The Role of PIF1 Helicase in Minisatellite Variation

By

Bradley Reinfeld

PI: Dr. Stephen Fuchs

BIO 199 Senior Honors Thesis

April 27, 2015

Abstract:

The largest subunit of the yeast RNA polymerase II complex, RPO21, contains a tandemly repeated heptapeptide sequence whose posttranslational modifications are key to properly orchestrate transcription. Sequencing of the RPO21 locus from 36 wild strains of budding yeast uncovered many polymorphisms that resulted in different lengths of its essential C-terminal domain (CTD) repeat. An alignment of the newly found variants demonstrated that over evolutionary time, there have been many expansions and contractions to the CTD resulting in the variety of alleles seen today. With this surprising finding of variability in an essential domain, it became important to find factors that could contribute to this observed variation. From analyzing the nucleotide sequence of RPO21, it was found that a G-quadruplex could form at the most C-terminal repeats of the CTD. Additionally this nucleotide region of the CTD has been found to be bound b ythe G4 speicifc helicas PIF1 by previous chromatin immunoprecipitation studies,. Using the tTa-dependent expression system developed in the Fuchs Lab with a *pif1-m2* mutant in a Luria-Delbrück fluctuation assay, it was found that PIF1 acts an inhibitor of expansion events in the CTD. This finding led to a larger survey of the yeast genome for other minisatellites that may behave like the CTD due to the presence of PIF1 binding as well as the possibility to form a G4. This analysis ultimately found that three key transcriptional enzymes SPT5, GAL11, and RPO21 all have G4 sequences in their variable minisatellites. However, future examination is required to understand why these three repeats demonstrate polymorphisms while other similar repeats do not.

Chapter 1: The role of PIF1 in the Expansion of the C-terminal domain of RPO21

Introduction:

Structure of RNA polymerase II and its CTD

Elucidating RNA polymerase II (Pol II) biology is central to understanding eukaryotic cell physiology. Pol II, an essential 12-member protein complex, is the sole enzyme responsible for transcribing messenger RNA (mRNA). The largest subunit of the Pol II complex, RPB1 (yeast gene name: RPO21), contains two crucial elements: the catalytic core, which is responsible for transcribing new mRNA and the repetitive C-terminal domain (CTD) that acts as an essential scaffold for many transcription associated proteins. The CTD repeat is comprised of the canonical heptapeptide sequence YSPTSPS. A comparative analysis of 114 RPO21 homologs reported that 87.7% of sequenced eukaryotes contain the characteristic repeat (Stump and Ostrozhynska, 2013). The repeat copy number, as well as the level of degeneracy, is dynamic across eukaryotes. In *S. cerevisiae,* this CTD repeat is highly conserved in that there are only 8 nonsynomous changes in the 26 copies of the heptapeptide repeat, while the human CTD is comprised of 52 heptapeptides where only 29 repeats contain the canonical YSPTSPS sequence. Additionally, the metazoan CTD contains frequent substitution where serine 7 is often replaced by a lysine, an arginine, or a glutamine.

Unique CTD sequence permits a plethora of posttranslational modification

The distinctive CTD repeat can be posttranslationally modified at all seven positions. The serines, tyrosines, and threonines can undergo a range of covalent posttranslational modifications (PTMs) due to their hydroxyl-containing side chains. The two proline residues in this repeat can also be modified by isomerases (Werner-Allen et al., 2011). Modifications to the serines (especially at positions 2 and 5) have been studied in depth due to their correlation to transcriptional state. Before transcription, the CTD is hypophosphorylated. At this stage in eukaryotes, serine 5 and 7 are O-GlcNAc glycosylated, which leads to the assembly of the preinitiation complex (Ranuncolo et al., 2012). The next key PTM is phosphorylation of serine 5, which leads to the release of Pol II from the initiation complex. After leaving the initiation complex, elongation is not immediately entered. Pol II stalls about 20-30 nucleotide into the new transcript until phosphorylation at serine 2 initiates elongation. As the transcript is elongated, serine 2 phosphorylation increases while serine 5 phosphorylation decreases. Transcription ends with the CTD returning to its hypophosphorylated state, therefore allowing Pol II to restart transcription at a transcriptional start site (Egloff and Murphy, 2008).

Functional consequences of CTD PTMs

The wide variety of PTMs in the CTD ultimately impacts transcriptional output through changes in mRNA processing and differential histone modifications. An example of the CTD's impact on mRNA stability comes from fission yeast where the activation of capping complex is dependent on serine 5 phosphorylation (Schwer and Shuman, 2011). Phosphoserine 5 is present at the beginning of the transcriptional cycle, allowing the nascent mRNA to be capped early and efficiently without the possibility of degradation

from cellular nucleases. Additionally, a key component of the yeast-splicing complex, PRP19, must interact with a phosphorylated CTD to perform its function (David et al., 2011). A third level of transcript regulation originates is due to the fact that PCF11, the yeast polyadenylation and cleavage enzyme, requires serine 2 phosphorylation for its activity (Licatalosi et al., 2002). Work by Fuchs and colleagues demonstrates that the methyltransferase, SET2, requires proper serine 2 and 5 phosphorylation to catalyze trimethylation of lysine residue 36 on histone H3 (Fuchs et al., 2012). These results suggest that the phosphorylation state of the CTD tethers specific chromatin modifications to stages of transcription.

Essential components of CTD structure

A multitude of kinases, phosphates, isomerases, and histone modifying enzymes require the repetitive CTD to perform its function. Previous work has demonstrated that eight wild type YSPTSPS CTD repeats is essential to support growth while ten or more repeats result in wild type growth (Nonet et al., 1987; West and Corden, 1995). More recent studies have uncovered that these repeats are recognized by CTD-associated proteins as units of multiple heptapeptides. Placing alanines in between direct heptapeptide repeats results in death, yet yeast can tolerate insertions of polyalanines between two heptapeptide repeats (Stiller and Cook, 2004). From this work, it is clear that the structure endowed by multiple heptapeptide repeats is crucial for transcription to occur faithfully and efficiently.

What may be most interesting about the CTD is the observed variability in this essential domain. As stated earlier, PTMs to the CTD recruit a host of proteins to

properly conduct the many steps of eukaryotic transcription. Surprisingly, Nonet and Young uncovered that this essential scaffold contains a copy number variant which results in a RPO21 subunit with 27 instead of 26 heptapeptide repeats in a wild strain of yeast (Nonet et al., 1987). The intraspecies variation of the CTD, illustrated by comparing CTD length from laboratory and wild yeast, is not restricted to fungi. There is a documented instance of an in frame loss of a YSPTSPS heptad in one subject in the 1000 Genomes Project.

The CTD as a minisatellite

The occurrence of variable length CTDs in many organisms demonstrates that RPO21 contains a repetitive DNA minisatellite that is prone to expansions and/or contractions. Work on non-essential genes with minisatellites like FLO1, NIS1, or DSN1, has demonstrated that these regions are variable across strains of wild budding yeast (Richard and Dujon, 2006; Verstrepen et al., 2005). Verstrepen and coworkers illustrate that copy number variants in FLO1 lead to differential phenotype and differential fitness (Smukalla et al., 2008; Verstrepen et al., 2005). Therefore, it could be hypothesized that differences in CTD length may allow for differential transcriptional efficiency. Additionally, the variation in the CTD across wild strains of yeast may contribute, in some part, to the phenotypic variation as seen when growing these wild strains on plates or in culture.

Previous analysis has examined the evolution of the CTD repeat across the tree of life. Chapman's perspective concludes that the CTD most likely arose multiple times throughout evolutionary history and that its degeneracy is due to purifying selection,

resulting in an optimization of length and tolerance for non-canonical repeats in each organism (Chapman et al., 2008). However, no group has ever extensively examined the variation of the RPO21 locus across multiple wild strains of budding yeast.

Unsurprisingly, work on my project found that there is significant variation in the CTD of Pol II across wild strains of yeast. Copy number variants, as well as single nucleotide polymorphisms that lead to both synonymous and non-synonymous mutations were uncovered in the CTD. With this knowledge, we became interested in the examining the factors that may contribute to the variation in the CTD.

Unique coding sequence of RPO21 causes G-quadruplex formation

Trying to uncover the basis of CTD expansion led to an analysis of the nucleotide sequence that encodes the CTD. From examining the sequence, a large imbalance between guanines and cytosines exists on a given strand of the CTD repeat. A G4 prediction algorithm predicted found this secondary occurred in the most c-terminal repeats of the CTD while another chromatin immunpreicitation study illustrated that PIF1, a G4 specific helicase, bound the aforementioned region of the CTD (Capra et al., 2010; Paeschke et al., 2011). G4 DNA is a DNA secondary structure that results from non-canonical Hoogsteen base pairing between four tracks of three of more guanines. This alternative form of DNA can form due to G tracks on one, two, or four strands of DNA and can act as a replicative barrier (Piazza et al., 2010; Voineagu et al., 2009). The orientation of the G4 structure in relation to an origin of replication appears to be crucial in terms of the G4's ability to impact repeat instability. When the G4 structure exists on the leading strand template, expansion or contraction events occur more often than when

in the opposite orientation (Lopes et al., 2011). The G4 DNA encoded by the CTD exists in the leading strand template orientation, which is known to cause instability. There is conflicting data that suggests that G4 DNA, regardless of its orientation, can promote recombination events. These types of events studied by the Zakian Lab happen at similar rates in both the leading and lagging strand template orientation (Paeschke et al., 2013; Paeschke et al., 2011). However, the events that occur in this lab's assay are not changes in copy number as seen in the Nicolas Lab CEB1 assay but are gross chromosomal rearrangements or direct mutations to the G4 encoding sequence (Paeschke et al., 2013; Paeschke et al., 2011).

PIF1, G4 specific helicase and inhibitor of telomere lengthening

PIF1, a helicase with homologs in all three domains of life, plays a unique role in unwinding these G4 structures (Bochman et al., 2011). This helicase has remarkable specificity for this secondary structure that is not shared by any other family of helicases (Paeschke et al., 2013). Chromatin immunoprecipitation (ChIP) found that PIF1 binds to approximately 25% G4 sites throughout the yeast genome, including the G4 sequence that is encoded by the signature CTD heptapeptide repeat (Paeschke et al., 2011). Yeast PIF1 was originally found because of its role as a negative regulator of telomere length, which is now known to be result of PIF1's ability to physically displace telomerase (Li et al., 2014). Telomeres share nucleotide characteristics with RPO21 in that they are comprised of GC imbalanced sequences, which form G4 DNA and are known to change lengths (Schulz and Zakian, 1994).

Mutations in PIF1 may have consequences on minisatellite stability in yeast cells. The deletion of *pif1Δ* leads to an increase in instability CEB1 (a human subtelomeric G4 containing minisatellite) when compared to a wild type background (Ribeyre et al., 2009). The rearrangements observed in the CEB1 assay were dependent on homologous recombination due to the fact that *pif1Δ, rad51Δ, or pif1Δ, rad52Δ* double mutants did not show this unstable phenotype. Interestingly, the instability disappears when the G tracks in the CEB1 repeat are mutated, demonstrating that the effect of *pif1Δ* is sequence dependent (Ribeyre et al., 2009). This combination of facts led to an investigation of PIF1's impact on the variation of the CTD.

Summary

Due to the observed variation in the RPO21 locus, combined with the presence of a predicted G4 structure and PIF1 binding, the role of this specialized helicase in CTD expansion was examined. To observe PIF1's role in the expansion of the CTD, we can take advantage of the knowledge that there is a minimum number of CTD repeats required for life (Nonet et al., 1987; West and Corden, 1995). By using the Fuchs Lab tTA-dependent expression system, yeast can solely transcribe a mutant copy of RPO21, which contains an insufficient number of CTD repeats to support growth. However, cells that survive this challenge have undergone mutations to the CTD that allow them to survive. Using a Luria-Delbrück fluctuation assay with a specific *pif1-m2* mutant, we can examine the change in mutation rate as well as the distribution of suppressor events to understand PIF1's role in the expansion of the CTD.

Methods:

DNA isolation using phenol chloroform isoamyl alcohol extraction

The 36 wild yeast strains from the NCYC SGRP set 1 (https://catalogue.ncyc.co.uk/sgrp-set-1) were grown in 5 mL overnight cultures at 30°C. These saturated cultures were spun down and washed. The remaining cell pellet was then lysed with equal volume breaking buffer (10 mM Tris-HCl Ph 8.0, 100 mM NaCl, 1 mM EDTA, 2% Triton X-100, and 1% SDS), prewashed glass beads, and a phenol-chloroform-isoamyl alcohol mixture in a tabletop vortexer at 4°C. The DNA was purified from this mixture by extracting the top aqueous layer, precipitating the DNA with ethanol, and then resuspending the pellet in TE. 1 µL of this genomic prep was used as PCR template.

PCR amplification of the CTD

The nucleotide region encoding the CTD was amplified using a primer with homology to the 3' untranslated region (UTR) of the RPO21 locus, and with a 5' primer with homology to the linker region upstream of the CTD's characteristic heptapeptide repeats. The PCR reaction contained 1 µL of template, 1 µL of 10uM forward primer, 1 µL of 10 µM reverse primer, 1 µL of 1 µM dNTPs, 0.2 µL NEB Taq Polymerase, 5 µL of 10X NEB Buffer and 40.8 µL of H_2O. The PCR conditions were one five minute denaturing step at 95°C, followed by 35 cycles of 30 seconds at 95°C, 45 seconds at 52°C and then 65 seconds at 68°C. There was one final ten-minute extension at 68° to ensure complete amplification. Samples were stored at either 4°C or -20°C. Agarose gel electrophoresis was used to verify PCR amplification. 1.0% gels were used for the

approximately 1 kb RPO21 CTD product.10 µL of PCR product were run with 2 µL of 6X loading dye at 120 V for 40 minutes. Gels were stained with 10,000X Sybr Safe and visualized in the blue light box. PCR products of the correct length were sent to Eton Biosciences for PCR cleanup and sequencing following specifications outlined on their website. The 3' reverse primer was used for this reaction. Samples, which required resequencing, were done so by using the forward primer. The sequences were in the 3'-5' direction in respect to the CTD. To make the analysis simpler, the reverse compliment of these sequences was taken using the program found at http://www.bioinformatics.org/sms/rev_comp.html. To create the multiple sequence alignment, any excess sequence from the 5' linker or the 3' UTR was thrown out. ClustalW was used to make a draft of multiple sequence alignments. Then, in Excel the repeats were aligned based on their polymorphic traits.

Isolation of pVS31

pVS31 was a generous gift from the Zakian Lab at Princeton University. This plasmid encodes the *pif1-m2* allele, which contains a variant of PIF1 where the methionine at position 39 is mutated to an alanine. Without this second start codon, no truncated nuclear PIF1 is transcribed, however, this specific mutation allows for mitochondrial maintenance due to the presence of the full length PIF1. This plasmid also contains a copy of URA3 to allow for transformation into yeast. Electrocompetent DH5α were transformed with 1 µL of this plasmid and plated on LB + 200 mg/ml Ampicillin. pVS31 was purified from overnight 5 mL LB + Amp cultures using Qiagen Miniprep kit. The Thermo Fisher Scientific Nanodrop 2000 Spectrophotometer was used to obtain

plasmid concentration and purity. Details about all plasmids and yeast strains used in this work can be found in supplemental tables S1 and S2.

Transformation of S. cerevisae with pVS31

The two-step gene replacement necessary to create the *pif1-m2* mutant followed a previously published protocol (Schulz and Zakian, 1994). A diagram of the necessary two-step gene replacement can be found in Figure 1. One µg of pVS31 was cut with restriction enzyme HindIII (NEB) overnight at 37°C in NEB Buffer 2.1. This digest was verified by running a portion of the reaction on a 1% agarose gel, and transformed into YSF1008 (a *ura3*Δ strain) following a "High Efficiency Yeast Transformation" (Gietz and Schiestl, 2007a). YSF1008, was incubated with the transformation mixture at 42°C for 40 minutes. Then, the cells were pelleted, washed, and plated on YPD media. After one day of growth, the cells were replica plated to SC-Ura plates to select for Ura^+ transformants.

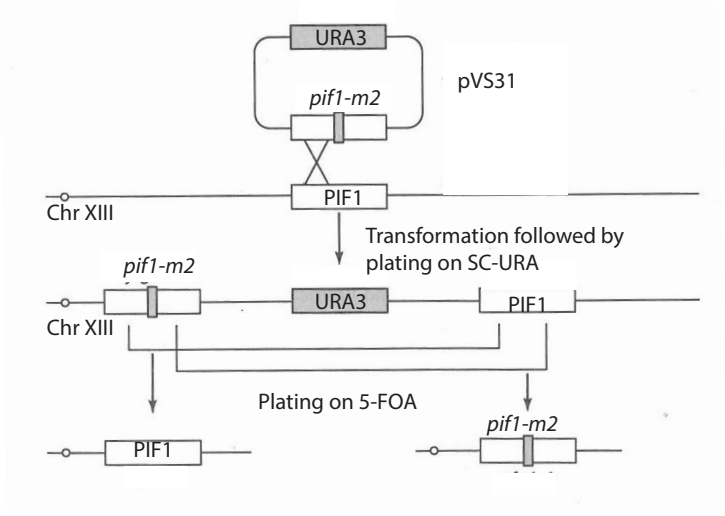


Figure 1: Overview of two-step gene replacement necessary to make *pif1-m2* mutant. First a Ura^- strain is transformed with pVS31 and then subsequent transformants are plated on 5-FOA to screen for colonies that solely contain the mutant allele.

Verification of transformants

To verify that the mutated version of PIF1 transformed appropriately into YSF1008, PCR primers were developed to amplify the first 600 nucleotides of the PIF1 locus. By mutating the second methionine at position 39, an Xho1 site was knocked into the sequence. This feature allows for efficient screening of the transformants that had both the wild type PIF1 allele, as well as the mutant *pif1-m2*. The resulting colonies were screned by amplifying the 5' region of PIF1 was using the primers described above. Then the subsequent PCR product was digested with Xho1. After digestion, the original uncut product and cut product were run on a gel to visualize this point mutation. Those PCR products that demonstrated a cut-banding pattern possessed the mutant allele. To start the screening process, the transformants were restreaked on SC-Ura plates. Then, genomic preps of the Ura^+ transformants were obtained using the previously described phenol-chloroform extraction. 1μL of prep was used as a template in the following PCR reaction. The PCR protocol for amplifying the 5' region of PIF1 was identical as described above except for a shorter 40-second extension time. Following amplification, the samples were digested with NEB's Xho1 in CutSmart™ Buffer for two hours at 37^{o}C and then run on a 1.5% agarose gel. The pVS31 plasmid was used as a positive control while a genomic prep from an unrelated GRY3019 strain was used as a negative control. Those colonies that demonstrated the correct banding pattern were grown overnight in 5mL YPD, plated on YPD, and additionally preserved as glycerol stocks.

Counter selection of the Ura$^+$ marker

To isolate colonies that only possessed the mutant *pif1-m2* allele, YBIR5 was plated onto 5-Fluoroorotic Acid (5-FOA). 5-FOA selected against the transformants that possessed both the wild type PIF1 allele and the mutant *pif1-m2.* Cells that grow on this medium contained only the mutant *pif1-m2* allele or the wild type allele. Colonies that appeared on day three and day five were screened as described previously to differentiate between PIF1 wild type and *pif1-m2* populations. The colony that demonstrated the appropriate Xho1 banding pattern after growth on 5-FOA was plated on SC-Ura to ensure proper counter-selection. YBIR6 is the strain containing solely the *pif1-m2* allele.

Mating pif1-m2 allele into GRY3019

The Strathern Lab graciously donated the strain GRY3019 for use in this project. This strain possesses a tetracycline operator in the promoter of the RPO21 locus, which contains a G418^r marker. Additionally, the Tet-transactivator, which binds to this operator is Ura$^+$ and integrated into the genome. Therefore, pVS31 could not be transformed directly into GRY3019 because it was already Ura$^+$. To construct a GRY3019 strain with the *pif1-m2* allele, YBIR6 (mating type alpha) had to be mated with GRY3019 (mating type a). Tetrad dissection was conducted and spores were replica plated onto 4 plates. The SC-Ura plate selected for spores with the Tet-transactivator, while YPD+G418 selected for spores with the Tet-promoter. To check for mating type, the tetrads were mated with the two mating tester strains (one * and one a) and then the diploids were selected for. This allowed for examination of the segregation of the mating type allele. Tetratypes that demonstrated 2:2 gene segregation at all 3 loci and had spores,

which were Ura^+ and $G418^r$, were screened for *pif1-m2* mutation using the same protocol as mentioned previously. Spore 13D had all the necessary markers and the mutant *pif1-m2* allele. This strain was renamed YBIR7.

Spotting assay to verify proper YBIR7 construction

To verify that YBIR7 was constructed correctly and results in the same phenotypes as GRY3019, a spotting assay was conducted. In this experiment, YBIR7 and GRY3019 were transformed with three Leu2 plasmids: a plasmid containing a wild type copy of RNA polymerase II (pRPO21), an empty Leu2 plasmid (pLeu), and a plasmid that encodes a copy of Rpo21 with only 8 functional heptapeptide repeats (pCTD8). The cells were transformed with 1 μg of appropriate plasmid DNA following the protocol outlined in "Quick and Easy Transformation" (Gietz and Schiestl, 2007b). Transformants were restreaked on SC-Leu plates to ensure Leu prototrophy. The spotting assays were performed following the protocol from Fuchs titled "Yeast Growth Assay" (Fuchs et al., 2012). The strains were plated onto two types of plates (SC-Leu plates and SC-Leu + 40 mg/ml doxycycline). Because of the tTA-dependent system used in the Fuchs Lab, cells in the presence of doxycycline (Dox) are only transcribing the plasmid copy of *rpo21,* while cells growing on media without the drug express both the genomic and plasmid copy.

Large scale fluctuation assay

Yeast transcribing solely pCTD8 will die due to insufficient number of heptad repeats in the CTD domain of RNA polymerase II. With this phenotype, a large-scale

fluctuation assay can be conducted to calculate the mutation rate of this event. In this assay, GRY3019 + pCTD8 and YBIR7 + pCTD8 were used. A single colony of each strain was diluted in 1 mL of water. Then, this mixture was diluted 1:1000 and 200 µL of that dilution was plated on YPD plates. These plates then grew for three days, allowing for mutational accumulation similar to the protocol outline in Shishkin (Shishkin et al., 2009). Then, 12 colonies from these plates were randomly selected and used in a fluctuation assay. An individual colony was placed in 250 µL of sterile water. 200 µL of this mixture was plated on SC-Leu + 40 mg/ml Dox. The colonies that appeared on these Dox plates were suppressors, which underwent mutations that allowed them to survive on these selective plates. Then the remaining 50 µL was diluted ten-fold four times in a 96-well plate. 200 µL of this 10^4 –fold dilution was plated on YPD to estimate the total number of cells originally plated. The YPD plates were counted at two days while the suppressor plates were counted at four days. To calculate mutation rate, the FALCOR program (found at http://www.keshavsingh.org/protocols/FALCOR.html) was used (Hall et al., 2009).

Evaluation of suppressors using PCR

DNA was purified from four suppressors from each plate, each using a modified version of the phenol-chloroform extraction called a “smash and grab”. 1 µL of this prep was used as template for a PCR reaction to characterize the mutation, which allowed the suppressor to successfully grow on Dox. The primers used to evaluate the type of suppressor mutation that occurred were the same primers used to sequence the RPO21 locus. Therefore, the same protocol was used as previously documented. However to

analyze the product, 2% agarose gels were run at 120 V for 60 minutes to get sufficient separation to visualize expansions. Three designations were given to the suppressors: extragenic mutations (G), homologous recombination events (HR), or Expansion events (E). Extragenic events showed no change in plasmid band length, homologous recombination events showed one full-length product, and expansion events showed two products, where the bottom band was larger than the control. The frequency of events was compared between YBIR7 and GRY3019.

Verification of suppressors

If the suppressor was found to be the result of an expansion or homologous recombination event, 1 μL of the "smash and grab" mixture obtained previously was transformed into electrocompetent DH5α cells. These cells were plated on LB + 200 mg/ml Amp plates to select for the modified plasmid. The resulting colonies were mini-prepped using a Qiagen Spin Miniprep kit, and the purified plasmid was sent to Eton Biosciences for sequencing. The returned sequences were analyzed for change in copy number (indication of expansion) or presence of wild type allele on the plasmid (indication of HR event).

Results:

Sequencing of Rpo21 illustrates variable nature of an essential domain

Sequencing the CTD of RPO21 from 36 strains of budding yeast demonstrated the underlying variability of this region. 17 different variants were observed with three general copy number variants. A majority of the sequences contained 26 heptapeptide

repeats (25/36), nine contained 25 heptapeptide repeats, and 1 contained 24 heptapeptide repeats (Figure 2B). The divergent nature of this heptapeptide repeat was seen when looking at 21 base nucleic acid sequence that encoded each individual YSPTSPS repeat. By aligning the individual DNA repeats from the common laboratory strain S288c and then making a phylogenetic tree, it appears that each amino acid repeat is encoded by a different nucleic acid repeat. Even though 20 repeats contain the consensus YSPTSPS sequence, only two of them arise from an identical nucleotide sequence. The divergence among individual repeats was highlighted by Nonet and Young, which allowed them to state which repeats in their system were the sources of the suppressor phentotype (Nonet and Young, 1989). A multiple sequence alignment of each repeat from S288c (Figure 3) demonstrates the divergent nature of the repeats because the clade distance between each group is not zero, indicating that each repeat is encoded by a unique sequence. The one exception to this pattern is the calde distance of zero between repeats three and four. It appears that repeat four arose from a duplication of repeat three due to the conservation of all 21 nucleotides. Conducting similar analysis to uncover the origins of each repeat is not possible due to the extreme sequence divergence.

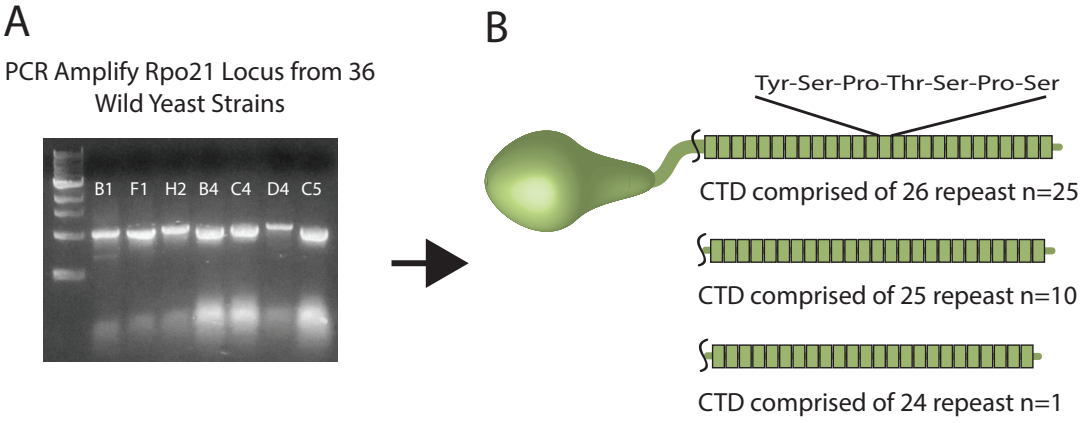


Figure 2: Sequencing *RPO21*from 36 wild yeast strains illustrates the variability of the essential CTD. **A)** Representative gel of PCR amplification of the RPO21 locus. **B)** Cartoon depiction of the RNAP II CTD illustrates copy number variations seen from sequencing results.

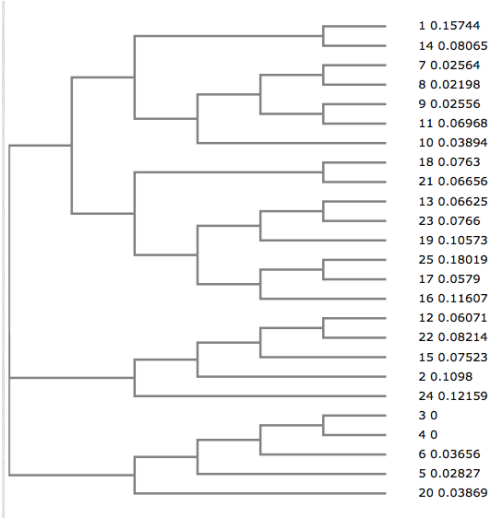


Figure 3: Multiple sequence alignment of the nucleotide sequence that encodes each YSPTSPS repeat from laboratory strain S288c. The non-zero distance between each branch demonstrates that no identical sequence is used to encode a given repeat. The only exception to this phenomenon is the complete conservation of sequence between repeats three and four, which indicates a duplication event.

Even though there are copy number variants in the CTD, this is not the only variable feature of this region. Since each repeat has its own unique and identifiable sequence, a map of the CTD can be constructed. This map is organized in such a way that each row is a different variant of the CTD, while each column is a different repeat. The colors change across the 26 columns, due to the divergent nature of each 21 base sequence. With this map (Figure 4), it is clear that mutations happen at different frequencies in different portions of the CTD. Insertions do occur in the middle of the arrary of repeats as shown by the insertion of the novel repeat 13B in strains A1 and B1. Additionally, deletions occur as well as evident by the loss of repeats, 9, and 10 in the strain F1 and the loss of repeat 16 in the strain F4. However, it appears that the 5' region is the most variable due to the fact that nine deletions occurs within the first six repeats as well as the duplication of repeat four from repeat three. Previous work illustrates the importance of the first 8-10 repeats for basal level growth, while the other 16 repeats do not give any noticeable growth advantage (Nonet et al., 1987; West and Corden, 1995).

This map raises an interesting question: why does the "non-essential" component of the CTD have few mutational events when it appears to have no functional role, while the essential portion of the CTD seems to be more variable?

Additionally, the map in Figure 4 allows for examination of single nucleotide polymorphisms (SNPs) that occur in the CTD. The SNPs are indicated by the textures overlaid on the colored blocks. There are 12 different polymorphisms observed. Nine of the SNPs are synonymous mutations while three of them cause point mutations, one of which mutates a serine 2 to a proline, another changes a proline 3 to a glutamine, and another that mutates a threonine 4 to a methionine. Similar to the trend noted earlier about insertions and deletions, a majority of the SNPs occur in the 5' portion of the tail. This adds more evidence to the fact that the 5' region may be more variable even though the literature has documented that this portion is necessary for function.

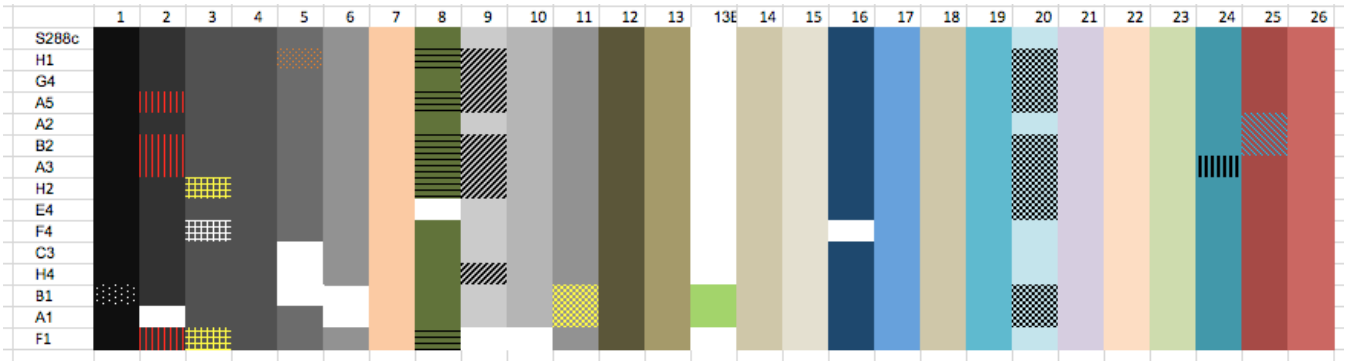


Figure 4: A representative map of the nucleotides encoding the CTD. Each variant is one row while each column is a 21 nucleotide repeat. The rows change colors due to the divergent sequence encoded in each repeat. Each texture represents a SNP while the empty white blocks illustrate deletions of repeats. It appears that the 5' region of the CTD is under different selective pressure then the C-terminal fragment. A majority of the rearrangements as well as the SNPs occur in the 5' region of the CTD.

YBIR7 functions like WT GRY3019 strain in spotting assay

The *pif1-m2* allele was successfully added to the GRY3019 background following the protocol from Schulz and Zakian (Schulz and Zakian, 1994). The data illustrating

proper mutant construction is in supplemental figures S1 and S2. With the successful construction of a strain with the desired properties, YBIR7 was used in spotting assays to ensure proper phenotype. In these assays, we can examine growth when transcribing different mutant *rpo21* alleles, which differ in the number of CTD repeats. This system is possible because of the Ura^{+} Tet-transactivator and $G418^{r}$ Tet-operator in the promoter of RPO21. When there is no Dox in the media, the Tet-transactivator binds to the Tet-operator in the promoter of RPO21, leading to transcription of the genomic copy of RPO21. In this situation, there is also transcription of the plasmid copy of mutant *rpo21,* which is under the control of the normal RPO21 promoter. When Dox is added to the media, the transactivator no longer binds to the operator, preventing transcription of the wild type RPO21. In this situation, only the mutant plasmid allele of *rpo21* is expressed. This system is illustrated in Figure 5A.

Using this tTA-dependent system, previous work in the Fuchs Lab has verified the results of Corden and West that cells with only 8 CTD repeats in RNAP II lead to minimal growth (West and Corden, 1995). YBIR7 was used in this assay to demonstrate that the Tet-off system works appropriately with the mutant *pif1-m2* allele. When YBIR7 harbors pRPO21, it grows sufficiently well on SC-Leu+Dox (data not shown). When YBIR7 harbors an empty Leu+ plasmid, the strain dies on SC-Leu+Dox plates (data not shown). Interestingly, the phenotype between the WT strain and YBIR7 on SC-Dox plates is noticeably different when transcribing a copy of RPO21 with only 8 CTD repeats. YBIR7 + pCTD8 has no noticeable growth while WT + pCTD8 grows slightly (Figure 5B). This difference in growth between wild type and *pif1-m2* was also found in

liquid culture (data not shown). However, when both strains harbor a plasmid with 10 CTD repeats or more, no discernable difference in growth is seen.

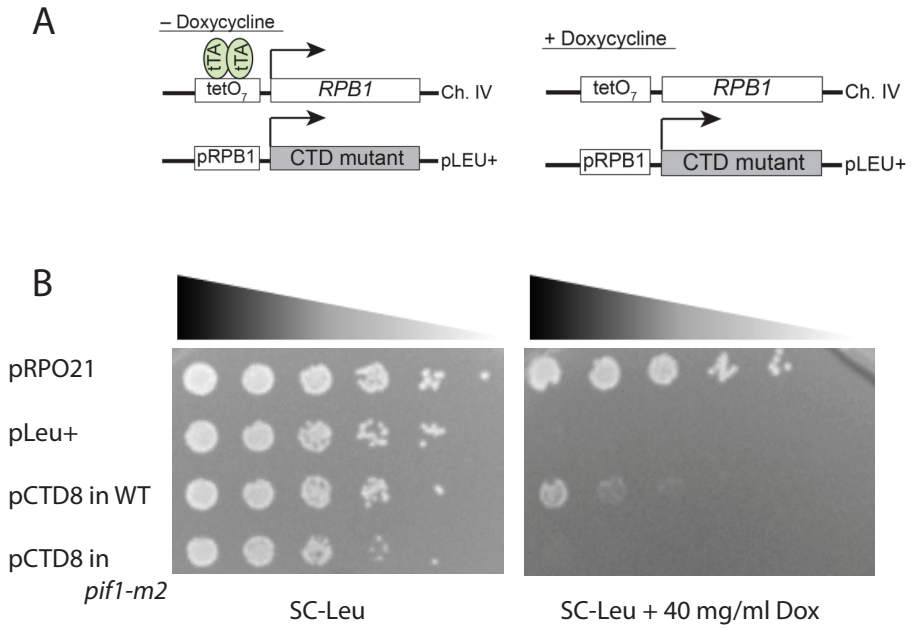


Figure 5: Spotting assays using *pif1-m2* YBIR7 strain. **A)** Depiction of the Tet-off system. Briefly, when cells are grown in normal media both the plasmid and genomic copies of RPO21 are expressed. However, when Dox is added to the media, only the plasmid copy is transcribed. **B)** Representative spotting assay where both the WT strain and YBIR7 contain a plasmid with only 8 CTD repeats. This is not enough to support wild type growth and these cells die when plated on media containing Dox. The SC-Leu plate acts as a control to ensure that the same number of cells was plated. The pRPO21 acts as a positive control, showing that plasmid expression of full length CTD can support cell growth. pLeu+ is the negative control showing that without a copy of RPO21 on a plasmid , the yeast will die when growing on Dox.

Fluctuation assay uncovers higher rate of mutation in pif1-m2 background

A Luria–Delbrück fluctuation experiment can be conducted to uncover the frequency of suppressor events based on the death phenotype seen in Figure 5B. Yeast growing on media in the presence of Dox while transcribing pCTD8 should die. Those colonies that appear on these plates are indicative of cells that underwent supressor mutations. An illustration of this process can be seen in Figure 6A . Figure 6B demonstrates the results of such an experiment comparing the mutation frequency of the

wild type against the mutation frequency of *pif1-m2*. The mutational frequency of *pif1-m2* cells is a $6.45*10^{-6}$ (95% CI=$3.97*10^{-6}$, $9.35*10^{-6}$, n=24) while the mutation frequency of the wild type GRY3019 is $1.38*10^{-6}$ (95% CI=$9.85*10^{-7}$, $1.82*10^{-6}$, n=24). This difference is significant due to the fact that the 95% confidence intervals, as calculated by FALCOR, do not overlap.

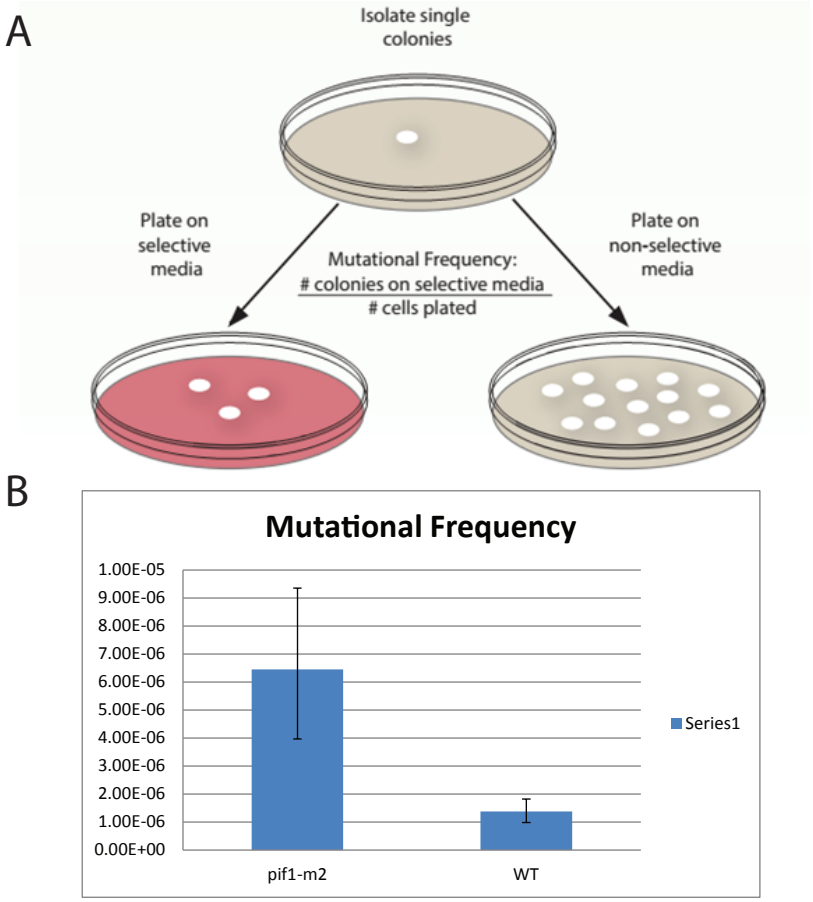


Figure 6: Large scale fluctuation assay demonstrates higher mutational frequency in *pif1-m2* background. **A)** Outline of fluctuation assay. One colony is plated on selective Dox media (red plate) allowing for the appearance of suppressors while a dilution of the colony is plated on nonselective media (off white color) to estimate total cells. **B)** Increase in mutational frequency in *pif1-m2* background. Difference between mutational frequencies is significant due to non-overlapping 95% confidence intervals.

Changes in suppressors frequency hints at PIF1 role in CTD expansion

To examine the mutation type of the suppressors that arose, PCR was used. Figure 7A and 7B are representative gels from this screening process. Two bands are present due to amplification of both the plasmid copy and the genomic copy of RPO21. Expansion events result in a bottom band that increases in size, homologous recombination events are visualuzed by only one genomic-sized product, and extragenic events are visualized by no change in either band.

When comparing the mutational landscape of these suppressors, PIF1's role in CTD variability appears. In the mutant *pif1-m2* background, more expansion events occur when compared to the wild type (56% vs. 32%) (Fig 7C). Additionally, it appears that homologous recombination events occur less frequently in the *pif1-m2* strain (2% vs 12%) (Figure 7C). Other unpublished data from the Fuchs Lab has demonstrated that when mutating other DNA repair related genes (i.e. YKU70 or POL32), there is no change mutational frequency or in suppressor mutational landscape. This data suggests that PIF1 is a negative regulator of CTD expansion.

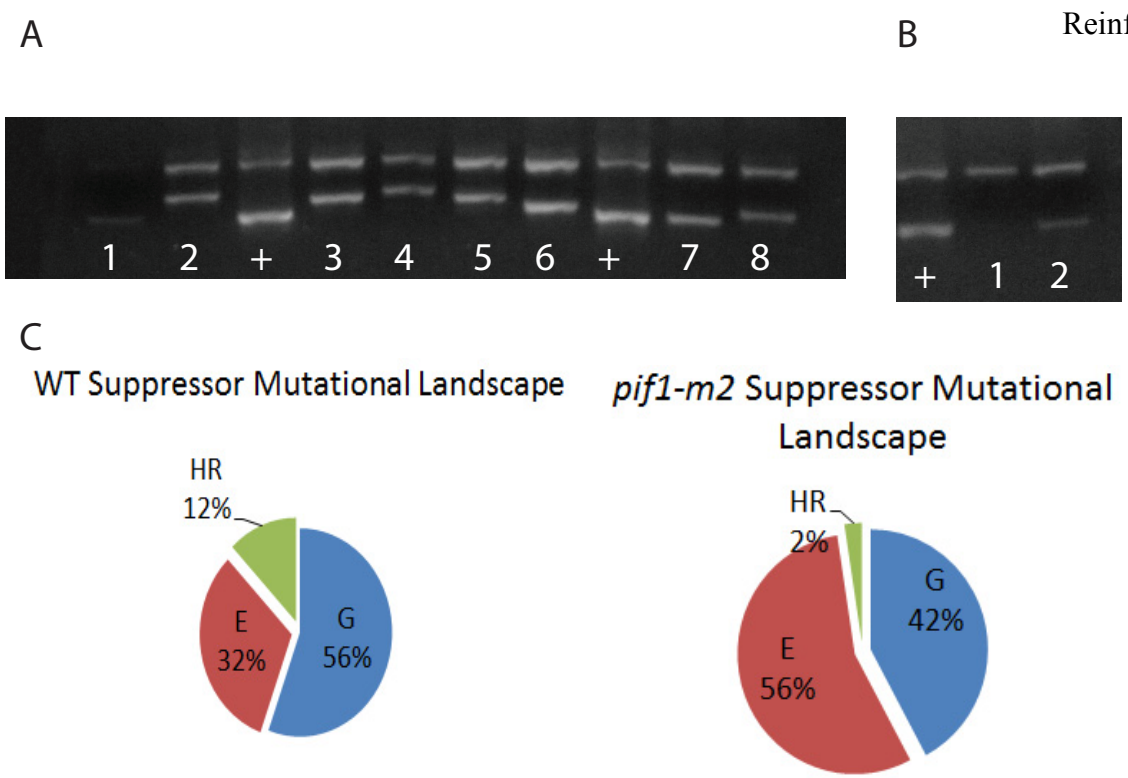


Figure 7: PIF1 may be a negative regulator of CTD expansion. **A)** and **B)** are representative gels from PCR evaluation of suppressors. The bottom band is the plasmid copy of *rpo21* while the top band is the full-length genomic copy. Lanes 2, 3, 4, 5, 6, and 8 in Gel **A** have expanded while Lane 1 in Gel B underwent a homologous recombination. **C)** After screening over 90 colonies from both strains there appears to be an increase in expansions in the *pif1-m2* background (56% vs. 32%). Additionally, there appears to be a dramatic decrease in homologous recombination events (2% vs 12%). E = Expansions, HR= Homologous Recombination, and G= Extragenic Vvents

Sequencing of suppressors demonstrate mutation free direct expansions of repeats

The plasmids obtained from suppressors were sequenced to verify their expansions as well as examine the inserted sequence. In all cases, the larger PCR product indicated insertion of additional repeats on the plasmid. Figure 8 graphically depicts the three different size expansions of 42, 84, and 126 nucleotides that were seen in YBIR7. These mutations added 2, 4, or 6 heptapeptides respectively. Additionally, the observed mutations were encoded by direct repeats of the synthetic sequence used to build the truncated CTD. Figure 9 demonstrates that the inserted sequence has 100% homology to the plasmid from which it arose.

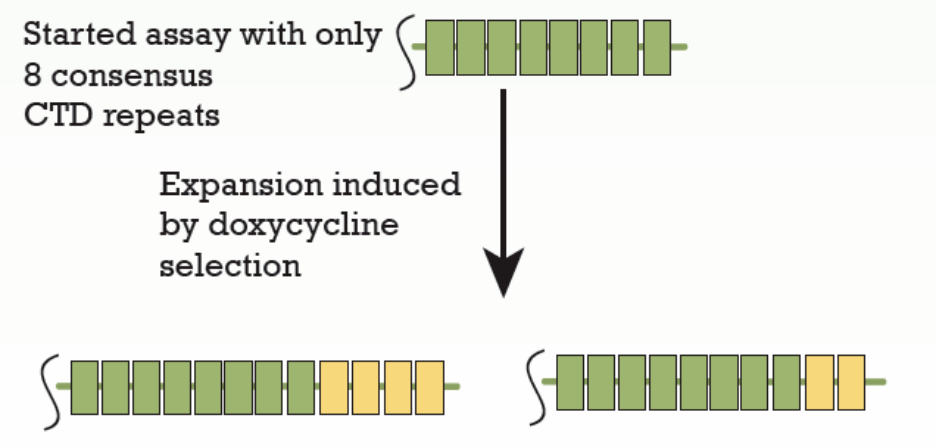


Figure 8: Purification and sequencing of *rpo21* recovered plasmid from suppressor colonies. Observed insertions of 42, 84, and 126 nucleotides to increase the number of heptapeptide repeats to 10,12, or 14 repeats respectively.

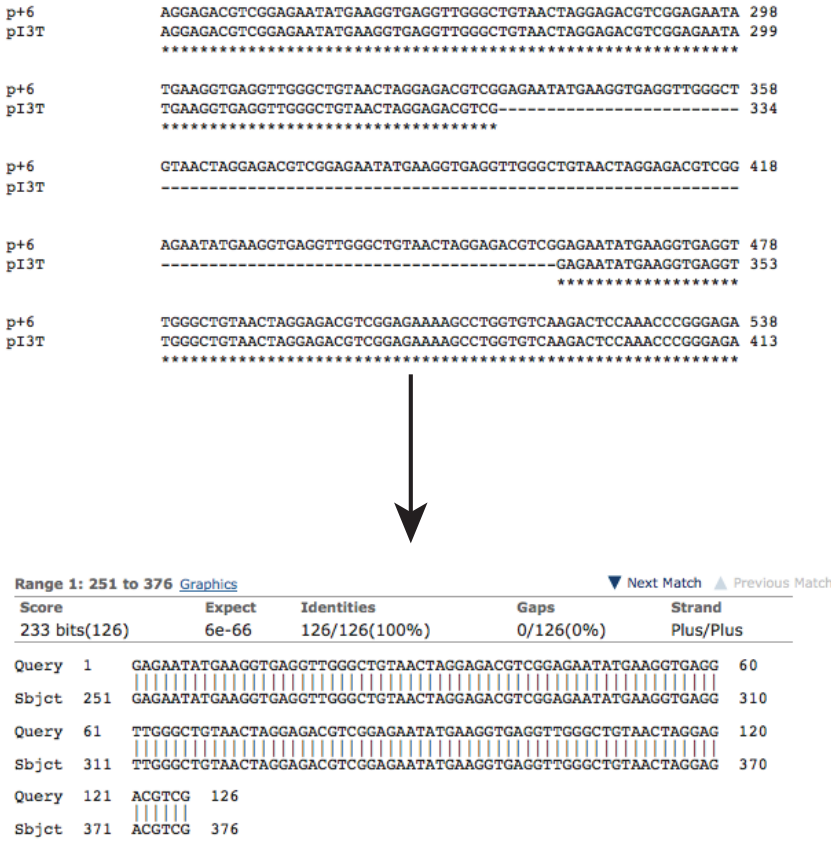


Figure 9: Inserted sequence is perfect duplication of synthetic CTD sequenced. Through a global alignment of suppressor plasmid to pCTD8 (as seen in the top panel), the inserted sequence was obtained. Then the insertion was locally aligned against the pCTD8 sequence. In all cases of expansions, there was perfect homology between the inserted sequence and the original plasmid.

Discussion:

Variability in the context of an essential domain of a necessary protein appears puzzling. In the case of a nonessential gene like FLO1, a cellular membrane glycoprotein involved in adherence and flocculation, variability seems to be adaptive. It was previously shown that differences in FLO1 lectin binding repeat could promote differential adherence to inert surfaces (Verstrepen et al., 2005). Currently, no difference in growth has been seen due to differences in CTD length after 10-consensus heptapeptide repeats (West and Corden, 1995). Yet, the fact that this region appears dynamic in similar ways to the previously mentioned non-essential FLO1, remains unclear. A massive DNA rearrangement is not deadly in the context of FLO1 and results in yeast with differential fitness for their environment. However, an alien insertion or an out of frame mutation could render RPO21 nonfunctional and kill the cell harboring that variant.

Therefore, to limit some of variability in this region, evolution has appeared to select for very divergent nucleotide sequences in the CTD. With the third position codon wobble as well as two different codon sets that can encode serine, there are many possible nucleotide sequences that could result in YSPTSPS repeat. Due to the inherent instability of repetitive DNA, it appears that evolution has favored using many of these possible combinations to encode each individual heptapeptide. Therefore, strains with more sequence divergence in between each 21 nucleotide repeat may be selected for limiting the amount of DNA rearrangements that can occur in this gene.

Although each repeat sequence is divergent, that does not explain the fact that the 5’ region appears to have experienced more mutational events then the 3’ region in *S.*

cerevisiae. One component of the 3' end that may be kept overtime is the aforementioned G4 encoding sequence. A fairly recent study by another research group found that, when comparing G4 conservation across fungal species, these DNA secondary structures were more often then by chance (Capra et al., 2010). Therefore, a G4 interacting protein may be recognizing the most 3' region of this repeat. Mutations to this sequence would eliminate a protein's ability to bind and/or to recognize this repeat, which could impact both the sequence and expression of RPO21.

The duplication event seen in repeat four can be used as a model for the overall evolution of this region. Over time, repeats expand or contract perfectly, but to prevent further copy number variation, the sequence that encodes the newly formed heptapeptide undergoes additional point mutations to differ from its neighboring repeats. This permits a maximal amount of variability while at the same time preventing large rearrangements that would be deleterious to the yeast cell. It remains unsolved why the most C-terminal repeats encode a secondary structure associated with fragility, yet appear more resistant to change.

The presence of G4 DNA in RPO21 raises a few questions. This structure is known to be associated with fragile sites as well as DNA damage (Capra et al., 2010; Paeschke et al., 2011). There is little reason for a crucial gene to be under such mutational pressure. Even though this seems unfavorable, there is some evidence that the location of this G4 is part of a larger currently unexplained trend seen across eukaryotic organisms. The Zakian Lab found that G4 DNA appear less often than chance in open reading frames, however the G4 DNA appears significantly more often (78% of the time)

on the template strand in fission and budding yeast as well as in humans (Sabouri et al., 2014).

Additionally, since the G4 structure in the CTD resides on the leading template strand for DNA replication, it exists in the unstable orientation. Lopes and coworkers showed that this orientation contributes to a significant increase in instability (Lopes et al., 2011). Therefore, the RPO21 locus, encoding an essential eukaryotic enzyme, contains a minisatellite poised to mutate more often then the rest of the genome.

The data compiled in the fluctuation assay illustrates that PIF1 may act as a negative regulator of expansion events in the CTD. With no functional nuclear PIF1, the frequency of expansion events increases. This is in agreement with data showing that PIF1 plays a role in negatively regulating *de novo* telomere lengthening as well as in instability in the CEB1 minisatellite (Ribeyre et al., 2009; Schulz and Zakian, 1994). It is worth noting that in three papers published by the Nicolas Lab use different *pif1* mutant alleles (Lopes et al., 2011; Piazza et al., 2010; Ribeyre et al., 2009). They found that the *pif1-m2* allele does not completely eliminate PIF1 unwinding in the nuclear genome (Ribeyre et al., 2009). Even though nuclear PIF1 cannot be detected by western blot, there is a significant decrease in CEB1 minisatellite instability. CEB1 still demonstrates greater instability in the *pif1-m2* mutant than observed in the wild type background, however both the *pif1Δ* strain, as well as the nonprocessive *pif1 K264A* mutant show larger increases in CEB1 instability. Therefore, a more dramatic phenotype may be observed when using either one of these mutants in this work when trying to uncover the basis of CTD expansions.

By looking at the work of Nicolas and colleagues who have studied the aforementioned human CEB1 minisatellite, it may be possible to gain some insight into the evolution of the CTD of RPO21 (Lopes et al., 2002; Lopes et al., 2011; Piazza et al., 2010; Ribeyre et al., 2009). This may be an appropriate comparison because the CEB1 repeat is encoded by a group of smaller 38-48 nucleotide repeats which are distinct from one another in the same way that the 21 nucleotides that encode an individual heptapeptide differ from one to the next. This lab similarly observed complete loss or gain of distinct repeats of the CEB1 minisatellite just as we see perfect expansion of our synthetic CTD sequence. Events characterized as complex events in the CEB1 assay contained both insertions and deletions in multiple locations in the tandem array (Ribeyre et al., 2009). This is reminiscent of one of the wild yeast CTD sequences from the strain B1 where it contained 25 heptapeptide repeats, but did so by losing repeats 9 and 10, and adding a new repeat in between 12 and 13.

PIF1 may be mediating these expansion events by two different mechanisms. One mechanism may be through PIF1's ability to unwind RNA/DNA hybrids (R-loops). To tease apart PIF1's relationship with CTD expansion, this assay could be replicated in an *rnh1*Δ *rnh201*Δ double mutant. These two genes encode different versions of Ribonuclease (RNase) H, which is a nuclease responsible for degrading RNA/DNA hybrids. PIF1 has a higher affinity for R-loop structures and has been shown to interact with these structures as a part of processing Okazaki fragments (Boule and Zakian, 2007; Pike et al., 2010). In this *rnh1*Δ *rnh201*Δ strain, there will be increased R-loops due to the depletion of RNase H. If this mutant demonstrates the same increase in expansion frequency as the *pif1-m2* data, increase in R-loop formation in the *pif1-m2* strain may be

contributing to the expansion of the CTD. This phenotype would suggest that PIF1's role in processing the R-loop flaps on the lagging strand may influence the expansion events. However, there is a possibility that this double mutant has no impact on the frequency of expansions. In that case, PIF1 may mediate these expansions due to its role in resolving G4 DNA as highlighted by both the Zakian and Nicolas labs (Lopes et al., 2011; Paeschke et al., 2011).

The data from the mutational analysis is also in agreement with recent work that uncovers PIF1's function in DNA repair. PIF1 appears to be key to proper resolution of the D-loop during homologous recombination (Wilson et al., 2013). The practical elimination of homologous recombination events, when using the *pif1-m2* mutant, validates the idea that this protein is key in allowing for faithful homologous recombination.

It is worth noting that further manipulation of the nucleotide sequence may uncover PIF1's role with more accuracy. The current sequence used in the pCTD8 plasmid is not imbalanced in terms of cytosines and guanines, nor does it encode a strong G4 motif. In the plasmid sequence, there are no guanine tetrads, however there is one stretch of three consecutive guanines as well as multiple instances of guanine couplets. The Fragile X trinucleotide (CGG) is a documented instance of an *in vitro* G4 DNA that is formed by a weak motif solely made up of guanine couplets (Fry and Loeb, 1994) Therefore, pCTD8 may not be a perfect recapitulation of the nucleotide composition of the CTD, but there is some possibility that this sequence can still form a G4 structure. A comparison of the G4 sequence from *RPO21* and the Fragile X minisatellite, as well as the possible G4 in the pCTD8 can be seen in Figure 10. It may be worthwhile to conduct

circular dichroism analysis on the oligonucleotides used to make the pCTD8 plasmid to assess whether or not this sequences leads to G4 formation *in vitro*. It is worth noting that the algorithm used by Capra would not have called the sequence in the plasmid or the sequence of the Fragile X repeat a G4 because neither of these sequences contains four tracks of three or more guanines (Capra et al., 2010).

Further work with other deletion strains may impart more information about the mechanism of minisatellite expansion. Work studying RAD27, a endonuclease responsible for trimming the flaps off of the Okazaki fragments, has shown that a deletion of this enzyme leads to both mutagenic and faithful change in minisatellite repeat copy number (Lopes et al., 2002; Serero et al., 2014). Other work has examined the frequency of FLO1's expansions/contractions, but no mechanism was investigated (Verstrepen et al., 2005). Using this fluctuation assay with the Tet-off system and pCTD8 in other mutants related to genomic instability (*msh2Δ tsa1Δ, rad27Δ),* it may be possible to more gain information about what proteins or pathways may contribute to the diversity seen in minisatellite of the CTD.

Fragile X G4
CGG**C**GG**C**GG**C**GG

p8CTD hypothesized G4
GGTGA**GG**TT**GGG**CTGTAACTA**GG**

RPO21 CTD G4
GGGCTGTAGCCTGGAGATGTT**GGG**GAGTAAGAAGGTGAAGTA**GGG**CTATAGTTTGGAGAGGT**GGG**

Figure 10: Sequences of G4 forming sequences. The pCTD8 demonstrates a weak G4 motif. This hypothetical motif is longer than the additionally weak G4 Fragile X repeat, which is known to form a G4. The G4 forming sequence in the CTD is included.

The data collected in this work combined with other data by Morrill (Fuchs Lab, unpublished results) illustrates that these expansions are most likely occurring post-

replication, which is in accordance with the literature (Lopes et al., 2011; Ribeyre et al., 2009). Other data from the lab shows that *rad52Δ* and *rad5Δ* result in an elimination of expansion events while this work indicates that the *pif1-m2* mutant results in an increase in expansions. With these facts, there is a high likelihood that these expansions are occurring after replication. This does raise a new issue as to how the G4 DNA can form when it is encoded on the leading strand. For this structure to form, single stranded DNA (ssDNA) is required. Therefore, it appears that the conditions of leading strand synthesis do not permit this structure from forming, because the replication fork would move through this region with too much speed, never giving rise to ssDNA. With our current understanding, if the G-rich strand cannot exist as ssDNA, the G4 structure should not form. Even though this inconsistency appears, the Nicolas Lab uncovered the same phenomenon that G4 DNA on leading strands appears to increase instability (Lopes et al., 2011). To resolve this discrepancy it may be hypothesized that discontinuous leading strand synthesis is occurring on the G-rich template of the CTD. *In vitro* data from Weitzmann and coworkers suggests that G-rich template strands in hypervariable minisatellites are difficult to faithfully replicate through (Weitzmann et al., 1997). Therefore, the inconsistent replication through the CTD allows for ssDNA to exist, which promotes formation of the G4. In the situation where nuclear PIF1 has been eliminated, the G4 in the CTD cannot be resolved properly, leading to instability of the repeat. The model described above is demonstrated pictorially below in Figure 11.

In future work to uncover the mechanism of CTD expansion, the heptapeptide repeat may need to be studied in a non-coding context, where it is integrated into the genome. Currently, the system used is completely plasmid based, which provides a

problem when trying to discern the mechanism of expansion. There is some evidence documenting that multiple plasmids can exist in a given suppressor. This may allow for the expansion or homologous recombination events to occur through using the excess plasmid as a template for these events. Additionally, there are constraints in this assay because the minisatellite in question encodes an essential protein repeat. Therefore, we cannot see contraction events because this type of mutation would result in fewer heptapeptide repeats and thus be lethal. The current method also cannot examine whether or not these events are mutagenic because those mutations would endow less fitness to a given suppressor. Therefore, using the nucleotides that encode RPO21 as well as the synthetic nucleotide repeat used in the Fuchs Lab in a non-coding reporter, may allow for a deeper understanding of the mechanism contributing to CTD diversity. A system like the one implemented in Shah's work where repetitive sequence is cloned in between GAL promoter and a copy of the CAN1 gene may allow for a better understanding of the factors that contribute to CTD instability (Shah et al., 2014).

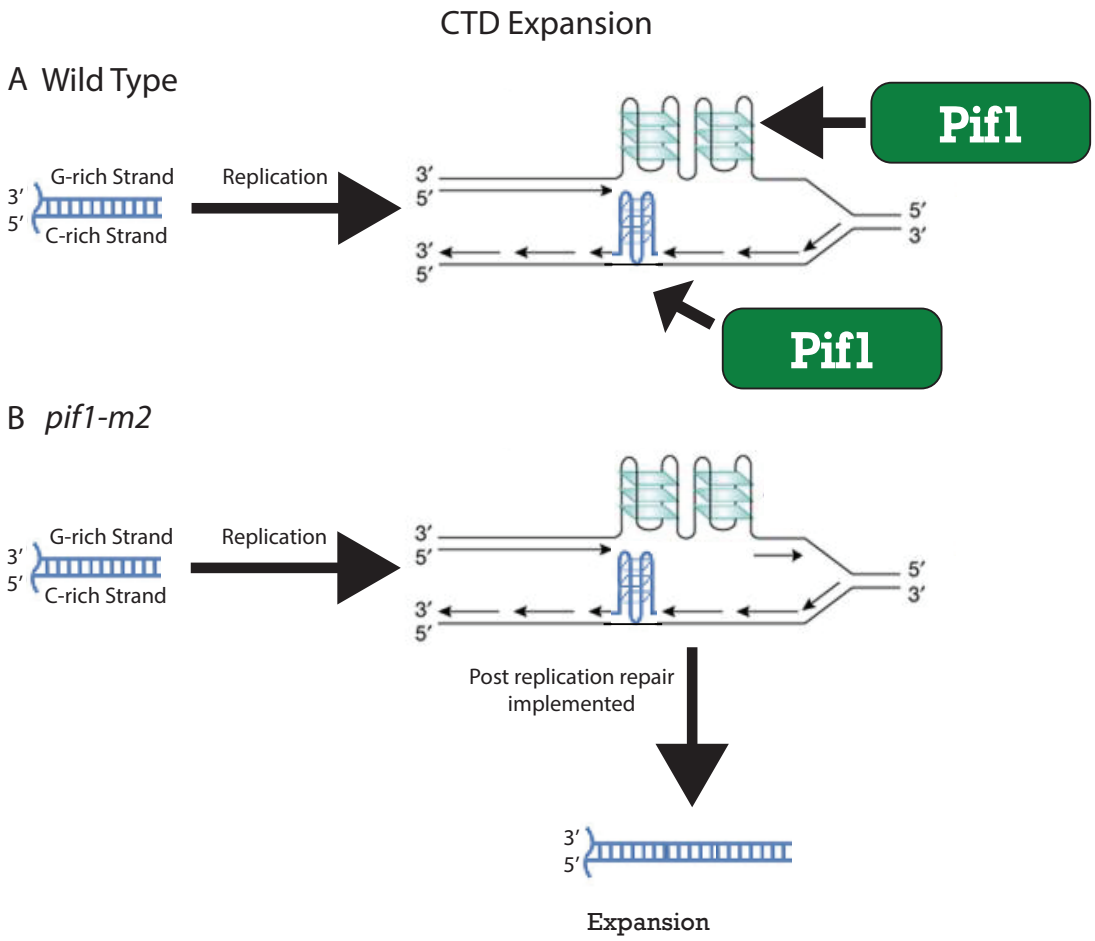


Figure 11: Possible Mechanism of CTD Expansion. The CTD is encoded by a G rich template, which is part of leading strand synthesis. A) Normal resolution of G4 structures that appear both in the template of the leading strand synthesis as well as lagging strand 5' flaps. With functional PIF1, these G4 DNAs are resolved and information is faithfully copied. B) Expansion of CTD occurs when nuclear PIF1 is absent. The G4 structures form as replication occurs discontinuously on leading strand. Replication continues on through the rest of the genome. G4 DNA still exists due to lack of PIF1. Post replication, the single stranded G4 DNA is resolved through a homologous recombination dependent repair mechanism. Figure is modified from Davis and Chung (Chung, 2014; Davis and Maizels, 2011)

Chapter Two: Implementing Bioinformatics to find minisatellites similar RPO21

Introduction:

Repetitive elements are ubiquitous features of eukaryotic genomes (46% of the human genome) that have long been considered junk DNA (Gemayel et al., 2010). Repetitive sequences can be located throughout the genome, including coding regions, non-coding regions, telomeres, and centromeres. Smaller repetitive sequences (with repeating units less than 10 nucleotides) are referred to as microsatellites. These regions have been studied extensively, especially in a disease context like Spinobulbar Muscular Atrophy and Huntington's Disease where the etiology of these diseases stems from trinucleotide expansions in coding regions. Larger repetitive sequences (of repetitive motifs between 10 and 150 nucleotides) are typically considered minisatellites. These elements are used to assemble diagnostic DNA fingerprints because of the variability of these regions. There is a third group of repetitive elements, megasatellites, which are repeats containing motifs greater than 150 nucleotides, that are important in the development of pathogenic *Candida albicans* (Rolland et al., 2010).

By examining coding minisatellites in the model genetic organism *S. cerevisiae*, a possible mechanism for minisatellite variability may be found. Previous work reported that coding minisatellites occur in 12.7% of yeast open reading frames (ORFs) and many of these repetitive elements have been found to be polymorphic (Gemayel et al., 2010; Richard and Dujon, 2006; Verstrepen et al., 2005). A signature of many yeast repetitive proteins is the extreme GC skew in their nucleotide sequence (Richard and Dujon, 2006). These large values are indicators of an imbalance between guanine and cytosine

nucleotides on a single strand of DNA. This value is calculated by subtracting the total amount of cytosine residues from guanine residues; then that value is divided by the combined total of guanines and cytosines. Positive values illustrate an enrichment of guanine nucleotides, while negative values illustrate a bias toward cytosines on a given strand of DNA. GC skew sign changes have been previously found to be associated with prokaryotic origins of replications (Lobry, 1996). Therefore, one could hypothesize that the abnormal GC skew in these repetitive regions may promote the formation of a certain secondary structure and/or recruit certain protein factors that lead to the instability of minisatellites. Even though Richard and Dujon (Richard and Dujon, 2006) uncovered this extreme trait, no link was found between GC skew and a factor that could contribute to the apparent variability of minisatellites.

The enrichment of GC-skewed DNA strands within minisatellites led to a prediction that G4 sequences may contribute to the known instability of coding minisatellites. Knowing that PIF1 unwinds G4 quadruplexes and that *pif1Δ* leads to minisatellite instability, I believe that that PIF1 may be involved in suppressing instability in all minisatellites, similar to its role as a negative regulator of telomere lengthening (Paeschke et al., 2013; Ribeyre et al., 2009; Schulz and Zakian, 1994). However, previous literature suggests that PIF1's function is specific for certain G4 DNAs (Ribeyre et al., 2009). The CEB1 minisatellite instability assay from the Nicolas Lab was repeated with minisatellites from the yeast open reading frames HKR1, FLO1, DAN4, or NUM1 (Ribeyre et al., 2009). Even though *pif1*Δ caused instability of a human G4 containing CEB1 repeat, it did not impact the frequency of expansion or contraction of any of the yeast minisatellites. A plausible explanation for this result is that PIF1 is not

recognizing *all* minisatellites, but rather it is recognizing *specific* minisatellites that contain G4 sequences. Even though these minisatellites are guanine rich, none of these polymorphic minisatellites encode a strong G4 motif as determined by Capra (Capra et al., 2010). The ChIP studies conducted on PIF1 and PFH1, the *S. pombe* homolog, support this model of high PIF1 target specificity due to the fact that these helicases were only found at 20-25% of all the respective predicted G4 structures (Paeschke et al., 2011; Sabouri et al., 2014).

It is worth noting that RAD27, the yeast FEN1 homolog, appears to be involved in some aspect of minisatellite evolution in budding yeast. This enzyme is responsible for degrading the overhang of nucleotides caused by Okazaki fragment displacement. The previously mentioned CEB1 minisatellite, as well as other human minisatellites such as MS1 and MS32, MS205 demonstrate increased instability in a *rad27Δ* strain (Lopes et al., 2002; Maleki et al., 2002). Evolving a *rad27Δ* strain through 100 bottlenecks also resulted in a significant increase in the amount of large insertions and deletions observed in coding minisatellites (Serero et al., 2014). The deletion of RAD27 is thought to have a drastic impact on these microsatellites because the improper flap degradation leads to strand invasion that ultimately causes changes in repeat copy number (Lopes et al., 2006). However, it appears that some genes, including RPO21, are not destabilized in the *rad27Δ* background, which suggests that the instability of RPO21 and repeats that share similar characteristics may be due to others mechanisms.

Using a bioinformatics-driven approach, the Fuchs Lab became interested in uncovering new elements of a subset of minisatellite variability. An aggregate of literature and data from the last chapter predicted that G4 structure and PIF1 binding may

impact the variability of some repetitive proteins. Therefore, I hypothesized that repeats that contain both PIF1 binding and predicted G4 sequences were more likely to be variable.

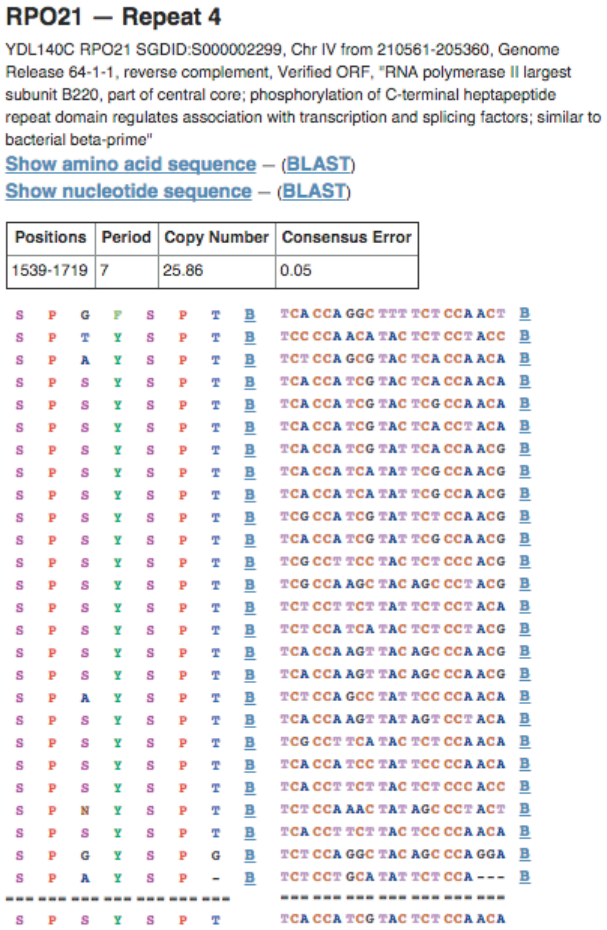
To evaluate this hypothesis, a tool was developed to easily examine the nucleotide sequence that encodes a given protein repeat. Previous tools have been designed to survey the amino acid sequence of any tandem repetitive protein, but no tool works directly to examine the DNA encoding these repetitive proteins. The Fuchs Lab, in conjunction with the Cowen Lab, developed a bioinformatics tool, XSTREAM with DNA correspondence, to do so. This program was built by implementing the already existing XSTREAM tandem protein finder developed by the Cooper Lab at USCB (Newman and Cooper, 2007). XSTREAM with DNA correspondence allows the user to view the repetitive nucleotide sequence side by side with the repetitive protein it encodes. After producing a list of repetitive proteins, the predicted G4 sites and the PIF1 binding locations were mapped to the *S. cerevisiae* genome (Capra et al., 2010; Paeschke et al., 2011). Using a multiple sequence alignment tool to find variable minisatellites, it became possible to assess the impact of G4 DNA as well as PIF1 binding on variability of minisatellites.

Methods:

XSTREAM with DNA correspondence development

XSTREAM with DNA correspondence was made in collaboration with the Cowen Lab from the Tufts Computer Science Department. With this tool, it became possible to look at both the repetitive amino acid and repetitive DNA of a given gene.

This program allows the user to change the parameters that ultimately define a repeat. The user can manipulate such aspects as minimum or maximum repeat copy number, period, minimum domain length and gaps. My work defined a repeat as an amino acid sequence that had at least a period of two (dipeptide), repeated itself at least twice, and had a minimum length of 10 amino acids. Therefore, a dipeptide had to repeat five times, and a pentaheptide had to repeat two times to be considered a repeat. Each repetitive gene's systematic name was saved for comparison with datasets from the Zakian Lab. An example of XSTREAM with DNA correspondence output is seen below in Figure 1.



RPO21 — Repeat 4

YDL140C RPO21 SGDID:S000002299, Chr IV from 210561-205360, Genome Release 64-1-1, reverse complement, Verified ORF, "RNA polymerase II largest subunit B220, part of central core; phosphorylation of C-terminal heptapeptide repeat domain regulates association with transcription and splicing factors; similar to bacterial beta-prime"

Show amino acid sequence – (BLAST)
Show nucleotide sequence – (BLAST)

Positions	Period	Copy Number	Consensus Error
1539-1719	7	25.86	0.05

S P G F S P T B TCA CCA GGC TTT TCT CCA ACT B
S P T Y S P T B TCC CCA ACA TAC TCT CCT ACC B
S P A Y S P T B TCT CCA GCG TAC TCA CCA ACA B
S P S Y S P T B TCA CCA TCG TAC TCA CCA ACA B
S P S Y S P T B TCA CCA TCG TAC TCG CCA ACA B
S P S Y S P T B TCA CCA TCG TAC TCA CCT ACA B
S P S Y S P T B TCA CCA TCG TAT TCA CCA ACG B
S P S Y S P T B TCA CCA TCA TAT TCG CCA ACG B
S P S Y S P T B TCA CCA TCA TAT TCG CCA ACG B
S P S Y S P T B TCG CCA TCG TAT TCT CCA ACG B
S P S Y S P T B TCA CCA TCG TAT TCG CCA ACG B
S P S Y S P T B TCG CCT TCC TAC TCT CCC ACG B
S P S Y S P T B TCG CCA AGC TAC AGC CCT ACG B
S P S Y S P T B TCT CCT TCT TAT TCT CCT ACA B
S P S Y S P T B TCT CCA TCA TAC TCT CCT ACG B
S P S Y S P T B TCA CCA AGT TAC AGC CCA ACG B
S P S Y S P T B TCA CCA AGT TAC AGC CCA ACG B
S P A Y S P T B TCT CCA GCC TAT TCC CCA ACA B
S P S Y S P T B TCA CCA AGT TAT AGT CCT ACA B
S P S Y S P T B TCG CCT TCA TAC TCT CCA ACA B
S P S Y S P T B TCA CCA TCC TAT TCC CCA ACA B
S P S Y S P T B TCA CCT TCT TAC TCT CCC ACC B
S P N Y S P T B TCT CCA AAC TAT AGC CCT ACT B
S P S Y S P T B TCA CCT TCT TAC TCC CCA ACA B
S P G Y S P G B TCT CCA GGC TAC AGC CCA GGA B
S P A Y S P - B TCT CCT GCA TAT TCT CCA --- B
--- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
S P S Y S P T TCA CCA TCG TAC TCT CCA ACA

Figure 1: A sample output of XSTREAM with DNA correspondence. The gene examined here is RPO21. The left panel is the amino acid repeat while the right panel is the nucleotide sequence. In the output, it is possible to see the previously reported extreme GC Skew trend (Richard and Dujon, 2006). The right panel illustrates the extreme bias for cytosines over guanines on the coding strand. The G4 encoding sequence can be seen at the very C-terminal repeats.

Mapping predicted G4 sites and PIF1 binding locations

The supplement from Paeschke contains the ChIP binding peaks associated with their myc-tagged *pif1k264a* strain. This mutation halts the enzyme's processivity by making a mutation in the Walker's A motif (Paeschke et al., 2011). This prevents the enzyme from unwinding its DNA sequence of interest, but not from binding at the appropriate location. Therefore, conducting ChIP with this mutant should accurately describe PIF1 binding sites. Additionally, the supplement from Capra published all the G4 sites in the *S. cerevisiae* genome as predicted by the algorithm developed in the Zakian Lab (Capra et al., 2010). The parameters of this algorithm define a G4 motif as a sequence containing four tracks of three of more guanines that are no further than 25 nucleotides away from each other. To assess the validity of this algorithm, circular dichroism as well as acrylamide gels were used to show that the predicted G4 sequences of varying nucleotide composition and length ultimately formed G4 structures *in vitro* (Capra et al., 2010).

To map the PIF1 binding sites well as the G4 sites to other annotated genome features, the Gbrowser program on yeastgenome.org was used. For both data sets, the same protocol was followed. Custom tracks were made by uploading the data from their respective supplements into the Gbrowser. Minor modification of the data had to be made for it to appear in the browser. The chromosome names had to be switched from chr1 (numerical character) to chrI (Roman numeral). Then, the genome browser was manually curated to find the regions of PIF1 binding or predicted G4 DNA.

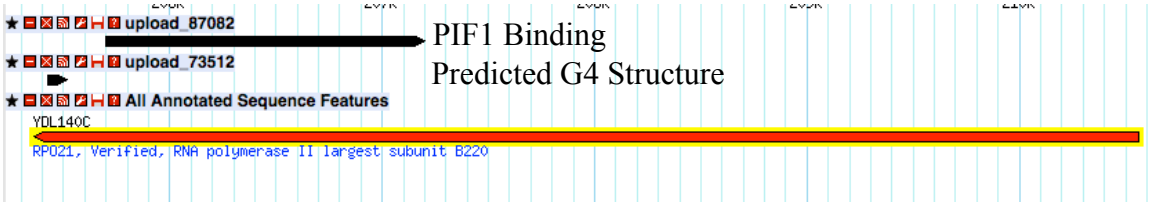


Figure 2: Mapping predicted G4 sites and PIF1 binding sites to the yeast genome. This example shows one of the rare instances of a spatial association of PIF1 binding with a predicted G4 structure in a coding minisatellite. In this example, the repeat is the CTD of RPO21. The 668 sites in the G4 genome as well as 1148 PIF1 binding sites were mapped by looking at the Gbrowser in a similar manner.

Implementing Excel to compare datasets

Using the auto-filter function in Excel, the overlap between datasets was found. This work looked for overlap between PIF1 binding sites, predicted and genes with ORFs with repeats. The G4 DNA data set was previously compared to the PIF1 binding peaks in the analysis of Paeschke (Paeschke et al., 2011). This study examined whether or not PIF1 was found bound to G4 sites *in vivo.* A comparison between the G4 sites that were found to have PIF1 binding and the repeats returned by XSTREAM with DNA correspondence was also conducted.

XSTREAM output as filter for G4 DNA

The multiple sequence alignment tool on yeastgenome.org is time consuming. Therefore, to ensure that the repeats of interest could form a G4 structure, the nucleotide output of XSTREAM was reanalyzed. If a track of three of more cytosines or guanines occurred in the nucleotide sequence of a repeat that existed in an ORF with predicted G4 DNA, the repeat was saved for future MSA. If there was no possibility of G4 formation in the repeat, the gene was no longer analyzed.

MSA on yeastgenome.org

The MSA program used can be found at the following link: http://www.yeastgenome.org/cgi-bin/FUNGI/alignment.pl. Alignments only were conducted on ORFs that contained a predicted G4, and a repeat that contained G or C tracks. The presence or absence of size polymorphisms was recorded with the relatively small subset of genes that contained both G4 DNA and a repeat. An overview of this workflow is seen below in Figure 3.

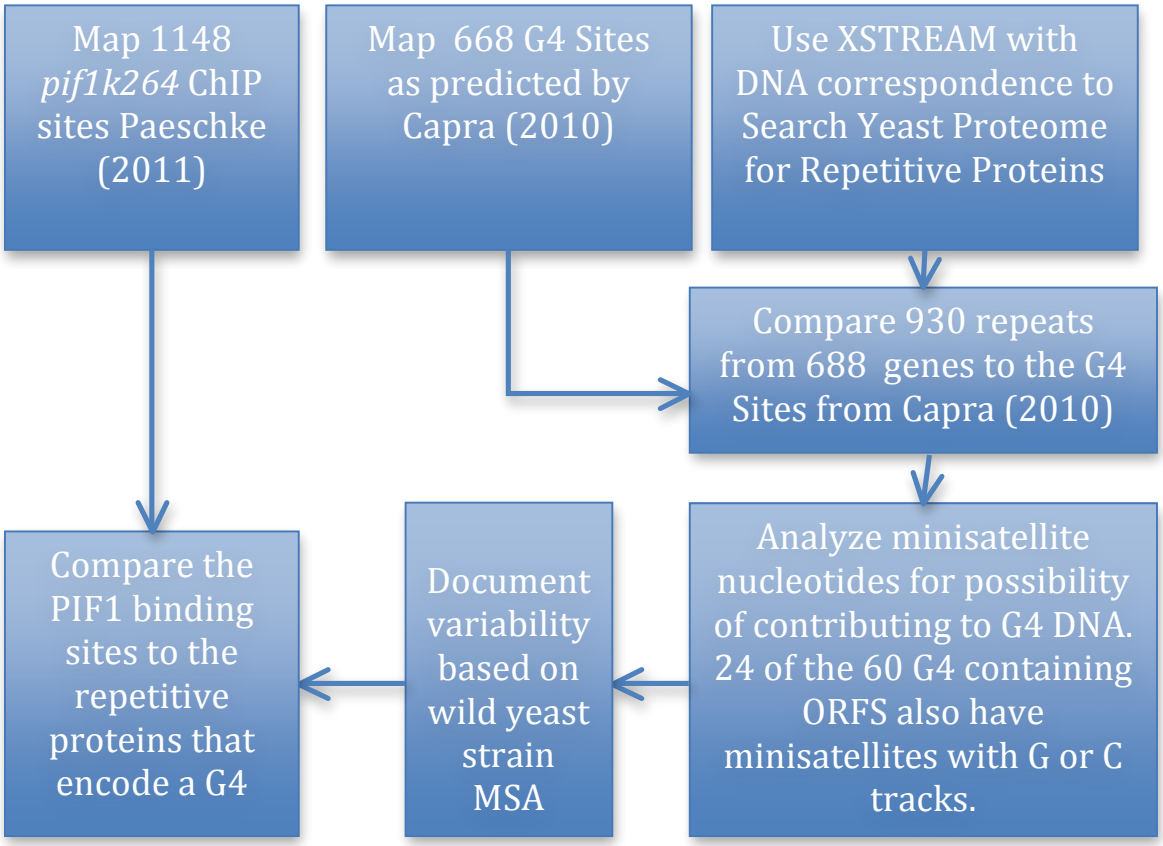


Figure 3: Flowchart used to find repeats that may behave similarly to RPO21. XSTREAM with DNA correspondence was used to find repetitive proteins in the yeast proteome. This data set was compared to the predicted G4 sites in the *S. cerevisiae*. Then, the XSTREAM nucleotide output of repeats with G4 sequence in the same ORF was analyzed for presence of poly-guanine or poly-cytosine tracks. If the repeat could potentially be a part of a larger G4 motif, it was analyzed for variability by examining for polymorphisms on the multiple strain alignment function at yeastgenome.org. This group of proteins was also curated for PIF1 binding[26].

Results:

9% of the proteome contains repeats larger than 2 amino acids

XSTREAM with DNA correspondence returned a total of 930 repeats from 688 genes with the settings described in the methods section. Unpublished data shows that when the filters used in this program become more stringent (i.e. more copies or less consensus error) the total number of repeats found decreases. Chromosome four, contained the most repetitive ORFs, which is not surprising given the fact that it is the largest chromosome (Figure 4). Pentapeptide repeats appeared the most often in this analysis (Figure 4). It appears that homopolymeric amino acid repeats represent a significant portion of repetitive coding elements. These repeats were ignored in this analysis. A general overview of the repetitive gene data set is seen in Figure 4.

In analyzing the XTSREAM with DNA correspondence data, the previously documented enrichment of large GC skew values in repeats became apparent (Richard and Dujon, 2006). The repeats did not have abnormally different GC content from the rest of the yeast genome. They did, however, possess incredibly different GC skews. When looking at a given gene, it is known that the GC skew starts negative and increases over the length of the gene (McLean and Tirosh, 2011). However, the repetitive regions appear to be very different in that they have dramatic localized skews. A perl script was written to calculate the GC skew for any given gene. When looking at the top 1 percent of GC skew hits from the entire genome (60 genes), 20% of these extreme values contain repeats, while only 9.3% of the genome was found to contain repeats as defined by this study (see methods for details).

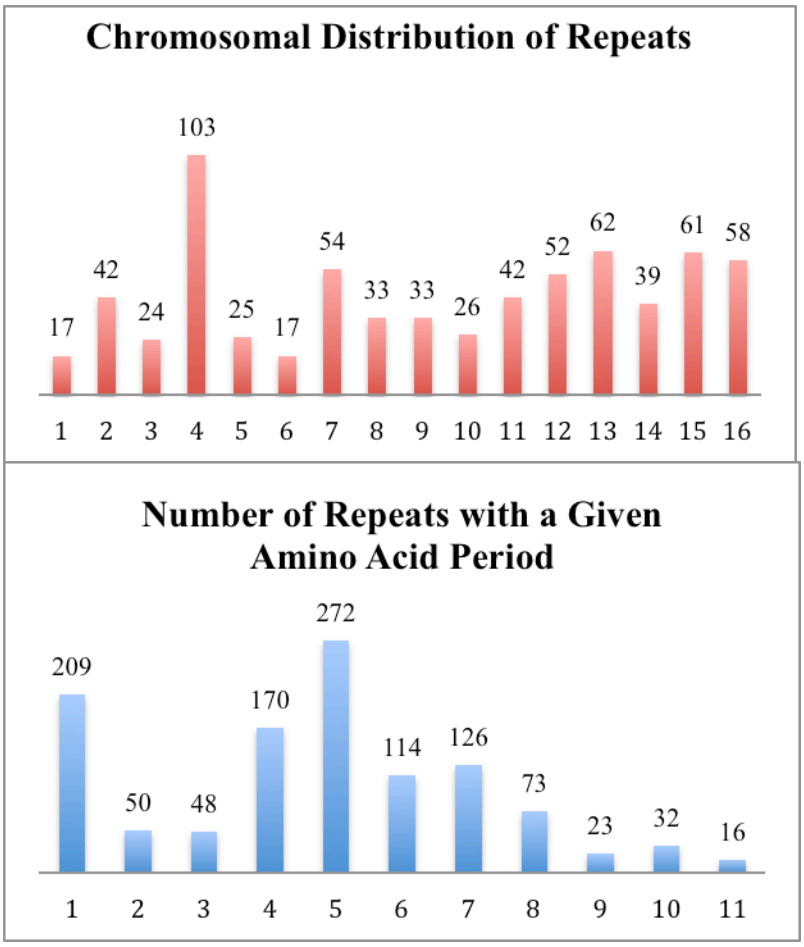


Figure 4: Overview of repeats found by XSTREAM with DNA. Chromosome IV contains the most repeats. XSTREAM with DNA correspondence returned many pentapeptide repeats as well homopolymeric repeats. However, the analysis conducted in this work excluded the single amino acid repeats because they were hypothesized to be variable through a mechanism different that RPO21.

This general analysis of GC skew may underestimate the intricacies of GC skew changes across an entire gene. Some genes mask the extreme localized skew of a repeat due to their large size. Therefore, a subset of repetitive genes have median GC skew values that hide their localized regions of very negative or very positive skews. One clear example of this is RPO21. The skew of the CTD is -0.68 (only one other gene has a value

higher than this) while the whole gene itself is 0.005. The median GC skew of genes from *S. cerevisiae* is 0.04, the mean is 0.05 with a standard deviation of 0.10.

Overlap of Repetitive Proteins with predicted G4 Sites and PIF1 binding locations

Out of the 688 predicted G4 sites, 60 (8.7%) G4 sites occurred in genes that had repeats as called by XSTREAM with DNA correspondence. There is the possibility that the G4 sequence occurs in a part of the gene, which does not encode the peptide repeat. Therefore, out of those 60 ORFs that had both a minisatellite and a predicted G4 structure, 24 genes (40%) had guanine or cytosine tracks in their respective minisatellites, that could allow for G4 formation. Of those 24 ORFs where it was possible for the DNA repeat to also encode G4, 12 (50%) contained predicted G4 structures in their repetitive regions. 14 (23%) of the original 60 genes were previously reported in the Paeschke study as genes where PIF1 binding occurred near or at predicted G4 structures (Paeschke et al., 2011). It is worth noting that 9 proteins found with G4 forming sequences contain multiple repeats. They were as follows: SPT5, NUP100, ENT2, ZDS1, BBC1, PAN1, WSC4, SGF73, and FIG2. Tables S3 and S4 (in the supplement) contain the information on the genes that resulted from the overlap of these three data sets.

Repetitive multiple strain alignment.

24 genes with 45 repeats in total were assessed for size variability. Single nucleotide polymorphisms (SNPs) that resulted in nonsynonymous mutations were not considered as variable in this analysis because SNPs did not appear to be a major source of variation in RPO21. A repeat was called variable when it was clear that the amino acid sequence of the repeat did not perfectly overlap in the MSA. An example of this

variability analysis is seen in Figure 5 where there are two sample MSAs from two different minisatellites. The repeat in Figure 5A, *SVL3,* was called variable due to the size polymorphism while *RFX1* in 5B was not called variable due to the fact that all strains contain a serine proline rich repeat with 10 amino acids. The variability of some repeats was unable to be assessed because of low quality sequence from yeastgenome.org. In genes with noticeably poor sequence, the MSA was unable to align stretches of hundreds of amino acids. This was a clear indicator that the original scaffold was assembled incorrectly and therefore the variability of that repeat could not be assessed. 6 of the 10 variable repeats found in this analysis also encoded G4 DNA. Those repeat were found in the following genes: SPT5, RPO21, BBC1, GAL11, DON1, SVL3. The repeats that were variable but did not contain G4 DNA were FIG2, WSC4, PAN1 (2 separate repeats). There were 5 additional repeats out of the total 45 examined that had G4 sequence but were not variable. Those genes were: RFX1, ATC1, IES1, GCR2, ENT2. PIF1 was found to bind to 4/6 variable proteins and 3/5 non-variable. Table 1 contains a summary of these proteins, which contain G4 sequence in their repeats.

Table 1: Summary of Genes that Contain Repeats that encode G4 DNA

Variable Repeats	Repeat	Copies	Total Size in AA Length	PIF1 Binding
SPT5 Repeat 2	GGASXW	3	18	Yes
Rpo21	YSPSTS	26	182	Yes
BBC1 Repeat 1	VPVPAAT	3	21	yes
GAL11	QA	31	62	No
DON1	KGKE	4.5	18	No
SVL3	QA	5	10	Yes
Non Variable repeats				
RFX1	SSPSP	2	10	Yes
ATC1	LSPXS	2	10	No
IES1	PKVTPXX	2	14	No

GCR2	NNG	3.67	11	Yes
ENT2	Poly Q repeat	N/a	21	No

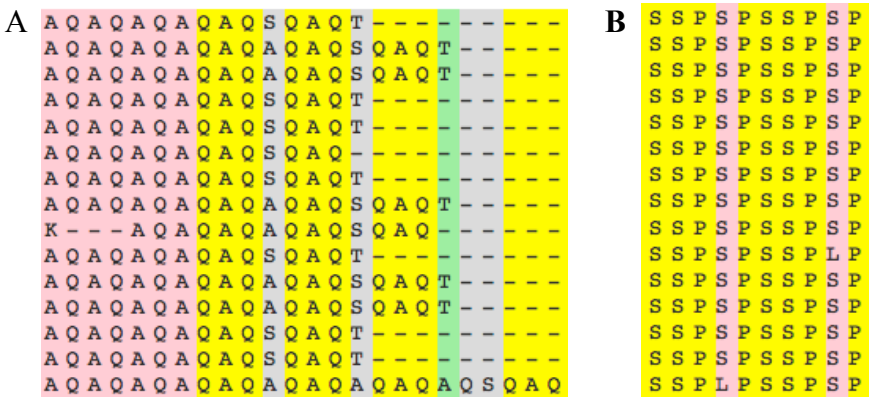


Figure 5: Using the multiple strain alignment tool to call repeat variability. **A)** MSA of SVL3. The AQ repeat from this gene clearly contains copy number variants. **B)** MSA of RFX1. The SP rich repeat contains a few instances of point mutations, but no size polymorphisms. Therefore it was classified as a non-variable repeat. The colors originate from the original analysis conducted by the yeastgenome.org alignment tool. The colors indicate the degree of sequence identity between strains.

Discussion

XSTREAM with DNA correspondence as a future tool to evaluate minisatellite sequence

Based on data from Chapter 1 and an extensive literature search, it was hypothesized that repeats containing G4 DNA that were also found to be binding sites for PIF1 would be variable, similar to the CTD of RPO21. This hypothesis that PIF1 and G4 DNA contribute to certain minisatellite variability is not ruled out from this study. Many aspects of this analysis must implement more rigorous filters to uncover the answers to this convoluted question. XSTREAM with DNA correspondence is only the beginning of manipulating the DNA sequence of larger repetitive minisatellites. With this tool, it illustrated the widespread nature of the GC skew imbalance in repetitive proteins that had been previously reported (Richard and Dujon, 2006). With more precise computational

methods, the relationship between these GC skewed repeats and variability can be better understood. A new project with a computer science student is currently focused on implementing a machine learning approach to understanding this irregular GC skew seen in these repeats. As McLean and Tirosh, observed, there is a GC skew signature across individual ORFs in all yeast genomes (McLean and Tirosh, 2011). However, as our lab and Richard and Dujon have previously documented, a minisatellite differs significantly from the rest of it own gene and the genome as a whole (Richard and Dujon, 2006). For example, the minisatellite in the middle of BSC1 results in a portion of the GC skew plot where the slope becomes dramatically negative. Given that GC skew increases across most genes in the yeast genome, BSC1 has a very different signature from other ORFs (McLean and Tirosh, 2011). Using this machine learning technique where the program develops its own canonical signature for ORFs without repeats, the new tool can easily find repetitive ORFs based on their deviation from the trend described by McLean and Tirosh (McLean and Tirosh, 2011). Not only can we implement this to find minisatellites in the yeast genome, it may also be possible to compare the diverse GC skew signatures of minisatellites to explain why some of the repeats demonstrate significant instability while others appear to rarely expand.

Additionally, the parameters of XSTREAM with DNA correspondence may not have been stringent enough, creating a list of too many target tandem repetitive proteins. A quick example of the sensitivity of XSTREAM with DNA correspondence is that by increasing the copy number from two to three, 2/3 of the hits are lost. Indeed, there are tandem repeats that only occur twice, but a ten amino acid stretch can also be repetitive by chance, or due to greater domain architecture. Also, the enrichment of pentapeptides

may indicate that the threshold length can bias the proteins returned by the program. This high representation of 5mer amino acid sequences may be an artifact of the original parameters because these relatively short regions would only have to repeat twice to be classified as repeats by XSTREAM with DNA correspondence. No analysis was conducted to evaluate whether changing the minimum repeat length would cause a biasing of another repeat size. Additionally, only the ORFs from S288c were used in the XSTREAM with DNA correspondence pipeline. S288c is the standard laboratory strain of budding yeast which is known to have less minisatellite variation than other wild strains seen at yeastgenome.org (Verstrepen et al., 2005). Therefore, some repeats may have fit the parameters when they were quarried on the wild strains, but because the program was run on the lab strain, these repeats were excluded.

This analysis resulted in many fewer repetitive ORFs than reported by the Gemeyal paper. This is due to the exclusion of homopolymeric amino acid repeats. When the extra 209 repeats are added to the original analysis, 13.6% of the proteome appears repetitive, while Gemayel reported 12.7% (Gemayel et al., 2010). These repeats were not analyzed because they have been extensively studied as trinucleotide repeats. The literature is extensive on different factors that contribute to trinucleotide variation, the impact of secondary structures on instability and fragility, and the many pathways involved in events which result in change of copy number in single amino acid repeats. One flaw with the less stringent XSTREAM with DNA correspondence search is that some of the trinucleotide repeats appear as larger minisatellites. An example of this phenomenon is LAS17 where it appears that this ORF contains a homopolymeric proline

repeat, however due to the insertion of alanines dispersed through this repeat it is listed as a repeat with a period of ten.

PIF1 may only be involved in negatively regulating instability of a small subset of ORFs

Between G4 structure and PIF1 binding, it appears that the conserved G4 helicase may have little impact on repeat variability. PIF1 binding appeared randomly distributed in the 11 variable and non-variable repeats that encoded G4 sequence. PIF1 may appear to have little or no effect due to limitation of ChIP as a high-resolution technique to demonstrate PIF1 binding. Therefore, using this data from Paeschke 2011 may not fully characterize PIF1 binding sites (Paeschke et al., 2013). It would be impossible to expect 100% occupancy however, PIF1 was only found at 25% of the predicted G4 structures. This deviation from high G4 occupancy could be a result of the fact that many predicted G4s do not actually form in physiological contexts. The original ChIP study was validated by ChIP/qPCR, which showed that G4 sequences that did not have PIF1 ChIP signal, also did not have qPCR signal. With the data put forth by this paper, it is reasonable to assume that the ChIP data provides an estimate of PIF1 occupancy across the yeast genome, but it may be imperfect.

With the data from the Nicolas Lab, it can be hypothesized that PIF1 may only contribute to the instability of a subset of repeats. As mentioned in the introduction, four yeast minisatellites, which all contain large GC skew values (*Num1*= -.51, *Hkr1*= -.289, *Flo1*=-.33, *Dan4*=-.89), did not show instability in a *pif1Δ* background (Ribeyre et al., 2009). This result conveys that PIF1 regulates the size of a subset of minisatellites that share more than just an enrichment of G or C rich sequence. PIF1's role may be more specific and is driven by its function in unwinding different DNA topologies.

The impact of G4 DNA on repeat variability

The presence of G4 DNA in repeats presents a possible mechanism for the observed size polymorphism in a subset of repetitive proteins. G4 DNA overlapped with 6 of the 10 variable repeats found in this study. The observed polymorphisms appear to have similarities with the G4 containing CEB1 human minisatellite from the work in the Nicolas Lab (Lopes et al., 2002; Lopes et al., 2011; Lopes et al., 2006; Piazza et al., 2010; Ribeyre et al., 2009). The work from this lab would suggest that the orientation of the G4 structure in respect to the origin of replication may be the reason why some of these minisatellites demonstrate CNVs while the other four repeats appear to not change in length (Lopes et al., 2011). Lopes found that when the G4 structure occurs on the leading strand template or the newly formed nascent lagging strand, there is an increase in instability (Lopes et al., 2011). In the case of RPO21, it is not difficult to assign G4 strand location, because the origin is 1.5 kb upstream of the gene. However, a gene like SPT5 is situated more than 50 kb away from its flanking origins of replication. In yeast, origins may fire at different times during S phase, continuing to complicate whether or not the G4 DNA is in the unstable orientation. Therefore, future work may want to uncover the relationship of the eleven previously listed genes with origins of replication to understand the implications of replication orientation on instability of G4 containing minisatellites.

A key future direction for this project is to use long read sequencing to understand the variability of the repeats deemed polymorphic by the MSA on yeastgenome.org. As

seen by the sequencing of the RPO21 locus, more variation may exist in the population of wild yeast than is currently accessible in online databases. Given the NCYC wild yeast library, the Fuchs Lab has the opportunity to examine the true variability of repeats that have sequence characteristics like RPO21. The short reads that are currently used for large scale yeast sequencing are successful at getting high coverage across the genome (Bergstrom et al., 2014; Serero et al., 2014; Stirling et al., 2014). However, with these short reads, it is difficult to properly classify repeat size. It is possible that many repeat sequences span multiple short reads. Therefore, when assembling the reads into contigs and scaffolds, it is very easy to call a duplication of sequence a limitation of the sequencing technology and not an additional repeat. As the next generation sequencing techniques continue to increase their read length and as single molecule DNA sequencing (ie Oxford nanopore) become more widely available, this will no longer be an issue. Implementing long read Sanger sequencing in the near future will help classify copy number variants as well as SNPs that occur in the aforementioned polymorphic repeats. Applying the same strategy developed in sequencing the CTD of RPO21 to sequence GAL11 and SPT5, may allow for the true characterization of the variability of these other key transcriptional genes.

This chapter laid the groundwork of finding a new set of minisatellites that share a G4 motif. A large-scale screen recently developed by Alver examined minisatellite instability during stationary phase. This work found when using different minisatellite sequences, dissimilar proteins were responsible for causing the observed instability (Alver et al., 2013). Therefore it may have been ambitious to think that XSTREAM with DNA correspondence would elucidate a key sequence factor that contributes to the

widespread minisatellite instability. However, with this focused approach laid out in the this chapter, it now may be possible to see how G4 structures impact instability of a subset of minisatellites.

Prospectus

Coding minisatellites have unique implications for protein biology. As seen below, in Figure P1, expansions and contractions of repetitive domains lead to differential protein size and differential patterns of possible PTMs. This variety in length and structure impacts a domain's ability to partake in protein-protein interactions, which ultimately affects many aspects of cell physiology. This unique hypervariable characteristic of coding minisatellites explains why they appear in almost all sequenced organisms and across all three kingdoms (Jernigan and Bordenstein, 2015). Verstrepen and colleagues demonstrated the physiological consequences of differential minisatellite and how minisatellite can impact evolutionary success (Smukalla et al., 2008; Verstrepen et al., 2005). Therefore, it seems that understanding the mechanisms of minisatellite instability may allow us to comprehend some of the phenotypic and biological variation that occurs in budding yeast and across the tree of life.

In hindsight, it seems naïve that with the development of XSTREAM with DNA correspondence, the mechanism of minisatellite expansion would appear by looking at DNA sequence alone. Each minisatellite contains its own unique characteristics, not just in nucleotide composition, but also in DNA topology. Examining sequence in isolation ignores the consequences of DNA secondary structure. In the one instance investigated in this work, it does appear that a DNA secondary structure, G-quadruplex, may ultimately impact coding minisatellite variability. However this does not appear to be a widespread trend because G4 DNA doesn't appear in many repetitive genes.

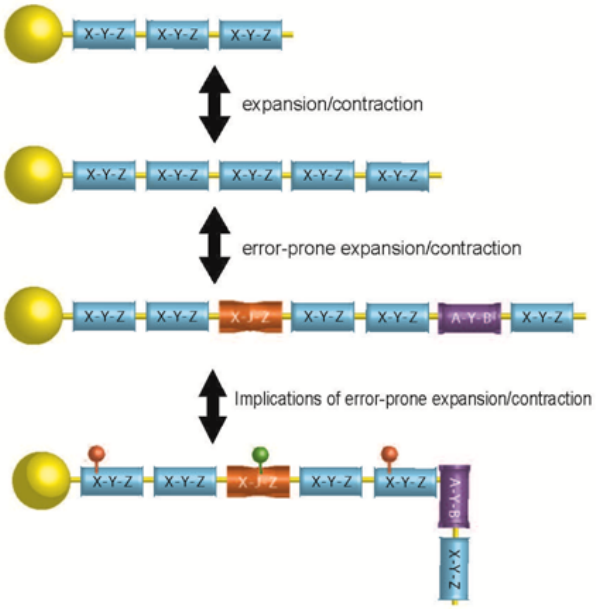


Figure P1: With expansion and contraction of minisatellites, new protein-protein interactions are possible through addition of new repeats. Additionally, mutations to these repeating units can allow for new PTMs. Combinations of differential repeat lengths and PTMs contribute to the possible diversity in these unstable and variable regions. This figure was modified from Fuchs (Fuchs, 2013).

It is interesting that G4 DNA exists in the repeats of three genes that work together to execute transcription properly. RPO21, as discussed earlier, is the enzyme responsible for transcribing all cellular mRNA, SPT5 is a key component of the splicing complex, and GAL11 plays an important role in the mediator complex to recruit transcriptional activators and repressors. The activity of both the mediator complex and the splicing complex relies on specific PTMs to the CTD of RPO21. Therefore, this may present a situation in which differences in copy number in all three genes could lead to variables phenotypes through differential transcriptional output.

In my work, it was found that two essential genes (RPO21 and SPT5) contain a G4 sequence, a DNA secondary structure that is associated with fragility as described by Paeschke and Sabouri (Paeschke et al., 2011; Sabouri et al., 2014). It is perplexing that

such necessary proteins would encode an inherently unstable element. If this region were to undergo a mutational event that led to an out of frame insertion or addition of non-consensus repeats, it would most likely lead to a decrease in organismal fitness. Looking at the location of G4 sequences across genomes illustrates that this trend, of G4 DNA forming in crucial regions of the genome, is not restricted to fungi.

G4 DNA has been found to be important in disease contexts. Surprisingly, this unstable element is found in many oncogene promoters like K-Ras and c-myc (Yang and Okamoto, 2010). In these locations, the DNA secondary structure acts like a transcriptional repressor. During oncogenesis, an upregulation of G4 unwinding proteins and/or an accumulation of mutations to the G4 sequence results in uncontrolled expression of the oncogenes downstream of the promoter (Yang and Okamoto, 2010). Additionally, a G4 sequence has been found in the coding region of the essential HIV protein Nef. A recent paper used G4 stabilizing ligands to inhibit HIV replication, showing that understanding the implications of G4 structures may present new therapeutic targets (Perrone et al., 2013). These two examples demonstrate the importance of understanding the biology surrounding G4 DNA because it may allow for an alleviation of disease burden.

Using minisatellite expansion and its relation to G4 DNA provides an opportunity to gain insight into the impact of this secondary structure on biological life. G4 DNA has a propensity to be associated with double strand break, however it also seems to control the mutations that occur in the most C-terminal region of the CTD repeat in RNA polymerase II. With more investigation into the inherent variability surrounding other genes with similar characteristics to RPO21 like SPT5 and GAL11, the roles of G4 DNA

on biological systems can be understood and applied to the many situations where this DNA secondary structure arises.

Supplement:

Construction of GRY3019+ pif1-m2

The protocol to create a *pif1-m2* mutation in GRY3019 followed the two-step gene replacement, illustrated in Figure 1 of the methods section of Chapter 1. The pVS31 plasmid was transformed into YSF1008 correctly, as seen in Figure S1A. By PCR amplifying the PIF1 locus and then digesting with Xho1, the presence of the mutant *pif1-m2* allele can be visualized. Therefore, the gel in Figure S1A indicates a successful transformation due to the presence of two bands. There are two bands because this strain is heterozygous at the PIF1 locus, containing both the mutant copy and the wild type copy. To select for a strain that only contains the *pif1-m2* allele, C4 was plated onto 5-FOA. The resulting colonies were screened in the same manner. Figure S1B indicates that C4 contained the wild type allele, while C5 possessed only the mutant allele. The undigested PCR product from C5 was purified and sequenced by Eton Biosciences. Figure S1C illustrates that this strain contained only the mutant copy, due to the presence of the Xho1 site (CTCGAG) with no double peak. Colony five was kept and used for mating.

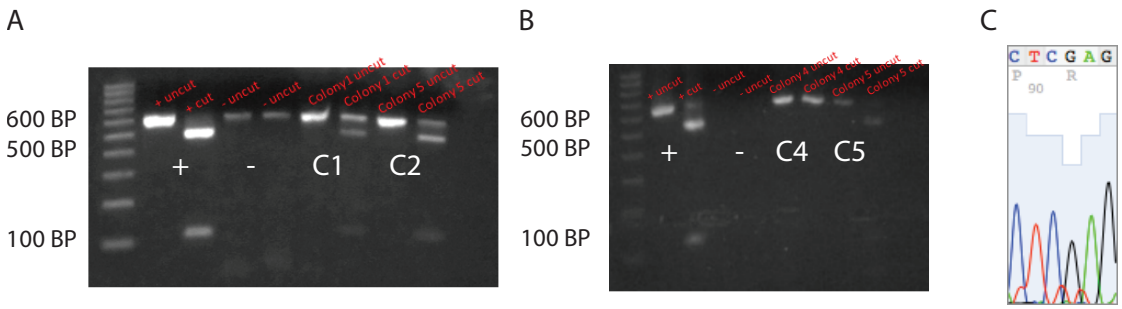


Figure S1: Construction of *pif1-m2* mutant **A)** Agarose gel depicting correct transformation due to the presence of both PIF1 WT and mutant alleles. **B)** Gel illustrating how 5-FOA selection of yeast leads to cells with only one PIF1 allele. In the 8th lane, only one shorter band is visible post Xho1 digestion, demonstrating loss of WT allele. This PCR product was sequenced. **C)** Correct sequencing traces indicate the presence of the Xho1 restriction site as well as no double peak. Therefore, this colony contains only the mutant *pif1-m2* allele and is not a heterozygote at the PIF1 locus.

Mating pif1-m2 allele into GRY3019

Colony five (Mat*, URA$^-$, G418$^-$, *pif1-m2*) was mated with GRY3019 (Mat A URA::CMV tTa, kanRPtetO7-TATA-RPB1, PIF1). 20 spores were dissected and plated on SC-Ura, YPD+G418 and plates that selected for alpha and A colonies. Tetrads that demonstrated 2:2 segregation of alleles were examined further for *pif1-m2* mutations. This mating resulted in tetrads that did not have four surviving spores due to the necessary co-segregation of Ura$^+$ Tet-transactivator and G418^r Tet-operator. If the Ura$^+$ transactivator segregates without the modified promoter, the cells survive because the wild type RPO21 promoter still functions properly. However, if the G418^r promoter segregates without the transactivator, there is no transactivator binding, and ultimately the expression of RNA polymerase II is halted. This is lethal due to the fact that global transcription will be stopped. An analysis of four tetrads appears in Figure S2A. Red boxes show spores that did not grow due to presence of the operator without the transactivator. The genotype of dead spores can be found based on the genotype of the

other three spores from the tetrad. Green boxes represent spores that contained the G418^{r} operator and Ura^{+} transactivator, two elements necessary for proper strain construction and viability. Those cells that had this phenotype were screened for *pif1-m2* mutations using the same PCR screening protocol seen in Figure S1A and S1B. Four of the seven spores screened in Figure S2B demonstrated the appropriate Xho1 banding pattern, indicating the presence of the *pif1-m2* allele. Spore 13D was saved and used in future work. This strain was renamed YBIR7. The genotype at the PIF1 locus for cells that contained zero elements of interest or contained only the Ura^{+} transactivator was not found because those spores would not be used in future work.

A

	Ura	G	Alpha	*	Genotype	Pif1
1a	-	+	+	-	A uG	
1b	+	-	-	+	* Ug	
1c	-	-	-	+	* ug	
1d	+	+	+	-	A UG	mutant
9a	+	-	+	-	A ug	
9b	-	-	-	+	* ug	
9c	-	+	-	+	* uG	
9d	+	+	+	-	A UG	mutant
12a	-	-	-	+	* ug	
12b	+	+	-	+	* UG	WT
12c	-	-	+	-	A ug	
12d	+	+	+	-	A UG	mutant
13a	-	+	+	-	A uG	
13b	+	+	+	-	A UG	mutant
13c	-	-	-	+	* ug	
13d	+	-	-	+	* Ug	

B

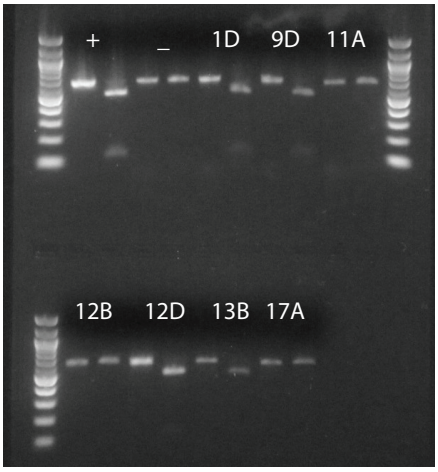


Figure S2: Mating of Colony Five with GRY3019. 20 tetrads were dissected. **A)** Examination of four of the dissected tetrads, which showed proper 2:2 gene segregation. Spores that contain both Ura^{+} transactivator and G418^{r} operator were screened for *pif1-m2* mutations. Green boxes demonstrate those colonies that had appropriate markers. Red boxes represent colonies that died due to segregation of the operator without the transactivator. A statement could be made on genotype of the dead cells genotype based on the other spores in the tetrad. The presence of *pif1-m2* mutation was screened by PCR followed by digestion with Xho1. **B)** Gel showing that DNA isolated from spores 1D, 9D, 12D, 13B contain the mutant *pif1-m2* allele while the other 3 examined spores (11A, 12B, 17A) have the PIF1 allele. 13D was used in future work and renamed YBIR7.

Table S1: Plasmids used in this study

Plasmid Name	Selection	Description
pCTD8	LEU2+	RPO21 with only 8 CTD repeats under promoter of RPO21
pLeu+	LEU2+	Empty vector used as a negative control in spotting assays
pRPO21	LEU2+	WT copy of RPO21 on a plasmid under normal RPO21promoter
pVS31	URA3+	Contains *pif1-m2*. Used in two step gene transfer to make specific nuclear *pif1* mutant

Table S2: Yeast strains used in this study

Strain Name	Important Genotype	Description
YSF1008	URA3Δ	Used to transform pVS31 with *pif1-m2*
YBIR5	YSF1008 with integrated pVS31	Contains properly transformed pVS31. No longer *ura3Δ*. Heterozygous at PIF1 locus
YBIR6	YSF1008+ *pif1-m2*	YBIR5 after counter selection on 5-FOA. No longer $URA3^{+}$. Only contains *pif1-m2*. Mated with GRY3019 to make YBIR7
YBIR7	GRY3019+*pif1-m2*	Used in fluctuation assay as well as in spotting assays. Has tTa dependent expression of RPO21 in a *pif1-m2* background
GRY3019	URA::CMV tTa, kanRPtetO7-TATA-RPB1,	Used in fluctuation assay as well as in spotting assays. Has tTa dependent expression of RPO21

Green boxes in tables S3, and S4 demonstrate that G4 sequence is related to variability, red boxes in these tables demonstrate that the variability occurs but there is no G4 sequence in the repeat.

Table S3: Summarizing overlap between XSTREAM with DNA correspondence with ORFs that only contain 1 repeat with predicted G4 from Capra 2010, and PIF1 binding (Capra et al., 2010; Paeschke et al., 2011)

Systematic name	Common Name	Repeat/G4 overlap	Repeat/Pif1 overlap	Variable
YDR006C	SOK1	N	N	N
YDR273	DON1	Y	N	Y
YEL017	GTT3	N	N	N
YHR042W	NCP1	N	N	N
YHR102W	KIC1	N	N	N
YKL038C	RGT1	N	N	N
YOL051W	GAL11	Y	N	Y
YOR140W	SFL1	N	N	N
YOR181W	LAS17	Y	Y	N
YPL032C	SVL3	Y	Y	Y
YLR176C	RFX1	Y	Y	N
YDR184C	ATC1	Y	N	N
YFL013C	IES1	Y	N	N
YNL199C	GCR2	Y	Y	N
YDL140C	RPO21	Y	Y	Y

Table S4: Summarizing overlap between XSTREAM with DNA correspondence with ORFs that only contain multiple repeat with predicted G4 from Capra 2010, and PIF1 binding (Capra et al., 2010; Paeschke et al., 2011)

Gene Common name	Repeat Number	Repeat/G4 Overlap	Repeat/Pif1 Overlap	Variable
Fig2	1	N	N	N
	2	N	Y	N/A-Bad Sequence
	3	CLOSE to repeat	N	Y
	4	N	Y	N
SGF73	1	N	N	N
	2	N	N	N
	3	CLOSE to repeat	Y	N
WSC4	1	N	N	N
	2	N	Y	N
	3	N	Y	Y
	4	N	Y	N
	5	CLOSE to repeat	Y	N
Pan1	1	N	N	Y
	2	N	N	N
	3	CLOSE to repeat	CLOSE	Y
	4	CLOSE to repeat	CLOSE	N
BBC1	1	YES	CLOSE	Y
	2a	CLOSE to repeat	Y	N/A- Bad Sequence
	2b	CLOSE to reapt	Y	N/A- Bad sequence
ZDS1	1	N	N	N
	2	N	N	N
ENT2	1	Y	Y	N
	2	N	N	Poly q
	3	N	N	Poly q
Nup100	1	N	N	N
	2	N	N	N
	3	N	N	N
	4	close	N	N
Spt5	1	N	N	N
	2	YES	Y	Y

References:

Alver, B., Jauert, P.A., Brosnan, L., O'Hehir, M., Vandersluis, B., Myers, C.L., and Kirkpatrick, D.T. (2013). A Whole Genome Screen for Minisatellite Stability Genes in Stationary Phase Yeast Cells. G3.

Bergstrom, A., Simpson, J.T., Salinas, F., Barre, B., Parts, L., Zia, A., Nguyen Ba, A.N., Moses, A.M., Louis, E.J., Mustonen, V.*, et al.* (2014). A high-definition view of functional genetic variation from natural yeast genomes. Molecular biology and evolution *31*, 872-888.

Bochman, M.L., Judge, C.P., and Zakian, V.A. (2011). The Pif1 family in prokaryotes: what are our helicases doing in your bacteria? Molecular biology of the cell *22*, 1955-1959.

Boule, J.B., and Zakian, V.A. (2007). The yeast Pif1p DNA helicase preferentially unwinds RNA DNA substrates. Nucleic acids research *35*, 5809-5818.

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 computational biology *6*, e1000861.

Chapman, R.D., Heidemann, M., Hintermair, C., and Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. Trends in genetics : TIG *24*, 289-296.

Chung, W.H. (2014). To peep into Pif1 helicase: multifaceted all the way from genome stability to repair-associated DNA synthesis. Journal of microbiology *52*, 89-98.

David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes & development *25*, 972-983.

Davis, L., and Maizels, N. (2011). G4 DNA: at risk in the genome. The EMBO journal *30*, 3878-3879.

Egloff, S., and Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends in genetics : TIG *24*, 280-288.

Fry, M., and Loeb, L.A. (1994). The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. Proceedings of the National Academy of Sciences of the United States of America *91*, 4950-4954.

Fuchs, S.M. (2013). Chemically modified tandem repeats in proteins: natural combinatorial peptide libraries. ACS chemical biology *8*, 275-282.

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. The Journal of biological chemistry *287*, 3249-3256.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual review of genetics *44*, 445-477.

Gietz, R.D., and Schiestl, R.H. (2007a). Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nature protocols *2*, 38-41.

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

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.
Jernigan, K.K., and Bordenstein, S.R. (2015). Tandem-repeat protein domains across the tree of life. PeerJ *3*, e732.
Li, J.R., Yu, T.Y., Chien, I.C., Lu, C.Y., Lin, J.J., and Li, H.W. (2014). Pif1 regulates telomere length by preferentially removing telomerase from long telomere ends. Nucleic acids research *42*, 8527-8536.
Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Molecular cell *9*, 1101-1111.
Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular biology and evolution *13*, 660-665.
Lopes, J., Debrauwere, H., Buard, J., and Nicolas, A. (2002). Instability of the human minisatellite CEB1 in rad27Delta and dna2-1 replication-deficient yeast cells. The EMBO journal *21*, 3201-3211.
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. The EMBO journal *30*, 4033-4046.
Lopes, J., Ribeyre, C., and Nicolas, A. (2006). Complex minisatellite rearrangements generated in the total or partial absence of Rad27/hFEN1 activity occur in a single generation and are Rad51 and Rad52 dependent. Molecular and cellular biology *26*, 6675-6689.
Maleki, S., Cederberg, H., and Rannug, U. (2002). The human minisatellites MS1, MS32, MS205 and CEB1 integrated into the yeast genome exhibit different degrees of mitotic instability but are all stabilised by RAD27. Current genetics *41*, 333-341.
McLean, M.A., and Tirosh, I. (2011). Opposite GC skews at the 5' and 3' ends of genes in unicellular fungi. BMC genomics *12*, 638.
Newman, A.M., and Cooper, J.B. (2007). XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC bioinformatics *8*, 382.
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.
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.
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.
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.
Perrone, R., Nadai, M., Poe, J.A., Frasson, I., Palumbo, M., Palu, G., Smithgall, T.E., and Richter, S.N. (2013). Formation of a unique cluster of G-quadruplex structures in the HIV-1 Nef coding region: implications for antiviral activity. PloS one *8*, e73121.

Piazza, A., Boule, J.B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M.P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. Nucleic acids research *38*, 4337-4348.
Pike, J.E., Henry, R.A., Burgers, P.M., Campbell, J.L., and Bambara, R.A. (2010). An alternative pathway for Okazaki fragment processing: resolution of fold-back flaps by Pif1 helicase. The Journal of biological chemistry *285*, 41712-41723.
Ranuncolo, S.M., Ghosh, S., Hanover, J.A., Hart, G.W., and Lewis, B.A. (2012). Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo. The Journal of biological chemistry *287*, 23549-23561.
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 genetics *5*, e1000475.
Richard, G.F., and Dujon, B. (2006). Molecular evolution of minisatellites in hemiascomycetous yeasts. Molecular biology and evolution *23*, 189-202.
Rolland, T., Dujon, B., and Richard, G.F. (2010). Dynamic evolution of megasatellites in yeasts. Nucleic acids research *38*, 4731-4739.
Sabouri, N., Capra, J.A., and Zakian, V.A. (2014). The essential Schizosaccharomyces pombe Pfh1 DNA helicase promotes fork movement past G-quadruplex motifs to prevent DNA damage. BMC biology *12*, 101.
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.
Schwer, B., and Shuman, S. (2011). Deciphering the RNA polymerase II CTD code in fission yeast. Molecular cell *43*, 311-318.
Serero, A., Jubin, C., Loeillet, S., Legoix-Ne, P., and Nicolas, A.G. (2014). Mutational landscape of yeast mutator strains. Proceedings of the National Academy of Sciences of the United States of America *111*, 1897-1902.
Shah, K.A., McGinty, R.J., Egorova, V.I., and Mirkin, S.M. (2014). Coupling transcriptional state to large-scale repeat expansions in yeast. Cell reports *9*, 1594-1602.
Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Molecular cell *35*, 82-92.
Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., Vinces, M.D., Jansen, A., Prevost, M.C., Latge, J.P*., et al.* (2008). FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. Cell *135*, 726-737.
Stiller, J.W., and Cook, M.S. (2004). Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. Eukaryotic cell *3*, 735-740.
Stirling, P.C., Shen, Y., Corbett, R., Jones, S.J., and Hieter, P. (2014). Genome destabilizing mutator alleles drive specific mutational trajectories in Saccharomyces cerevisiae. Genetics *196*, 403-412.
Stump, A.D., and Ostrozhynska, K. (2013). Selective constraint and the evolution of the RNA polymerase II C-Terminal Domain. Transcription *4*, 77-86.
Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nature genetics *37*, 986-990.
Voineagu, I., Freudenreich, C.H., and Mirkin, S.M. (2009). Checkpoint responses to unusual structures formed by DNA repeats. Molecular carcinogenesis *48*, 309-318.

Weitzmann, M.N., Woodford, K.J., and Usdin, K. (1997). DNA secondary structures and the evolution of hypervariable tandem arrays. The Journal of biological chemistry *272*, 9517-9523.
Werner-Allen, J.W., Lee, C.J., Liu, P., Nicely, N.I., Wang, S., Greenleaf, A.L., and Zhou, P. (2011). cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. The Journal of biological chemistry *286*, 5717-5726.
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.
Wilson, M.A., Kwon, Y., Xu, Y., Chung, W.H., Chi, P., Niu, H., Mayle, R., Chen, X., Malkova, A., Sung, P.*, et al.* (2013). Pif1 helicase and Poldelta promote recombination-coupled DNA synthesis via bubble migration. Nature *502*, 393-396.
Yang, D., and Okamoto, K. (2010). Structural insights into G-quadruplexes: towards new anticancer drugs. Future medicinal chemistry *2*, 619-646.

Addendum

Gene name	Consensus	Period	NT	Variable
PAN1-A	QPTQPV	7	126	yes
Pan1-B	PQTTGMM	3.5	73.5	yes
Pan1C	GLQSQLT	1.5	31.5	no
Pan1D	AGIP	2	24	no
PTK1	APS	3	27	yes
FIG2	needs to be resequenced multiple repeats			
SOK1	NPLSL	2	30	no
SGF73	needs to be resequenced/ repeat are poly N and D			
NCP1	VALGL	2	30	no
KIC1	NNSGP	2	30	no
ZDS1A	PVQASA	4	72	no
ZDS1B	SSSp	3	36	no
SFL1	AP	5	30	no
LAS17	PAPPPPP	2	42	no
WSC4A	ST	7	42	yes
WSC4B	SSSST	2	30	no
WSC4C	YQSKY	2	30	no
WSC4D	NSNTT	2	30	no
WSC4E	STSTTP	2	36	no
Nup100A	GLFGQNN	3	63	no
Nup100B	GSLFG	2	30	no
NUP100C	SSNQG	2	30	no
NUP100D	GSNLSF	2	36	no

Analysis of additional variable repeats. It is clear that many of the small tandem repeats found by XTSREAM are not variable. There is not one instance where the minimum repeat (length of five and period of two) results in a variable repeat. It appears that repeats that have larger periods are more likely to be variable. This is evident by the fact that all the repeats that are variable have periods greater than three. This analysis indicates that it may be important to throw out those small repeats and to further examine the relationship between repeat period and variability.