycline was introduced to repress transcription of the genomic copy of *RPB1*, which was under the control of a tetracycline-responsive promoter (**Figure 3A**). Under these conditions we could force the cell to rely on the Rpb1p variant for transcription and monitor these effects by growth on SC-leucine plates containing doxycycline. As expected, the protein produced from our RPB1-4stop mutant harbours only 7 CTD repeats and cells show extremely compromised growth comparable to a RPB1–CTD_8, a CTD variant that contains only 8 repeats (**Figure 3C**). Cells were grown to large colonies to allow for the accumulation of spontaneous mutations. When plated on media

containing doxycycline, after 3–4 days we observed the presence of colonies that were able to bypass the slow growth conferred by our reporter constructs. We expected these fast growing suppressors to result from four types of events: 1) contractions where the introduced stop codons were removed from the suppressor plasmids, 2) expansions where the region encoding the first 8 repeats was expanded to yield a protein product that had more than 8 repeats, 3) homologous recombination events where the mutant plasmid copy of *RPB1* underwent a rearrangement with the doxycycline-regulated copy of *RPB1* in the genome, and 4) mutations elsewhere in the genome. We performed a fluctuation analysis to measure the mutation frequency of this reporter and determined the relative proportion of different mutagenic events based on PCR analysis of numerous fast-growing suppressors. In order to avoid potential bias for jackpot mutations, cells plated on each fluctuation plate were derived from individual yeast colonies. Furthermore, we analysed no more than 4 colonies from any one individual plate. Data are summarized in **Table 3**. Nearly half of the events we observed were the result of contractions that remove the stop codons (**Figure 4A, 4B**). Interestingly, for all contractions the product yielded a protein sequence with the expected YSPTSPS sequence at all repeats. To confirm that the improved growth was due to the change in the plasmid, several suppressor plasmids were retransformed into GRY3019 and monitored for improved growth (**Figure 4A**), CTD coding region length (**Figure 4C**), and Rpb1p protein size by western blot (**Figure 4D**).

Truncated CTD variants undergo spontaneous expansion – We expected our pRPB1-4stop reporter construct to reveal both expansion and contractions within the CTD coding region, however, expansions were difficult to detect with this system due to the high frequency of contractions. We therefore set out to design a second reporter construct to measure expansion by simply removing repeats from the C-terminal end of the CTD. First we determined the CTD length required to support growth in our tet-off system. Under these conditions, we see that CTD variants with as few as 10 repeats grow normally on +DOX plates, whereas cells with 8 repeats show impaired growth (**Figure 5A and 5B**). We therefore measured the appearance of suppressor colonies using this pRPB1-CTD_8 plasmid. For pRPB1-CTD_8, we observed the appearance of fast growing suppressor colonies after three or four days of growth at 30˚C. We performed a fluctuation assay as described above to characterize the events that allowed cells to bypass our CTD truncation mutants (**Figure 6**). We used colony PCR to characterize suppressor events as expansions, homologous recombination, or extragenic mutations, and determined the frequency of each event (**Table 3**). Using this construct, we observed that expansions occurred in about one third of all colonies and were more common than homologous

recombination with the genomic copy. Subsequent sequencing of expansion and recombination events revealed that the CTD undergoes expansion from 8 to up to 30 repeats – longer than the canonical wild-type sequence (**Figure 6B and 6C**).

G_4-DNA in the CTD coding region– Expansion and contraction of trinucleotide repeat DNA is often mediated by slipped strands resulting in regions with DNA secondary structure (e.g. hairpins), which in turn involves the recruitment of various proteins associated with DNA repair processes. In yeast, the CTD coding consists of an imperfect 21bp repeat where (for strain S288C) only two repeats share the same sequence (**Figure 7**). Therefore it is not intuitive how secondary structure might form to promote expansion and contraction processes. Capra *et al.* recently reported G-quadruplex (G_4-DNA) near the 3' end of the *RPB1* coding sequence (24). G_4-DNA is known to promote DNA rearrangements in many organisms, including in the promoter of MYC in humans (8,25). While the rest of the coding region was not predicted to form additional G-quadruplex structures, we uncovered an unusual sequence feature within the CTD coding region, namely an increased frequency of cytosine bases on the coding strand (**Figure 8A**). Conversely, the non-coding strand would be rich in guanine nucleotides, a feature known to promote secondary structure formation (26). Interestingly, this discrepancy in G:C ratio is restricted to the CTD coding region (**Figure 8B**).

To test whether these sequences may be forming secondary structure, we used the QGRS Mapper (27) to identify candidate regions from the CTD coding region. We synthesized several single-stranded oligonucleotides corresponding to regions of the CTD coding sequence that have four consecutive tracks of at least two guanines and loop lengths of fewer than 10 bases. Indeed, by circular dichroism these oligonucleotides formed stable structures (**Figure 8C**) with a broad maxima between 265-300 nm and a minimum near 240 nm in the presence of potassium ions. In our hands this is consistent with what has been reported for complex mixed parallel/antiparallel G_4-DNA structures (28) such as that formed by human telomeres (**Figure 8C**). G4-DNA structures generally require the presence of a cation (K^+ > Na^+). As such, we performed equivalent experiments in the absence of K^+ (but still 2 mM Na^+ from the buffer). Under these circumstances all CTD-derived DNAs still exhibited some structure, but there was little to no similarity among any of the DNA sequences under these conditions.

The DNA helicase Pif1 suppresses CTD rearrangement - The Zakian lab recently showed that the G_4-DNA containing region within the CTD coding region also overlaps with genome association by the G-quadruplex unwinding helicase Pif1p (29). To test the role of Pif1p in mediating CTD expansion, we measured expansion

and contractions in a *pif1-m2* mutant. This mutant disrupts Pif1p activity in the nucleus and thus the mutant is predicted to have more difficulty resolving G_4-DNA structures (11). Indeed, the percentages of suppressors using our pRPB1-4stop and pRPB1-CTD_8 constructs were both increased in *pif1-m2* cells (**Table 4**). In particular, mutating *PIF1* increased expansion in pRPB1-CTD_8 from 38% to 56% of all suppressor events.

Rad52 is required for expansion of CTD repeats – Much like expansions in microsatellite DNA, we hypothesized that CTD repeat length variation resulted from DNA damage repair pathways, primarily homologous recombination (HR). Accordingly, we measured expansions and contractions in recombination-deficient *rad52*Δ cells (**Table 4**). Expansions were essentially absent when *RAD52* was deleted. Surprisingly, in our contraction assay we found that nearly all of the recovered pRPB1-4stop suppressors exhibited contraction, though the total number of suppressors identified was decreased in *rad52Δ* cells. We interpret these data to mean that HR is the preferred cellular pathway leading to suppressor mutants. However, HR only leads to length polymorphism when it results in an expansion. Conversely, contractions in our system must result from a recombination-independent mechanism that competes with HR.

DISCUSSION

The CTD of RNAPII has been the subject of scientific fascination for many years. Not only does it exhibit a highly conserved and highly repetitive protein sequence, but it also is the substrate for numerous post-translational modifications that dictate its function. The CTD has received much attention from an evolutionary perspective due to the fact that the number and composition of 7-amino acid repeats differs so greatly between organisms (20,30-32). Of particular interest to us was the fact that model organisms, such as budding yeast, have CTDs that are considerably longer than necessary for laboratory growth. Furthermore, historical observations by our lab and many others suggested that the CTD region, though essential, was also very prone to mutation. From these two observations we reasoned the non-essential CTD repeats may serve a function as templates for DNA repair. Here we explore that possibility by identifying some of the factors that contribute to CTD instability and variability.

We resequenced 36 yeast strains to properly characterize the underlying variability within the CTD coding region. Our initial evaluation of the sequence alignment suggested a high propensity of single-nucleotide polymorphism (SNP) across this CTD coding region. However, careful re-examination of the DNA encoding each repeat following Sanger sequencing revealed numerous contraction events and evidence of the addition of at least one novel repeat. Repetitive

minisatellites are prone to expand and contract (33-35). However, what differentiates the CTD of RNAPII is that this variable region encodes a conserved, essential region within an essential protein. Furthermore, the DNA encoding these repeats is in fact relatively dissimilar even though it encodes the same peptide sequence **(Figure 7)**. These factors make its instability worthy of additional study.

Our analysis of the SGRP sequencing data supports a model where the CTD coding region spontaneously contracts or expands resulting in a CTD length that could potentially alter fitness. Growth experiments from our lab and others clearly show that cells expressing very short CTDs (8-10 repeats) are less fit than cells with what is considered a full-length CTD (3,36). Nonetheless, this initial finding suggested that yeast have the ability to explore different CTD repeat lengths. Previous studies from Corden and Young and others clearly demonstrate that cells expressing CTDs with at least 12 repeats show near wildtype growth under a range of growth conditions. Therefore it is interesting that the CTD copy number is variable in yeast but has fixed at a number of repeats much geather than what is required for viability.

We tested our model by developing two different genetic reporters – one that predominantly measured CTD contraction (pRPB1-4stop) and the other that measures expansion (pRPB1-CTD_8). Both took advantage of a very powerful doxycycline-regulated *RPB1* system developed by the Strathern lab, and utilized the inability of Rpb1p with short CTDs to support normal yeast growth. This system turns out to be extremely important for the quantitation of suppressor events. Previous studies with CTD mutants have been conducted using the *URA3* shuffle system (4,36,37), which requires a round of 5-FOA selection to cure cells of a wildtype *RPB1* plasmid. During this time, cells are prone to fast-growing suppressor mutations. By using a small-molecule controlled system, we can switch off wildtype expression in already growing cells, reducing the likelihood of spurious mutation events.

Using our reporters we were able to identify three broad classes of fast-growing suppressors – expansions/contractions (changes in length of the CTD as measured by PCR), homologous recombination events with the genomic copy of *RPB1*, or mutations elsewhere in the genome that supported growth but have yet to be characterized. We were able to show by reintroduction of isolated suppressor plasmids that rearrangements of the DNA that led to constructs expressing more than 8 wildtype CTD repeats were sufficient to confer improved growth. This additional large class of mutations elsewhere in the genome can presumably bypass our selection system through a number of mechanisms including altering *RPB1* expression (or copy number), altering sensitivity to DOX, or facilitating transcription elongation. In fact, Young and colleagues used a similar system to initially identify members of the Mediator complex

more than twenty years ago (38). The increased flexibility and sensitivity of our DOX system may facilitate future identification of additional regulators of transcription.

The rate of suppressor accumulation for both reporters was more than 1 $x10^{-6}$ per cell per cell division, with the rate in our contraction assay being 10-fold higher than our expansion assay. Even though the DNA repeat of the CTD is a degenerate 21 bp repeat, this preference for contraction over expansion is consistent with other reports of instability within repetitive DNA sequences (39). This also explains the low frequency of expansions seen with the pRPB1-4stop construct, as it would likely be necessary to screen several hundred colonies to identify relatively rare expansion events amongst the more common contractions.

Mainly from studies of microsatellite DNA, several models have been developed to explain DNA expansion and contraction (34,40). DNA secondary structure formation is important for many types of microsatellite instability including GAA and CAG repeats. In these models, stable secondary structures block processes, such as replication, leading to a break in the DNA that must be repaired. It is during this repair that sequence is lost (contraction) or reamplified (leading to expansion). Most of these studies have been performed on repeat sequences of fewer than 6 bp. The sequence encoding the CTD is unusual in that it is comprised of a 21 bp repeat and is non-identical in the native sequence. Nonetheless, we noted that there is a feature of this sequence that supports secondary structure formation – namely an imbalance in the number of cytosine bases relative to guanine bases (225:44 C:G) on the coding strand. The high density of guanine bases in the CTD noncoding strand is able to form a stable secondary structure that resembles G_4-DNA structures by circular dichroism (**Figure 8**). This phenomenon was first described in minisatellite regions of the human genome (41), and G-rich strands from microsatellites such as the CGG repeat from *FMR1* (Fragile X mental retardation 1) and the CAGGG repeat from Ms6-hm were subsequently shown to form secondary structure (26,42). Likewise, these regions exhibited extremely high instability with mutation rates greater than 1% (26,43). Interestingly, many but not all imbalanced minisatellites studied to date have been in non-coding regions. Furthermore, their base repetitive unit is generally less than 10 base pairs in length. How the CTD coding sequence may have evolved to have such strong imbalance remains a mystery. However, it is intriguing that each repeat encodes two prolines, ensuring a regular repeating CCN sequence every 6 or 9 bp, invariably across all 26 repeats and heavily skewing the GC balance. Loop length between poly-guanine tracks has been shown to be crucial for the structure and stability of G-quadruplexes with shorter and uniform loop lengths being preferred. We measured the

structure of several oligonucleotide sequences derived from the CTD-coding region using circular dichroism. Nearly all formed structures despite non-uniform loop length and at least one loop of more than 7 bp. The exact structure of these oligonucleotides is still unknown. However, due to the large number of G-tracks within the CTD, there is a possibility of secondary structure formations at dozens of sites within the CTD region.

This leads to a model for instability wherein the strong bias in single strand nucleotide composition within a region, rather than repetitive DNA, could promote instability. As expected, mutation of *PIF1*, a G_4-DNA specific helicase, indeed increased the percentage of expansions in our system. It also drastically decreased the percentage of homologous recombination events with the genomic copy of *RPB1*. This can be explained based on recent work in which Pif1p is required for efficient recombination-coupled DNA repair (44). This also suggests HR as a probable pathway mediating CTD coding region instability. To begin to test this we deleted *RAD52* – a central player in several branches of DDR. As expected, expansion events were nearly eliminated in *rad52*Δ, confirming a role for recombination in this process. Deletion of *RAD52* decreased the overall number of suppressors isolated, but increased the percentage of contraction events. We interpret these results to mean that homologous recombination is necessary for expansion but competes for the mutagenic process of contraction. This also means contraction must occur through a recombination-independent pathway. It is not uncommon for multiple DNA repair processes to be recruited to the same site, as this has been shown for minisatellite instability in yeast (45) and double strand break repair in *Drosophila* (46). Sequence analysis of suppressor mutants isolated from the RPB1-4stop construct show that the deletions tend to unite sites of microhomology on both sides of the stop codon (**Figure 4B**). There are many types of microhomology-mediated endjoining (47) and thus future studies may be able to decipher which repair processes and additional helicases contribute to observed differences in expansion and contraction within the CTD coding region.

In conclusion, we believe the results described here highlight the instability inherent within the RNAPII CTD coding sequence. Furthermore, we provide a rationale for how both DNA secondary structure and DNA repair processes may contribute to this variability. This phenomenon is particularly significant as the CTD of *RPB1* is essential for viability in yeast. We suspect few other essential genes show such nucleotide-level variability across individual members within the same species. Future studies will determine if other tandem repeat-containing proteins are regulated by similar mechanisms.

ACKNOWLEDGMENTS

The authors would like to thank J. Strathern, J. Corden, A. Chilkoti, S. Mirkin, and C.A. Freudenreich for strains and plasmids. The authors also acknowledge technical assistance from D. Kulla and F. Blake for construction of *RPB1* vectors and preliminary data. The authors would also like to thank L. Cowen, A. Gallant, and Y. Bentov in the computational analysis of the CTD sequence. Lastly, the authors would also like to thank S. Mirkin, C.A. Freudenreich, M. McVey, G. Dixit, C. Moore, and other members of the Fuchs lab and Tufts Biology department for numerous helpful discussions in the preparation of this manuscript.

This work was supported by startup funds and a Tufts Collaborates! grant from Tufts University to S.M.F. S.A.M. was partially supported by a Russell L. Carpenter scholarship and B.I.R. was partially supported by the Tufts Summer Scholars program.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest with the contents of this article.

AUTHOR CONTRIBUTIONS

SMF conceived and coordinated the study and wrote the paper. SAM designed, performed, and analysed the experiments shown in Figures 5 and 6 and Tables 3 and 4, contributed to the preparation of the figures and the writing and editing of the original manuscript. AE performed and analyzed the experiments shown in Figure 4 and 6 and Tables 3 and 4. MB performed and analyzed experiments associated with Figures 4 and 6. BIR designed, performed, and analyzed the experiments shown in Figures 1 and 2 as well as Table 4. All authors contributed to the preparation and editing of the manuscript.

REFERENCES

1. Egloff, S., and Murphy, S. (2008) Cracking the RNA polymerase II CTD code. *Trends Genet* **24**, 280-288
2. Fuchs, S. M., Laribee, R. N., and Strahl, B. D. (2009) Protein modifications in transcription elongation. *Biochim Biophys Acta* **1789**, 26-36
3. Nonet, M., Sweetser, D., and Young, R. A. (1987) Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell* **50**, 909-915
4. Nonet, M. L., and Young, R. A. (1989) Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of Saccharomyces cerevisiae RNA polymerase II. *Genetics* **123**, 715-724
5. Liu, G., Chen, X., Bissler, J. J., Sinden, R. R., and Leffak, M. (2010) Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. *Nat Chem Biol* **6**, 652-659
6. Wang, G., and Vasquez, K. M. (2014) Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* **19**, 143-151
7. Orr, H. T., and Zoghbi, H. Y. (2007) Trinucleotide repeat disorders. *Annu Rev Neurosci* **30**, 575-621
8. Paeschke, K., Bochman, M. L., Garcia, P. D., Cejka, P., Friedman, K. L., Kowalczykowski, S. C., and Zakian, V. A. (2013) Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **497**, 458-462
9. Malagon, F., Kireeva, M. L., Shafer, B. K., Lubkowska, L., Kashlev, M., and Strathern, J. N. (2006) Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. *Genetics* **172**, 2201-2209
10. Janke, C., Magiera, M. M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., and Knop, M. (2004) A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* **21**, 947-962
11. Schulz, V. P., and Zakian, V. A. (1994) The saccharomyces PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation. *Cell* **76**, 145-155
12. Tong, A. H., and Boone, C. (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171-192
13. McDaniel, J. R., Mackay, J. A., Quiroz, F. G., and Chilkoti, A. (2010) Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes. *Biomacromolecules* **11**, 944-952
14. Sikorski, R. S., and Hieter, P. (1989) A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122**, 19-27
15. Gietz, R. D., and Schiestl, R. H. (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**, 31-34
16. Aksenova, A. Y., Greenwell, P. W., Dominska, M., Shishkin, A. A., Kim, J. C., Petes, T. D., and Mirkin, S. M. (2013) Genome rearrangements caused by interstitial telomeric sequences in yeast. *Proc Natl Acad Sci U S A* **110**, 19866-19871
17. Hall, B. M., Ma, C. X., Liang, P., and Singh, K. K. (2009) Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* **25**, 1564-1565

18. Akada, R., Murakane, T., and Nishizawa, Y. (2000) DNA extraction method for screening yeast clones by PCR. *Biotechniques* **28**, 668-670, 672, 674
19. Fuchs, S. M., Kizer, K. O., Braberg, H., Krogan, N. J., and Strahl, B. D. (2012) RNA polymerase II carboxyl-terminal domain phosphorylation regulates protein stability of the Set2 methyltransferase and histone H3 di- and trimethylation at lysine 36. *J Biol Chem* **287**, 3249-3256
20. Chapman, R. D., Heidemann, M., Hintermair, C., and Eick, D. (2008) Molecular evolution of the RNA polymerase II CTD. *Trends Genet* **24**, 289-296
21. Yang, C., and Stiller, J. W. (2014) Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. *Proc Natl Acad Sci U S A* **111**, 5920-5925
22. Simonti, C. N., Pollard, K. S., Schroder, S., He, D., Bruneau, B. G., Ott, M., and Capra, J. A. (2015) Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. *BMC Evol Biol* **15**, 35
23. Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J., van Oudenaarden, A., Barton, D. B., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. (2009) Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341
24. Capra, J. A., Paeschke, K., Singh, M., and Zakian, V. A. (2010) G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS Comput Biol* **6**, e1000861
25. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**, 11593-11598
26. Weitzmann, M. N., Woodford, K. J., and Usdin, K. (1997) DNA secondary structures and the evolution of hypervariable tandem arrays. *J Biol Chem* **272**, 9517-9523
27. Kikin, O., D'Antonio, L., and Bagga, P. S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**, W676-682
28. Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R. A., and Yang, D. (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res* **34**, 2723-2735
29. Paeschke, K., Capra, J. A., and Zakian, V. A. (2011) DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678-691
30. Stump, A. D., and Ostrozhynska, K. (2013) Selective constraint and the evolution of the RNA polymerase II C-Terminal Domain. *Transcription* **4**, 77-86
31. Liu, P., Kenney, J. M., Stiller, J. W., and Greenleaf, A. L. (2010) Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. *Mol Biol Evol* **27**, 2628-2641
32. Kishore, S. P., Perkins, S. L., Templeton, T. J., and Deitsch, K. W. (2009) An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. *J Mol Evol* **68**, 706-714
33. Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005) Intragenic tandem repeats generate functional variability. *Nat Genet* **37**, 986-990

34. Kelly, M. K., Brosnan, L., Jauert, P. A., Dunham, M. J., and Kirkpatrick, D. T. (2012) Multiple pathways regulate minisatellite stability during stationary phase in yeast. *G3 (Bethesda)* **2**, 1185-1195
35. Ribeyre, C., Lopes, J., Boule, J. B., Piazza, A., Guedin, A., Zakian, V. A., Mergny, J. L., and Nicolas, A. (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* **5**, e1000475
36. West, M. L., and Corden, J. L. (1995) Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics* **140**, 1223-1233
37. Liu, P., Greenleaf, A. L., and Stiller, J. W. (2008) The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. *Mol Biol Evol* **25**, 719-727
38. Thompson, C. M., Koleske, A. J., Chao, D. M., and Young, R. A. (1993) A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell* **73**, 1361-1375
39. Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**, 445-477
40. Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M. P., Foiani, M., and Nicolas, A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J* **30**, 4033-4046
41. Jarman, A. P., and Wells, R. A. (1989) Hypervariable minisatellites: recombinators or innocent bystanders? *Trends Genet* **5**, 367-371
42. Zumwalt, M., Ludwig, A., Hagerman, P. J., and Dieckmann, T. (2007) Secondary structure and dynamics of the r(CGG) repeat in the mRNA of the fragile X mental retardation 1 (FMR1) gene. *RNA Biol* **4**, 93-100
43. Mitani, K., Takahashi, Y., and Kominami, R. (1990) A GGCAGG motif in minisatellites affecting their germline instability. *J Biol Chem* **265**, 15203-15210
44. Wilson, M. A., Kwon, Y., Xu, Y., Chung, W. H., Chi, P., Niu, H., Mayle, R., Chen, X., Malkova, A., Sung, P., and Ira, G. (2013) Pif1 helicase and Poldelta promote recombination-coupled DNA synthesis via bubble migration. *Nature* **502**, 393-396
45. LeClere, A. R., Yang, J. K., and Kirkpatrick, D. T. (2013) The role of CSM3, MRC1, and TOF1 in minisatellite stability and large loop DNA repair during meiosis in yeast. *Fungal Genet Biol* **50**, 33-43
46. Preston, C. R., Engels, W., and Flores, C. (2002) Efficient repair of DNA breaks in Drosophila: evidence for single-strand annealing and competition with other repair pathways. *Genetics* **161**, 711-720
47. Sfeir, A., and Symington, L. S. (2015) Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem Sci* **40**, 701-714
48. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797

FIGURES LEGENDS

Figure 1. Variability within the CTD coding sequence. Genomic DNA from the 36 strains used in the Saccharomyces Genome Resequencing Project (SGRP) was sequenced by Sanger sequencing, and a multiple sequence alignment was created using MUSCLE(48). The alignment was manually adjusted to correct for repeat length changes misannotated as single nucleotide polymorphisms. Each strain is represented as a circle on the phylogenic tree. The schematics shows how unique 21 bp blocks (shown as colored rectangles) are arranged to give rise to a protein product consisting of 24–26, 7 amino acid units. Blocks shown in green represent the consensus sequence for that block position (see **Figure 2**), orange blocks indicate the presence of a SNP within that sequence, the offset block (blue) indicates a 21 bp insertion unique to five *S. cerevisiae* strains, and open gaps in the block structure indicate that a 21 bp repeat has been lost from a particular strain.

Figure 2. Sequence alignment of the CTD region from 36 yeast strains from SGRP. Strains are grouped by identity and numbered according to vertical position on the tree in Figure 1. Alignments were made using MUSCLE and hand corrected to account for misaligned repeats.

Figure 3. A genetic system for measuring changes in CTD repeat length. A) Tet-off system for examining mutations within the CTD. In the absence of doxycycline, the tet transactivator (tTA) binds to $tetO_7$ sites upstream of the genomic copy of *RPB1* allowing transcription. In the presence of doxycycline, the genomic copy is effectively off and cells rely on a plasmid-based copy of *RPB1* under the control of its endogenous promoter. B) Schematic of mutant CTD constructs. Each block indicates the presence of one 21 bp repeat encoding the seven amino acid CTD consensus sequence. pRPB1-CTD_8 encodes a sequence with 8 CTD repeats, whereas pRPB1-4stop encodes 26 repeats, but is interrupted by 4 stop codons (in red). pRPB1-4stop encodes a protein with only 7 functional CTD repeats. C) Spotting assay demonstrating the effectiveness of our tet-off system. Both pRPB1-CTD_8 and pRPB1-4stop produce a protein product that results in poor yeast viability in the presence of doxycycline. Spotting assays are representative examples of at least three independent trials.

Figure 4. The CTD coding region undergoes spontaneous contraction. A) Schematic depicting an *RPB1* mutant that spontaneously arises in cells with pRPB1-4stop. Plasmids isolated from fast-growing colonies of GRY3019 transformed with pRPB1-4stop were retransformed into new isolates of GRY3019 to demonstrate genetic changes to the plasmid were responsible for the phenotype. Retransformation of the Δ4 repeat suppressor into new cells resulted in improved growth on media containing DOX. Spotting assays are representative examples of at least three independent trials. B) Sequence alignment of pRPB1-4stop and pΔ4repeats reveals that the contraction occurred between two regions of microhomology flanking the stop codons (underlined). C) PCR analysis of the same plasmids shown in A shows that DNA encoding pΔ4repeats is indeed smaller as observed on agarose gel. D) Western blot of extracts grown from the

cells in A in the presence of Dox. The contraction event seen in the Δ4 mutation results in a protein product that is larger than RPB1-4stop. Key to visualizing small molecular weight differences in CTD mutants is the non-standard acrylamide:bisacrylamide ratio (see Experimental Procedures). Data shown are a representative blot of three independent trials.

Figure 5. Yeast viability is dependent on CTD length. A) Schematic depicting *RPB1* mutants with variable length CTDs. Each 7 amino acid (21 base pair) repeat is shown as a single green box. B) Strains expressing CTD mutants with as few as 10 repeats show near wild type growth, however a mutant with only 8 repeats is severely compromised as reported by others using the classic *URA3* shuffling system. Spotting assays are representative examples of at least three independent trials.

Figure 6. Spotting analysis of several suppressors isolated from strains originally transformed with pRPB1-CTD_8. A) Reintroduction of the plasmids from these suppressors resulted in drastically improved growth on media containing doxycycline, confirming genetic changes to pRPB1-CTD_8 were responsible for the phenotype. Spotting assays are representative examples of at least three independent trials. B) PCR analysis of the plasmids from A shows that DNA encoding suppressors of pRPB1-CTD_8 contain more CTD repeats. This was confirmed by Sanger sequencing (data not shown). C) Western analysis demonstrates that expression of *RPB1* from these suppressor plasmids results in protein products that are higher molecular weight than the original CTD_8 product. Key to visualizing small molecular weight differences in CTD mutants is the non-standard acrylamide:bisacrylamide ratio (see Experimental Procedures). Data shown are a representative blot of three independent trials.

Figure 7. Alignment of sequences for the individual CTD repeats. Repeats are numbered according to their location in the DNA sequence. Bases that vary from the consensus (at top) are shaded in grey. Polymorphisms that result in an amino acid change are underlined.

Figure 8. GC imbalance marks the CTD coding region and contributes to secondary structure formation. A) nucleotide statistics for the core of *RPB1* and the CTD region. The CTD coding strand has reduced guanine bases and a high concentration of cytosine (grey). B) Plot of nucleotide differences for a 20 kilobase region of yeast chromosome IV including *RPB1*. Data were calculated from 150 base pair windows with 50 base pairs of overlap between adjacent windows and plotted in Excel. The CTD region has a very negative value of G-C. A-T for this same region is less variable. C) Circular dichroism for G-rich DNA from the non-coding strand of the CTD identified as potential G4-DNA forming sequences using QGRS mapper and a known G4-DNA forming region of human telomeres (Tel26). Data were collected in the presence (solid line) and absence (dotted line) of 100 mM KCl. In the presence of potassium ions, all three oligonucleotides derived from the CTD coding sequence exhibit a minimum near 240 nm and a broad positive peak between 260 and 300 nm.

TABLES

Table 1: Oligonucleotides used in this work

sequence 5'→3'	description
AGAAAAGCCTGGTGTCAAGACTCCAAACCCGGGTTGG	CTD upstream forward for RDL
AACCCGGGTTTGGAGTCTTGACACCAGGCTTTTCTCC	CTD upstream reverse for RDL
AGAATATGAAGGTGAGGTTGGGCTGTAACTAGGAGACGTCGG	CTD repeat forward for RDL
GACGTCTCCTAGTTACAGCCCAACCTCACCTTCATATTCTCC	CTD repeat reverse for RDL
AAGAGCTCATCTGGAATTTTCATTTTCATTATGCTTTTGTTCGTCTTGCTTTGGAGAATATG	CTD terminator forward for RDL
GACGTCTCCTGCATATTCTCCAAAGCAAGACGAACAAAAGCATAATGAAAATGAAAATTCCA	CTD terminator reverse for RDL
TTTACTAGCGCCGTTGGTTT	*RPB1* reverse PCR primer
GATCGATGAGGAGTCACTGG	*RPB1* forward PCR primer
ATATGCGTCAGGCGACCTCT	*RPB1* forward PCR primer
AATGCAAACAAGGAGGTTGC	*RAD52* forward primer
CGAGTACGAGATGACCACGA	NAT reverse primer
GCCAAAGTGGATAAGATCAAC	*PIF1* forward primer
TGCAATTCAGTGAAGCTAGGTC	*PIF1* reverse primer
CACCAACGTCACCATCATAATGGCCAACGTCACCATCATAT	quikchange stop codon forward
ATATGATGGTGACGTTGGCCATTATGATGGTGACGTTGGTG	quikchange stop codon reverse
TAATACGACTCACTATAGGG	T7 forward
TATGCTAGTTATTGCTCAG	T7 reverse
TGGTGATGTTGGGGAATAGGCTGGAGACGTTGGGT	CTD1
TTGGGGAATAGGCTGGAGACGTTGGGC	CTD2
TGGTGATGTTGGCGAGTACGATGGTGATGTTGGT	CTD3
AAAGGGTTAGGGTTAGGGTTAGGGAA	Tel26

Table 2: Plasmids used in this work

Plasmid	Ref.	Notes
pL-RPB1	(9)	
pRS315	(14)	Referred to as pLEU in the text
pVS31	(11)	Plasmid for making pif1-m2 mutation
pJMD2	(13)	Plasmid for recursive directional ligation used to make CTD mutants
pSMF2	This work	pL-RPB1 with *XmaI* and *SacI* sites flanking the CTD coding region. Referred to as pRPB1 in the text
pRPB1-CTD_8	This work	pSMF2 with 8 CTD repeats
pRPB1-CTD_{10}	This work	pSMF2 with 10 CTD repeats
pRPB1-CTD_{14}	This work	pSMF2 with 14 CTD repeats
pRPB1-CTD_{26}	This work	pSMF2 with 26 CTD repeats
pRPB1-4stop	This work	pSMF2 with repeats 8-10 interrupted with 4 repeats containing stop codons

Table 3: Analysis of CTD rearrangements

The rates of suppressor formation were determined by measuring the frequencies of fast growing colonies on SC-LEU plates containing doxycycline. Rates and 95% confidence intervals (in parentheses) were calculated using the MSS-Maximum Likelihood Estimator method (16,17).

Reporter	rate x 10^{-6} / cell / generation	Total	Expansion	Contraction	Recombination	Other	% Expansion	% Contraction
pRPB1-4stop	26 (19-42)	115	1	59	n.d.*	55	< 1	51
pRPB1-CTD_8	4.5 (2.5-6.9)	169	65	0	15	89	38	0

* Homologous recombination events cannot be distinguished from extragenic suppressors by colony PCR so they are included into the “other” category.

Table 4: Influence of Pif1p and Rad52p on CTD variability

Characterization of fast growing suppressors based on DNA sequence for two *RPB1* reporter constructs.

pRPB1-CTD_8

	Expansions	Recombination	Other	Total	% Expansions
WT	65	15	89	169	38
pif1-m2	44	1	33	78	56
*rad52*Δ	6	0	115	115	5

pRPB1-4stop

	Contraction	Expansion	Other	Total	% Contractions
WT	59	1	55	115	51
pif1-m2	47	0	35	82	57
*rad52*Δ	51	0	1	52	98

Figure 1

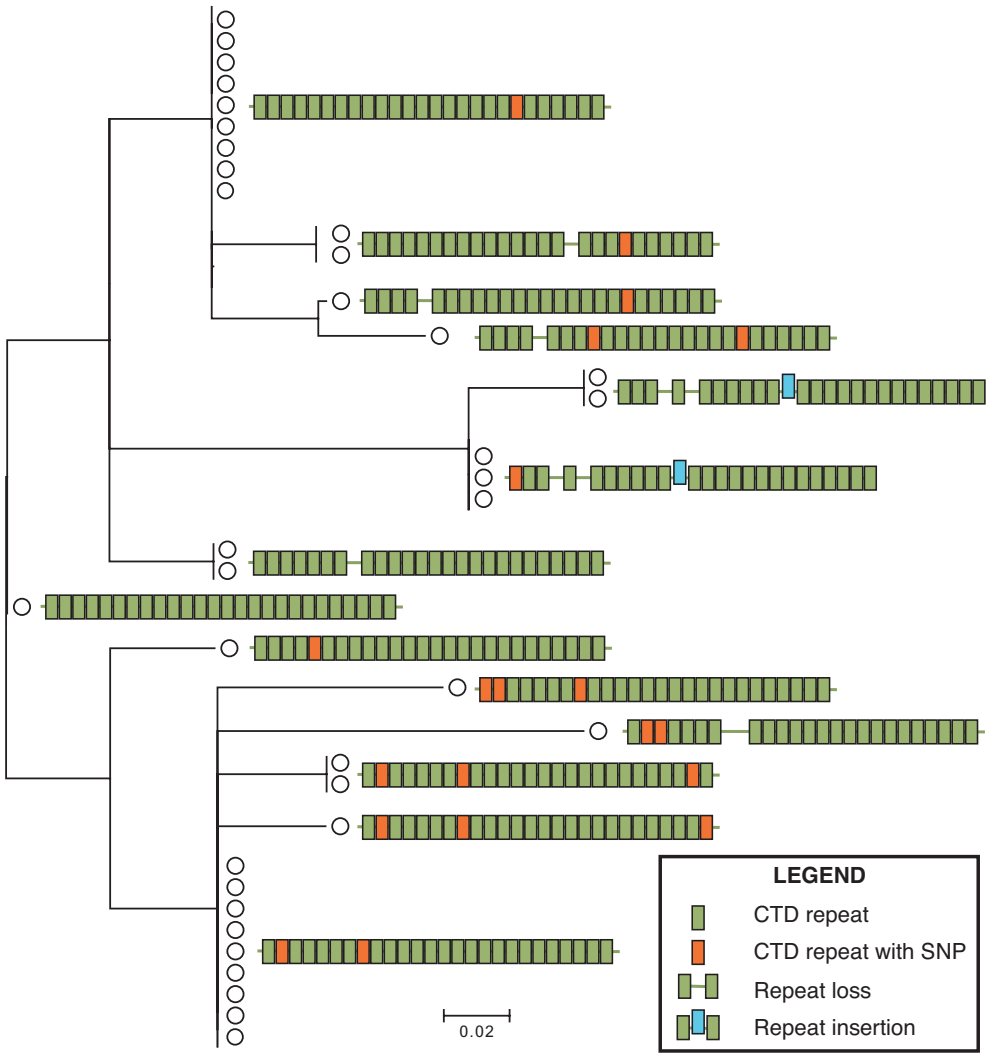

LEGEND
CTD repeat
CTD repeat with SNP
Repeat loss
Repeat insertion
0.02

Figure 2

Strains are grouped by identity and numbered according to position on tree in Figure 1

```
Group 1  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 2  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCGTCG TACTCACCAACATCACCATCG
Group 3  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 4  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 5  TCACCAGGC TTCTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG ----------
Group 6  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG ----------
Group 7  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 8  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 9  TCACCAGGC TTTTCTCCAACTTCCCCAACA TACTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 10 TCACCAGGC TT-TCTCCAACTTCCCC-ACA TA-TCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 11 TCACCAGGC TTTTCTCCAACTTCCCCAACA TATTCTCCTACCTCTCCAGCG TACTCACAAACATCACCATCG TACTCACCAACATCACCATCG
Group 12 TCACCAGGC TTTTCTCCAACTTCCCCAACA TATTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 13 TCACCAGGC TTTTCTCCAACTTCCCCAACA TATTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG
Group 14 TCACCAGGC TTTTCTCCAACTTCCCCAACA TATTCTCCTACCTCTCCAGCG TACTCACCAACATCACCATCG TACTCACCAACATCACCATCG

Group 1  TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCGCCAACGTCGCCATCG
Group 2  TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCGCCAACGTCGCCATCG
Group 3  ---------- TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCGCCAACGTCGCCATCG
Group 4  ---------- TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCACCAACGTCGCCATCG
Group 5  TACTCGCCAACATCACCATCG ---------- TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCGCCAACGTCGCCATCG
Group 6  TACTCGCCAACATCACCATCG ---------- TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCGCCAACGTCGCCATCG
Group 7  TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA ----------
TATTCGCCAACGTCGCCATCG
Group 8  TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCA
TATTCACCAACGTCGCCATCG
Group 9  TACTCGCCAACATCACCATCC TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCG
TATTCACCAACGTCGCCATCG
Group 10 TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCG
TATTCACCAACGTCGCCATCG
Group 11 TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGT-------- ------------
---------
Group 12 TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCG
TATTCACCAACGTCGCCATCG
Group 13 TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCG
TATTCACCAACGTCGCCATCG
Group 14 TACTCGCCAACATCACCATCG TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA TATTCGCCAACGTCACCATCG
TATTCACCAACGTCGCCATCG

Group 1  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 2  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 3  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 4  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 5  TATTCTCCAACGTCACCATCG TATTCGCCAATGTCGCCTTCC TACTCTCCCACGTCGCCAAGC TACAGCCCTACGTCGCCAAGC
TACAGCCCTACGTCTCCTTCT
Group 6  TATTCTCCAACGTCACCATCG TATTCGCCAATGTCGCCTTCC TACTCTCCCACGTCGCCAAGC TACAGCCCTACGTCGCCAAGC
TACAGCCCTACGTCTCCTTCT
Group 7  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 8  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 9  TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 10 TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 11 -------------CACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 12 TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 13 TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT
Group 14 TATTCTCCAACGTCACCATCG TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC ----------
TACAGCCCTACGTCTCCTTCT

Group 1  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 2  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT ---------- TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 3  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 4  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 5  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 6  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 7  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 8  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
```

```
Group 9  TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 10 TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 11 TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 12 TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 13 TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT
Group 14 TATTCTCCTACATCTCCATCA TACTCTCCTACGTCACCAAGT TACAGCCCAACGTCACCAAGT TACAGCCCAACGTCTCCAGCC
TATTCCCCAACATCACCAAGT

Group 1  TATAGTCCTACATCGCCTTCA TACTCTCCAACATCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 2  TATAGTCCTACATCGCCTTCA TACTCTCCAACATCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 3  TATAGTCCTACATCGCCTTCA TACTCTCCAACATCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 4  TATAGTCCTACATCGCCTTCA TACTCTCCAACATCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 5  TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 6  TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 7  TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 8  TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 9  TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 10 TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 11 TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 12 TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 13 TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT
Group 14 TATAGTCCTACATCGCCTTCA TACTCTCCAACGTCACCATCC TATTCCCCAACATCACCTTCT TACTCTCCCACCTCTCCAAAC
TATAGCCCTACTTCACCTTCT

Group 1  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 2  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 3  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 4  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 5  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 6  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 7  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 8  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 9  TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 10 TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 11 TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 12 TACTCCCCAACATCTCCAGGT TACAGCCCAGGATCTCCTGCATATTCTCCA
Group 13 TACTCCCCAACATCTCCAGGC TACAGTCCAGGATCTCCTGCATATTCTCCA
Group 14 TACTCCCCAACATCTCCAGGC TACAGCCCAGGATCTCCTGCATATTCTCCA
```

GROUPS

Group 1 S288C, BC187, Y9, YS9, Y12,
UWOPS83-787.3, K11, 378604X, YPS128
Group 2 UWOPS03-461.4, UWOPS05-227.2
Group 3 YS4
Group 4 Yllc17_E5
Group 5 SK1, UWOPS05-217.3
Group 6 DBVPG6765, DBVPG6044, NCYC110
Group 7 W303, 273614N'
Group 8 DBVPG6040
Group 9 DBVPG1853
Group 10 UWOPS87-2421
Group 11 DBVPG1788
Group 12 NCYC361, 322134S
Group 13 YPS606
Group 14 DBVPG1373, L-1528, L-1374,
DBVPG1106, YJM978, Y55, YS2 YJM975, YJM981

Figure 3

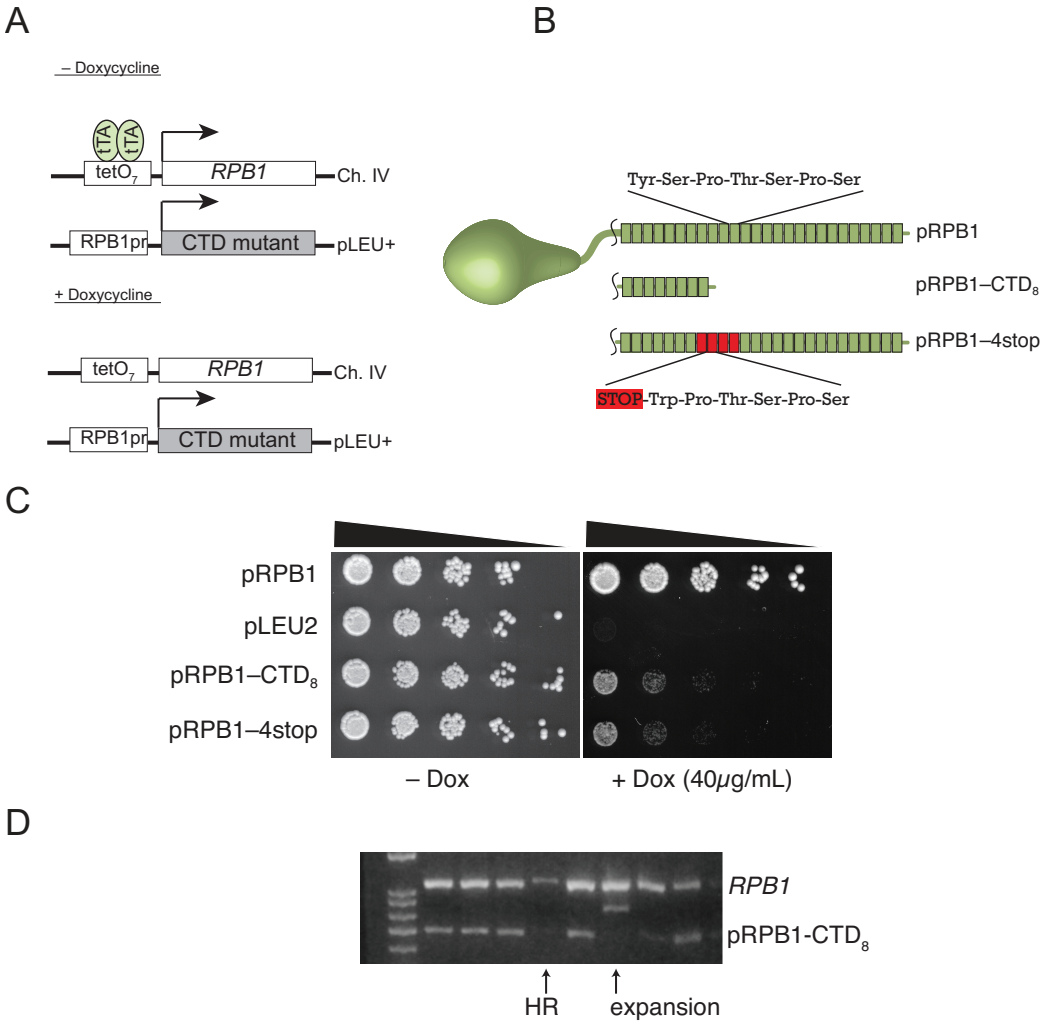

A
– Doxycycline
tTA
tTA
tetO7
RPB1
Ch. IV
RPB1pr
CTD mutant
pLEU+
+ Doxycycline
tetO7
RPB1
Ch. IV
RPB1pr
CTD mutant
pLEU+
B
Tyr-Ser-Pro-Thr-Ser-Pro-Ser
pRPB1
pRPB1–CTD8
pRPB1–4stop
STOP-Trp-Pro-Thr-Ser-Pro-Ser
C
pRPB1
pLEU2
pRPB1–CTD8
pRPB1–4stop
– Dox
+ Dox (40μg/mL)
D
RPB1
pRPB1-CTD8
HR
expansion

Figure 4

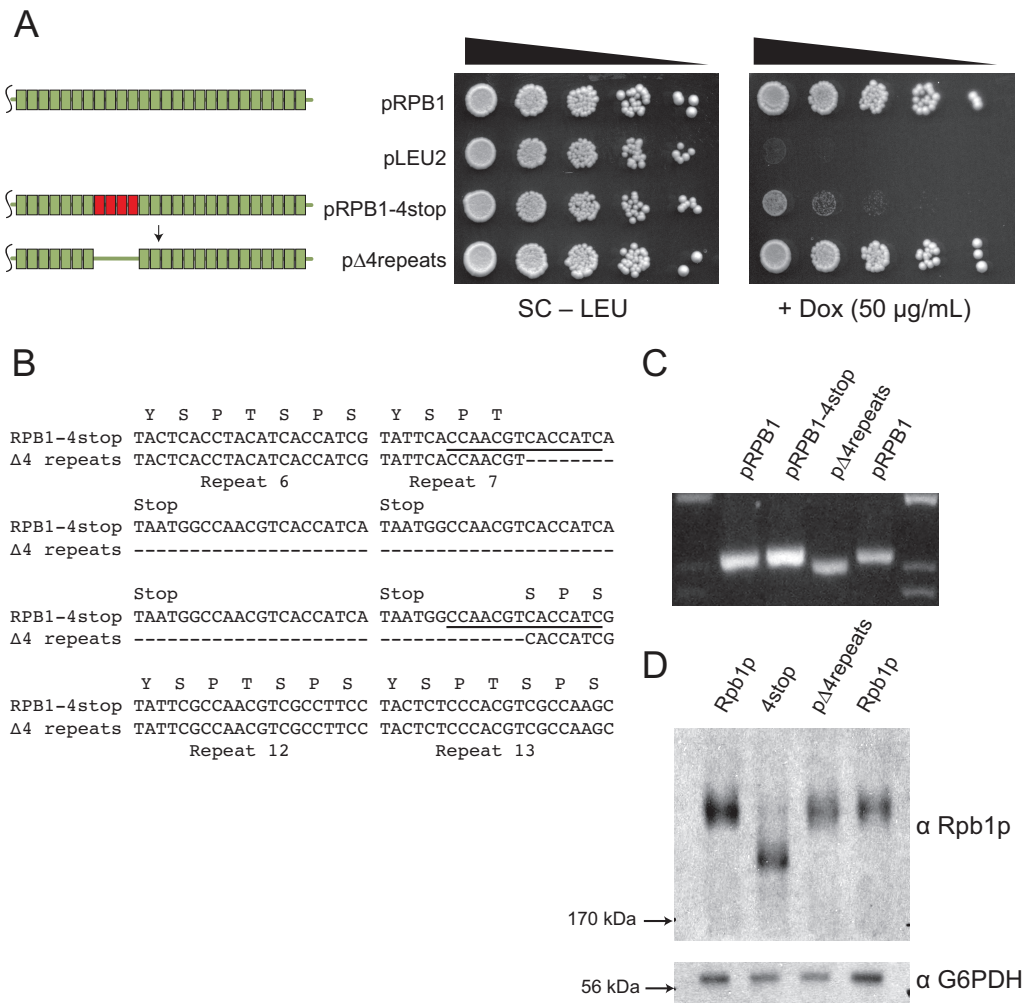

A
pRPB1
pLEU2
pRPB1-4stop
pΔ4repeats
SC – LEU
+ Dox (50 µg/mL)
B
Y S P T S P S Y S P T
RPB1-4stop TACTCACCTACATCACCATCG TATTCACCAACGTCACCATCA
Δ4 repeats TACTCACCTACATCACCATCG TATTCACCAACGT--------
Repeat 6 Repeat 7
Stop Stop
RPB1-4stop TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCA
Δ4 repeats ---------- ----------
Stop Stop S P S
RPB1-4stop TAATGGCCAACGTCACCATCA TAATGGCCAACGTCACCATCG
Δ4 repeats ---------- ----------CACCATCG
Y S P T S P S Y S P T S P S
RPB1-4stop TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC
Δ4 repeats TATTCGCCAACGTCGCCTTCC TACTCTCCCACGTCGCCAAGC
Repeat 12 Repeat 13
C
pRPB1
pRPB1-4stop
pΔ4repeats
pRPB1
D
Rpb1p
4stop
pΔ4repeats
Rpb1p
α Rpb1p
170 kDa
56 kDa
α G6PDH

Figure 5

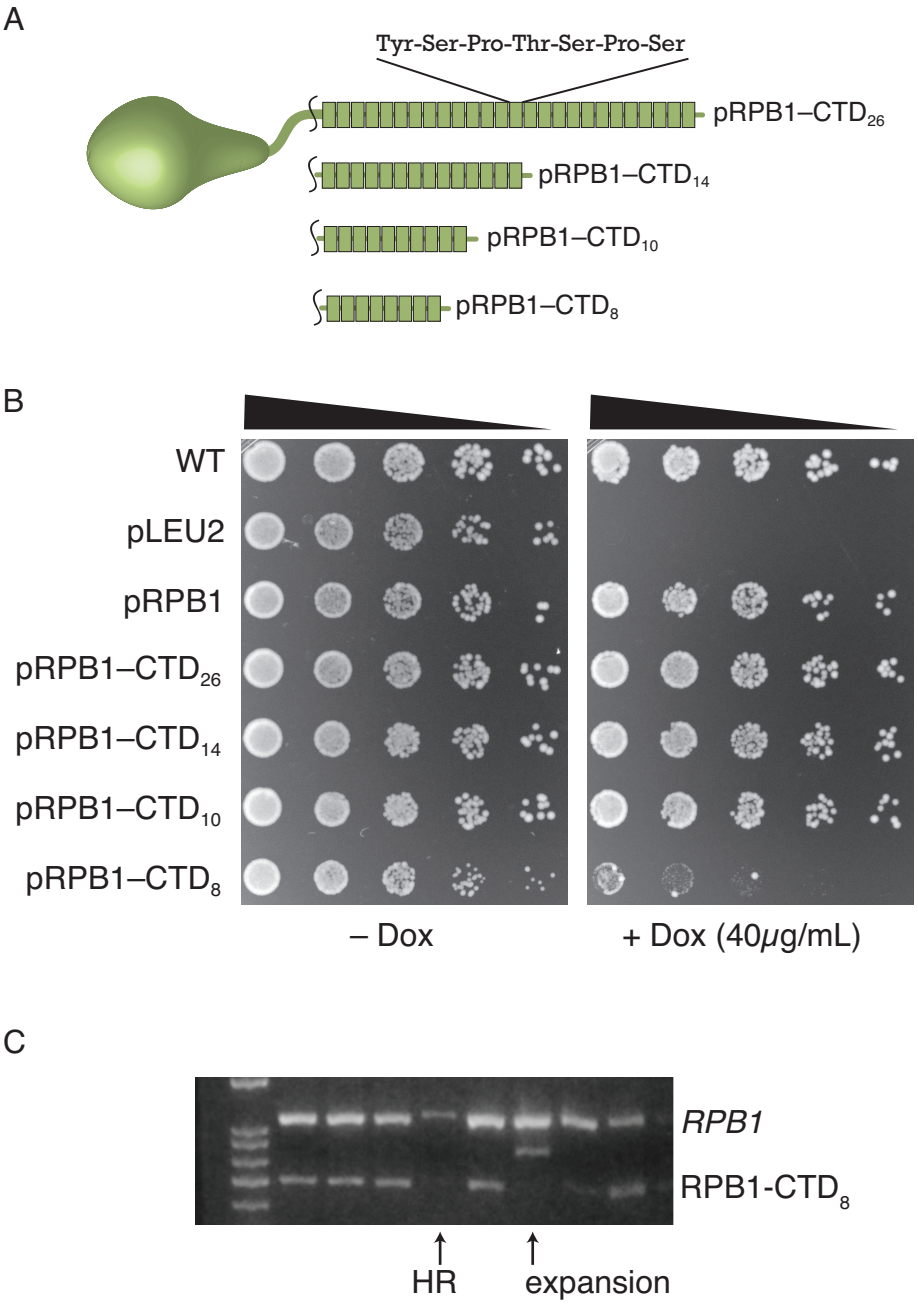
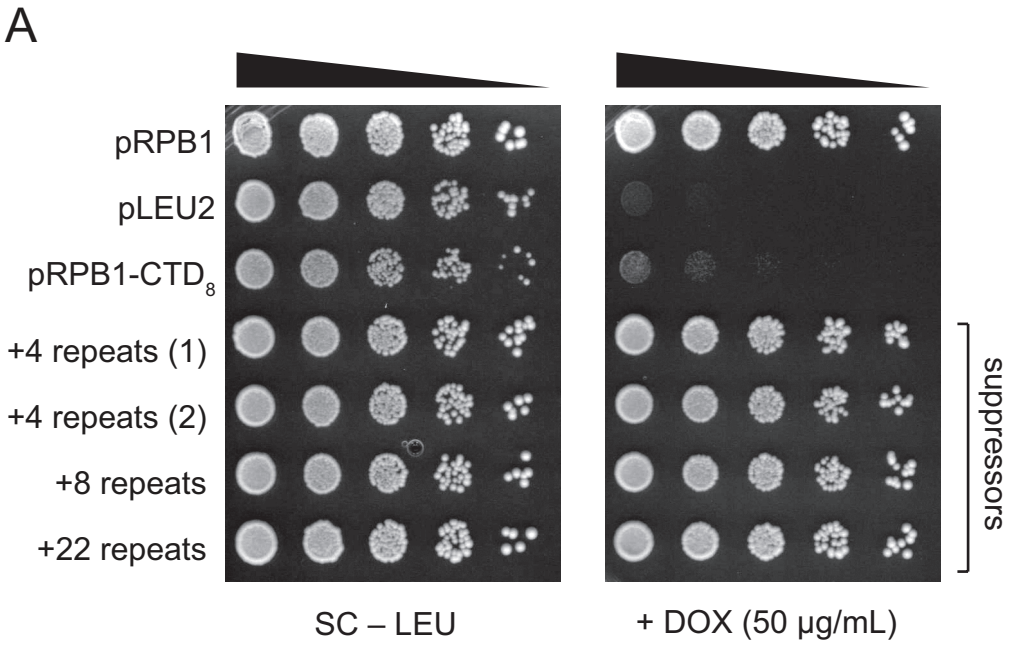
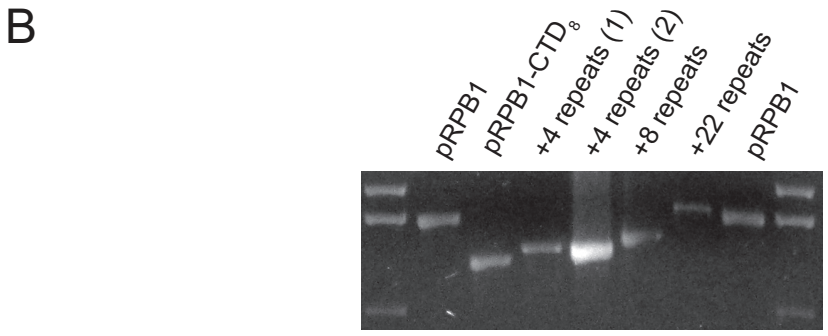

A
Tyr-Ser-Pro-Thr-Ser-Pro-Ser
pRPB1–CTD26
pRPB1–CTD14
pRPB1–CTD10
pRPB1–CTD8
B
WT
pLEU2
pRPB1
pRPB1–CTD26
pRPB1–CTD14
pRPB1–CTD10
pRPB1–CTD8
– Dox
+ Dox (40μg/mL)
C
RPB1
RPB1-CTD8
HR
expansion

Figure 6


A
pRPB1
pLEU2
pRPB1-CTD8
+4 repeats (1)
+4 repeats (2)
+8 repeats
+22 repeats
suppressors
SC – LEU
+ DOX (50 µg/mL)


B
pRPB1
pRPB1-CTD8
+4 repeats (1)
+4 repeats (2)
+8 repeats
+22 repeats
pRPB1

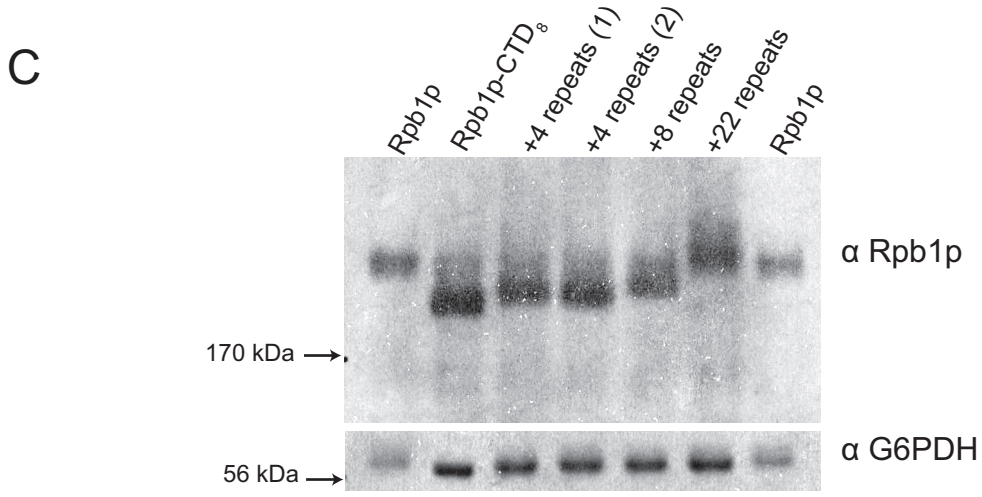

C
Rpb1p
Rpb1p-CTD8
+4 repeats (1)
+4 repeats (2)
+8 repeats
+22 repeats
Rpb1p
α Rpb1p
170 kDa
α G6PDH
56 kDa

Figure 7

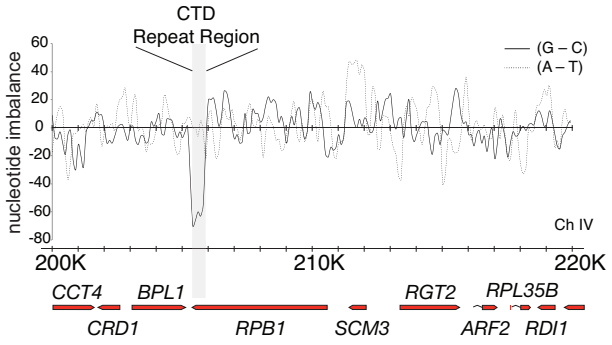
	TAC	TC–	CCA	ACG	TCA	CCA	TC–	Consensus
1	TTC	TCT	CCA	ACT	TCC	CCA	ACA	
2	TAC	TCT	CCT	ACC	TCT	CCA	GCG	
3	TAC	TCA	CCA	ACA	TCA	CCA	TCG	
4	TAC	TCA	CCA	ACA	TCA	CCA	TCG	
5	TAC	TCG	CCA	ACA	TCA	CCA	TCG	
6	TAC	TCA	CCT	ACA	TCA	CCA	TCG	
7	TAT	TCA	CCA	ACG	TCA	CCA	TCA	
8	TAT	TCG	CCA	ACG	TCA	CCA	TCA	
9	TAT	TCG	CCA	ACG	TCG	CCA	TCG	
10	TAT	TCT	CCA	ACG	TCA	CCA	TCG	
11	TAT	TCG	CCA	ACG	TCG	CCT	TCC	
12	TAC	TCT	CCC	ACG	TCG	CCA	AGC	
13	TAC	AGC	CCT	ACG	TCG	CCA	AGC	
14	TAC	AGC	CCT	ACG	TCT	CCT	TCT	
15	TAT	TCT	CCT	ACA	TCT	CCA	TCA	
16	TAC	TCT	CCT	ACG	TCA	CCA	AGT	
17	TAC	AGC	CCA	ACG	TCA	CCA	AGT	
18	TAC	AGC	CCA	ACG	TCT	CCA	GCC	
19	TAT	TCC	CCA	ACA	TCA	CCA	AGT	
20	TAT	AGT	CCT	ACA	TCG	CCT	TCA	
21	TAC	TCT	CCA	ACG	TCA	CCA	TCC	
22	TAT	TCC	CCA	ACA	TCA	CCT	TCT	
23	TAC	TCT	CCC	ACC	TCT	CCA	AAC	
24	TAT	AGC	CCT	ACT	TCA	CCT	TCT	
25	TAC	TCC	CCA	ACA	TCT	CCA	GGC	
26	TAC	AGC	CCA	GGA	TCT	CCT	GCA	

Figure 8

A

	Ch IV	*RPB1* core	*RPB1* CTD
G%	19.0	22.0	8.6
A%	31.1	30.2	27.1
T%	31.0	29.9	24.7
C%	18.9	17.9	39.6
GC%	37.9	39.9	48.2
GA%	50.1	52.2	35.7

B



C