Variation and specialization of repeats in budding yeast

A dissertation

submitted by

Michael Babokhov

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Biology

Tufts University

August 2018

Advisor: Dr. Stephen M. Fuchs

Abstract

Tandem repeats are unique DNA sequences that consist of repeating motifs arranged end-to-end in the genome. When present in open reading frames, tandem repeats can significantly impact protein function through several unique mechanisms. Although previously dismissed as selfish or junk DNA, tandem repeats are now well-appreciated for their role in gene and protein regulation. However, studies up until now have mostly focused on only a subset of tandem repeats and a more complete understanding of repeats requires further detailed examination of their structure and function. The work presented in this thesis leverages the power of yeast genetics to address two fundamental questions in the field of repeat biology: 1) what is the extent and purpose of repeat variation? 2) what is the extent of specialization versus redundancy in tandem repeats? To address the first question, high quality genomic sequences of almost 100 strains and species of budding yeast were examined to look for instances of repeat variation. Tandem repeats were found to be variable, highly conserved and correlated with regions of intrinsic disorder. To answer the second question, a repeat in the largest subunit of RNA polymerase II was investigated in depth to look for instances of repeat specialization. Certain repeats demonstrated specific functions and these functions were tied to genetic and physical interactions with key co-transcriptional factors. Taken together, these results characterize surprising new aspects of tandem repeat function and lay the foundation for further studies into repeat variation and specialization. Future perspectives are also discussed, highlighting promising new directions in the study of tandem repeats and their impact on protein structure and function.

Acknowledgements

First and foremost my thanks go to my wonderful parents and my big brother. Their constant love and encouragement supported me at every step of my PhD studies.

Thank you to my committee chair and advisor, Dr. Stephen M. Fuchs for his guidance on the project and allowing me free reign to pursue research fellowships in Japan.

Thank you to my committee members Dr. Susan Ernst, Dr. Juliet Fuhrman and Dr. Kelly McLaughlin for their guidance and advice over the years.

Special thanks to Dr. Jason Kuehner for agreeing to sit on my committee as an outside examiner for this thesis.

Thank you to coauthors Bradley I. Reinfeld, Kevin Hackbarth, Yotam Bentov, Mohammad M. Mosaheb and Dr. Richard W. Baker for their expertise and contribution to the publications contained within.

Thank you to Dr. Takeshi Urano and Dr. Hiroaki Kato from Shimane University Medical School and Dr. Yota Murakami and Dr. Takuya Kajitani from Hokkaido University for hosting me during my research fellowships in Japan.

Finally, thanks to former and present labmates and colleagues at Tufts Biology for their friendship and stimulating scientific conversations.

Table of Contents

Chapter 1

Intrinsic disorder and tandem repeats in models of protein function

Abstract

Genetic information encodes proteins that perform essential functions in the cell. Protein structure determines function and understanding the connection between structure and function is a fundamental aim of molecular biology. Significant advances in the 20th century established a standard model of structural biology centered on stable and discretely folded protein domains. However, recent experimentation has demonstrated that extensive swathes of proteins are disordered and/or repetitive, opening the way for new models to explain protein function. A more complete grasp of the mechanisms underlying protein function will further enable scientists to manipulate these factors for applications in medicine and biotechnology. In this chapter, I will introduce the principles of disorder and repetitive sequences in proteins and their influence on the standard model of structural biology. Additionally, I will provide background on one particular protein, RNA polymerase II, and highlight the features that make it an ideal model system to study the roles of disorder and repetitive sequences on protein function. Finally, I will present the rationale for the research in this thesis and briefly outline the content and aims of the remaining chapters. I propose that studying disorder and repeats, particularly in RNA polymerase II, is a powerful approach to expand our appreciation of the impact of protein structure on essential biological processes.

The standard model of structural biology

Proteins play an essential role in the processes that sustain life in the cell, owing to the chemical diversity of their constituent amino acid building blocks. The twenty canonical amino acids contribute to a given protein's structure and its ability to perform biochemical reactions. Consequently, understanding how a protein's structure governs its function is a central goal in the field of molecular biology. The present view of this structure/function relationship is anchored in the one-gene, one-polypeptide model. In this model, one gene encodes information to produce one independently folding polypeptide (protein) that either has function on its own or as part of a protein complex. Gene duplications or even whole genome duplications can increase the number of genes available for natural selection to lead to new functional proteins (Ohno 1999). Mutational events at existing genes can change the amino acid sequence of the resulting protein, further contributing to structural and functional diversity. This relationship between genetic information and the encoded proteins serves as the bedrock for the life sustaining processes of the cell.

Once they are properly synthesized, proteins need to be able to perform their required functions in the cell. I will be referring to the prevailing model of how proteins are able to carry out these functions as the standard model of structural biology. The standard model was established in the latter half of the 20th century primarily thanks to advances in X-ray crystallography that enabled scientists to visualize sub-microscopic protein structures (Shi 2014). Further technological developments in cryogenic electron microscopy also drove the observation of larger protein complexes at increasingly finer

resolutions (Bai et al. 2015). Visualizing proteins under various experimental conditions allowed scientists to link structural changes to outcomes from complementary investigations and establish how protein structure was related to function. This method proved powerful, as protein structures were then used to rationally inform experimental design (Choi and Roush 2017). Protein structures were also important in comparative evolutionary approaches, as organisms with similar amino acid sequences were predicted to have similar 3D structures and functions (Miliara and Matthews 2016). Therefore, both technological and methodological advances were crucial in establishing the standard model of structural biology.

The first expression of the standard model was in the so-called lock and key model of protein function (Figure 1A). This model was established to primarily explain enzymatic activity but can also be applied to the interactions of structural proteins (Habchi et al. 2014). In the lock and key model proteins must be able to fit properly with their interacting partners to carry out a function, much like a key must fit its lock in order to open a door. The “fit” of a protein is determined by its 3D structure, tightly linking the concepts of structure and function. A stable and properly fitted structure, seen by the crystallized form, is necessary for function under the lock and key model; both the lock and the key need to be made out of metal to open the door. Disrupting or disordering the protein structure eliminates the activity, although function can be restored if the conditions are returned to normal and the amino acid sequence is still intact (Anfinsen 1973). In this way, the lock and key model provided a powerful framework, backed by X-

ray crystallography data, to explain the relationship between protein structure and function.

Further refinement of the standard model came in the form of the inducible fit model of protein function (Figure 1B). While the lock and key model was an effective conceptual framework, it did not fully explain the biological properties of proteins. Unlike metal locks and keys, proteins are in a state of constant motion, vibrating or twitching on their own and showing more significant conformational changes while active (Koshland 1958). To better fit the data an inducible fit model was proposed whereby an interaction, for instance an enzyme binding its substrate, would induce a structural change in the protein that enabled the relevant function to occur. Accordingly, mutations that prevented internal motion of the protein in addition to bulk disruptions of structure would inhibit function. The induced changes to structure can be quite dramatic, involving large rearrangements of protein structure that lead to the final function (Kuser et al. 2008). The induced fit model better accounted for the biological activity of proteins in the cell while still maintaining the fundamental concepts of the lock and key model. A stable protein structure was still required, albeit with the possibility of folded protein domains to move in relation to each other. As a result, the induced fit model is successfully able to explain the structure/function relationship of many proteins in their biologically-relevant context.

There are many factors under the standard model, both at the level of the genome and the proteome, which can fine tune the structure and therefore the function of proteins. At

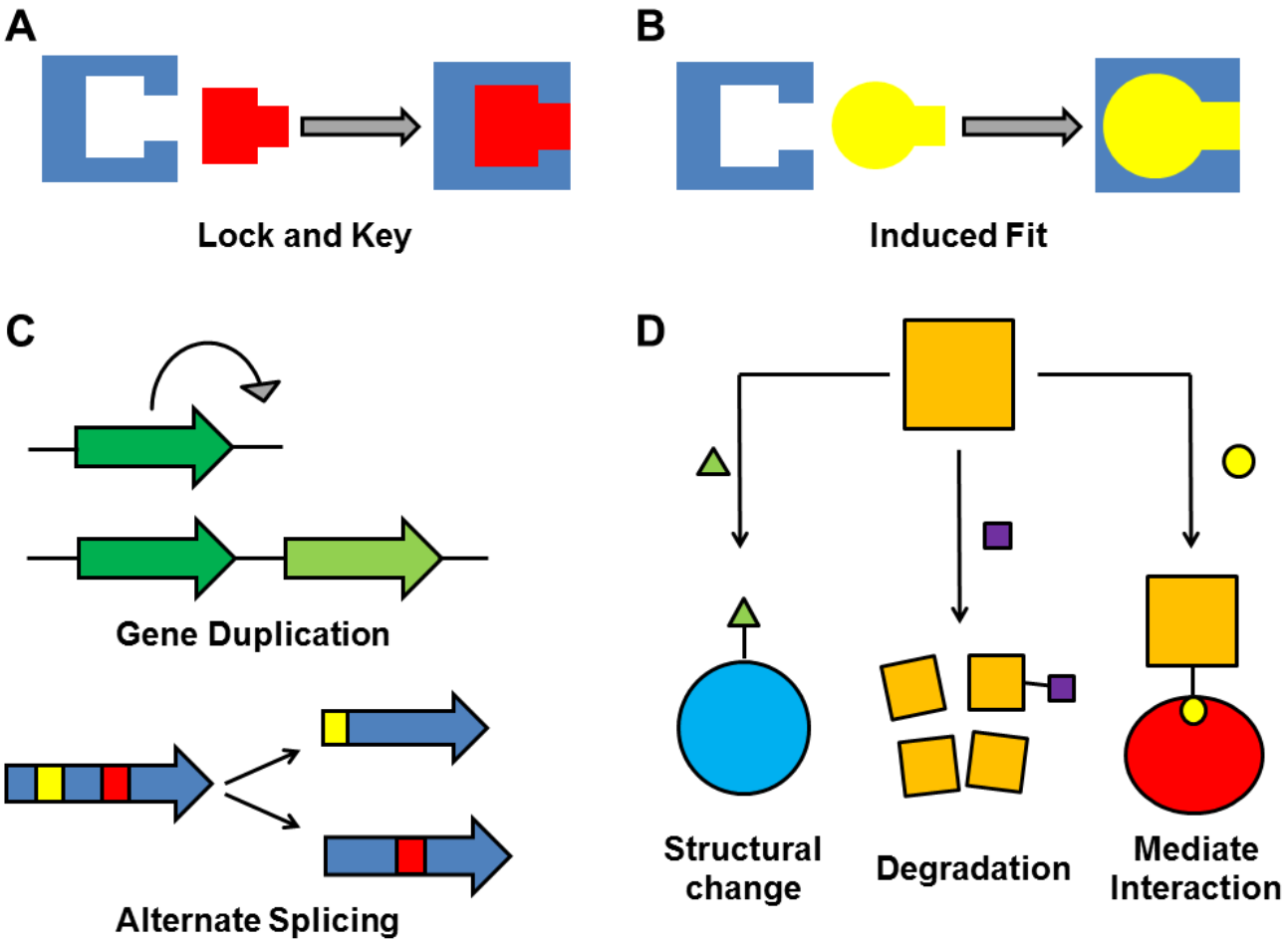


Figure 1. Models of protein structure and function. **A**) The lock and key model explains how protein structure is required to properly fit with interacting small molecules and proteins. **B**) Under the induced fit model, subtle conformational changes are vital to enable protein structural interactions. **C**) Gene duplication and alternative slicing are well-accepted mechanisms to increase genetic diversity, leading to increased numbers of protein functions. **D**) Post-translational modification can greatly increase the functional attributes of a given protein through the addition of pathway-specific reversible modifications. Figure 1D Adapted from (Fuchs 2013).

the level of the genome different functions can arise from the creation of paralogs, second copies of a gene that result from segmental or whole genome duplications (Innan and Kondrashov 2010). Gene duplications provide the raw material of evolution, as the paralog is free to accumulate mutations and potential new functions while the original copy maintains the original function (Figure 1C). Genetic mutations in the paralog can give rise to structural changes in the amino acid sequence which would then lead to different function. Whole genome duplications, for instance as a result of a hybridization event, can especially change protein function at a massive scale (Scannell et al. 2006). In addition to creating new genes, protein structural diversity can come from a single gene in the form of alternative splicing. Differential retention of exons in the final transcript can lead to protein products with different structures, often at the level of whole folded domains (Figure 1C). These different versions of proteins from the same gene are known as isoforms, and display a range of related functions that allow for the fine tuning of a given biological process (Naftelberg et al. 2015). Genetic processes are therefore an important determinant in not only producing proteins but also regulating their functions.

Modulation of protein structure and function also occurs at the levels of the proteins themselves, broadly referred to as post-translational modifications (PTMs). A dazzling amount of different PTMs have been identified that can regulate all aspects of protein structure and function (Figure 1D). A major function of PTMs is to directly change protein structure to either activate or inhibit the function of the protein. One of the most dramatic examples of structural change is through regulated cleavage of stored forms of

enzymes, known as zymogens, to produce the final active protein (Huber and Bode 1978). More subtle changes are caused by PTMs that affect the packing and folding of proteins through electrostatic or hydrophobic interactions to enable conformational shifts. PTMs also modulate protein function by promoting or inhibiting the interactions of important binding partners. Protein-protein interactions are frequently mediated through structural changes wrought by PTMs such as phosphorylation or methylation (Kouzarides 2007). Some PTMs are maintained through several cell cycles and are hypothesized to comprise a type of epigenetic memory (Patel and Wang 2013). A vital aspect of PTMs is that they are predominantly reversible – they can be both added and removed from a target protein. This reversibility allows proteins to experience a range of activated or deactivated states throughout their lifetimes, all governed through PTM-induced structural changes. All of these different mechanisms of action enable PTMs to precisely regulate the activities of proteins in the cell.

The standard model of structural biology is an effective framework to explain the behavior of proteins and direct experimental design. The model takes into account how the flow of information from the genome to active processes in the cell takes place. As a result, structural data is now frequently used to explain the function of proteins in tandem with other molecular biology approaches. Aside from the explanatory power, the standard model also enables design of experiments based on structural data. Applications include probing the activity of conserved proteins and rational design of pharmaceuticals and of enzymes used in biofuel production. However, the model does not fully explain all aspects of protein structure and function. As I will address in the next

section, there is a significant portion of protein structural and functional space that the standard model does not cover that warrants further consideration.

Gaps in the standard model: disorder and repeats

While the standard model was very effective at explaining the stably folded structures obtained from X-ray crystallography experiments, not all proteins fall into this category. It was recognized early on that there were gaps, sometimes very long stretches of amino acids, in crystal structures that would not crystallize and therefore could not be visualized. These were the first acknowledged instances of disordered sequences that do not adopt a single stable conformation that can be captured by structural biology techniques. In addition, it was always known that certain proteins mysteriously avoided all attempts at crystallization. Initially, missing regions of disorder were dismissed as spacer elements with little functional significance of their own (Sickmeier et al. 2007). However, increasingly larger disordered stretches were detected in functionally important regions of proteins and bioinformatics approaches were developed to analyze the properties of these disordered regions (Melamud and Moult 2003). The building consensus at the end of the 20th and start of the 21st centuries was that disordered regions could indeed have biologically relevant functions, something that the standard model proved unable to explain.

In parallel to the developments in disordered proteins, the genetic basis of the standard model was challenged by the emerging appreciation of repetitive sequences. One of the

most surprising findings of the human genome project was the unexpectedly low number of protein coding genes identified. Only around 20,000 protein coding genes were confirmed, down from previous estimates of 100,000 (Ezkurdia et al. 2014). Instead, a significant portion of the human genome, and that of other organisms, was found to consist of repetitive DNA sequences (Liang et al. 2015). Although repeats were initially considered as “selfish DNA” (Orgel and Crick 1980) further research uncovered a number of important properties of these sequences. The presence of these repeats explained why genome size was often not correlated with organism complexity – some simple organisms just had a very large repeat component of their genomes (Hartl 2000). Repetitive sequences were found both between and within genes and could therefore be present in the coding sequences of proteins. Repetitive sequences are evolutionarily unstable compared to the rest of the genome, and their expansion and contractions rapidly change protein structure compared to the canonical gene or genome duplication events (Gemayel et al. 2010). The changing repeat units are also frequently much smaller than the typical folded protein domains that are central to the standard model. Repetitive sequences can have significant consequences for protein function in both healthy and disease states (Lopez Castel et al. 2010) and warrant considerable future study.

The emerging consensus in the study of protein structure and function is that both disordered and repetitive sequences have biologically relevant roles in determining protein function. Once dismissed as rare exceptions to the standard model, disordered regions and repeats are now appreciated as an important component of the genome

and proteome. Disordered regions can account for 20% of all proteins and repeats can comprise around 50% of a genome, depending on the organism examined (Tompa et al. 2006). Disordered regions and repetitive sequences come with unique mechanisms of behavior that impart proteins with functions that could not be possible with the canonical folded domain. Furthermore, the properties of disordered regions and repeats frequently overlap, leading to additional complex interactions (Simon and Hancock 2009). These interesting functional consequences of disordered and repetitive protein sequences justify further research into their composition and activity. Indeed, both fields have surged ahead recently, in great part due to rapid advances in computational approaches to recognize and predict disordered and repetitive sequences. In the following two sections I will introduce disordered regions and repetitive sequences in depth and identify outstanding questions of interest in both fields that deserve further investigation.

Intrinsically disordered regions

Disordered regions of proteins have frequently been referred to as the “black box” of structural biology due to their inability to be visualized by X-ray crystallography. With no 3D structures available to determine the effects of experimental perturbation, the typical procedures of the standard model cannot be applied to study disordered regions. However, advances in structural biology techniques outside of crystallography, combined with computational approaches, have started to decode the black box of disordered regions (Habchi et al. 2014). In this section, I will provide an overview of the properties and behavior of intrinsically disordered regions (IDRs) of proteins (Figure

2A). The “intrinsically” part of the name refers to the amino acid sequence of the region itself as causing the disorder, as opposed to disorder or unfolding brought about by environmental or chemical means. Furthermore, the processes described below apply both to IDRs flanked by structured protein and to proteins that are predominantly disordered, which are aptly referred to as intrinsically disordered proteins. The overall picture that emerges is of a highly dynamic system that can perform important functions that are otherwise impossible for stable folded proteins.

As mentioned previously, the existence of IDRs was first inferred from the missing sequences of X-ray crystallography structures (Sickmeier et al. 2007). However, the intrinsic structural determinants of disorder make IDR identification possible due to their unique behavior in comparison to structured regions in traditional biochemical assays. The unfolded nature and sequence composition of IDRs cause them to run abnormally in gel electrophoresis and size-exclusion chromatography when compared to structured proteins of the same size. Due to their unstructured state, IDRs are also highly resistant to heat and chemical denaturation that precipitates out folded proteins (Uversky 2002). Finally, IDRs are observed to be highly vulnerable to digestion by proteolytic enzymes (Iakoucheva et al. 2001). These unique behaviors allow IDRs to be tentatively identified and subjected to more specialized techniques like NMR to characterize and quantify the degree of disorder (Kosol et al. 2013). As a result, both traditional biochemical and specialized structural biology approaches can be applied to identify and characterize IDRs from wider collections of proteins.

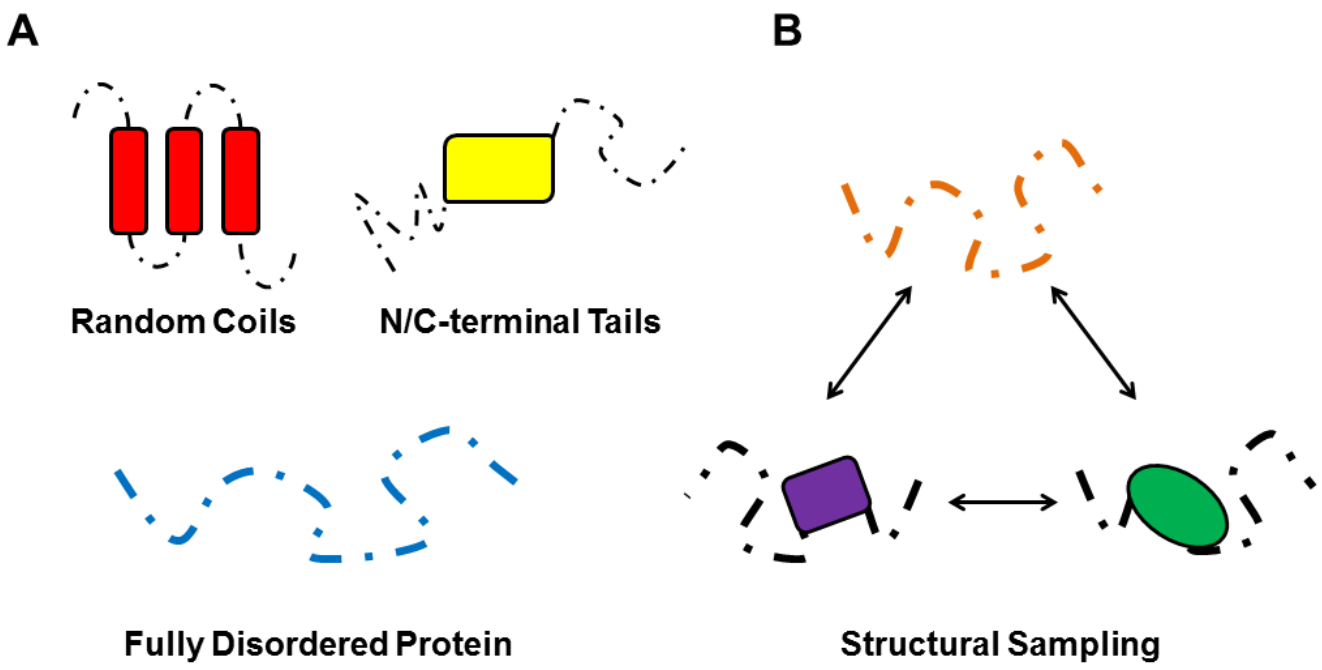


Figure 2. Modes of function of disordered proteins. **A**) Intrinsically disordered regions can either act as spacers between folded domains or are located as flexible tails at the ends of proteins. Disordered proteins also exist that are either dependent on their fully disordered nature in solution or when bound to an interacting partner. **B**) Disordered regions do not fully lack structure, rather they dynamically shift between various partial or fully structured states. This shift can be strongly influenced by binding interactions or by post-translational modification.

The increasing number of identified IDRs allowed scientists to characterize the sequence requirements that led to disorder. Studying the sequences of IDRs revealed a heavy bias towards amino acids that promote disorder and a bias away from amino acids that promote stable folding (Campen 2008). Electrostatic and hydrophobic repulsion are the two main processes that determine disorder. IDRs are enriched in amino acids with charged side chains, giving the regions a high net charge that encourages electrostatic repulsion and discourages folding. On the other hand, IDRs are depleted of both small and bulky amino acids that are hydrophobic, discouraging the formation of stable hydrophobic core that is characteristic of stable folded proteins. Additionally, IDRs are enriched for the amino acids glycine and proline, which are known to disrupt stable folding and structure formation (Campen 2008). This unique profile of amino acids gives IDRs their unique properties and leads to the behaviors observed in the biochemical assays described above. Initially, the unstructured nature of IDRs was hypothesized to be an artifact of the diluted *in vitro* experimental conditions. However, studies that tracked behavior of IDRs found that they maintained their disorder regardless of molecular crowding, further supporting the importance of protein sequence (Szasz et al. 2011). IDR sequence therefore emerged as an effective metric to identify and characterize disorder.

One of the most important advances in studying IDRs came from the development of algorithms to predict disorder in whole proteomes. Establishing the sequence requirements of IDRs as well as collecting databases of known disordered regions proved vital to developing these computational approaches. IDR prediction algorithms

act on either of two basic principles: predicting disorder *de novo* based on sequence properties or based on known IDR sequences in proteins (Habchi et al. 2014). Predicting disorder based on sequence takes in to account either the amino acid content or the predicted resulting biophysical properties of sequences. Algorithms weigh the content of disorder-promoting amino acids with metrics of electrostatic repulsion and hydrophobicity to craft an order/disorder prediction (Prilusky et al. 2005). These algorithms have the advantage of being based on fundamental biophysical and biochemical principles and ignore biases present in current databases of characterized IDRs. Conversely, this approach takes many assumptions on the behavior of disorder that might not reflect the biological reality of actual IDRs collected in databases. The second approach uses machine learning algorithms trained on actual datasets of IDRs to make future predictions of disorder. Given a novel proteome, these algorithms will base their predictions on the thousands of previous IDRs that have been identified to date. Machine learning algorithms are able to predict IDRs using actual data confirmed by biochemical and structural experiments, however the selection of the data set to train the algorithm will greatly affect the resulting prediction (Peng 2006). As a consequence, most computational studies of IDRs will use multiple algorithms in parallel to arrive at a concordant prediction of disorder in their proteins of interest. The use of these algorithms has proved to be an essential tool in mechanistic and comparative studies of IDR function.

The drastic consequence of IDRs' unique structural properties is the unique functions disordered protein sequences are able to perform in the cell. An important consideration

of IDR function is that they are disordered and not unstructured. IDRs are able to shift between many transient structured states, or else undergo folding following a specific event such as binding to another protein or small molecule (Figure 2B, 3). This flexibility of IDR structure underlies their functional roles and sets them apart from other stably folded proteins. The flexibility of IDRs allows them to either bind many of one kind of interactor or bind one at a time to variety of different partners (Peti et al. 2012). IDRs are also capable of making highly specific yet weak interactions that come together quickly and can be rapidly reversed (Oldfield et al. 2005). As mentioned above, IDRs can adopt a stable conformation when bound but amazingly there are also instances of regions that remain disordered in a “fuzzy” state even upon binding (Savvides et al. 2004). These unique structural arrangements are a far cry from the relatively static interactions of the standard model and studies into IDR function have vastly expanded our appreciation of the diverse biochemical processes mediated by disorder (Figure 3).

The functional diversity of IDR interactions is significantly expanded by the reversible addition of post-translational modifications (PTMs). The wide range of available PTMs and their regulated and reversible addition serves to further amplify the diversity of the already dynamic IDR activities. The PTM modification state can mediate the many-to-one or one-to-many binding modes of IDRs. For example, the disordered N and C-terminal tails of histone proteins are targeted by a variety of site-specific PTMs that govern the binding of interacting proteins to regulate chromatin packing and transcription (Rothbart and Strahl 2014). Similarly, the disordered C-terminal domain of RNA polymerase II is heavily modified to orchestrate essential protein factors

throughout the transcription cycle (Corden 2013). More broadly, PTMs can act as regulators of the disorder/order transitions of IDRs, altering the level of structure to influence reversible or fuzzy binding states (Habchi et al. 2014). As with structure proteins, post-translational modification adds an additional layer of regulation to IDRs to enable finer tuning of their unique activities.

Recent developments in the field of IDR biology have increased our appreciation for the role of disorder in protein function, however many questions remain concerning IDR structure and function. As more IDRs continue to be identified, questions regarding their conservation and evolution will need to be addressed. IDRs are known to evolve at a more rapid rate than structured sequences, although the mechanism and its relation to disorder is still poorly understood. IDRs have also been demonstrated to form protein aggregates under certain conditions and the process of aggregation has implications for prion formation and epigenetic memory, as well as neurodegenerative diseases involving amyloid-like plaques (Lopez Castel et al. 2010). The disease angle is especially important as several IDRs have been identified as therapeutic targets through unknown mechanisms. The overall disordered content of a proteome is also an intriguing field of study as the amount of IDRs is correlated with organismal complexity (Dunker et al. 2000). The broad role of disordered sequences in promoting complexity is not well understood but could yield insight into how disorder affects protein function across wide stretches of evolutionary time. These questions and more are expected to be addressed with increasingly sophisticated molecular and computational approaches

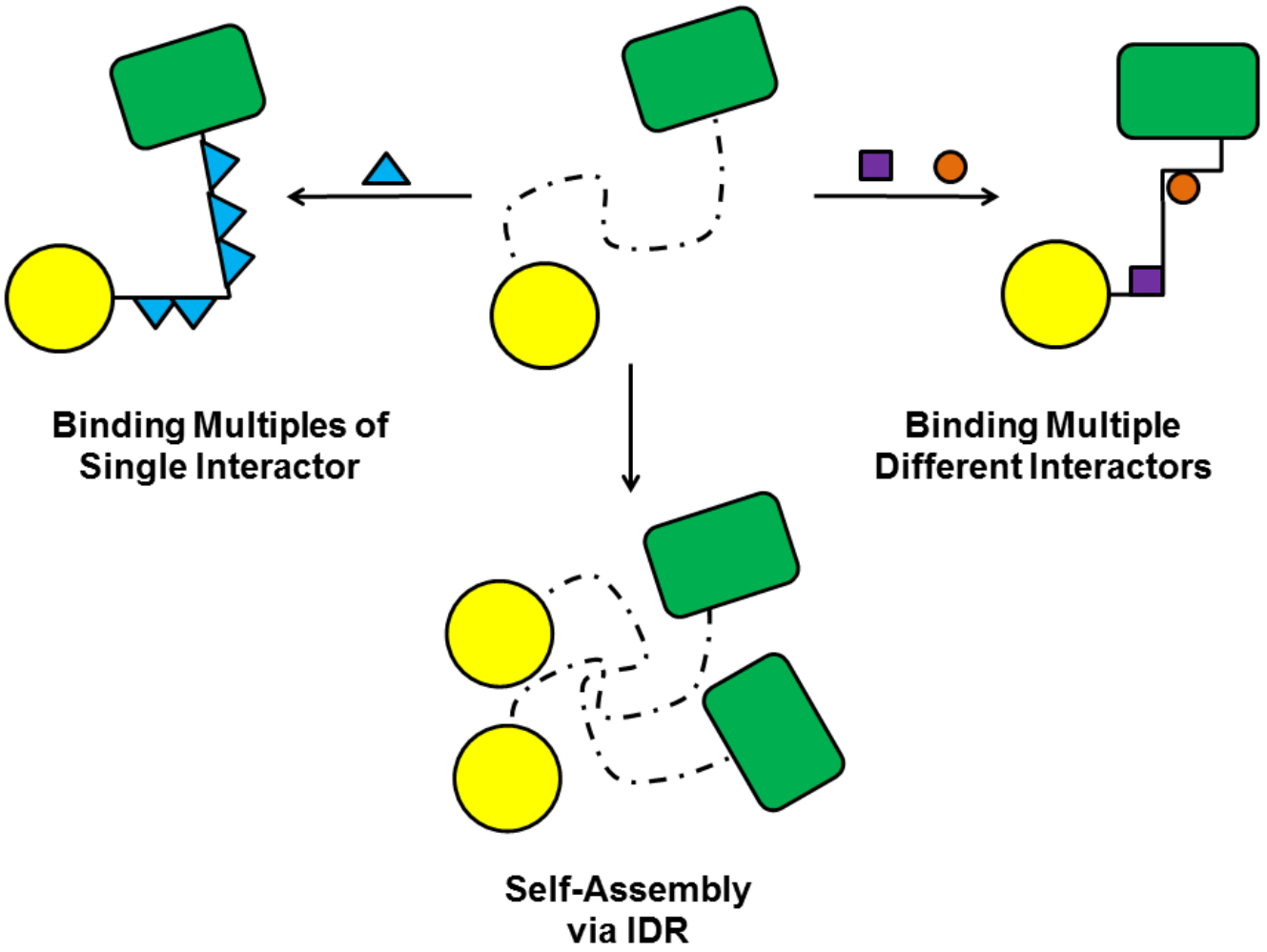


Figure 3. Protein-protein interactions influence disorder state. Intrinsically disordered regions enable a range of unique interactions that are not possible for fully structured proteins. A disordered region can adopt a more defined structure (indicated by the straight lines) when binding either many of a single interactor, or multiple different interactors. Disordered regions can also self-aggregate and may remain disordered in a "fuzzy" state even following binding.

to further grasp how intrinsically disordered regions regulate cellular functions through their shifting and versatile structure.

Repetitive sequences

The era of genomes heralded an explosion in the availability of genetic sequence data and uncovered many surprising mechanisms behind the regulation and expression of genetic information. One of the most unexpected findings was the extent of repetitive sequences within the genomes of eukaryotes, with the human genome containing close to 50% repetitive content (Gemayel et al. 2010). Repetitive sequences were known to biology previously, having been identified as satellite sequences, so called because they appeared as extra satellite bands in genomic DNA preparations. However, repeats were initially labeled as selfish or junk DNA and were thought to have no purpose other than their own replication (Orgel and Crick 1980). Repeats were overlooked even in the modern era of genomics, as their repetitive sequences could not be effectively aligned and they were typically left out of the final genome assembly. As a consequence, the biological relevance of repetitive sequences has been underappreciated and overlooked in favor of discrete non-repetitive genes and regulatory elements. More recently, in-depth investigations into repeat function have begun to uncover important roles for repetitive sequences and their expansion and contraction (Verstrepen et al. 2005, Gemayel et al. 2017). In this section, I will review the types of repetitive sequences and their functions with particular emphasis on repeats located in protein sequences. Variation of these repetitive sequences can significantly alter protein function and can drive functional diversity in relatively short evolutionary time frames.

By far the most prevalent repetitive elements are interspersed repeats derived from retrotransposon or viral sources. Interspersed repeats are so named because expansions of these sequences are placed throughout the genome and not next to the original sequence as in tandem repeats (Figure 4A). The DNA sequences of these repeats varies by the type but frequently code for little more than the proteins required for their maintenance and propagation in the genome (Feschotte 2008). This observation led to the hypothesis at the time that repeats were merely genomic parasites with no function on their own. However, more detailed studies revealed several mechanisms by which interspersed repeats could contribute to genomic diversity and adaptive evolution. One such example is the presence of transcription factor binding sites located within interspersed repeats. Repeat expansion at the proper genomic location could then pave the way for novel genetic expression networks to emerge. Indeed, this mechanism is believed to play an important role in the divergence of humans from other primates (Lee et al. 2015). This and many other mechanisms are aided by the rapid rate at which interspersed repeats expand throughout the genome (Kidwell and Lisch 2001). As with progress in the disorder field, advances in repeat analysis have been greatly aided by developments in computational approaches for identifying and tracking repeat expansions. Thus, we can see the progress of interspersed repeats from useless junk DNA to a crucial element of genomic regulation and function.

In contrast to interspersed repeats, tandem repeats either expand or contract next to each other in a particular genomic locus (Figure 4B). Tandem repeating units are

typically much smaller than interspersed repeats and their instability is driven by the intrinsic properties of their repetitive sequences (Fan and Chu 2007). These repeats are further categorized by repeat size as either microsatellites or minisatellites. Although the exact size cutoff varies within the literature, microsatellites are typically defined as between three and nine nucleotides (e.g. CAG coding for a polyQ repeat) and minisatellites are defined as greater than ten nucleotides (Gemayel et al. 2010). When they are not excluded from genomic sequences, tandem repeats are identified using algorithms that look for statistically significant repetitions of a particular sequence. Computational prediction of tandem repeats is complicated by the selection of the cutoffs used to identify something as repetitive and by the presence of degenerate motifs that may not be counted as a repeating unit (Merkel and Gemmell 2008). Tandem repeats are present in the open reading frames of protein coding genes and can code for repetitive sequences with important functional consequences for the final protein. This tandem repeat variation is especially significant for protein function and its relevance will be the focus for the remainder of this section.

The instability of tandem repeat sequences leads to often extensive expansion or contraction of the repeating motif that can have drastic consequences when connected to protein expression or function. Like with interspersed repeats, tandem repeats have a highly elevated rate of mutation when compared to non-repetitive sequences (Brinkmann et al. 1998). However, unlike interspersed repeats, tandem repeat instability is caused by the repetitive sequence itself and not by coded replication proteins. Tandem repeat instability occurs either through replication coupled mechanisms or

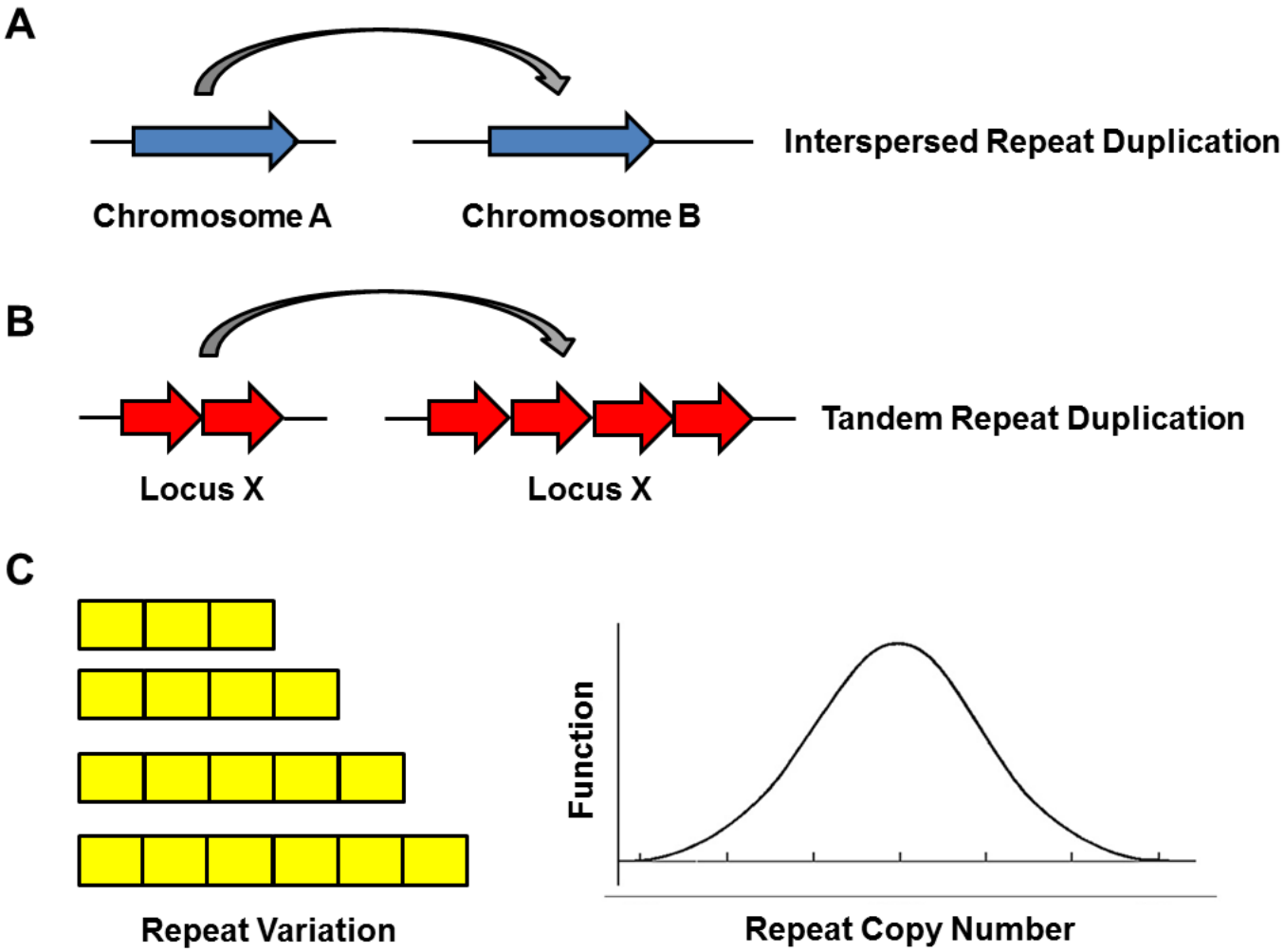


Figure 4. Genetic and functional diversity of repetitive sequences. **A**) Interspersed repeats code for the factors required for their replication and integration into distant genomic sites. **B**) Tandem repeats expand or contract based upon intrinsic sequence factors and do so next to each other in a given locus. **C**) Tandem repeat variation can lead to different functional outcomes for a protein if they are located in the open coding frame. Changes in repeat number can provide a fine tuning mechanism of protein function that leads to a gradation of function.

homologous recombination related mechanisms, with the former typically associated with microsatellites and the latter with minisatellites (Fan and Chu 2007). Replication coupled expansions or contractions occur when either the replicating or replicated strand aberrantly pairs with one of the many repetitive motifs, looping out the sequence and changing the repeat number as a result (Paques et al. 1998). A similar mismatch between the repetitive motifs also underlies misalignment during homologous recombination and crossing over, resulting in expansions or contractions (Richard and Paques 2000). The explanation given above is a very simplified version of events and researchers continue to study the precise mechanisms and their contributions to microsatellite and minisatellite instability. The resulting tandem repeat instability is a significant source of genetic variation that is especially important in the context of the coding sequence of proteins.

Tandem repeat variation by the mechanisms described above has been linked to surprising consequences for protein function. The most well studied examples to date are microsatellites that are linked to neurodegenerative diseases such as Huntington's disease. In these cases, microsatellite expansions of polyQ or other repeat tracts caused by replication slippage events can lead to increasingly worse disease states through disruption of regulatory elements, introns or the protein product itself (La Spada and Taylor 2010, Albrecht and Mundlos 2005). Aside from a role in disease states, tandem repeat variation is also a powerful adaptive force in evolution. Tandem repeats are especially prevalent in prokaryotes involved in virulence when compared to their benign cousins, suggesting that the environment and life history can select for repeat

content (Mrazek et al. 2007). Variation of tandem repeats in proteins has been linked to a number of functional outcomes between different copy number variants. To give one example, variation in the copy number of tandem repeats in the polyubiquitin gene led to differential response to environmental stress (Gemayel et al. 2017). Tandem repeat variation has been found to be important across a variety of different contexts for both microsatellites and minisatellites and further insights are expected as more tandem repeats are tested.

As is the case with disordered proteins, tandem repeats located in proteins are often associated with post-translational modifications. The repeating protein motifs have a strong amplifying effect on PTMs, as PTM-mediated interactions can be increased or decreased following repeat expansion or contraction. Repeat expansion can also free up PTM modified motifs to accumulate mutations and develop novel functions while the original sequence maintains the necessary function (Fuchs 2013). Tandem repeat PTMs are important in the variable repeats of the *FLO* genes in yeast that are responsible for flocculation (yeast aggregation) as a response to environmental conditions. The variable repeat units are modified by glycosylation PTMs that are required for cell to cell interactions (Verstrepen et al. 2005). Repeat variation therefore not only changes the protein structure directly, but can increase the PTM-mediated interactions as well. Modified tandem repeats also occur in proteins important for transcription. RNA polymerase II was mentioned in the intrinsically disordered section previously and its heavily modified C-terminal domain consists of repeating units (Corden 2013). Another important protein is the transcription elongation factor Spt5p,

which contains a similar repetitive C-terminal region that is modified by PTMs such as phosphorylation. These modifications mediate the binding of other protein factors to both Spt5 and to the CTD of RNA polymerase II to properly execute transcriptional processes (Mbogning et al. 2015). Tandem repeat variation, in combination with post-translational modification can therefore serve as a potent regulator of protein structure and function.

Taken together, recent advances in our understanding of tandem repeats establishes them as a potent force in protein function. The overall effect of repeats on protein function can be summarized by a dimmer switch model (Figure 4C). Expansion or contraction of tandem repeats in protein sequences enables a spectrum of protein function much like a dimmer switch controls light brightness. The whole range of protein function can then be sampled by selective forces to enable organisms to adapt to a particular condition or to buffer against multiple environmental pressures. Post-translational modification of these repeats provides even finer regulation of the dimmer switch. Despite recent advances in the field, a number of questions regarding tandem repeat structure and function remain to be addressed. While a number of instances of repeat copy number variation have been identified, more general principles of the effects of repeat variation are still elusive. Additionally, the large number of repeating units for some proteins begs the question as to the necessity of so many repeats. Are all of the repeating units in particular sequence equivalent in function or are there instances of specialization despite the identical or similar sequences? How might this functional specialization occur at the level of proteins and their interacting partners?

Further work will be required to answer these and other important questions to continue to explain the roles of tandem repeats in the function of proteins.

RNA polymerase II as a model system of disorder and repeats

Both regions of disorder and repetitive sequences can have a substantial influence on protein structure and function. As outlined in the previous two sections, there have been a multitude of studies on various proteins that have established the roles of either IDRs or tandem repeats. Many questions remained unaddressed, in particular the interesting overlap between IDRs and tandem repeats. IDRs and repeats share many of the same amino acid biases in their sequences and are enriched for similar biological processes (Campen 2008, Gemayel 2010). While there are some comparative studies that address this overlap for a selection of well-known proteins (Simon and Hancock 2009, Jorda et al. 2010), in-depth analyses of IDR and repeat function is still lacking. To address this gap, I propose that the disordered and repetitive C-terminal domain (CTD) of the enzyme RNA polymerase II makes an ideal model system to study IDR and repeat function in detail and uncover the interplay between these two regulators of protein function.

The CTD of RNA polymerase II has several advantages that make it an ideal model protein. The CTD is both disordered and highly repetitive, with repeat copy number variation across and within species. There is a wealth of functional studies of the CTD repeats due to their essential role in orchestrating co-transcriptional processes. The essential requirement of the CTD means that mutations made to the repetitive

sequence can have highly visible phenotypes that enable conclusions to be drawn on the functional requirements of particular aspects of repeat structure. The repetitive motif of the CTD is also highly conserved across eukaryotes and allows research done in genetically tractable model organisms such as yeast to be applied to humans and other organisms of interest (Corden 2013). The CTD is also one of the few large minisatellite repeats that has been well studied and further work would help correct the balance in a field that predominantly focuses on microsatellite repeats (Tompa 2003). In the following two sections I will introduce RNA polymerase II and the CTD in greater depth and further highlight the features that make it a powerful model system to study IDRs and tandem repeats.

RNA polymerase II and the transcription cycle

Transcription is the fundamental biological process where genetic information in the form of DNA is read and RNA is synthesized. In eukaryotes, transcription is carried out by three main RNA polymerases: I, II and III. RNA polymerase I performs the bulk of a cell's transcriptional activity by synthesizing ribosomal RNA while RNA polymerase III synthesizes transfer RNA and other small RNAs (Grummt and Langst 2013, Arimbasseri 2018). RNA polymerase II synthesizes all of the protein coding messenger RNA (mRNA) in addition to a number of noncoding RNAs. Plants also have two more RNA polymerases, IV and V, that regulate epigenetic silencing of specific loci (Zhou and Law 2015). All eukaryotic RNA polymerases are orthologous to each other and are descended from a single ancestral polymerase similar to the extant RNA polymerase of prokaryotes. Two of the core subunits of the three main eukaryotic polymerases have

homology to bacterial RNA polymerase, while the remaining subunits have evolved to have specific functions. Notably, RNA polymerase II is the only enzyme to have evolved a CTD that is required to regulate gene expression (Corden 2013). For this reason, I will focus the discussion of transcription specifically on mechanisms of mRNA synthesis by RNA polymerase II.

RNA polymerase II is a large 12 subunit complex that must interact with a number of other complexes to carry out mRNA synthesis (Figure 5). The largest subunit is Rpb1p that contains both the catalytic site for mRNA synthesis and the CTD to regulate co-transcriptional processes. Both Rpb1p and Rpb2p (the second largest subunit) are homologous to the bacterial *β* and *β'* RNA polymerase subunits, respectively (Allison et al. 1985). Out of the remaining ten subunits, five are specific to RNA polymerase II while the other five are shared between all three RNA polymerases (Cramer 2002). Two of the RNA polymerase II specific subunits, Rpb4p and Rpb7p, form a detachable subcomplex with specific roles in stress response and gene regulation (Choder and Young 1993). All RNA polymerase II subunits with the exception of Rpb4p and Rpb9p are required for viability, highlighting the essential role of complex composition in transcription. Both RNA polymerase II assembly in the cytoplasm (Boulon et al. 2010) and import into the nucleus (Gomez-Navarro et al. 2017) are highly regulated processes to ensure proper levels of the complex are present for transcription. Once inside the nucleus, RNA polymerase II interacts with a number of other protein complexes of equal size or larger to perform transcription.

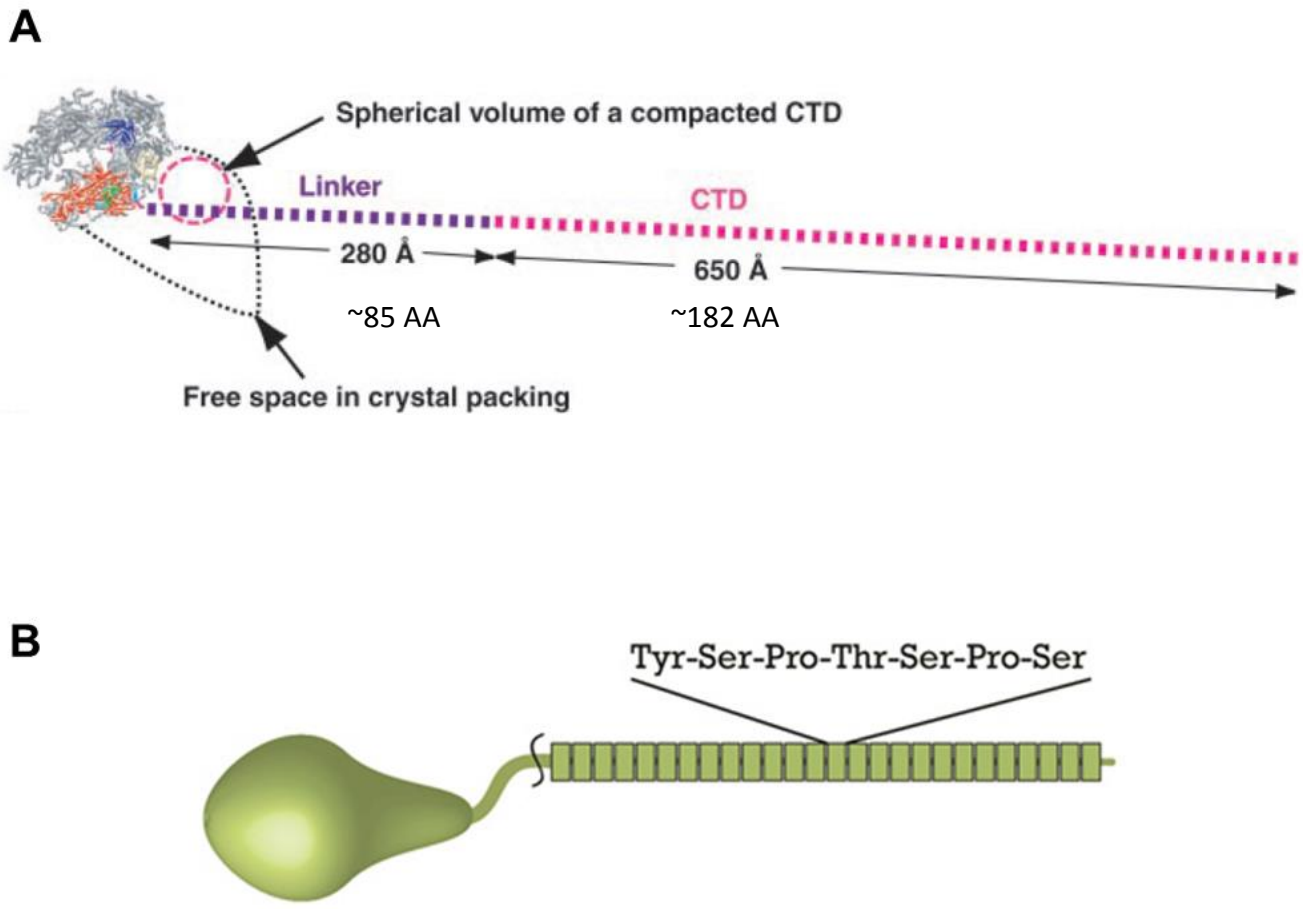


Figure 5. The disordered and repetitive CTD of RNA polymerase II. **A**) As an intrinsically disordered region, the CTD is not visible in crystal structures of RNA polymerase II. The hypothetical length of the CTD and its flexible linker is many times the diameter of the RNA polymerase II complex. While it is unlikely that the CTD is fully extended under physiological conditions, post-translational modification of key residues can change the conformation of the CTD. Figure adapted from (Cramer et al. 2001). **B**) Outline of the consensus sequence and repeat number in the budding yeast CTD. Figure adapted from (Babokhov et al. 2018).

The first step in the transcription cycle is the initiation of transcription at the promoter of a gene (Figure 6A). Transcription initiation is carried out with the help of general transcription factors that select the start site and guide RNA polymerase II to begin transcription. Promoter sequences are selected for by the TATA binding protein and the general transcription factor TFIID, itself a large protein complex of multiple subunits (Hantsche and Cramer 2017). RNA polymerase II is recruited to the site of initiation by the combined action of TFIIA, TFIIB and TFIIF (Kuras et al. 2000). At this point RNA polymerase II is also associated with an essential regulatory complex known as Mediator that is also recruited to the initiation site (Allen and Taatjes 2015). RNA polymerase II bound to the general transcription factors forms the pre-initiation complex that is ready to accept the signal to begin transcription. TFIIE and TFIIH unwind the promoter and allow the Rpb1p subunit to access the DNA to begin mRNA synthesis (Holstege et al. 1996). After a short window where transcription can be prematurely halted, RNA polymerase II leaves the pre-initiation complex and proceeds to the elongation phase.

The second step of the transcription cycle is the elongation phase where the full length transcript is actively synthesized (Figure 6A). As with initiation, RNA polymerase II works with a number of regulatory proteins known as elongation factors to efficiently synthesize mRNA during elongation. Elongation factors are identified by the positive role they have on elongation rate and typically follow RNA polymerase II through the body of a gene, binding to the CTD of Rpb1p or to some other surface of the complex. The general transcription factor TFIIS assists RNA polymerase II when elongation

stutters or halts by alieving the block and resetting transcription (Lisica et al. 2016). Additional elongation factors like Spt4/5p and the Paf1 complex facilitate elongation in tandem with other co-transcriptional events (Jonkers and Lis 2015). In metazoans, elongation is highly regulated at a step just after initiation known as promoter-proximal pausing. Pausing is widely prevalent phenomenon where RNA polymerase II is detected at a spot just after the promoter where it waits for a signal to begin elongation. This pausing is seen as way to regulate transcription to quickly produce necessary transcripts in response to an environmental condition, the classical example being heat shock stress response (Bunch 2017). The RNA polymerase II elongation complex is in a paused state until it is activated by the kinase P-TEFB which phosphorylates a number of targets including an elongation repressor which turns into an elongation factor (Jonkers and Lis 2015). These mechanisms serve to control elongation rate at a variety of levels to produce the necessary amount of transcript.

Following the formation of the transcript, the last step of the cycle is to terminate transcription and export the mRNA to be synthesized into protein (Figure 6A). As the elongating polymerase approaches the end of the gene, termination factors recognize signals in the elongating complex and the nascent RNA that trigger the termination process. The CPF-CF complex is a conserved set of proteins responsible for terminating the transcription of protein coding genes and enabling nuclear export of the resulting mRNA (Porrua and Domenico 2015). Following the termination of transcription the 3’ end of the mRNA is modified with a polyadenosine tail that controls mRNA stability and export (Dunn et al. 2005). Aside from completing mRNA transcription,

termination also has a role in finishing noncoding RNA transcription and preventing unregulated transcription of the genome. Noncoding RNAs are terminated mostly by a different pathway catalyzed by the NNS complex, although additional termination pathways have also been identified (Porrua and Domenico 2015). The NNS complex is also required to terminate undesired cryptic transcription and direct the resulting RNAs to the exosome for degradation (Tudek et al. 2014). The role of pervasive cryptic transcription and its termination by the NNS pathway is currently an exciting field of study in transcription biology to determine how much of this so-called “leaky” transcription is functionally relevant. Following the completion of termination, RNA polymerase II is released from the elongation complex and is either degraded or repurposed for another round of transcription. The transcription cycle of initiation-elongation-termination is a carefully orchestrated process involving a number of protein complexes to ensure smooth and regulated production of transcripts.

RNA polymerase II activity is also involved with a number of co-transcriptional processes that produce the mature transcript. Shortly after transcription initiation, the 5’ end of the emerging nascent transcript is capped with a modified guanosine by capping enzymes that bind to the RNA polymerase II complex. This capping event is timed to coincide with the emergence of the native transcript and serves to protect the transcript from degradation and eventually promote its export to the nucleus (Cho et al. 1997). The mRNA transcript is further modified throughout transcription by the action of a ribonucleoprotein complex known as the spliceosome. The spliceosome excises intron sequences in parallel with transcription to include only exons in the exported mRNA

(Morris and Greenleaf 2000). Spliceosome activity is particularly important in exon selection during alternative splicing that greatly increases the diversity of proteins that can be produced from a single genetic sequence (Naftelberg et al. 2015). As mentioned previously, the 3' end of the mRNA is modified with a polyadenosine tail that mediates transcript stability and export (Dunn et al. 2005). In addition to these well-known processes, there are a number of other RNA modifications that are thought to occur co-transcriptionally to regulate translation dynamics of the modified RNA (Roundtree et al. 2017). All of these co-transcriptional processes described above are carefully tuned to the appropriate stages of the transcription cycle through interactions with RNA polymerase II and illustrate the regulation necessary to produce transcripts at the appropriate levels.

RNA polymerase II must not only read DNA, but has to also navigate the chromatin context of the genome. DNA is packaged tightly with histone proteins to make a complex known as chromatin that protects genetic information and regulates its expression. Post-translational modification of the histone proteins marks genes as either active or inactive and the placement of these marks is both interpreted and influenced by the activity of RNA polymerase II. Histone modifications are present at all stages of the transcription cycle and act by both directly influencing histone packing as well as mediating recruitment of other protein factors (Zentner and Henikoff 2013). The arrangement of histones at the promoter, tied to histone acetylation, is altered to enable the general transcription factors and RNA polymerase II to access the DNA and proceed with initiation (Rundlett et al. 1996). As the elongating RNA polymerase II

complex passes through the body of the gene, histone proteins are displaced and returned by the actions of histone chaperones that travel together with RNA polymerase II (Venkatesh and Workman 2015). The interaction of the transcription machinery with chromatin can also enable longer term changes to chromatin structure to promote or inhibit further transcription. The importance of histone modifications on RNA polymerase II activity demonstrates the importance of the chromatin context in controlling transcription. As I will discuss in the following section, all steps of RNA polymerase II activity including the interaction with chromatin are precisely regulated by the disordered and repetitive C-terminal domain of the catalytic Rpb1p subunit.

The essential function of the CTD in transcription

The C-terminal domain (CTD) is a tail-like extension of the largest subunit of RNA polymerase II, Rpb1p that is in the center of the network of protein-protein interactions that enable progress through the transcription cycle (Figure 5A). The CTD was first discovered through observations that RNA polymerase II had three different forms when separated by gel electrophoresis. The three forms were found to be RNA polymerase II with an unmodified and modified CTD and a final form that was missing the CTD altogether (Corden 2013). Sequencing the CTD found that the domain consisted of heptad repeats of the sequence tyrosine-serine-proline-threonine-serine-proline-serine or YSPTSPS (Figure 5B). The CTD was not only repetitive, it was disordered as well and the whole CTD and most of its linker region was absent from crystal structures of the RNA polymerase II complex. The CTD is not required for RNA synthesis *per se* as transcription of RNA in vitro is possible without the CTD, however the cell requires

between one third and half of the repeats for viability (Eick and Geyer 2013). Consequently, understanding the function of the repeats of the CTD is essential to appreciate the regulation of transcription by RNA polymerase II as a whole.

The wealth of genomic sequences has allowed the CTDs of various different eukaryotes to be studied to uncover the evolutionarily-important aspects of the CTD structure. The canonical YSPTSPS heptad sequence is highly conserved among the budding and fission yeast model systems where the majority of CTD studies are performed. The mammalian CTD, which is amazingly well-conserved among all mammals sequenced so far, also contains the consensus heptad repeat although the latter half of the repeats contain more substitutions, especially at the S7 position (Simonti et al. 2015). The fruit fly CTD presents an interesting example wherein almost all of the heptad repeats have some sort of residue substitution, although altogether the CTD repeats follow the consensus sequence (Eick and Geyer 2013). Many other single celled eukaryotes do not follow the exact heptad consensus sequence, but appear to preserve the spacing of the two SP motifs, suggesting that this is the primary function of the CTD in these organisms. Other simple organisms lack this spacing of SP motifs and instead have C-terminal extensions that are enriched for the amino acids of the heptad repeat without any periodicity (Yang and Stiller 2014). The evolutionary history of the CTD repeat structure hints at the functions the CTD is able to perform and the selection between these functions among the various eukaryotes that have been studied to date.

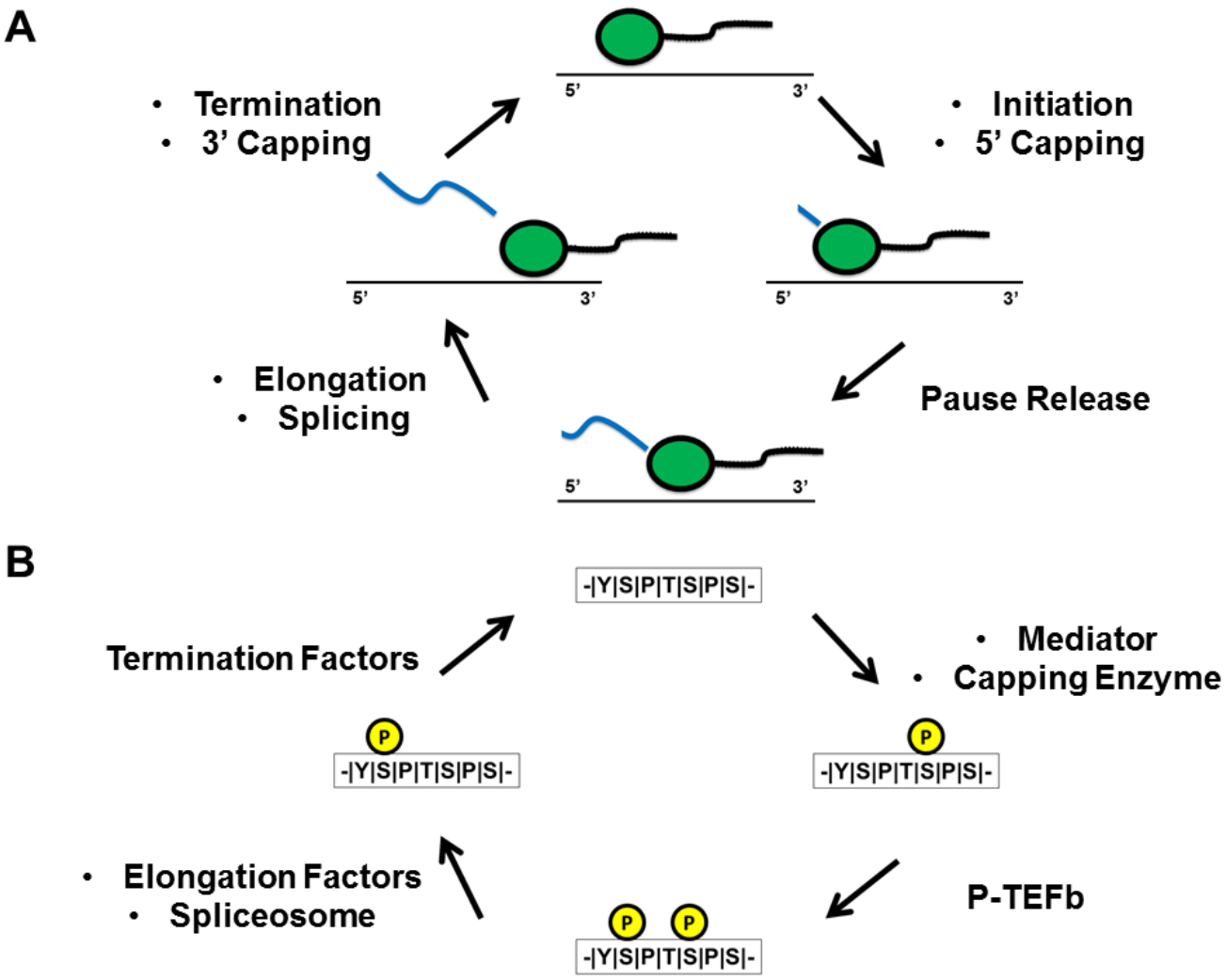


Figure 6. RNA polymerase II and CTD modifications during the transcription cycle. **A**) As RNA polymerase II travels through the body of the gene, key steps of mRNA synthesis are coordinated with co-transcriptional processes and chromatin regulation to properly produce the mature transcript. **B**) CTD serine phosphorylation is the best understood modification throughout the transcription cycle. The addition and removal of phosphorylation at serine 5 and serine 2 recruit key co-transcriptional factors at the appropriate point in the transcription cycle to enable regulated transcription. Figure adapted from (Eick and Geyer 2013).

While the CTD itself does not synthesize RNA, it is absolutely essential for transcription by acting as a scaffold to bind proteins involved in co-transcriptional processes. The CTD's disordered nature, as well as its long length relative to the RNA polymerase II complex, enables it to sample a wide range of the transcription site (Portz et al. 2017). This wide range allows the CTD to localize co-transcriptional processes to the transcribing polymerase and to the local chromatin environment (Figure 6B). Depending on the modification state of the CTD (discussed below), various protein factors are recruited to the CTD where they have access to RNA polymerase II and the emerging transcript. This activity is characteristic of IDRs that mediate protein-protein interactions and is known as the "fly-casting" mechanism (Huang and Liu 2009). The large number of heptad repeats that are modified suggests that, in principle, multiple protein factors can bind the CTD simultaneously to carry out several co-transcriptional processes at once. However, most CTD interactions are studied one at a time using peptide fragments of the heptad repeats and the big picture of protein factor binding to the CTD is currently unclear. What is apparent is that both the repetitive and disordered nature of the CTD is necessary to coordinate co-transcriptional processes with the activity of RNA polymerase II.

Co-transcriptional protein factor recruitment to the CTD and the synchronization of their binding to the transcription cycle is mediated by extensive post-translational modification of the heptad repeats. Easily the most well-studied modification is phosphorylation of the serine residues of the heptad repeat, especially serine 2 (S2) and serine 5 (S5), although phosphorylation of the tyrosine and threonine residues has

also been reported (Heidemann and Eick 2012). During the initiation phase of transcription, unmodified serine residues bind the Mediator complex to enable coordination of RNA polymerase II with promoter and enhancer elements. Phosphorylation of S5 by the Kin28p kinase component of TFIIH inhibits the Mediator-CTD interaction and leads to promoter escape and elongation by RNA polymerase II (Jeronimo and Robert 2014). The S5 mark is predominant early on in the transcription cycle and recruits factors active at this early stage such as the capping enzyme complex (Cho et al. 1997). The elongation phase is marked by S2 as well as S5 phosphorylation that recruits factors including Set2p and Spt6p that assist with elongation and polymerase passage through chromatin (Fuchs et al. 2012, Yoh et al. 2008). S5 phosphorylation is progressively removed throughout the elongation phase resulting in primarily an S2 phosphorylation signal during termination. This S2 phosphorylation signal is then recognized by polyadenylation and termination factors to complete transcription and export the resulting mRNA (Dunn et al. 2005). In addition to these two main residues, Y1, T4 and S7 phosphorylation has also been reported and is thought to influence the modification state of the S2 and S5 residues (Heidemann and Eick 2012). Furthermore, S7 phosphorylation in metazoans is required to recruit the Integrator complex to process snRNA transcription, indicating that additional phosphorylated residues have specific functions in more complex organisms (Simonti et al. 2015). Taken together, CTD heptad phosphorylation demonstrates the important role PTMs play in aligning the timing of protein factor binding to the appropriate stage of the transcription cycle.

There are a number of other CTD post-translational modifications in addition to phosphorylation that regulate protein factor binding to the heptad repeats. The two proline residues of the heptad repeat are subject to reversible isomerization, switching their conformation between *trans* and *cis*, greatly affecting the geometry of the CTD. CTD-binding proteins often prefer either the *cis* or the *trans* conformation, and proline isomerization serves as a switch to enable modification of other important residues (Albert et al. 1999). In one well-characterized example, the proline isomerase Ess1p switches the conformation of P6 from *trans* to *cis*, enabling the phosphatase Ssu72p to bind and dephosphorylate S5, marking the transition from elongation to termination (Werner-Allen et al. 2011). CTD repeats can also be glycosylated at many of the same residues that are phosphorylated and this modification is thought to act as a competitor to phosphorylation or to prevent aberrant modification of important residues (Lu et al. 2016). Finally, heptads in metazoans that contain serine to lysine substitutions at position seven have been shown to be modified by methylation and acetylation that further expand the regulatory potential of the CTD in complex organisms (Voss et al. 2015). These examples illustrate PTM cross-talk on the CTD and demonstrate how multiple PTMs work together to mediate progression through the transcription cycle.

PTM cross-talk also occurs between the modifications of the CTD and histone proteins and works to coordinate transcription in the context of chromatin. Around the stage of initiation S5 phosphorylation recruits the Set1p subunit of the COMPASS complex to methylate Histone H3 at the lysine 4 residue (H3K4me) (Ng et al. 2003). This histone methylation controls acetylation levels at the promoter region and also helps to recruit

the NSS complex to terminate cryptic transcription. Another methyltransferase, Set2p, is recruited by a combination of S2 and S5 phosphorylation during elongation to methylate Histone H3 at the lysine 36 residue (H3K36me) (Li et al. 2005). Set2p activity is also related to regulating histone acetylation levels, preventing cryptic transcription from firing in areas that RNA polymerase II has passed through. Set2p activity is performed in tandem with the histone chaperone and elongation factor Spt6p, which recognizes S2 phosphorylation and the histone deacetylase Rpd3p which recognizes S2 and S5 phosphorylation (Youdell et al. 2008, Govind et al. 2010). This tight interplay leads to simultaneous regulation of both transcriptional activity and chromatin structure through the post-translational modification of both the CTD and histone proteins.

The repetitive nature of the CTD is essential to its function, although curiously not all of the repeats are required for viability. Studies in yeast have demonstrated that only 12 out of the 26 repeats are necessary for normal growth, and as little as 8 repeats can support yeast viability under laboratory conditions (West and Corden 1995). Repeat requirements are also similar in mammalian CTDs, where the first 25 out of the total 52 repeats are sufficient for viability (Bartolomei et al. 1988). While the entire CTD does not appear to be required for growth, wildtype CTD repeat numbers are strongly selected for in nature, indicating that there are other roles for the CTD repeats aside from just maintaining growth and survival. Additionally, the actual functional unit of the CTD has been found to be two consecutive heptad repeats that have a proper tyrosine 1, serine 2 and serine 5 periodicity (Liu et al. 2008). This finding was confirmed by numerous structural studies that show CTD binding factors interacting with two or more heptad

repeats (Eick and Geyer 2013). These findings shed light on the specific structural properties of the CTD repeats and how they relate to its role as a scaffold for co-transcriptional processes.

Functional studies of the CTD have been instrumental in explaining how the various co-transcriptional processes are orchestrated in step with the transcriptional cycle to regulate gene expression. However, many questions remain concerning CTD structure and the role of the heptad repeats in coordinating transcription. One of the main questions concerns the structural arrangement of the CTD and how the many protein factors are aligned along the CTD to perform their functions. While some interactions are known to be sequential, whether these proteins occupy the CTD at the same time and wait their turn or bind in sequence is currently unknown. Understanding the arrangement of protein factors is complicated by the intrinsically disordered nature of the CTD, which makes it difficult to visualize the structure in the context of transcriptional complexes. Modeling the CTD interactions based upon genetic and biochemical data is therefore necessary to get a sense of the structural interactions taking place. The question of the seemingly redundant CTD repeats will also need to be addressed to explain the strong selection for the wildtype CTD lengths observed across many species. Finally, the role of sequence replacements in the heptad repeats will need to be examined in more depth to determine the role of specific repeats in the function of the CTD. Addressing these and other questions will be important to expand the field of CTD study from reports of single protein-protein interactions to a more global understanding of the CTD and of RNA polymerase II.

Overview of thesis aims and content

As summarized in the preceding sections, recent advances in the understanding of IDRs and repeats have greatly expanded our knowledge of the mechanisms of protein function. Previously dismissed as small linker sequences and junk DNA, IDRs and repeats have now emerged as prevalent players in the regulation of protein function for a number of important biological processes. The CTD of RNA polymerase II in particular is a prime example of the properties of IDRs and repeats at work to regulate the complex molecular mechanisms of transcription. However, a number of questions remain regarding the variation, structure and function of IDRs and tandem repeats. The overall aim of the work presented in this thesis is to address these structure/function relationships of IDRs and tandem repeats using the budding yeast *Saccharomyces cerevisiae* as a model organism. The big picture approaches used within shed light on general organizing principles of IDR and repeat function to guide the design and execution of further experiments.

The following three chapters of this thesis detail the research on first IDR and then repeat function in budding yeast. Chapter 2 presents an analysis of IDR variation among 93 strains of wild and laboratory budding yeast. The majority of IDR variation was found to be associated with tandem repeats, including both microsatellites and minisatellites. Further characterization of these variable repetitive IDRs demonstrates that they are highly diverse and conserved among budding yeasts, suggesting an evolutionarily conserved function for these repeats. This work characterizes extensive IDR variation in budding yeast and suggests a tandem repeat-based mechanism for the

diversity observed in the wild. Chapters 3 and 4 concern the disordered and repetitive CTD of RNA polymerase II. The heptad repeats of the budding yeast CTD were found to have region specific roles and were not redundant as had been previously assumed. Chapter 3 presents the genetic basis for the model of region specificity while chapter 4 details follow-up work to establish a mechanism of the specific functions of the CTD repeats. Finally, chapter 5 introduces the perspectives of this work and outlines several directions of future study to further develop our understanding of intrinsically disordered regions and tandem repeats.

Chapter 1 Literature Cited

Ahmad Y, Neel H, Lamond AI, Bertrand E. HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. Mol Cell. 2010 Sep 24;39(6):912-924.

Albert A, Lavoie S, Vincent M. A hyperphosphorylated form of RNA polymerase II is the major interphase antigen of the phosphoprotein antibody MPM-2 and interacts with the peptidyl-prolyl isomerase Pin1. J Cell Sci. 1999 Aug;112 (Pt 15):2493-500.

Albrecht A, Mundlos S. The other trinucleotide repeat: polyalanine expansion disorders. Curr Opin Genet Dev. 2005 Jun;15(3):285-93.

Allen BL, Taatjes DJ. The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol. 2015 Mar;16(3):155-66.

Allison LA, Moyle M, Shales M, Ingles CJ. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell. 1985 Sep;42(2):599-610.

Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973 Jul 20;181(4096):223-30.

Arimbasseri GA. Interactions between RNAP III transcription machinery and tRNA processing factors. Biochim Biophys Acta. 2018 Apr;1861(4):354-360.

Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015 Jan;40(1):49-57.

Bartolomei MS, Halden NF, Cullen CR, Corden JL. Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. Mol Cell Biol. 1988 Jan;8(1):330-9.
Boulon S, Pradet-Balade B, Verheggen C, Molle D, Boireau S, Georgieva M, Azzag K, Robert MC,

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet. 1998 Jun;62(6):1408-15.

Bunch H. RNA polymerase II pausing and transcriptional regulation of the HSP70 expression. Eur J Cell Biol. 2017 Dec;96(8):739-745.

Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15(9):956-63.

Cho EJ, Takagi T, Moore CR, Buratowski S. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev. 1997 Dec 15;11(24):3319-26.

Choder M, Young RA. A portion of RNA polymerase II molecules has a component essential for stress responses and stress survival. Mol Cell Biol. 1993 Nov;13(11):6984-91.

Choi JY, Roush WR. Structure Based Design of CYP51 Inhibitors. Curr Top Med Chem. 2017;17(1):30-39.

Corden, J. L., 2013 RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem Rev 113: 8423-8455.

Cramer P. Multisubunit RNA polymerases. Curr Opin Struct Biol. 2002 Feb;12(1):89-97.

Cramer P, Bushnell DA, Kornberg RD. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science. 2001 Jun 8;292(5523):1863-76.

Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71.

Dunn EF, Hammell CM, Hodge CA, Cole CN. Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. Genes Dev. 2005 Jan 1;19(1):90-103.

Eick D, Geyer M. The RNA polymerase II carboxy-terminal domain (CTD) code. Chem Rev. 2013 Nov 13;113(11):8456-90.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 2014 Nov 15;23(22):5866-78.

Fan H, Chu JY. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics. 2007 Feb;5(1):7-14.

Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008 May;9(5):397-405.

Fuchs SM, Kizer KO, Braberg H, Krogan NJ, Strahl BD. RNA polymerase II carboxyl-terminal domain phosphorylation regulates protein stability of the Set2 methyltransferase and histone H3 di- and trimethylation at lysine 36. J Biol Chem. 2012 Jan 27;287(5):3249-56.

Fuchs SM. Chemically modified tandem repeats in proteins: natural combinatorial peptide libraries. ACS Chem Biol. 2013 Feb 15;8(2):275-82.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445-77.

Gemayel R, Yang Y, Dzialo MC, Kominek J, Vowinckel J, Saels V, Van Huffel L, van der Zande E, Ralser M, Steensels J, Voordeckers K, Verstrepen KJ. Variable repeats in the eukaryotic polyubiquitin gene ubi4 modulate proteostasis and stress survival. Nat Commun. 2017 Aug 30;8(1):397.

Gómez-Navarro N, Peiró-Chova L, Estruch F. Iwr1 facilitates RNA polymerase II dynamics during transcription elongation. Biochim Biophys Acta. 2017 Jul;1860(7):803-811.

Govind CK, Qiu H, Ginsburg DS, Ruan C, Hofmeyer K, Hu C, Swaminathan V, Workman JL, Li B, Hinnebusch AG. Phosphorylated Pol II CTD recruits multiple HDACs, including Rpd3C(S), for methylation-dependent deacetylation of ORF nucleosomes. Mol Cell. 2010 Jul 30;39(2):234-46.

Grummt I, Längst G. Epigenetic control of RNA polymerase I transcription in mammalian cells. Biochim Biophys Acta. 2013 Mar-Apr;1829(3-4):393-404.

Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev. 2014 Jul 9;114(13):6561-88.

Hantsche M, Cramer P. Conserved RNA polymerase II initiation complex structure. Curr Opin Struct Biol. 2017 Dec;47:17-22.

Hartl DL. Molecular melodies in high and low C. Nat Rev Genet. 2000 Nov;1(2):145-9.

Heidemann M, Eick D. Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code. RNA Biol. 2012 Sep;9(9):1144-6.

Holstege FC, van der Vliet PC, Timmers HT. Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. EMBO J. 1996 Apr 1;15(7):1666-77.

Huang Y, Liu Z. Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism. J Mol Biol. 2009 Nov 13;393(5):1143-59.

Huber, R. & Bode, W. Structural basis of the activation and action of trypsin. Acc Chem Res 1978 11, 114–122

Iakoucheva LM, Kimzey AL, Masselon CD, Bruce JE, Garner EC, Brown CJ, Dunker AK, Smith RD, Ackerman EJ. Identification of intrinsic order and disorder in the DNA repair protein XPA. Protein Sci. 2001 Mar;10(3):560-71.

Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010 Feb;11(2):97-108.

Jeronimo C, Robert F. Kin28 regulates the transient association of Mediator with core promoters. Nat Struct Mol Biol. 2014 May;21(5):449-55.

Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015 Mar;16(3):167-77.

Jorda J1, Xue B, Uversky VN, Kajava AV. Protein tandem repeats - the more perfect, the less structured. FEBS J. 2010 Jun;277(12):2673-82.

Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution. 2001 Jan;55(1):1-24.

Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A. 1958 Feb;44(2):98-104.

Kosol S, Contreras-Martos S, Cedeño C, Tompa P. Structural characterization of intrinsically disordered proteins by NMR spectroscopy. Molecules. 2013 Sep 4;18(9):10802-28.

Kouzarides T. Chromatin modifications and their function. Cell. 2007 Feb 23;128(4):693-705.

Kuras L, Kosa P, Mencia M, Struhl K. TAF-Containing and TAF-independent forms of transcriptionally active TBP in vivo. Science. 2000 May 19;288(5469):1244-8.

Kuser P, Cupri F, Bleicher L, Polikarpov I. Crystal structure of yeast hexokinase PI in complex with glucose: A classical "induced fit" example revised. Proteins. 2008 Aug;72(2):731-40.

La Spada AR, Taylor JP. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet. 2010 Apr;11(4):247-58.

Li M, Phatnani HP, Guan Z, Sage H, Greenleaf AL, Zhou P. Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. Proc Natl Acad Sci U S A. 2005 Dec 6;102(49):17636-41.

Liang KC, Tseng JT, Tsai SJ, Sun HS. Characterization and distribution of repetitive elements in association with genes in the human genome. Comput Biol Chem. 2015 Aug;57:29-38.

Lisica A, Engel C, Jahnel M, Roldán É, Galburt EA, Cramer P, Grill SW. Mechanisms of backtrack recovery by RNA polymerases I and II. Proc Natl Acad Sci U S A. 2016 Mar 15;113(11):2946-51.

Liu P, Greenleaf AL, Stiller JW. The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. Mol Biol Evol. 2008 Apr;25(4):719-27.

Lee HE, Ayarpadikannan S, Kim HS. Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates. Genes Genet Syst. 2015;90(5):245-57.

López Castel A, Cleary JD, Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol. 2010 Mar;11(3):165-70.

Lu L, Fan D, Hu CW, Worth M, Ma ZX, Jiang J. Distributive O-GlcNAcylation on the Highly Repetitive C-Terminal Domain of RNA Polymerase II. Biochemistry. 2016 Feb 23;55(7):1149-58.

Mbogning J, Pagé V, Burston J, Schwenger E, Fisher RP, Schwer B, Shuman S, Tanny JC. Functional interaction of Rpb1 and Spt5 C-terminal domains in co-transcriptional histone modification. Nucleic Acids Res. 2015 Nov 16;43(20):9766-75.

Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins. 2003;53 Suppl 6:561-5.

Merkel A, Gemmell N. Detecting short tandem repeats from genome data: opening the software black box. Brief Bioinform. 2008 Sep;9(5):355-66.

Miliara X, Matthews S. Structural comparison of yeast and human intra-mitochondrial lipid transport systems. Biochem Soc Trans. 2016 Apr 15;44(2):479-85.

Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A. 2007 May 15;104(20):8472-7.

Naftelberg S, Schor IE, Ast G, Kornblihtt AR. Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu Rev Biochem. 2015;84:165-98.

Ng HH, Robert F, Young RA, Struhl K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell. 2003 Mar;11(3):709-19.

Ohno S. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol. 1999 Oct;10(5):517-22.

Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry. 2005 Sep 20;44(37):12454-70.

Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980 Apr 17;284(5757):604-7.

Patel DJ, Wang Z. Readout of epigenetic modifications. Annu Rev Biochem. 2013;82:81-118.

Pâques F, Leung WY, Haber JE. Expansions and contractions in a tandem repeat induced by double-strand break repair. Mol Cell Biol. 1998 Apr;18(4):2045-54.

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006 Apr 17;7:208.

Peti W, Nairn AC, Page R. Folding of Intrinsically Disordered Protein Phosphatase 1 Regulatory Proteins. Curr Phys Chem. 2012 Jan;2(1):107-114.

Portz B, Lu F, Gibbs EB, Mayfield JE, Rachel Mehaffey M, Zhang YJ, Brodbelt JS, Showalter SA, Gilmour DS. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nat Commun. 2017 May 12;8:15231.

Porrua O, Libri D. Transcription termination and the control of the transcriptome: why, where and how to stop. Nat Rev Mol Cell Biol. 2015 Mar;16(3):190-202.

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics. 2005 Aug 15;21(16):3435-8.

Richard GF, Pâques F. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. 2000 Aug;1(2):122-6.

Rothbart SB, Strahl BD. Interpreting the language of histone and DNA modifications. Biochim Biophys Acta. 2014 Aug;1839(8):627-43.

Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA Modifications in Gene Expression Regulation. Cell. 2017 Jun 15;169(7):1187-1200.

Rundlett SE, Carmen AA, Kobayashi R, Bavykin S, Turner BM, Grunstein M. HDA1 and RPD3 are members of distinct yeast histone deacetylase complexes that regulate silencing and transcription. Proc Natl Acad Sci U S A. 1996 Dec 10;93(25):14503-8.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 2006 Mar 16;440(7082):341-5.

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. 2007 Jan;35(Database issue):D786-93.

Simon M, Hancock JM. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol. 2009;10(6):R59.

Simonti CN, Pollard KS, Schröder S, He D, Bruneau BG, Ott M, Capra JA. Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. BMC Evol Biol. 2015 Mar 10;15:35.

Shi Y. A glimpse of structural biology through X-ray crystallography. Cell. 2014 Nov 20;159(5):995-1014.

Szasz CS, Alexa A, Toth K, Rakacs M, Langowski J, Tompa P. Protein disorder prevails under crowded conditions. Biochemistry. 2011 Jul 5;50(26):5834-44.

Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. Bioessays. 2003 Sep;25(9):847-55.

Tompa P, Dosztanyi Z, Simon I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. J Proteome Res. 2006 Aug;5(8):1996-2000.

Tudek A, Porrua O, Kabzinski T, Lidschreiber M, Kubicek K, Fortova A, Lacroute F, Vanacova S, Cramer P, Stefl R, Libri D. Molecular basis for coordinating transcription termination with noncoding RNA degradation. Mol Cell. 2014 Aug 7;55(3):467-81.

Uversky VN. What does it mean to be natively unfolded? Eur J Biochem. 2002 Jan;269(1):2-12.

Zentner GE, Henikoff S. Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol. 2013 Mar;20(3):259-66.

Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol. 2015 Mar;16(3):178-89.

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nat Genet. 2005 Sep;37(9):986-90.

Voss K, Forné I, Descostes N, Hintermair C, Schüller R, Maqbool MA, Heidemann M, Flatley A, Imhof A, Gut M, Gut I, Kremmer E, Andrau JC, Eick D. Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II. Transcription. 2015;6(5):91-101.

Werner-Allen JW, Lee CJ, Liu P, Nicely NI, Wang S, Greenleaf AL, Zhou P. cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. J Biol Chem. 2011 Feb 18;286(7):5717-26.

West ML, Corden JL. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics. 1995 Aug;140(4):1223-33.

Yang C, Stiller JW. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A. 2014 Apr 22;111(16):5920-5.

Yoh SM, Lucas JS, Jones KA. The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. Genes Dev. 2008 Dec 15;22(24):3422-34.

Youdell ML, Kizer KO, Kisseleva-Romanova E, Fuchs SM, Duro E, Strahl BD, Mellor J. Roles for Ctk1 and Spt6 in regulating the different methylation states of histone H3 lysine 36. Mol Cell Biol. 2008 Aug;28(16):4915-26.

Zhou M, Law JA. RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules. Curr Opin Plant Biol. 2015 Oct;27:154-64.

Chapter 2

Tandem repeats drive variation of intrinsically disordered regions in budding yeast[1]

Michael Babokhov, Bradley I. Reinfeld, Kevin Hackbarth, Yotam Bentov and Stephen M. Fuchs

Author contributions:

Michael Babokhov performed the analysis of IDR and repeat variation and wrote the manuscript.

Bradley I. Reinfeld conceived experiments and worked with the XSTREAM algorithm to obtain a list of repetitive sequences.

Kevin Hackbarth contributed to the analysis of IDR variation.

Yotam Bentov analyzed repeat sequences and created the list of IDRs.

Stephen M. Fuchs conceived and directed the study and contributed to writing the manuscript.

[1]Manuscript submitted for publication at FEBS Journal

Abstract

Copy-number variation in tandem repeat coding regions is more prevalent in eukaryotic genomes than current literature suggests. We have reexamined the genomes of nearly 100 yeast strains looking to map regions of repeat variation. From this analysis we have identified that length variation is highly correlated to intrinsically disordered regions (IDRs). Furthermore, the majority of length variation is associated with tandem repeats. These repetitive regions are rich in homopolymeric amino acid sequences but nearly half of the variation comes from longer-repeating motifs. Comparisons of repeat copy number and sequence between strains of budding yeast as well as closely related fungi suggest selection for and conservation of IDR-related tandem repeats. In some instances, repeat variation has been demonstrated to mediate binding affinity, aggregation, and protein stability. With this analysis, we can identify proteins for which repeat variation may play conserved roles in modulating protein function.

Introduction

Understanding how proteins carry out diverse functions and how these functions are regulated by the cell is a critical challenge of biology. Most proteins are produced as a linear polymer of amino acids that fold into three-dimensional structures and this structure is generally thought to determine protein function. Intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) within proteins are sequences that generally do not adopt a single defined configuration. Recent evidence has now demonstrated that conformational disorder plays an important regulatory role in tuning protein interactions and stability (Habchi, Tompa et al., 2014, Tompa, 2012).

Our lab has been interested in a subset of IDRs that consist of repetitive amino acid sequences such as the repetitive C-terminal domain of RNA polymerase II Rpb1p (Babokhov, Mosaheb et al., 2018, Morrill, Exner et al., 2016). IDRs are generally enriched for amino acids that are not structure promoting (Campen, Williams et al., 2008). Curiously, these same amino acids are also often enriched in repetitive amino acid sequences in proteins (Simon & Hancock, 2009). A unique aspect of repetitive amino acid sequences is that they are often encoded by repetitive DNA. Repetitive DNA sequences are known to be genetically unstable, often resulting in expansions and contractions within the genomic sequence (Gemayel, Vinces et al., 2010, Richard & Dujon, 2006). Studies of repetitive regions have generally focused on trinucleotide repeat sequences (Albrecht & Mundlos, 2005, La Spada & Taylor, 2010) but our group recently showed that DNA encoding longer repetitive sequences also showed genetic instability (Morrill et al., 2016). The genetic diversity that could result from repeat instability is now being realized as a potential important player in complex traits. Additionally, previous studies have hinted at significant overlaps between repetitive sequences and IDRs (Jorda, Xue et al., 2010, Simon & Hancock, 2009, Tompa, 2003). Thus, the primary goal of the work described below was to determine whether repetitive sequences are a general feature of IDRs and how they might function to tune IDR function.

We hypothesize that genetic instability in repetitive regions generally contributes to population-level genetic variation. In this work, we analyzed existing genomic sequencing data from 93 *Saccharomyces cerevisiae* genomes (Strope, Skelly et al.,

2015) and additional available data from related yeast species (Scannell, Zill et al., 2011) to determine whether repeat variation may be a significant contributor to the function of IDR domains. In brief, we found length polymorphisms in nearly 10% of yeast IDR domains. The vast majority of this variation derives from copy number variation in amino acid tandem repeats contained within these IDRs. Copy number variation is extensive and most commonly found in homopolymeric amino acid repeats and larger oligopeptide (>5 amino acid) repeat motifs. Lastly, variable repeats within IDRs are highly conserved across yeast species, suggesting an important biological function for these sequences. We propose that the genetic variation caused by repetitive sequences would further expand the regulatory features of IDRs in proteins.

Methods

Prediction of disordered and tandem repeat regions

Variation within intrinsic disordered regions was measured using the data available from 93 recently sequenced *S. cerevisiae* genomes (Strope et al., 2015). Genomic sequences were acquired and the open reading frames of 5,860 annotated genes were aligned in Geneious v. 10.2.3 [Biomatters Ltd.] using MAFFT v. 7.308 with the default settings (Katoh, Misawa et al., 2002, Katoh & Standley, 2013) and corrected manually when necessary. From this data set we removed regions associated with retrotransposons and dubious open reading frames. Disordered regions in proteins were compiled using the VSL2B disorder prediction algorithm using *S. cerevisiae* reference genome release 63.3 Saccharomyces Genome Database and was restricted to regions that were at least 30 amino acids in length (Oates, Romero et al., 2013, Peng

et al., 2006). Each IDR was given a unique identifier based on its start position and assigned as either variable or non-variable based on visual inspection of the alignments.

Tandem repeats were acquired using the XSTREAM repeat prediction algorithm (Newman & Cooper, 2007). We used XSTREAM parameters were set as: minimum character identity, I = 0.7; minimum consensus match, I = 0.8; maximum consecutive gaps, g = 3; minimum period, MinP = 1; minimum length, L = 5; any other settings were set to default. These settings were chosen to be very inclusive in order to identify both short perfect repeats and longer degenerate repeats in the reference proteome. A minimum overall length for the repeat region was set at five amino acids (Chavali et al., 2017). Using the data from XSTREAM each IDR was assigned as either containing a tandem repeat (Y) or not (N). Note in supplemental tables that some longer IDRs contained more than one tandem repeat region as called by XSTREAM. Variation in each unique repeat was recorded separately (Table S1).

Assessment of IDR variation

The presence or absence of repeat length variation at each individual IDR was assessed using the MAFFT alignments of the 100 yeast genomes. Alignments of each gene in the Geneious browser were examined manually for the presence of gaps that indicated sequence variation between the different yeast strains. IDRs that had at least one gap within their sequence range were scored as variable and the number of variable regions and the variable protein motifs were recorded. Only gaps in the

alignment from insertions and deletions (indels) were examined. Variation due to single nucleotide substitution was excluded from the current study. The range of variation present at a given IDR was calculated as the difference in amino acid length between the longest and shortest forms of the variable sequence. We also calculated the length variance for each variable IDR. Consequently, the repetitiveness of each variable IDR was also annotated as either repetitive or not and the number repeats per IDR as well as their sequences were noted.

A number of highly variable repeats were realigned using nine kilobases of flanking sequence in order to get accurate frequency counts. Even after realignment, a subset of 31 hypervariable repeats still did not yield sufficient alignments to get frequency data and were marked as variable but excluded from further analysis (Appendix 2.A). The frequency data for each variable repeat was plotted and compared to a theoretical Poisson distribution for an equivalently sized dataset with mean in GraphPad prism by a chi-squared test. Variable tandem repeats were then categorized as either having a Poisson or a non-Poisson distribution.

Determination of variable repeat conservation across Saccharomyces sensu stricto

Alignments of annotated genes from high-quality genomes of four *Saccharomyces* species closely related to budding yeast (Scannell et al., 2011) were downloaded and compared to the alignments from the 100 yeast genomes resource. The corresponding positions of each IDR identified as variable in *S. cerevisiae* were examined manually in the alignments of the other four species to look for variable repeats. If the tandem

repeat was present in both the *S. cerevisiae* and at least one of the *sensu stricto* alignments, then the repeat was classified as conserved. Occasionally, the repeat sequence was not conserved but there was a different sequence that was still repetitive at the corresponding location. In these cases, the repetitiveness was classified as conserved while the sequence was not. A subset of variable repeats could not be classified because there was no corresponding alignment file in the *sensu stricto* dataset.

Results and Discussion

A pipeline to identify IDR variation

Given the previously hinted association between tandem repeats and IDRs (Jorda et al., 2010, Simon & Hancock, 2009, Tompa, 2003), we developed a pipeline to characterize IDR and repeat variation in budding yeast (Figure 1). A list of IDRs was compiled using the VSL2B algorithm for 5,860 annotated genes to yield 7,531 predicted IDR sequences using a minimum IDR length of 30 amino acids to approximate long IDRs (Peng, Radivojac et al., 2006). High quality genomic sequences from the 100 yeast genomes project (Strope et al., 2015) were aligned and the predicted IDRs were scanned for length variation. IDRs were identified as either variable or non-variable and the presence of tandem repeats was noted and compared to predictions based on the XSTREAM algorithm (Newman & Cooper, 2007). The resulting dataset characterizes the variation and repetitiveness for all IDRs in our pipeline (supplementary Table S1 online). IDRs identified as both variable and repetitive were then further analyzed to determine selection and conservation.

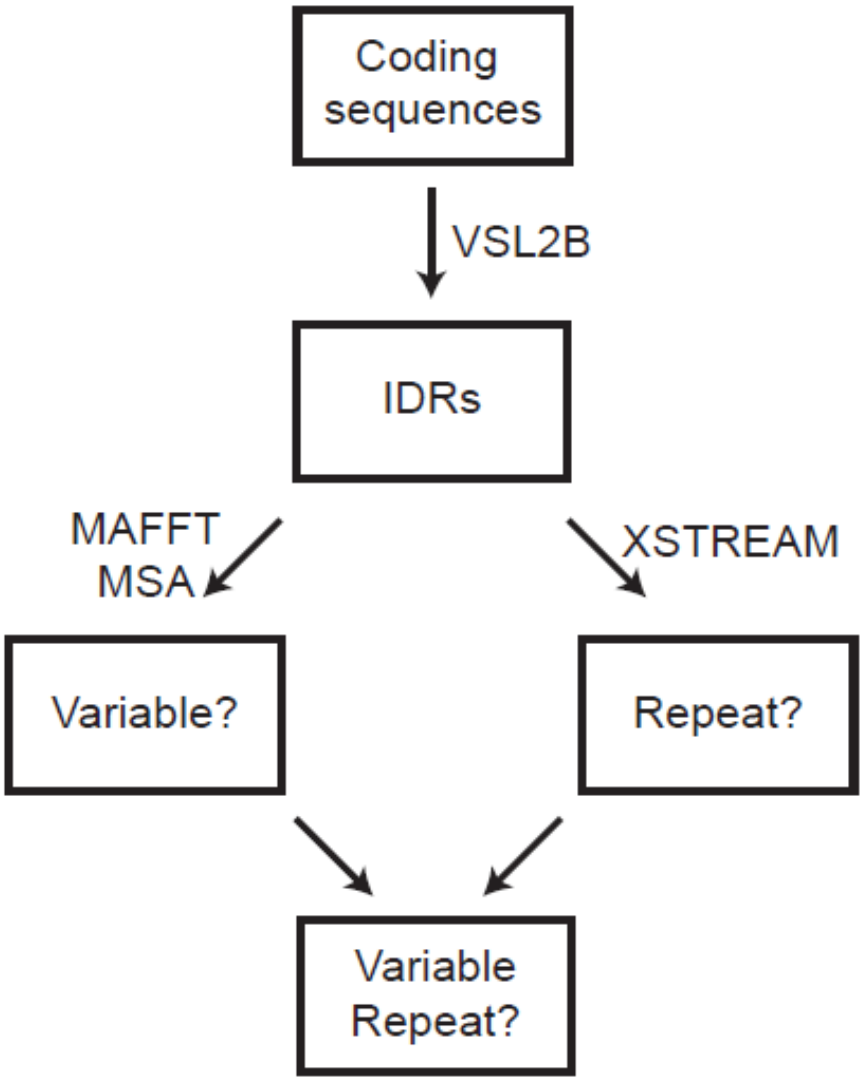


Figure 1. Flowchart of variable IDR identification. Coding sequences for all open reading frames of *S. cerevisiae* were inputted into VSL2B disorder predictor (Peng et al., 2006) to obtain a list of IDRs. The resulting sequences were aligned using MAFFT multiple sequence alignment (Katoh & Standley, 2013) to uncover length polymorphisms. In parallel, the repeat-finding algorithm XSTREAM (Newman & Cooper, 2007) was used to create a list of all repetitive sequences within IDRs. The overlap between these two data sets was then curated to identify variable repetitive sequences.

IDR variation is associated with tandem repeats

Using the pipeline described above, we identified 899 IDRs that exhibited length polymorphisms in at least one of the 93 genomes examined (Figure 2A). Many of these polymorphic IDRs varied by only a single amino acid where others exhibited differences of more than 100 amino acids between the 93 examined sequences. Many proteins contained multiple IDRs and of these, 105 contained more than one polymorphic IDR. These findings are in line with previous studies that show genetic instability in IDR coding regions (Brown, Takayama et al., 2002, Nilsson, Grahn et al., 2011) and indicate the widespread variation that exists within natural and laboratory strains of budding yeast.

IDRs and tandem repeats in proteins share many similarities including enrichment for small polar amino acids and high rates of genetic mutation. We therefore examined whether tandem repeat variation was responsible for the polymorphisms observed in IDRs. Using XSTREAM we identified 645 variable tandem repeats within the 899 polymorphic IDRs (~72%) (Figure 2A). This number is likely conservative as we identified tandem repeats as being at least five consecutive amino acids for homorepeats (Chavali, Chavali et al., 2017) and greater than two repetitions of larger repeats. Overlooked motifs with smaller copy numbers or sequence substitutions may represent degenerated repeats that could still retain functional significance. The remaining 254 variable IDRs were generally short duplications of sequence that were observed in a small subset of the 93 strains for a given IDR. This variation may

represent genuine polymorphisms or may simply be an artifact of short-read next generation sequencing.

Close to half of all variable IDRs contained polymorphic homorepeats, commonly polyQ or polyN, which were highly polymorphic in length across the 93 strains. These homorepeats most frequently demonstrated poly-Q, -N, -D and -E sequences (Figure 2B and Table S1), in line with previous findings on both repeat and IDR-enriched amino acids (Campen et al., 2008, Gemayel et al., 2010). Repeat motifs of homorepeats were also differentially enriched between variable and non-variable IDRs. Homorepeats in variable IDRs were heavily skewed towards polyQ and polyN, while non-variable IDRs had many more polyS and polyK repeats (Figure 2C). Further analysis of these different enrichments would help uncover why particular repeats are variable within IDRs.

Trinucleotide repeats consisting of repeating sequences of a single codon are well known to be genetically unstable and we generally expected that this mechanism would be responsible for the majority of the homorepeat IDR variation genome-wide (Usdin, House et al., 2015). Interestingly, we uncovered several variable homorepeats that were not encoded by trinucleotide repeats signifying variation in these regions is governed by more complex mechanisms than was previously suggested. Appendix 2.B shows the codon distribution for long and short forms of polyQ repeats. In addition to these low-complexity homorepeats we found that repeating sequences made up of more than a single amino acid comprised nearly half of the variable repeats (Figure 2B). Larger motifs containing five or more amino acids were most frequently found to be variable in

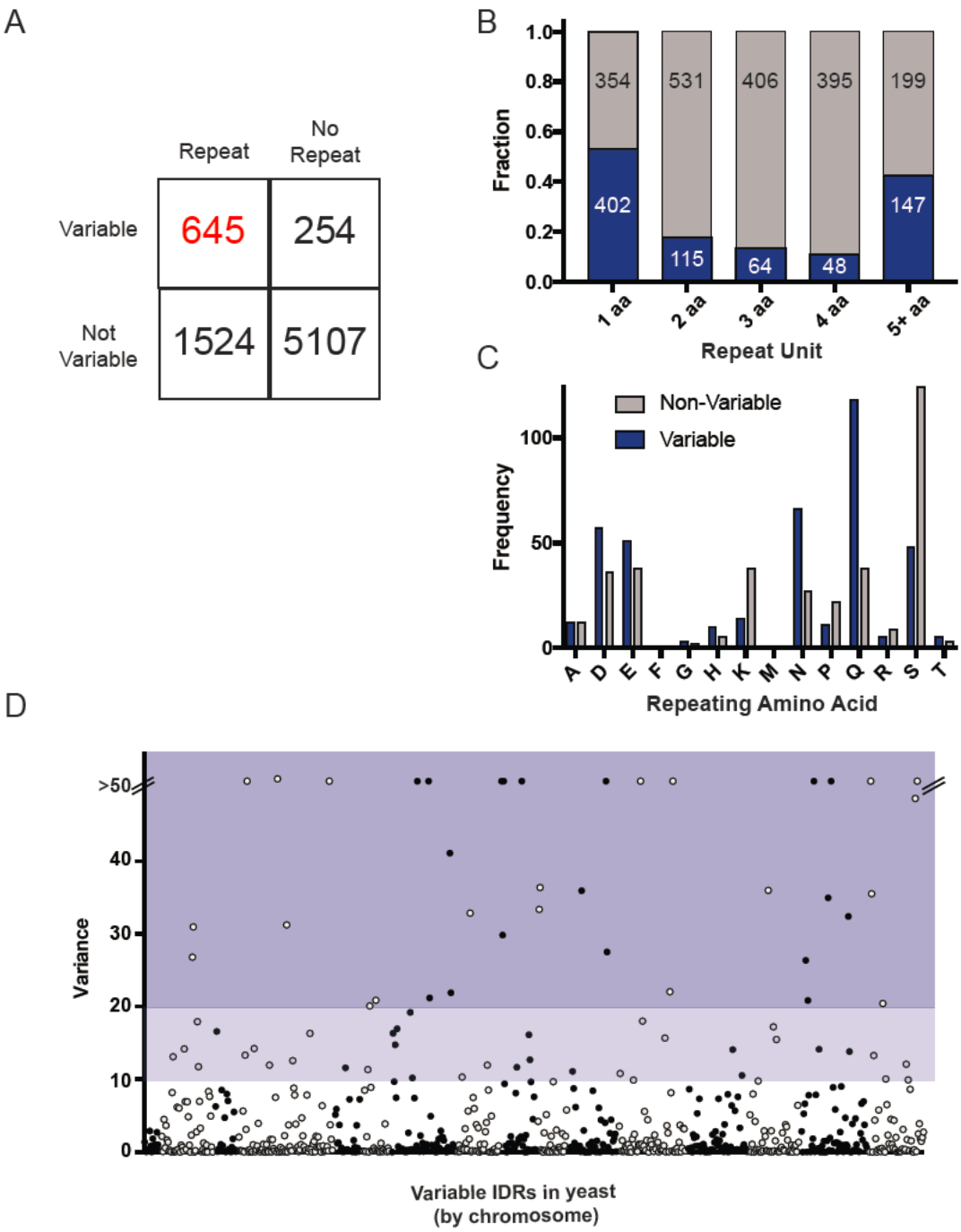


Figure 2. Characteristics of variable repeats in IDRs. **A**) Total numbers of identified IDRs based on repetitiveness and variability. **B**) Fraction of variable and non-variable repeats for each class of repeat unit length. **C**) Total numbers of variable and non-variable single amino acid repeats broken down by the repeated residue. **D**) Manhattan plot of length variances of variable IDRs by chromosome. The two shaded regions represent the top 5% (dark) and 10% (light) of IDRs with respect to variance.

length, while motifs of two to four amino acids were more likely to be non-variable. Finally, we also identified small linear duplications that created new repeating motifs leading to IDR length variation (Appendix 2.C). These sequences may be examples of nascent repeats and it would be interesting to further explore these motifs to determine whether there are specific genomic features that contribute to the genesis of new repeats.

As stated above, we found many IDRs for which there was a wide variance in length, due to polymorphism in a repetitive region. In some cases we also found several length polymorphisms but they were restricted to just a few individuals. We therefore calculated the sample variance of repeat length to represent both the repeat length polymorphisms and the frequency of a given variant within the population (Table S1). Visualizing the data on a Manhattan plot shows the variances for all 645 variable tandem repeat IDRs across the 16 chromosomes of budding yeast (Figure 2D). Any variation in an IDR may be relevant for protein function but we stratified this data to highlight the IDRs with the greatest variance (top 5 and 10% as shaded boxes in Figure 2D). To further ensure the variation reported in the next-generation sequences were true representations of the genomic sequence we PCR amplified repeat regions and subjected them to Sanger sequencing. As an illustrative example, we present the variation of the C-terminal repeat of the *MNN4* gene (Figure 3). *MNN4* is known to be polymorphic in populations (Carvalho-Netto, Carazzolle et al., 2013) as seen by PCR amplification of genomic DNA, and encodes a protein important for mannosylphosphorylation of yeast cell wall proteins (Kim, Kang et al., 2017) and

represents an interesting class of polyampholyte IDR domains (Das & Pappu, 2013, Sickmeier, Hamilton et al., 2007). Alignment of the resulting Sanger sequencing revealed extensive rearrangements of the repetitive motif and residue substitutions among four sample stains (Figure 3B), with the overall distribution of copy number variation seen in Figure 3C. We have performed similar analysis on several other IDRs, confirming to us that the variation reported in the next-generation sequence data largely represents true genetic polymorphism and not errors from sequencing or genome assembly.

Variable and repetitive IDRs are conserved across Saccharomyces species

We and others have proposed that variation in repetitive regions is an important driver of protein diversification (Fuchs, 2013, Gemayel, Yang et al., 2017, Morrill et al., 2016, Rogers, McConnell et al., 2017, Verstrepen, Jansen et al., 2005, Zhao, Strope et al., 2014). However, it is also possible that some or many of these repeats have randomly risen within the genome and thus are unlikely to contribute to protein function. Conservation of a repeat across different species would suggest it plays some functional role. To test this assertion, we examined evolutionary conservation of variable repeats across four closely related species of the *Saccharomyces sensu stricto* (*SSS*) genus using existing high quality genome data (Scannell et al., 2011). We found that 573 of 645 variable IDRs (~89%) were found in at least one other species of the *SSS*. Such a high level of conservation argues that the repeat regions within IDRs are not the consequence of random mutagenesis but have important biological functions within budding yeasts. In addition to the conservation of the repeat sequence, we also

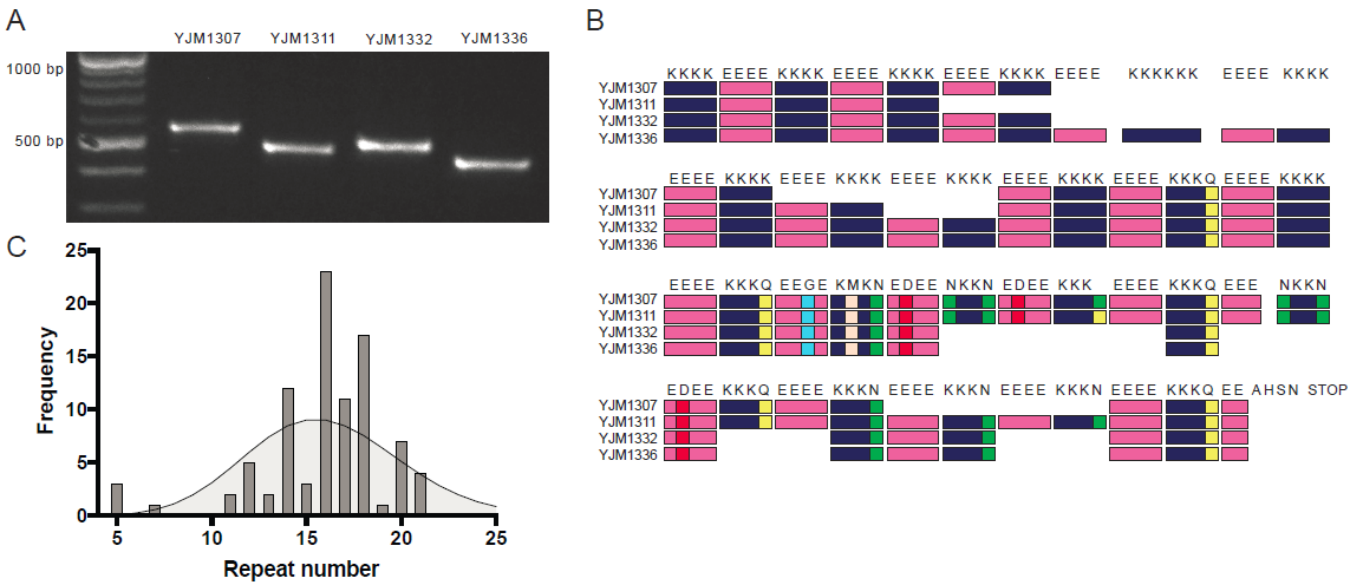


Figure 3. Variation in the C-terminal repeat of *MMN4*. **A**) Agarose gel of an amplified region of *MMN4* showing the length variation in the C-terminal repeat region of four representative strains. **B**) Diagram of *MMN4* repeat structure alignments derived from Sanger sequencing of the PCR products from A. **C**) Frequency distribution of all 93 *MMN4* repeats organized by repeat copy number. The curve represents a theoretical Poisson distribution calculated from the observed mean copy number.

observed copy number differences of tandem repeats across the *SSS*, indicating that repeat variation may play an important role in protein function across species as well as evolution/speciation (Table S3). However, we currently don't have enough representative samples to know the full range of repeat lengths that might exist for a given IDR in these other species. Further genomic sequencing and long-read next generation sequencing of *sensu stricto* strains and other fungi will be required to fully appreciate tandem repeat variation across evolutionary time.

In addition to completely conserved repeats, we found 18 IDRs that preserved the tandem repeat nature of the loci while having a different consensus sequence. A prime example of these repeats is located in an IDR of Pan1, where the 6-mer PIQPVQ repeat in *S. cerevisiae* is replaced with either a PAQ or PVQ trimer motif in other *sensu stricto* organisms (Appendix 2D). In all 18 cases, the difference in repeat motif between *SSS* sequences are conservative and thus we predict them to not differ greatly in function. We also identified 31 repeats that did not have corresponding sequences in the other *sensu stricto* genomes, and thus were novel to *S. cerevisiae*. All but three of these non-conserved repeats contained repeating motifs with longer periods (2 or more amino acids) and were absent from the other *SSS* species (Appendix 2D). These non-conserved repeats may have been acquired recently or may be required for an adaptation specific to *S. cerevisiae*. Lastly, we note that we were unable to characterize 52 variable repeats because the corresponding genes were absent in the *SSS* genomic sequences (Table S4). Overall, our analysis of tandem repeats across the *SSS*

demonstrates that variable repeats located within IDRs are highly conserved in budding yeasts, implicating them in biologically relevant functions.

Potential role for genetic selection in repetitive IDRs

Tandem repeat variation is thought to be a mechanism for rapid evolution of protein coding sequences (Marcotte, Pellegrini et al., 1999). As many variable IDRs are conserved, this suggests they may play some role in protein function. Therefore, we might expect two scenarios: the first is a situation where there is genetic selection for repeat length such that a few discrete repeat lengths would dominate the population. This might be the expectation of an IDR that functions to link two folded domains, for example, where the length is highly constrained by the function of the neighboring domains. The second scenario would be one where length polymorphism is neutral or even evolutionarily advantageous. In this scenario, we would expect to see some distribution of repeat lengths for the sample tested. We looked at the distribution of repeat lengths for each gene and determined that nearly all genes conformed to the first scenario, *i.e.* they showed variation but a high level of genetic selection for a preferred repeat length. To illustrate this, we present the length distribution of two orthologs in yeast which both harbor variable-length repetitive IDRs in their C-terminal domain, *VHS3* and *HAL3* (Figure 4). *VHS3* is among the most highly polymorphic IDRs we examined with repeat lengths ranging from 21 to 64 consecutive aspartic acid residues (Figure 4A) whereas *HAL3* had, on average, a much shorter polyD region and the distribution of lengths was more tightly clustered (Figure 4B).

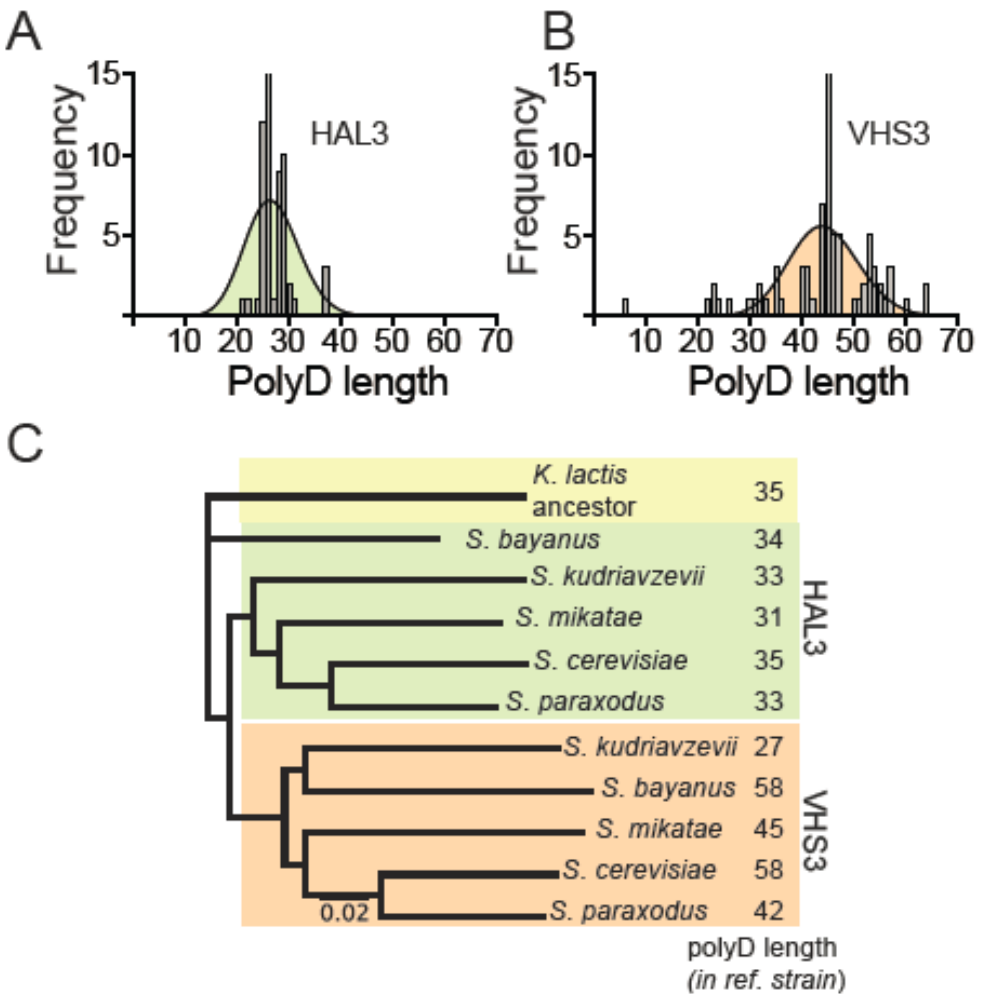


Figure 4. Comparison of polyD variation in paralogs *HAL3* and *VHS3*. **A**) Frequency distribution of HAL3 polyD homorepeat for all 93 yeast strains together with the theoretical Poisson distribution curve. **B**) Same frequency distribution for the VHS3 polyD homorepeat. **C**) Phylogenetic tree of genetic distances between the *Saccharomyces sensu stricto* species and the pre-whole genome duplication *Kluyveromyces lactis* ancestor. Distances were calculated based off of changes to *HAL3* and *VHS3* polyD homorepeat copy number listed to the right.

In order to estimate selective pressures on these two IDRs we compiled the polyD copy number for all of the *sensu stricto* yeasts and used repeat variation as a proxy for genetic distance between the species and *Kluyveromyces lactis*, an evolutionary ancestor that precedes the whole genome duplication seen in the *sensu stricto* clade (Wolfe & Shields, 1997). We saw that the polyD repeat encoded by both *VHS3* and *HAL3* is found in all organisms, including *K. lactis*. Overall, we found that the *VHS3* and *HAL3* variation among the *SSS* species recapitulated the pattern in *S. cerevisiae*: the *HAL3* polyD was tightly clustered while the *VHS3* polyD showed greater variation (Figure 4C). Interestingly, the genetic distances that we established based on repeat variation closely resembled the overall phylogenic relationship between the *sensu stricto* (Scannell et al., 2011), suggesting that variation in IDR repetitive sequences tracks with evolutionary changes.

Conclusion

In summary, we have uncovered extensive IDR length polymorphisms across 93 wild and laboratory strains of *S. cerevisiae*. The majority of the length differences between strains can be explained by copy number variation in both single amino acid and oligopeptide tandem repeat sequences. We have exhaustively characterized the variation that we observed and presented several illustrative examples of the trends within repetitive variable IDRs. Two key patterns have emerged as a result of our analysis. The first is that repeat variation is more complex than was previously appreciated, covering many different kinds of motifs at both the DNA and the protein level. The second is that although repetitive sequences are variable within *S. cerevisiae*,

they are highly conserved across budding yeast species, arguing for an important biological function for the repeats and perhaps their copy number variants. Consequently, our analysis of repeat length variation provides a foundation to further study both the mechanisms that lead to variation within IDRs and investigate the impact of length variation on protein function.

Appendix 2.A

Table 2.A-1. List of highly variable IDRs.

Gene	Nucleotide start	Nucleotide stop	Description
YAL063C	780	3912	Flo9
YAL065C	3	288	Similar to Flo genes
YAR050W	765	4554	Flo1
YBL007C	2817	3732	Sla1
YCR076C	303	750	Fub1
YCR089W	2364	2694	Fig2, Aga1 paralog
YDR080W	3	336	Vps41, vacuole protein
YDR420W	84	3708	Hkr1
YDR534C	156	345	Fit1 cell wall protein
YDR534C	378	552	
YDR534C	591	747	
YDR534C	789	1497	
YFL021W	276	987	Gat1 transcription factor
YFL067W	57	438	
YGL092W	3	1389	Nup145, nuclear pore protein
YHL050C	534	1014	
YHR211W	768	3171	Flo5
YIL115C	1179	3498	Nup159, nuclear pore protein
YIL169C	1785	2823	Css1, similar to HPF1
YIL169C	45	939	
YIL169C	1125	1395	
YIR019C	543	4023	Flo11
YJL159W	69	972	Hsp150, cell wall protein
YJR151C	321	3192	Dan4, cell wall protein
YKL164C	153	765	Pir1, cell wall protein
YKR092C	3	1011	Srp40
YKR102W	849	3066	Flo10
YMR164C	753	2274	Mss11, Flo gene transcription factor
YMR173W	3	1290	Ddr48
YMR317W	3	2532	
YMR317W	2574	3360	
YNR044W	219	2112	Aga1, cell wall protein
YOL155C	66	1029	HPF1, involved in aggregation
YOL155C	1833	2802	

Appendix 2.B

Table 2.B-1. Codon use of short and long variants of polyQ repeats.

	Shortest variant				Longest variant			
IDR	#Qs	#CAG	#CAA	%CAG	#Qs	#CAG	#CAA	%CAG
YBR016W	4	4	0	100.0	5	5	0	100.0
YBR108W	7	4	3	57.1	19	10	9	52.6
YBR112C	4	1	3	25.0	5	1	4	20.0
YBR135W	16	9	7	56.3	19	6	13	31.6
YBR212W	8	8	1	100.0	15	14	1	93.3
YBR289W ~200	6	1	5	16.7	14	1	13	7.1
YBR289W ~610	7	5	2	71.4	11	8	3	72.7
YBR289W ~790	30	14	16	46.7	46	23	23	50.0
YCR084C	9	6	3	66.7	16	10	6	62.5
YCR093W	5	0	5	0.0	6	0	6	0.0
YDL048C	4	1	3	25.0	8	3	5	37.5
YDL161W	6	1	5	16.7	10	2	8	20.0
YDL186W	6	2	4	33.3	7	2	5	28.6
YDR099W	13	3	10	23.1	20	5	15	25.0
YDR122W	5	5	0	100.0	9	9	0	100.0
YDR130C	6	4	2	66.7	9	6	0	66.7
YDR145W	5	5	0	100.0	6	6	0	100.0
YDR228C	12	9	3	75.0	24	13	11	54.2
YDR505C	10	1	9	10.0	19	4	15	21.1
YEL036C	20	7	13	35.0	29	8	21	27.6
YER109C	6	4	2	66.7	19	17	3	89.5
YER111C	4	0	4	0.0	7	0	7	0.0
YER177W	5	0	5	0.0	9	0	9	0.0
YFL024C	9	5	4	55.6	24	11	13	45.8
YFR008W	11	7	4	63.6	14	7	7	50.0
YFR019W	7	4	3	57.1	8	5	3	62.5
YGL025C	5	5	0	100.0	15	12	3	80.0
YGL036W	3	3	0	100.0	5	5	0	100.0
YGL066W	10	6	4	60.0	16	9	7	56.3
YGL237C	14	8	6	57.1	19	9	10	47.4
YGR009C	4	2	2	50.0	8	4	4	50.0
YGR119C	7	6	1	85.7	8	7	1	87.5
YGR249W	7	6	1	85.7	14	13	1	92.9
YHL002W	2	1	1	50.0	5	4	1	80.0
YHL015W	5	0	5	0.0	6	0	6	0.0
YHL020C	15	8	7	53.3	27	13	14	48.1
YHL025W	1	1	0	100.0	6	4	2	66.7
YHL027W	4	3	1	75.0	10	10	0	100.0
YHR030C	10	6	4	60.0	22	14	8	63.6
YHR135C	5	2	3	40.0	18	7	11	38.9
YHR149C	6	4	2	66.7	7	5	2	71.4
YHR161C	9	3	6	33.3	23	8	15	34.8
YHR186C	9	3	6	33.3	11	3	8	27.3
YHR200W	6	2	4	33.3	7	3	4	42.9
YIL105C	15	8	7	53.3	20	9	11	45.0

YIL152W	5	5	0	100.0	10	10	0	100.0
YIL156W	5	5	0	100.0	10	7	3	70.0
YIR006C	7	4	3	57.1	13	8	5	61.5
YIR023W	12	6	6	50.0	27	15	12	55.6
YJL019W	5	5	0	100.0	12	12	0	100.0
YJL141C	10	6	4	60.0	18	12	6	66.7
YJL162C	7	4	3	57.1	22	15	7	68.2
YJR086W	7	3	4	42.9	9	5	4	55.6
YJR127C	5	1	4	20.0	14	1	13	7.1
YKL032C	6	1	5	16.7	12	2	10	16.7
YKL054C	7	3	4	42.9	11	4	7	36.4
YKL088W	5	5	0	100.0	6	5	1	83.3
YKR045C	6	4	2	66.7	16	13	3	81.3
YLL013C	10	2	8	20.0	22	6	18	27.3
YLR095C	9	5	4	55.6	17	9	8	52.9
YLR177W	5	3	2	60.0	17	13	4	76.5
YLR207W	6	0	6	0.0	10	0	10	0.0
YLR228C	6	0	6	0.0	9	0	9	0.0
YLR256W	6	3	3	50.0	16	9	7	56.3
YLR278C	4	2	2	50.0	5	2	3	40.0
YLR437C	11	6	6	54.5	18	8	11	44.4
YML103C	5	1	4	20.0	8	1	7	12.5
YML113W	8	1	7	12.5	12	6	6	50.0
YMR002W	5	0	5	0.0	8	0	8	0.0
YMR016C	7	3	4	42.9	8	4	4	50.0
YMR043W	11	5	6	45.5	23	6	17	26.1
YMR047C	2	0	2	0.0	7	3	4	42.9
YMR124W	5	4	1	80.0	10	9	1	90.0
YMR164C	23	9	14	39.1	40	18	22	45.0
YNL016W	7	1	6	14.3	8	1	7	12.5
YNL154C	7	3	4	42.9	9	3	6	33.3
YNL161W	15	7	8	46.7	35	12	23	34.3
YNL298W	4	3	1	75.0	11	9	2	81.8
YNR052C	10	4	6	40.0	20	10	10	50.0
YOL051W	11	7	4	63.6	28	12	16	42.9
YOR113W	4	3	1	75.0	9	4	5	44.4
YOR267C	11	1	10	9.1	29	2	27	6.9
YOR329C	7	2	5	28.6	9	1	8	11.1
YOR359W	12	7	5	58.3	15	9	6	60.0
YOR372C	13	3	10	23.1	21	2	19	9.5
YPL016W	5	3	2	60.0	6	4	2	66.7
YPL026C	6	1	5	16.7	20	2	18	10.0
YPL049C	3	2	1	66.7	4	3	1	75.0
YPL190C	8	3	5	37.5	18	3	15	16.7
YPL229W	7	6	1	85.7	17	11	6	64.7
YPR022C	9	4	5	44.4	23	3	20	13.0
YPR065W	7	4	3	57.1	9	4	5	44.4
YPR154W	4	0	4	0.0	12	0	12	0.0
YPR185W	11	8	3	72.7	19	14	5	73.7

Appendix 2.C

Table 2.C-1. List of nascent repetitive sequences**.**

Gene	Nucleotide start	Nucleotide stop	Duplicated Sequence
YBL029W	174	327	SNNN
YBR083W	234	375	AT
YDL203C	3	708	QEKVVRT
YDR151C	834	975	SLAP
YDR206W	2055	2406	ASMPPS
YDR227W	3636	3873	TDSNT
YDR310C	2022	2433	NDTESA
YDR507C	1095	3129	DQEK
YEL037C	135	777	AEQPSTAATTA
YER008C	756	1737	LE
YER024W	561	726	DP
YER025W	3	264	QET
YFR046C	492	780	IS
YGL228W	1506	1731	ERK
YGR112W	783	951	EEHTRN
YGR162W	324	1227	AAGS, ST
YGR229C	1011	1515	EEMNKK
YHL008C	966	1719	LPHN
YHL020C	483	654	EQVNAS
YHR062C	714	879	GGGSGN
YHR066W	1044	1359	GE
YHR103W	3	975	RNTIK
YHR177W	456	894	QQQHR
YHR216W	1224	1326	GVPCMADGG
YIL101C	321	528	SATPH
YIR003W	3	2037	RS
YIR033W	2490	2862	YSISRK
YKL020C	933	1521	SNSSVSTS
YKL089W	3	819	ND
YLR260W	3	372	NISRTSFQS
YML065W	591	1149	KKEIKRGPQ, N
YMR031C	3	192	TSSI
YMR133W	426	873	SSNR
YNR054C	3	345	KQEEKEDV
YOL141W	3	2085	VDGS
YOR264W	1110	1290	DDSDDG
YPL157W	798	945	EKEELSS

Appendix 2.D

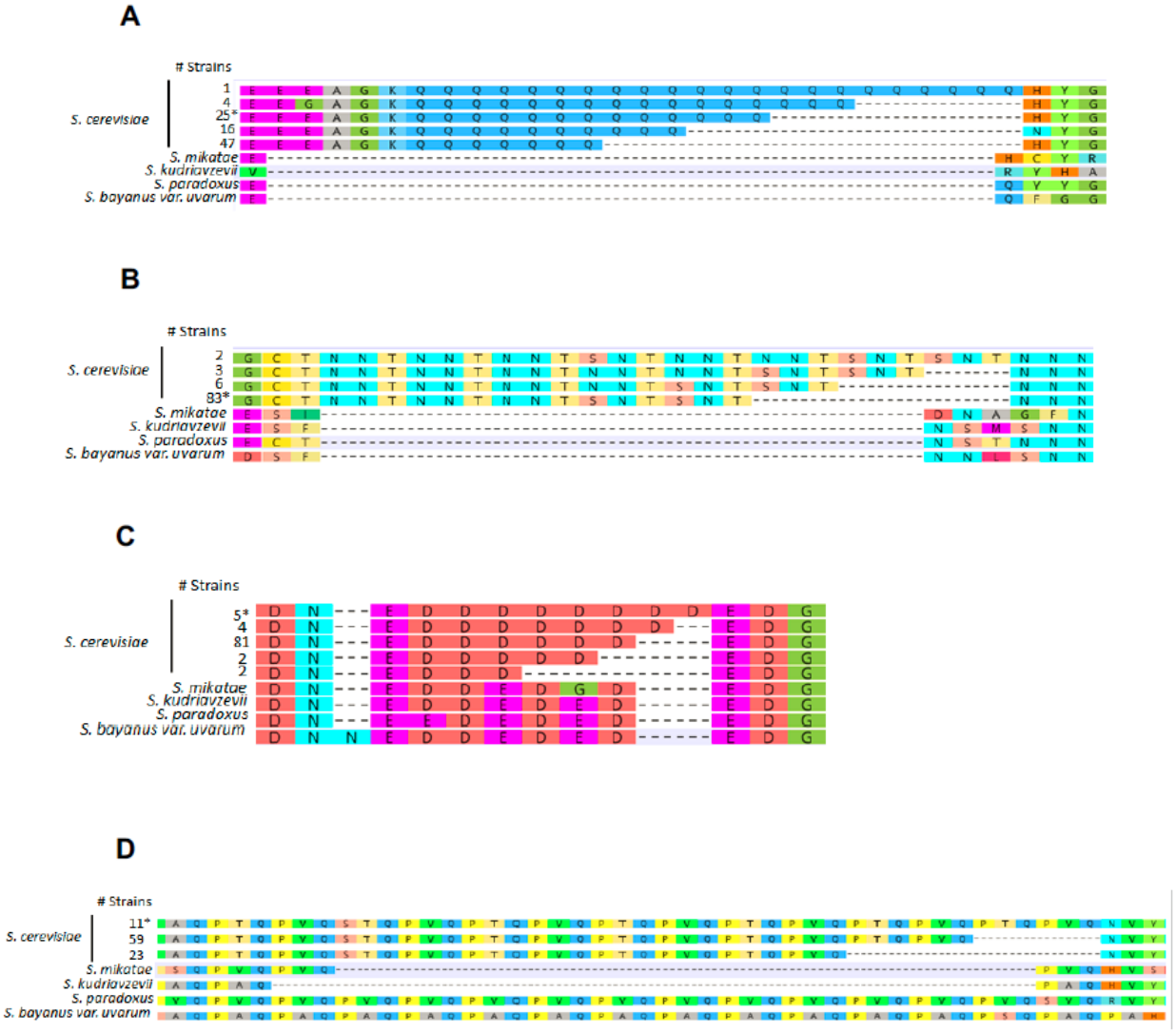


Figure 2.D-1. Examples of non-conserved and sequence-variable repeats in the *Saccharomyces sensu stricto.* **A**) MAFFT alignment of the five length variants of the JJJ2p polyQ homorepeat in *S. cerevisiae* with the corresponding sequences in the SSS that lack the repeat. The number of *S. cerevisiae* strains for each copy number variant is listed to the left with the starred number representing the copy number in the S288C reference genome. **B**) Alignment of the *YPR003C* NNT repeat. **C**) Alignment of the *KAP104* polyD homorepeat in *S. cerevisiae* showing a mixed polyD/polyE repeat in the other SSS species. **D**) Alignment of the *PAN1* repeat showing different consensus sequences of the repeat depending on the species of the *Saccharomyce sensu stricto* that is examined.

Chapter 2 Literature Cited

Albrecht A, Mundlos S (2005) The other trinucleotide repeat: polyalanine expansion disorders. Curr Opin Genet Dev 15: 285-93

Babokhov M, Mosaheb MM, Baker RW, Fuchs SM (2018) Repeat-Specific Functions for the C-Terminal Domain of RNA Polymerase II in Budding Yeast. G3 (Bethesda) 8: 1593-1601

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol 55: 104-10

Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett 15: 956-63

Carvalho-Netto OV, Carazzolle MF, Rodrigues A, Braganca WO, Costa GG, Argueso JL, Pereira GA (2013) A simple and effective set of PCR-based molecular markers for the monitoring of the Saccharomyces cerevisiae cell population during bioethanol fermentation. J Biotechnol 168: 701-9

Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, Latysheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, Babu MM (2017) Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol 24: 765-777

Das RK, Pappu RV (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A 110: 13392-7

Fuchs SM (2013) Chemically modified tandem repeats in proteins: natural combinatorial peptide libraries. ACS Chem Biol 8: 275-82

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-77

Gemayel R, Yang Y, Dzialo MC, Kominek J, Vowinckel J, Saels V, Van Huffel L, van der Zande E, Ralser M, Steensels J, Voordeckers K, Verstrepen KJ (2017) Variable repeats in the eukaryotic polyubiquitin gene ubi4 modulate proteostasis and stress survival. Nat Commun 8: 397

Habchi J, Tompa P, Longhi S, Uversky VN (2014) Introducing protein intrinsic disorder. Chem Rev 114: 6561-88

Jorda J, Xue B, Uversky VN, Kajava AV (2010) Protein tandem repeats - the more perfect, the less structured. FEBS J 277: 2673-82

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-66

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-80

Kim YH, Kang JY, Gil JY, Kim SY, Shin KK, Kang HA, Kim JY, Kwon O, Oh DB (2017) Abolishment of N-glycan mannosylphosphorylation in glyco-engineered Saccharomyces cerevisiae by double disruption of MNN4 and MNN14 genes. Appl Microbiol Biotechnol 101: 2979-2989

La Spada AR, Taylor JP (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet 11: 247-58

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D (1999) A census of protein repeats. J Mol Biol 293: 151-60

Morrill SA, Exner AE, Babokhov M, Reinfeld BI, Fuchs SM (2016) DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. J Biol Chem 291: 11540-50

Newman AM, Cooper JB (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinformatics 8: 382

Nilsson J, Grahn M, Wright AP (2011) Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol 12: R65

Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Res 41: D508-16

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7: 208

Richard GF, Dujon B (2006) Molecular evolution of minisatellites in hemiascomycetous yeasts. Mol Biol Evol 23: 189-202

Rogers DW, McConnell E, Miller EL, Greig D (2017) Diminishing Returns on Intragenic Repeat Number Expansion in the Production of Signaling Peptides. Mol Biol Evol 34: 3176-3185

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT (2011) The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda) 1: 11-25

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the Database of Disordered Proteins. Nucleic Acids Res 35: D786-93

Simon M, Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol 10: R59

Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, Dietrich FS, McCusker JH (2015) The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res 25: 762-74

Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 25: 847-55

Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37: 509-16

Usdin K, House NC, Freudenreich CH (2015) Repeat instability during DNA repair: Insights from model systems. Crit Rev Biochem Mol Biol 50: 142-67

Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005) Intragenic tandem repeats generate functional variability. Nat Genet 37: 986-90

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708-13

Zhao Y, Strope PK, Kozmin SG, McCusker JH, Dietrich FS, Kokoska RJ, Petes TD (2014) Structures of naturally evolved CUP1 tandem arrays in yeast indicate that these arrays are generated by unequal nonhomologous recombination. G3 (Bethesda) 4: 2259-69

Chapter 3

Repeat-specific functions for the C-terminal domain of RNA polymerase II in budding yeast[2]

Michael Babokhov, Mohammad M. Mosaheb, Richard W. Baker, and Stephen M. Fuchs

Author contributions:

Michael Babokhov designed and performed spotting assays with all CTD constructs, RT-PCR and suppressor mapping experiments and wrote the manuscript.

Mohammad M. Mosaheb performed recursive directional ligation to create all region-specific and serine residue-specific CTD plasmids. He also performed preliminary inositol auxotrophy spotting assays and isolated initial suppressors of region-specific mutants with CTD rearrangements.

Richard W. Baker created the structural models of the Mediator-PIC complex and modeled the likely path of the CTD in the completed alignment.

Stephen M. Fuchs conceived and directed the study.

[2]Manuscript accepted for publication at G3: Genes, Genomes, Genetics

Abstract

The C-terminal domain (CTD) of the largest subunit of RNA polymerase II (RNAPII) is required to regulate transcription and to integrate it with other essential cellular processes. In the budding yeast *Saccharomyces cerevisiae*, the CTD of Rpb1p consists of 26 conserved heptad repeats that are post-translationally modified to orchestrate protein factor binding at different stages of the transcription cycle. A long-standing question in the study of the CTD is if there are any functional differences between the 26 repeats. In this study, we present evidence that repeats of identical sequence have different functions based on their position within the CTD. We assembled plasmids expressing Rpb1p with serine to alanine substitutions in three defined regions of the CTD and measured a range of phenotypes for yeast expressing these constructs. Mutations in the beginning and middle regions of the CTD had drastic, and region-specific effects, while mutating the distal region had no observable phenotype. Further mutational analysis determined that Ser5 within the first region of repeats was solely responsible for the observed growth differences and sequencing fast-growing suppressors allowed us to further define the functional regions of the CTD. This mutational analysis is consistent with current structural models for how the RNAPII holoenzyme and the CTD specifically would reside in complex with Mediator and establishes a foundation for studying regioselective binding along the repetitive RNAPII CTD.

Introduction

RNA polymerase II (RNAPII) is a 12-subunit complex responsible for the transcription of all mRNA in eukaryotes. The largest subunit of RNAPII, Rpb1p, contains a conserved C-terminal domain (CTD) connected to the catalytic core by a flexible linker. The CTD is required for RNAPII activity *in vivo*, acting as a binding site for proteins involved in transcription initiation and other essential co-transcriptional processes (CORDEN 2013; EICK AND GEYER 2013). The coordination of all of these processes by the CTD throughout the transcriptional cycle continues to be a topic of intensive research.

The CTD consists of tandemly repeating peptide units with the consensus sequence of $Y_1S_2P_3T_4S_5P_6S_7$. Budding yeast Rpb1p contains 26 repeats that adhere closely to the consensus sequence while vertebrates have 52 repeats with the distal half containing many repeats degenerate at the Ser7 position. Protein factor binding to the CTD is mediated by extensive post-translational modification of the five hydroxylated amino acids and isomerization of the two prolines within each repeat (LEE AND GREENLEAF 1989; FEAVER *et al.* 1991; ZHANG AND CORDEN 1991; VALAY *et al.* 1995; FUCHS *et al.* 2009). Ser2 and Ser5 phosphorylation are the most commonly studied modifications and are evenly distributed across the CTD repeats (SUH *et al.* 2016). Dynamic patterns of CTD modifications are proposed to form a CTD code that directs progress through the transcription cycle.

Extensive research has uncovered many of the essential functional elements within the repetitive CTD. Mutations to the modifiable residues in the CTD cause drastic

phenotypes, although the specific residue requirements can differ between organisms (HSIN *et al.* 2011; SCHWER AND SHUMAN 2011). Although the consensus sequence is a heptad repeat, the functional unit of the CTD is actually defined by two consecutive repeats that contain properly spaced Tyr1, Ser2 and Ser5 residues (STILLER AND COOK 2004; LIU *et al.* 2008). This functional unit is in agreement with structural studies of CTD binding factors that show surface interactions with two or more repeats (KUBICEK *et al.* 2012; ROBINSON *et al.* 2012). Furthermore, while the wildtype number of repeats is strongly selected for in nature (NONET AND YOUNG 1989), less than half of the total repeats appear to be required for normal growth (BARTOLOMEI *et al.* 1988; WEST AND CORDEN 1995; SCHNEIDER *et al.* 2010). These findings demonstrate that there are additional determinants of CTD function beyond just the linear sequence of heptad repeats.

The large number of CTD repeats, beyond those needed to support growth, raises the possibility of a division of function between the different repeats. For example, in mammalian cells, the elongation factor Spt6p requires the N-terminal half of the CTD (YOH *et al.* 2008) while splicing and 3' processing require the C-terminal half (FONG AND BENTLEY 2001). These differences could be explained either by the different heptad sequences found in the two halves of the mammalian CTD (CORDEN 2013), the distance of the region from the core of the polymerase holoenzyme, or a combination of these considerations. Work in the budding yeast CTD similarly uncovered functional differences between the two halves of the CTD (WEST AND CORDEN 1995; WILCOX *et al.* 2004). However, unlike the mammalian CTD, the yeast CTD consists almost exclusively

of consensus repeats. Therefore, in the absence of extensive sequence differences there must be additional determinants, such as distance from the holoenzyme, that lead to functional specialization in the budding yeast CTD.

Previously we developed a genetic system to investigate instability and repeat number in the budding yeast CTD (MORRILL *et al.* 2016). Here, we use this system to examine the effects of position specific repeat mutation on cellular survival and gene expression. We find that serine to alanine mutations within blocks of repeat units have profoundly different effects on cell survival and several other phenotypes (e.g. salt stress, inducible growth, and 6-azauracil sensitivity (POWELL AND REINES 1996)), dependent on their location within the CTD. In particular, we found that mutations within the middle third of the CTD resulted in generally poor growth whereas mutations to the first eight repeats had growth defects specific to inositol auxotrophy. In contrast, mutations in the last eight repeats had no discernable effect in any conditions tested. The repetitive coding sequence of the CTD makes it prone to spontaneous mutagenesis (MORRILL *et al.* 2016). We exploited this property to identify and analyze plasmid-based spontaneous suppressors that would bypass the poor growth of our CTD mutants. From these suppressors, we identified two discreet windows within the CTD that are required for viability in the presence and absence of inositol. Based on existing structural models of RNAPII and the Mediator complex, we propose that these regions are responsible for coordinating CTD interactions with Mediator.

Materials and Methods

Yeast Strains and Plasmids

Yeast strains were cultured in standard media and grown at 30° C except where otherwise noted. All of the reported strains are derivatives of GRY3019 (MATa *his3Δ*, *leu2Δ*, *lys2Δ*, *met15Δ*, *trp1Δ*::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1) provided by the Strathern lab (MALAGON *et al.* 2006) or from the yeast deletion collection (WINZELER *et al.* 1999). Gene tagging cassettes were created using PCR and integrated by homologous recombination (JANKE *et al.* 2004). The full set of strains used in this study is listed in Table 1. Selection was performed in synthetic complete (SC) media or plates lacking the appropriate amino acid for auxotrophic strains (ADAMS *et al.* 1997). Dominant drug resistance markers KanMX6, HphNT1 and NatNT2 were selected for using 50 µg/mL of geneticin (G418), hygromycin B and nourseothricin (ClonNAT), respectively. Ammonium sulfate was replaced with 1 g/L of monosodium glutamate as a nitrogen source whenever these drugs were used for selection in liquid media or plates.

Region-specific mutants, and serine-specific CTD variant plasmids were made using the recursive directional ligation by plasmid reconstruction method (MCDANIEL *et al.* 2010). The construction of full length consensus and truncated CTD plasmids was described previously (MORRILL *et al.* 2016). To build CTD plasmids with repeat specific mutations, oligonucleotides that coded for two repeat blocks of the sequence $(PTAPAYA)_2$ were recursively ligated together using two base pair overhangs until the desired CTD sequences were obtained. Similar oligonucleotides were used to create the serine-specific constructs (e.g. PTAPSYS for S5A). These CTD sequences were then cloned

into pRPB1 using *Sac*1 and *Xma*1 restriction sites and verified by sequencing as described previously (MORRILL *et al.* 2016).

Spotting Assays

Cells were grown overnight at 30° C in SC–LEU media and diluted to OD_{600} 0.2 in fresh SC–LEU in the morning. Yeast were allowed to divide at least two times (OD_{600} 0.8 – 1.0) before being collected and washed twice in sterile water. Cell number was estimated by spectrophotometry (OD_{600} = 1 ~ $1x10^7$ cells/mL) and suspensions were transferred to a 96 well plate (250µL of ~$1x10^7$ cells/mL) and serially diluted 5-fold in sterile water. The dilutions were then spotted onto the appropriate plates using a 48-pin replicator. The plates were grown at 30° C (except where indicated) and photographed daily starting at two days. All spotting experiments were performed a minimum of three times from independent plasmid transformations and a representative image was selected to display.

Western Blotting

Western analysis of the block mutants was performed as described previously (MORRILL *et al.* 2016) with the following changes. Proteins were separated on an 8% SDS-PAGE gel made with a standard 37.5:1 acrylamide:bis-acrylamide ratio and transferred to PVDF. Membranes were incubated with the primary antibodies raised against: Rpb1p (Y-80, Santa Cruz), phosphorylated Ser2 (a generous gift from Dirk Eick), phosporylated Ser5 (clone 3E8 from Active Motif) and G6PDH loading control (Sigma A9521).

INO1 expression

RNA extracts to assay *INO1* expression were prepared by growing cultures at 30° C in SC–LEU media with 50 µg/mL doxycycline (DOX) (Alfar Aesar) to mid log phase (OD_{600} 0.6 – 0.8). Cells were harvested, washed twice in sterile water to remove any remaining inositol and resuspended in SC–LEU+DOX media that lacked inositol (SC–LEU–INO+DOX). Cultures were grown for two hours without inositol to induce *INO1* gene expression. After induction, cells were harvested, washed and stored at -80° C. Total RNA was extracted using an Illustra RNAspin Mini kit (GE Healthcare), following the manufacturers protocol for yeast. RNA extracts were quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and 50 ng of total RNA was primed with poly-(dT) primers to obtain cDNA with a SuperScript First-Strand Synthesis kit (Invitrogen). One microliter of cDNA was used as a template to amplify the *INO1* gene and the resulting bands were quantified in ImageJ and normalized to *ACT1* levels. RT-PCR was performed with RNA from three independent cultures for controls and six independent cultures for –INO+DOX experimental samples. The mean is reported in the text and a two-way ANOVA was performed to assess significance using Graphpad Prism software.

Suppressor mutant screen

Spontaneous suppressor mutations in block mutant plasmids were obtained using cultures grown in a 96 well plate. Individual colonies were taken from a fresh plasmid transformation and suspended in 1 mL of SC–LEU media in a deep well plate. Plates were grown at 30° C with occasional shaking for one day. The cultures were diluted to an OD_{600} 0.8 – 1.0 using fresh SC–LEU media and plated to SC–LEU plates with and

without DOX and inositol using a 48-pin replicator. Plates were incubated at 30° C between three and five days until fast-growing colonies appeared. Colonies were screened by PCR with primers flanking the CTD coding region and loaded on a 1% agarose gel to identify plasmid-based mutational events as indicated by a change in band size relative to the amplified genomic *RPB1* CTD coding region. Mutated plasmids were extracted, sequenced and retransformed into GRY3019 to confirm their ability to support growth on media containing DOX.

Structural Modeling

A cryoEM structure of the full PIC-Med complex from *S. cerevisiae* was recently determined to a resolution of ~20 Å ((ROBINSON *et al.* 2016), EMD-8308). This structure includes RNAPII, TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and TFIIS, and the full Mediator complex, including the Head, Middle, and Tail modules. Robinson et al. were able to make a molecular model for RNAPII, Mediator Head and Middle, and subsets of the general TFs (ROBINSON *et al.* 2012). To build on this model, we aligned the molecular model for the complete human TFIIH from a recent high resolution cryoEM structure (5of4.pdb) (GREBER *et al.* 2017). TFIIH was aligned using XPD and CXPD, the human homologs of Rad25 and Rad3. We also aligned a crystal structure of *Schizosaccharomyces pombe* RNAPII, which includes around 30 amino acids of the Rpb1p linker that makes contacts with the RNAPII subunits Rvb1p and Rvb7p ((SPAHR *et al.* 2009), 3h0g.pdb). All docking, alignment, and figure making was done using UCSF Chimera (PETTERSEN *et al.* 2004) or PyMol (The PyMol Molecular Graphics System, Version 1.7, Schrodinger, LLC).

Reagent and Data Availability

The complete list of plasmids used in this is found in Appendix 3.A and pertinent plasmids have been deposited to Addgene. Additional data regarding the CTD constructs are presented in Appendices 3.B–3.E. All reagents and data are available upon request.

Results

Repeat number requirements in phenotypes related to CTD function

Previous studies of RNAPII examined a number of phenotypes to dissect CTD repeat function (NONET *et al.* 1987; ARCHAMBAULT *et al.* 1996). We first determined the phenotypic consequences of varying CTD repeat number requirements in our TET-off system. Briefly, the addition of the antibiotic doxycycline (DOX) to growth media prevents expression of the genomic wildtype copy of *RPB1*. This leaves the plasmid-based copy containing our CTD constructs as the only source of Rpb1p for the cell. We tested a series of CTD truncation mutants ranging in length from 8 repeats to 26 repeats (Figure 1A). Spotting serial dilutions of actively growing yeast cultures then allowed us to score the growth of the truncated CTD mutants relative to wildtype controls.

In the absence of any stress, all cells grew equally well without DOX and this condition served as a loading control for our spotting assays. Strains labeled WT had the native budding yeast CTD sequence while the pRPB1-CTD_{26} construct had a synthetic full length CTD consisting of all perfect consensus repeats. Under all conditions tested the WT and CTD_{26} plasmids grew equally well and were considered equivalent. Addition of

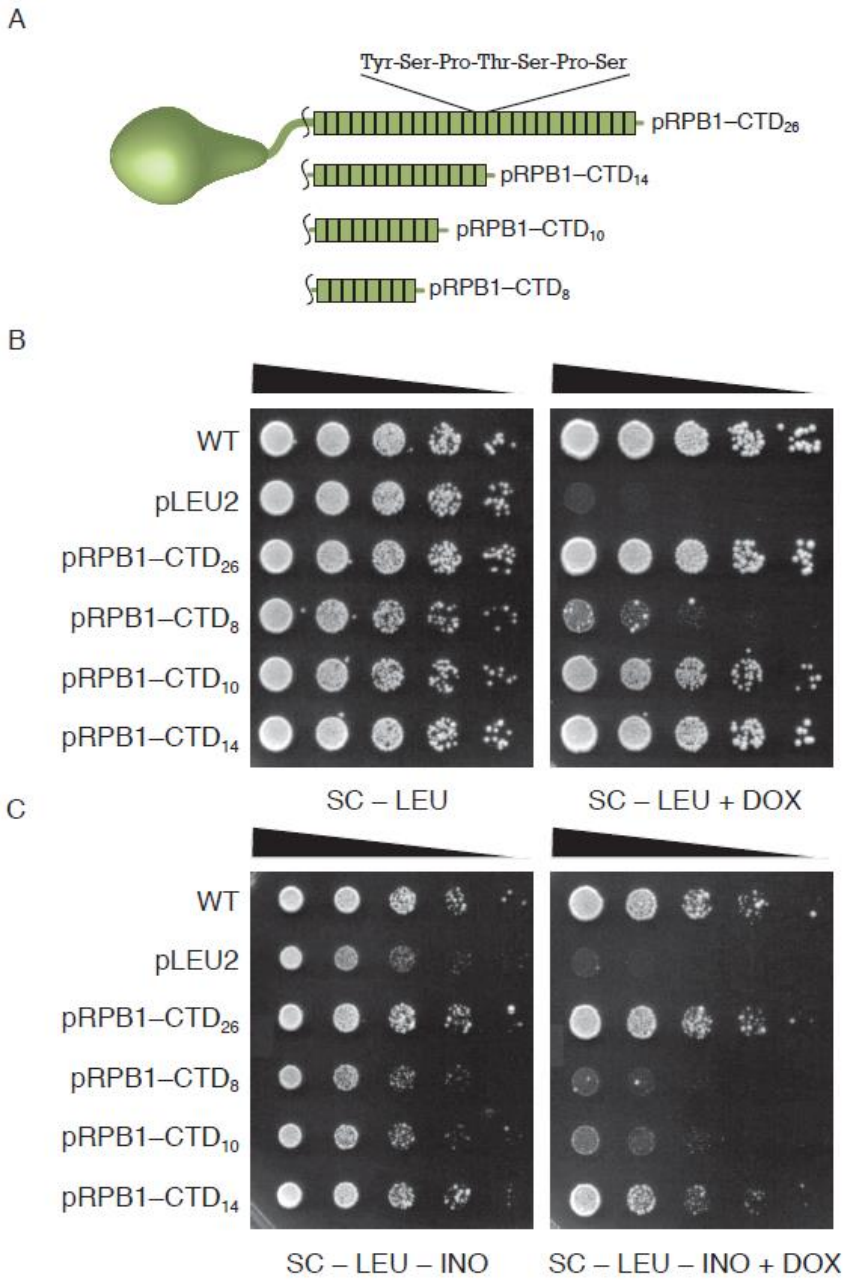


Figure 1. Length dependence of RNAPII CTD in TET-off system. **A**) CTD truncation mutants tested in this study. Each block represents a single seven amino acid heptad repeat sequence. Constructs are labeled based on the number of total CTD repeats. **B**) Spotting assay measuring the dependence of CTD length on yeast viability. In the absence of doxycycline (DOX) both the genomic copy of *RPB1* and the *LEU2* plasmid copy of *RPB1* harboring different length CTD regions are expressed. When DOX is present, only the plasmid copy is transcribed (MALAGON *et al.* 2006; MORRILL *et al.* 2016). **C**) Spotting assay measuring the dependence of CTD length on yeast viability in media lacking inositol (INO).

DOX led to a severe growth defect in the pRPB1–CTD_8 construct, while the constructs with 10 and 14 repeats grew at wildtype levels (Figure 1B), in line with results from our previous work and the studies of other groups (NONET *et al.* 1987; MORRILL *et al.* 2016). We also assessed the ability of these truncation mutants to tolerate a number of nutrient and environmental stresses. pRPB1-CTD_{10} and pRPB1-CTD_{14} exhibited growth comparable to the full length CTD under all conditions except on media lacking inositol (Figure 1C and Appendix 3.B). *INO1* is induced upon inositol starvation and both the CTD and its associated Mediator complex are required for this process (ARCHAMBAULT *et al.* 1996). The pRPB1–CTD_8 mutant was inviable under the –INO+DOX condition while the pRPB1–CTD_{10} and pRPB1–CTD_{14} showed decreased growth relative to wildtype. In fact, even in the absence of DOX (when both the plasmid and genomic copies of RPB1 are expressed), the truncation plasmids have a slight effect on growth (compare Figure 1B left to Figure 1C left). Based on these observations we focused in on the inositol auxotrophy that results from mutating specific regions of the CTD.

Positional requirements of CTD repeats in inositol auxotrophy

The graduated inositol auxotrophy of CTD mutants led to two hypotheses: growth on media lacking inositol requires either: 1) more than 8 CTD repeats, and approximately 14 repeats to achieve levels comparable to wild-type; or 2) CTD repeats in a particular linear position within the CTD sequence. Previous analysis of the CTD suggested that serine mutations in the consensus sequence had different effects if they were placed in proximal or distal repeats (WEST AND CORDEN 1995). We expanded on this work by creating a series of plasmids harboring serine to alanine (S>A) substitutions at different

linear regions within the CTD (Figure 2A). The constructs each contained eight mutated repeats in a series of three windows while all maintaining 18 consensus repeats in various arrangements. The three block mutants were both expressed at similar levels and had comparable bulk Ser2 and Ser5 phosphorylation to the consensus plasmid pRPB1–CTD_{26} as measured by western blotting. Furthermore, overall Rpb1p levels in all constructs were reduced upon addition of DOX, reflecting the expected lack of expression from the DOX-regulated genomic copy of *RPB1* (Appendix 3.C).

In the absence of any stress, the $pCTD_{26}$–$S>A_{18-25}$ mutant behaved identically to the full-length consensus plasmid pRPB1–CTD_{26} (Figure 2B). Both $pCTD_{26}$–$S>A_{2-9}$ and $pCTD_{26}$–$S>A_{10-17}$ showed slower growth in the presence of DOX and in the absence of inositol (Figure 2B, 2C). In both the presence and absence of inositol, $pCTD_{26}$–$S>A_{10-17}$ was nearly inviable and we found this to be true for a number of additional phenotypic conditions as well. Specifically, the $pCTD_{26}$–$S>A_{10-17}$ mutant showed varying degrees of sensitivity to the drug 6-azauracil, galactose media and osmotic stress while the other two mutants were unaffected (Figure 3). However, for $pCTD_{26}$–$S>A_{2-9}$, while there was as slight growth defect on DOX, this mutant exhibited the same slow growth on media lacking inositol as seen for pRPB1–CTD_{10} and pRPB1–CTD_{14}.

Mutations in the CTD repeats 2–9 and 10–17 lead to impaired INO1 expression

The inositol auxotrophy phenotype has been extensively studied and mutations in the RNA polymerase II holoenzyme fail to induce expression at the *INO1* locus (ARCHAMBAULT *et al.* 1996). We suspected that poor growth of $pCTD_{26}$–$S>A_{2-9}$ and

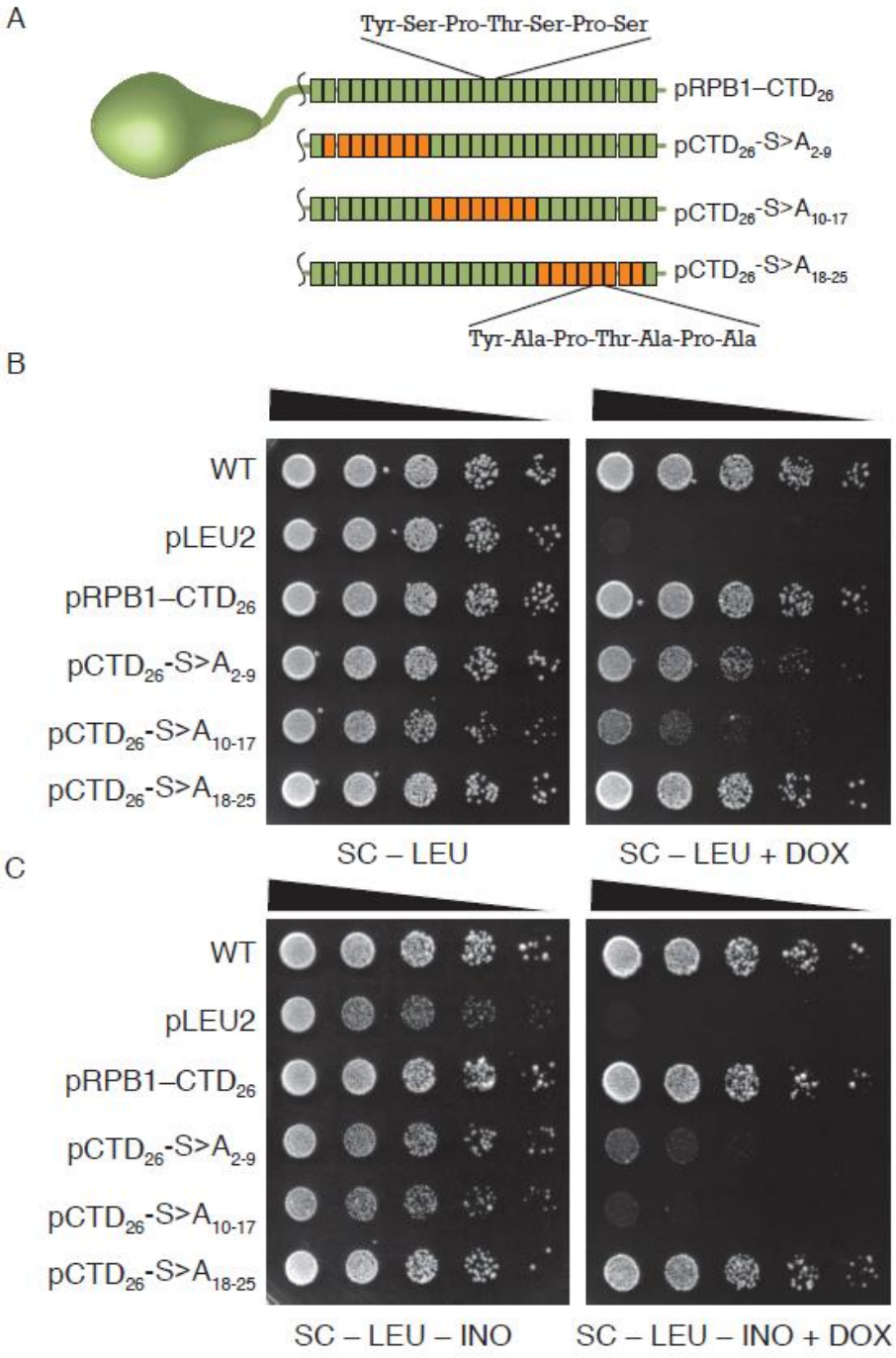


Figure 2. Position-specific phenotypes of CTD mutants. **A**) CTD mutants were created that harbored Ser>Ala substitutions at precise positions within the CTD sequence as noted by the subscripts in the name. Repeats with wildtype sequence are colored in green with mutant repeats in orange. Spotting assay measuring the dependence of CTD position on yeast viability (**B**) and inositol auxotrophy (**C**).

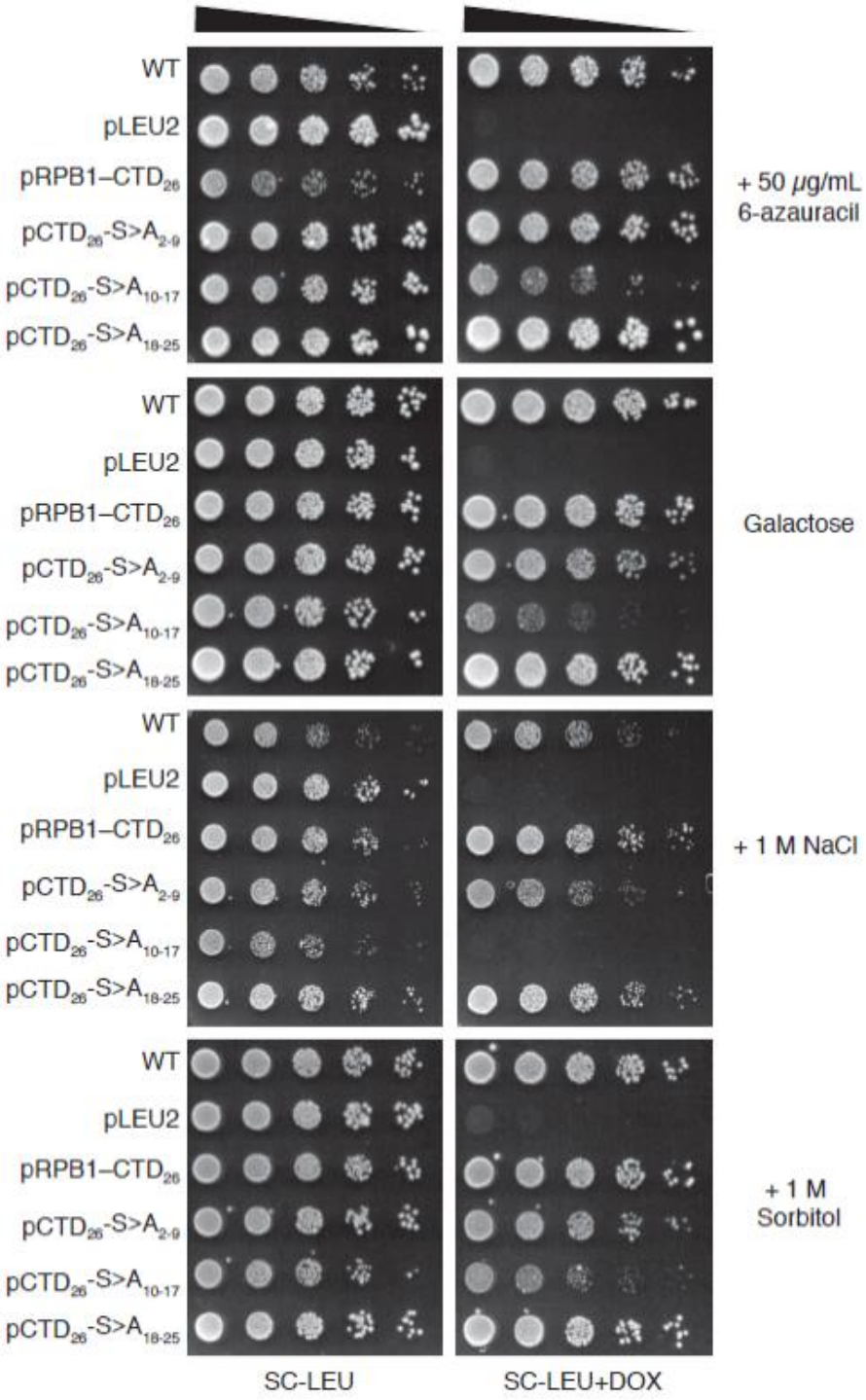


Figure 3. Additional phenotypes of position-specific CTD mutants. Preparation of the spotting assay and ordering of the mutants is the same as in Figure 2. CTD constructs were assayed on additional stresses including: 50 µg/mL of 6-Azauracil (6AU), plates with galactose as the only sugar (SC-GAL) and osmotic stress in the form of 1M NaCl and 1M sorbitol.

pCTD$_{26}$–S>A$_{10-17}$ on –INO+DOX was due to a similar lack of *INO1* induction. Therefore, we measured the induction of this gene in our positional mutants using endpoint RT-PCR with *ACT1* as a reference gene (Figure 4A). When strains were grown in the presence of inositol (+INO), *INO1* expression was completely suppressed while *ACT1* remained constant in both the –DOX and +DOX conditions. Under inducing conditions without DOX (–INO–DOX) all strains were able to induce *INO1*. Adding DOX to the inducing media (–INO+DOX) led to a sharp loss of *INO1* expression in both the pCTD$_{26}$–S>A$_{2-9}$ and the pCTD$_{26}$–S>A$_{10-17}$ mutants while *ACT1* expression remained constant. Representative gels are shown in Figure 4A while quantification of the RT-qPCR data for at least three independent cultures is shown in Figure 4B.

Serine 5 is solely responsible for pCTD$_{26}$–S>A$_{2-9}$ inositol auxotrophy

The binding of protein factors to the CTD is determined, in part, by the modification state of defined residues in the heptad repeat (PHATNANI AND GREENLEAF 2006; WERNER-ALLEN *et al.* 2011). In order to determine which serine residue in the regions-specific mutants contributed to the observed phenotypes we created a series of residue-specific mutants. These mutants have Ser2, Ser5 or Ser7 mutated to Ala within pCTD$_{26}$–S>A$_{2-9}$ (Figure 5A). Spotting the serine-specific mutants in the absence of stress led to wildtype levels of growth for all strains except pCTD$_{26}$–S5A$_{2-9}$ which showed the same slight growth defect as the original mutant (Figure 5B). Spotting the strains on –INO+DOX plates revealed that, indeed, pCTD$_{26}$–S5A$_{2-9}$ mirrored the slow growth phenotype of the pCTD$_{26}$–S>A$_{2-9}$ (Figure 5). Neither pCTD$_{26}$–S2A$_{2-9}$, nor pCTD$_{26}$–S7A$_{2-9}$, showed a growth defect without inositol despite having a similar number of serines mutated.

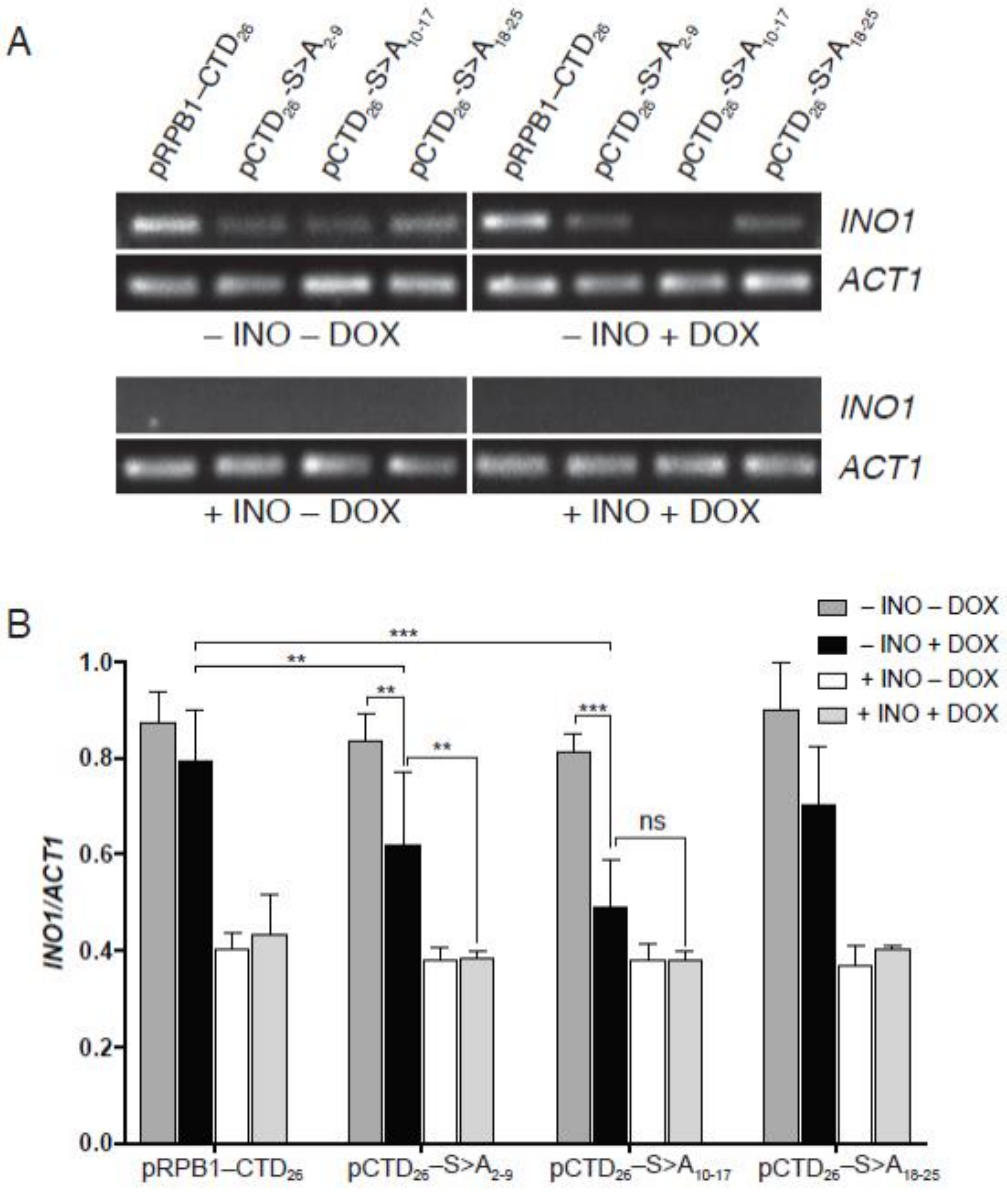


Figure 4. Effect of CTD position on *INO1* expression. **A**) Representative agarose gels of RT-PCR reactions using primers specific for *INO1* and *ACT1* as a loading control. **B**) Quantification of the effects of CTD mutation on *INO1* expression. Signal from agarose gels was quantified by densitometry using ImageJ and data are plotted as the ratio of the *INO1* band intensity to the *ACT1* band intensity. Two-way ANOVA was used to measure significance of interactions, and a subset of significant interactions are indicated as (**, adjusted P-value < 0.05; ***, adjusted P-value <0.01).

Therefore, both the position of the serine with the heptad repeat, and its location within the linear CTD sequence is important for growth on media lacking inositol.

Suppressor mapping reveals two essential windows on the CTD

When plating $pCTD_{26}$–S>$A_{2\text{-}9}$ and $pCTD_{26}$–S>$A_{10\text{-}17}$ we observed that both yielded fast-growing suppressors when spotted on plates with DOX. Plasmid sequencing revealed two types of plasmid-based suppressors: homologous recombination with the genomic copy of *RPB1* or rearrangements of the repetitive CTD coding sequence itself to remove mutated repeats. We predicted that we could use the sequences of these suppressors to map important functional regions of the CTD. To screen for suppressors, 48 independent colonies of both the S>$A_{2\text{-}9}$ and the S>$A_{10\text{-}17}$ block mutants were grown overnight in a *rad52*Δ background, to bias towards rearrangements and away from homologous recombination with the genomic *RPB1* (MORRILL *et al.* 2016). Cells were spotted on +DOX and –INO+DOX plates and large, fast-growing colonies were isolated and analyzed by colony PCR and Sanger sequencing. Sequencing of over 30 independent contraction events revealed that the most common suppressors deleted either four or six mutant repeats or removed all variant repeats (resulting in truncations similar to Figure 1). Unexpectedly, we also recovered one suppressor of $pCTD_{26}$–S>$A_{10\text{-}17}$ which contained both a deletion and a duplication of variant repeats (Figure 6A and Appendix 3.D).

After confirming the identity of the suppressors, we isolated the plasmids and transformed them back into the original GRY3019 strain. These transformants were

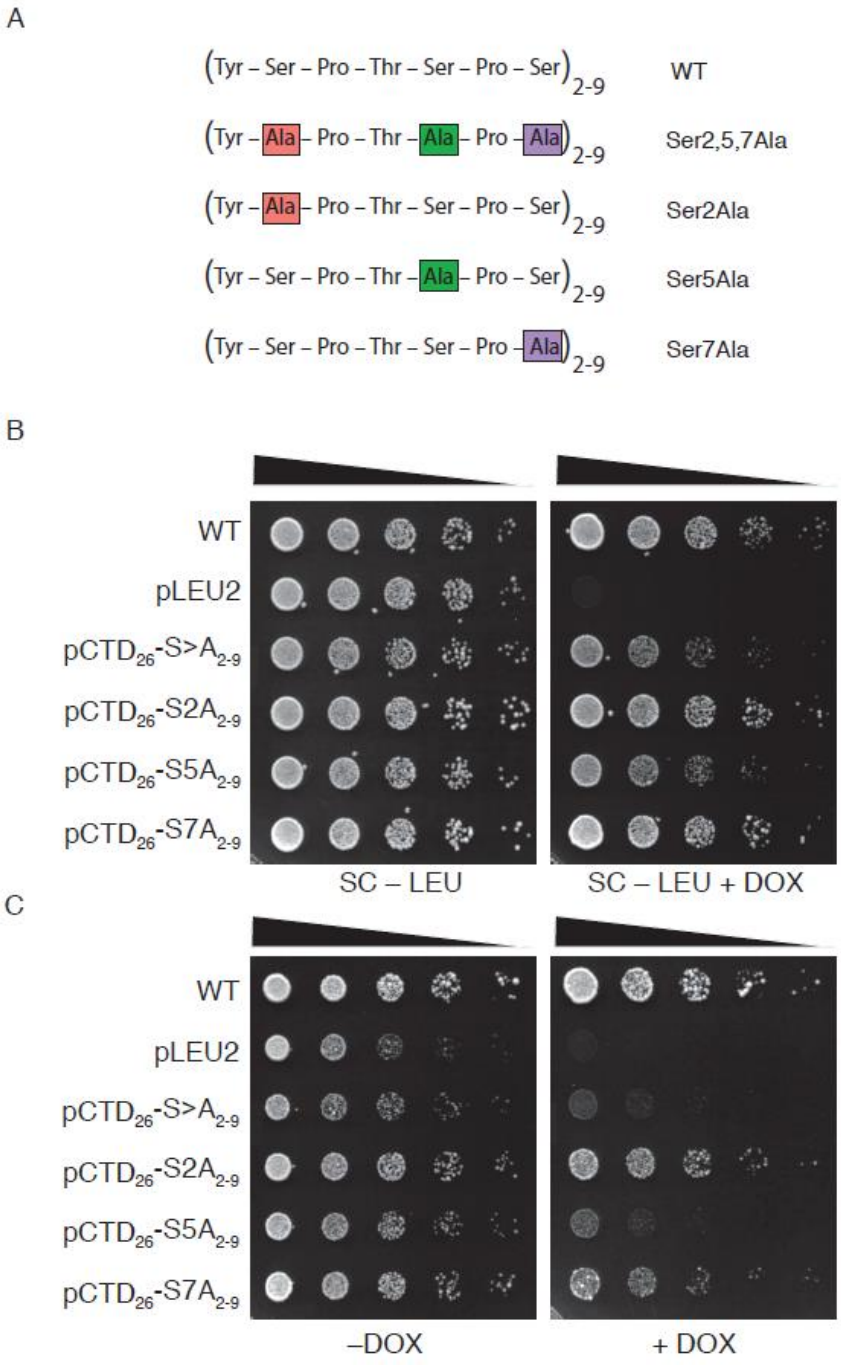


Figure 5. Influence of Ser>Ala substitutions on inositol auxotrophy in proximal CTD repeats. **A**) CTD mutants expressing one or more Ser>Ala substitutions at discrete positions within repeats 2–9 of the RNAPII CTD. The position of the Ser>Ala substitution is marked in pink, noted in the name, and is carried by all 8 repeats within this region. Spotting assays measured the dependence of Ser position on yeast viability (**B**) and inositol auxotrophy (**C**).

plated and their growth was scored on media containing DOX and/or inositol (Figure 6A and Appendix 3.D). Retransforming these plasmids confirmed that the improved growth is from a change in CTD sequence rather than another acquired mutation within the strain. Deleting four mutant repeats was sufficient to restore growth for both $pCTD_{26}$–$S>A_{2\text{-}9}$ and $pCTD_{26}$–$S>A_{10\text{-}17}$ when suppressors were spotted on +DOX plates. In contrast, deleting six mutant repeats from either block was necessary for a significant improvement in growth on –INO+DOX plates (Figure 6 and Appendix 3.E). These repeat requirements enabled us to identify two repeat windows along the CTD that are necessary for viability and growth on media lacking inositol (Figure 6B).

Structural modeling of CTD positional requirements

Our finding that the RNAPII CTD has two unique regions necessary for growth suggests that there exists at least two, non-overlapping binding sites for essential CTD-associating protein factors. Repeats 12-14 seem to be most important for growth and based on the strong phenotypes from mutations in this region it is difficult to predict which of the essential CTD-associated activities may be recruited to this site. However, both the importance of Ser5 and the inositol auxotrophy of $pCTD_{26}$–$S>A_{2\text{-}9}$ suggested to us that this region of the CTD may be important for recruitment of the Mediator complex. Deletion of some non-essential Mediator components leads to impaired growth on media lacking inositol (Appendix 3.F). Thus to understand the geometric restraints of the CTD and gain insight into potential binding partners, we turned to recent cryo-electron microscopy (cryoEM) and x-ray crystal structures (ROBINSON *et al.* 2012; ROBINSON *et al.* 2016). Combining our suppressor data and current structural models,

we built a to-scale representation of how the CTD tail might bind Mediator complex and a generic transcriptional regulator (Figure 6C). Based on the physical proximity between the core of Rpb1p and the Mediator-CTD binding site (only ~50 Å), we propose that Mediator would most likely interact with the first CTD window, repeats 4-9.

Discussion

The CTD of RNAPII is known to physically interact with a number of factors that are critical for cellular function including the capping enzyme complex, the Mediator complex, mRNA processing machinery, and termination factors such as Pcf11p (PHATNANI AND GREENLEAF 2006). It interacts with numerous additional non-essential factors such as the histone methyltransferase Set2p (KIZER *et al.* 2005). How protein factors organize on the RNAPII CTD during transcription has been of great interest for several decades. Early research dissected the importance of the heptad repeat sequence and the role of post-translational modifications toward the recruitment of factors (HSIN *et al.* 2011; SCHWER AND SHUMAN 2011). Mainly, the pattern of phosphorylation is known to change during different phases of transcription and this allows for the correct temporal recruitment of factors. Later mutational analysis revealed that the functional unit of the CTD consisted of two heptad repeats, with many protein factors binding the Ser5 region of the first repeat and the Ser2 region of the next (STILLER AND COOK 2004; LIU *et al.* 2008). Due to the repetitive nature of the CTD, it has been difficult to determine whether different repeats within the linear sequence had specific functions. Early synthetic mutants showed variable phenotypes depending on the location of the mutation but these mutants were neither systematic nor uniform in

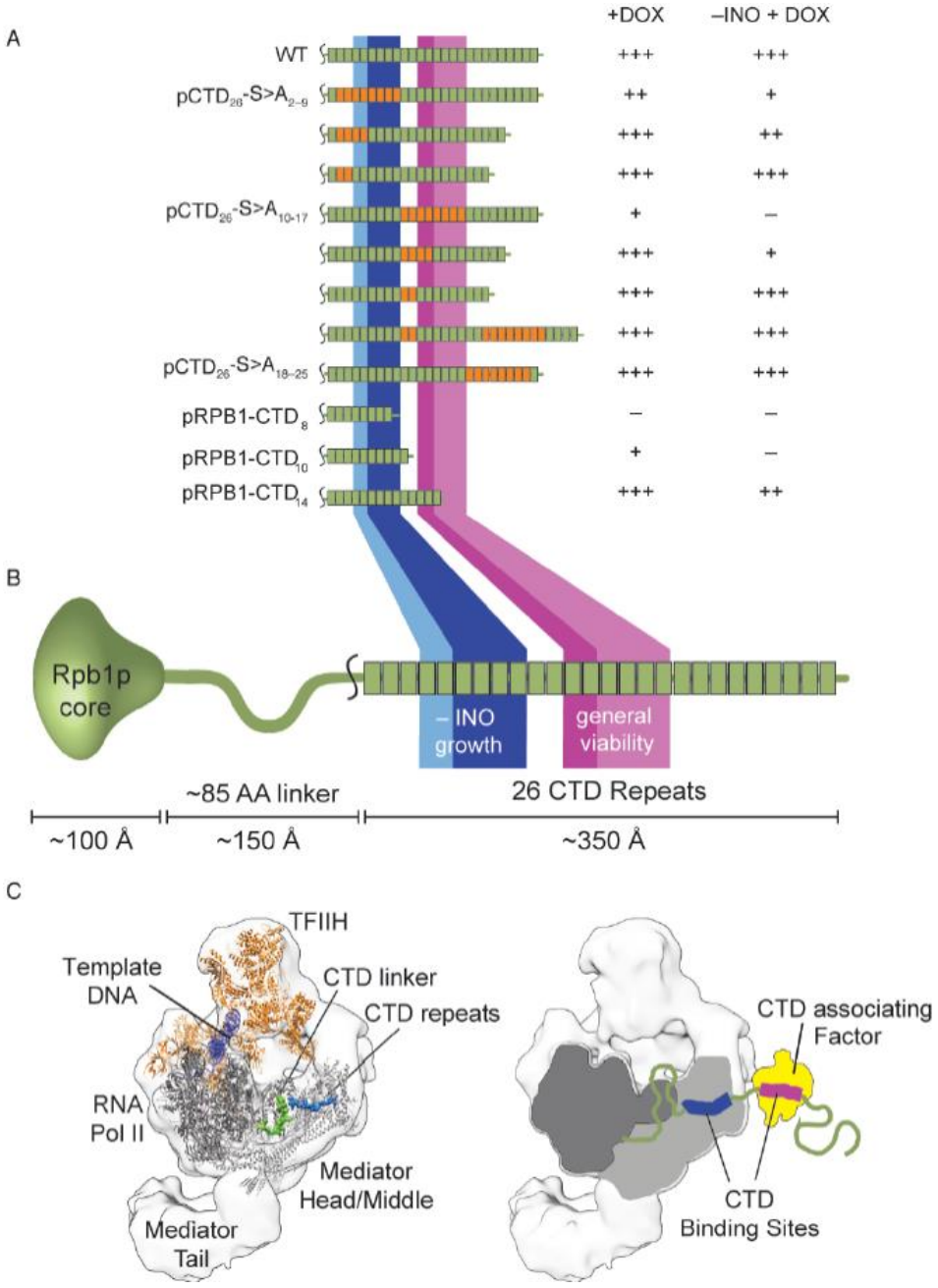


Figure 6. Mapping functional regions of the yeast CTD. **A**) Summary of plasmid-based spontaneous suppressor mutants and their growth characteristics on general growth (+DOX) and inositol deficient (–INO+DOX) media. No growth is scored as (–) with poor, moderate, and unimpaired growth scored as +, ++, and +++, respectively. **B**) Based on the growth of different constructs in (A), Regions essential for general growth (purple), and important for growth in –INO media (blue) were mapped to the 26 repeats of the yeast CTD. Intensity of the color correlates with importance of a repeat for a particular phenotype. The scale bar represents approximate length of the CTD tail in a fully extended conformation. **C**) A model based on existing structures for the RNAPII in complex with Mediator proposing that Mediator interacts with repeats in the proximal region of the CTD, most likely repeats 5–9 (blue).

length (WEST AND CORDEN 1995). In mammalian cells, the binding of certain factors has been reported to be specific to either the N or C-terminal halves of the repeats (FONG AND BENTLEY 2001; YOH *et al.* 2008). However, the seventh residue in the C-terminal repeats is frequently degenerate, therefore both sequence and spatial positioning may determine function. Currently, we collectively know a lot about how different factors can recognize a region of the CTD but there is still little known about how they might simultaneously interact with the full-length CTD. For example, are all repeats functionally equivalent, or is there undescribed specificity built into the linear CTD sequence? Additionally, is factor binding controlled by both phosphorylation state and competition with other factors? Or are there higher-order structural interactions that provide targeting of particular factors to specific regions of the CTD?

We recently reported on an improved TET-off system to investigate CTD mutants (MORRILL *et al.* 2016). With this system, it became possible to make precise mutations to different regions of the CTD. Here, we utilized our TET-off system to explore whether specific repeats had essential CTD functions. We constructed three different block mutants containing Ser to Ala mutations in the heptad repeats and uncovered different effects for all three regions. If all the CTD repeats had identical function as implied by their primary amino acid sequence, then we would expect all three of our block mutants to behave equally. Indeed, our block mutants had similar expression levels and serine phosphorylation profiles to each other and to the pRPB1–CTD_{26} control (Appendix 3.C). However, we found that all three blocks still behaved differently. Critically, while all three mutants contained 18 wildtype repeats in various arrangements (Figure 2A) the first two

blocks yielded different phenotypes. Given that 14 wildtype repeats permit growth under all tested conditions (Figure 1B, C), we can rule out bulk repeat number as the only determinant for CTD function. Instead we propose that our mutations are interrupting positional cues within the CTD sequence that coordinate the binding of protein factors.

We examined a number of different phenotypes using our block mutants and found that mutating repeats 10–17 caused a general sensitivity to stress conditions. The sensitivity was comparable to that of the pRPB1–CTD_8 mutant and even led to complete inviability in the presence of 1M NaCl (Figure 3). One result from our screen that stood out in particular was the unique sensitivity of the $pCTD_{26}$–S>$A_{2\text{-}9}$ mutant to the inositol auxotrophy phenotype (Figure 2C). Inositol auxotrophy has been a commonly observed phenotype in transcription mutants (VILLA-GARCIA *et al.* 2011). In particular, mutations in both the CTD and the Mediator complex have been shown to cause inositol auxotrophy (ARCHAMBAULT *et al.* 1996; SINGH *et al.* 2006). Although disruptions in other physiological pathways can lead to inositol auxotrophy (YOUNG *et al.* 2010), our $pCTD_{26}$–S>$A_{2\text{-}9}$ and $pCTD_{26}$–S>$A_{10\text{-}17}$ mutants fail to express *INO1*, demonstrating that these regions are required for inducible transcription (Figure 4). Based on the specific inositol auxotrophy when repeats 2–9 were mutated, we further probed this region using a set of residue-specific mutants. Mutating Ser5 within repeats 2–9 resulted in the same inositol auxotrophy observed when all three serine residues within these repeats were replaced with alanine (Figure 5B). This requirement for Ser5 is consistent with a number of known CTD binding proteins including Kin28 and the Mediator complex (ROBINSON *et al.* 2012; WONG *et al.* 2014) as well as the 5' capping enzyme (FABREGA *et al.* 2003).

An analysis of the spontaneous suppressors we found allowed us to further define the regions of the CTD important for function. Mapping these suppressors identified two regions, repeats 6–9 and 14–17, that were required for growth on +DOX plates. Most CTD-binding factors use two repeats to bind while the largest known interaction is with three repeats and the Mediator complex (KUBICEK *et al.* 2012; ROBINSON *et al.* 2012). Therefore, it is highly unlikely that both regions of four repeats each are bound by a single protein factor. Additionally, the different phenotypes observed when repeats 2–9 or 10–17 are mutated (Figure 2B, C) support at least two independent binding events that are required at repeats 6–9 and 14–17. These two regions expand to require repeats 4–9 and 12–17 for growth on media lacking inositol.

Intriguingly, repeats 10 and 11, which reside between the two CTD regions defined in this work, are dispensable for growth under all tested conditions. Previous mutational analysis of the CTD demonstrated that spacers of two or five Ala residues could still maintain viability provided they were inserted between every functional diheptad (STILLER AND COOK 2004). Consequently, non-functional sequences are tolerated provided the spacing of essential repeats is not disrupted. Thus, repeats 10 and 11 may be acting as natural structural spacers that help align the essential regions in repeats 4–9 and 12–17 for separate binding events.

To address the possibilities raised by our genetic data we attempted to use existing RNAPII and Mediator structural data to model how Mediator may interact with specific CTD repeats. The best structural evidence available suggests that Rpb1p and the

known CTD-Mediator binding site are only 50 Å apart. While the length of the linker would allow Mediator to bind any CTD repeat, we propose that CTD repeats 4–9 are the primary site of Mediator association. This arrangement is consistent with the requirement of more than eight CTD repeats for viability (NONET *et al.* 1987; WEST AND CORDEN 1995; MORRILL *et al.* 2016) and more than 12 repeats for normal growth. If the Mediator binding site is confined to the 4–9 window, this would allow a second binding event within the 13-CTD tail, and could explain why further truncations are inviable. Mediator interactions require an unphosphorylated Ser5 within the CTD for binding (JERONIMO AND ROBERT 2014), consistent with our $pCTD_{26}$–$S5A_{2\text{-}9}$ mutant under inositol auxotrophy (Figure 5C). Mediator is required for growth without inositol and a number of nonessential Mediator subunits demonstrate inositol auxotrophy when deleted (SINGH *et al.* 2006; YOUNG *et al.* 2010) and (Appendix 3.F). Although we found that both the 4–9 and 12–17 windows are required to survive without inositol (Figure 6A), the second window is also required for viability under a wide range of stresses (Figure 2C, Figure 3). Thus, we reason that the 4–9 window harbors an exclusive Mediator binding site while the 12–17 window is used for other essential CTD-related pathways.

The model we present shows one conformation of Mediator and other factors bound to the two essential regions we identified in our genetic screen. However, the dynamic nature of both the CTD (ZHANG *et al.* 2010) and the Mediator complex (SENNETT AND TAATJES 2014; WANG *et al.* 2014) means that alternative binding conformations are also possible. Residues as far away as the very tip of the CTD are able to make contacts with Mediator subunits (NOZAWA *et al.* 2017), raising the prospect that either of our two

windows might bind to Mediator. Mapping suppressors revealed that repeats 12–17 are also required for viability under inositol auxotrophy (Figure 6A) and may represent a possible binding site for Mediator. The Mediator complex, or some other factor required for *INO1* induction, may either bind this 12–17 window or possibly sample both the 4–9 and 12–17 windows to properly coordinate transcription. Interestingly, while many deletion mutants of non-essential Mediator subunits demonstrate inositol auxotrophy (SINGH *et al.* 2006; YOUNG *et al.* 2010), the severity of the defect varies based on the subunit deleted. We found that while some mutants (e.g. *srb5Δ*, *rox3Δ*) were inviable when grown without inositol, other mutants (e.g. *srb2Δ*, *soh1Δ*) showed only a reduced growth rate (Appendix 3.F). These growth differences recall the different phenotypes of our $pCTD_{26}$–S>$A_{2\text{-}9}$ and $pCTD_{26}$–S>$A_{10\text{-}17}$ mutants and raise the possibility that Mediator complexes or subcomplexes of differing subunit composition may selectively bind either of the two essential CTD windows.

In addition to the Mediator complex, there are a number of other possible CTD-binding proteins that may specifically occupy the CTD windows at repeats 4–9 and 12–17. Similarly to Mediator subunits, CTD kinases Ssn3p and Ctk1p have been identified in screens for inositol auxotrophy (YOUNG *et al.* 2010). Our Ser to Ala substitution mutants in the 4–9 window could be preventing proper phosphorylation at these repeats even if Mediator is productively bound at the 12–17 window. While we did not detect any differences in Ser2 and Ser5 phosphorylation in our CTD mutants (Appendix 3.C), a region-specific defect in phosphorylation may be too subtle to be detected or restricted to a certain stress condition. Alternatively, other essential co-transcriptional processes

such as mRNA 5’ capping have also been shown to require properly modified Ser5 (FABREGA *et al.* 2003) and capping enzyme may bind to repeats 4–9 while mediator occupies the 12–17 site. The elongation factor Spt4p has also been associated with the inositol auxotrophy phenotype (YOUNG *et al.* 2010) and could potentially bind to one of the two CTD windows in a sequential manner following Mediator or a CTD kinase. Discriminating between these multiple possible binding configurations will require biochemical characterization of the RNA polymerase II complexes across our various region-specific CTD constructs.

Most fundamentally, the analysis here demonstrates conclusively that, although they have the same amino acid sequence, different heptads of the CTD have specific cellular functions. Repeats 12–17 are important for growth on a range of phenotypes whereas repeats 4–9 are required specifically for growth in the absence of inositol which is consistent with these repeats being important for Mediator binding. This solidifies CTD repeat location, in addition to CTD phosphorylation, as an important factor in determining how CTD-associating proteins interact with the CTD during transcription. Using our approach and growing panel of site-specific mutants, it should be possible to probe even more specific CTD interactions with factors ranging from the RNA capping and processing machinery to chromatin modifying enzymes.

Appendix 3.A

Table 3.A-1. List of plasmids used in chapter 3.

Name	Reference	Description
pJMD2	McDaniel et al. 2010	Vector used for CTD plasmid construction by recursive directional ligation.
pRS315	Sikorski RS and Hieter P 1989	Empty LEU2 vector. Referred to as pLEU2 in the text.
pRPB1	Morrill et al. 2016	Plasmid with copy of RPB1 containing 26 wildtype CTD repeats, referred to as "WT" in the text.
pRPB1-CTD$_8$	Morrill et al. 2016	pRPB1 with 8 consensus CTD repeats.
pRPB1-CTD$_{10}$	Morrill et al. 2016	pRPB1 with 10 consensus CTD repeats.
pRPB1-CTD$_{14}$	Morrill et al. 2016	pRPB1 with 14 consensus CTD repeats.
pRPB1-CTD$_{26}$	Morrill et al. 2016	pRPB1 with 26 consensus CTD repeats.
pCTD$_{26}$-S>A$_{2-9}$	This work	pRPB1 with all S to A mutations in only repeats 2-9.
pCTD$_{26}$-S>A$_{10-17}$	This work	pRPB1 with all S to A mutations in only repeats 10-17.
pCTD$_{26}$-S>A$_{18-25}$	This work	pRPB1 with all S to A mutations in only repeats 18-25.
pCTD$_{26}$-S2A$_{2-9}$	This work	pRPB1 with only S2 to A2 mutations in repeats 2-9.
pCTD$_{26}$-S5A$_{2-9}$	This work	pRPB1 with only S5 to A5 mutations in repeats 2-9.
pCTD$_{26}$-S7A$_{2-9}$	This work	pRPB1 with only S7 to A7 mutations in repeats 2-9.
pCTD$_{26}$-S>A$_{2-9}$ Δ4	This work	Suppressor of pCTD$_{26}$-S>A$_{2-9}$ where 4 repeats have been deleted.
pCTD$_{26}$-S>A$_{2-9}$ Δ6	This work	Suppressor of pCTD$_{26}$-S>A$_{2-9}$ where 6 repeats have been deleted.
pCTD$_{26}$-S>A$_{10-17}$ Δ4	This work	Suppressor of pCTD$_{26}$-S>A$_{10-17}$ where 4 repeats have been deleted.
pCTD$_{26}$-S>A$_{10-17}$ Δ6	This work	Suppressor of pCTD$_{26}$-S>A$_{10-17}$ where 6 repeats have been deleted.
pCTD$_{26}$-S>A$_{10-17}$ Δ12-21 ^2-17	This work	Suppressor of pCTD$_{26}$-S>A$_{10-17}$ where repeats 12 to 21 have been deleted and repeats 2 to 17 have been added.

Appendix 3.B

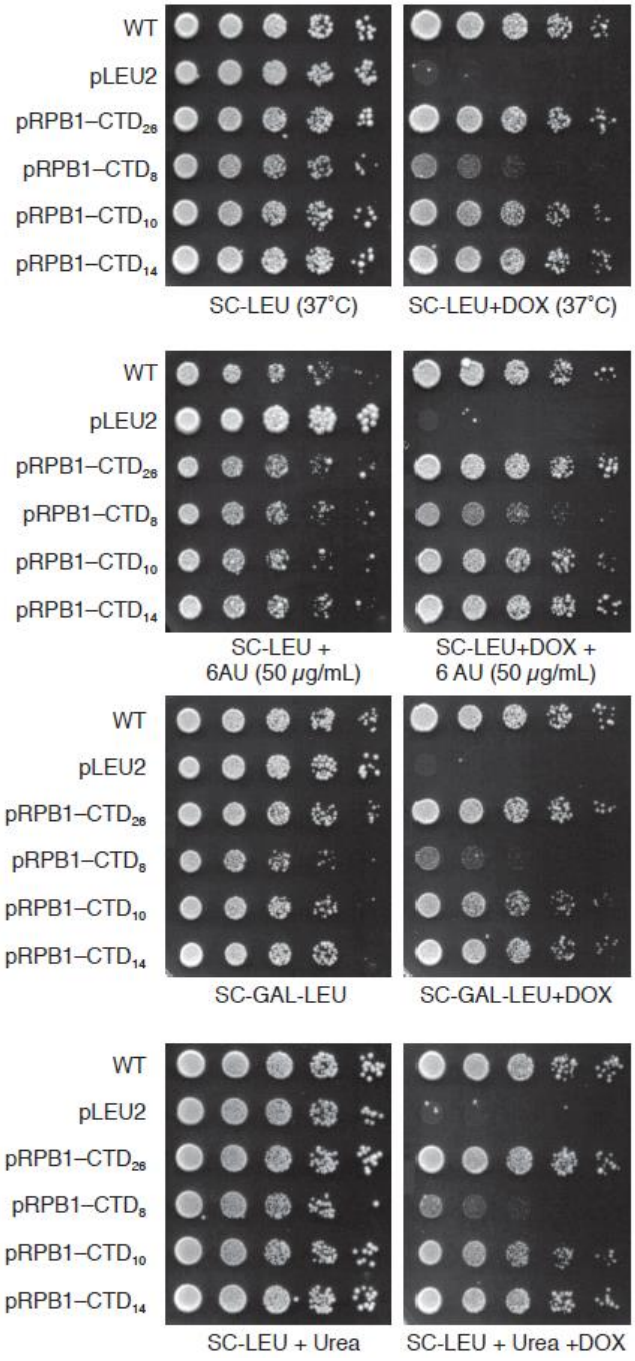


Figure 3.B-1. Additional phenotypes of CTD truncation mutants. Preparation of the spotting assay and ordering of the mutants is the same as in Figure 1. Conditions tested are: 37° C, 50 µg/mL of 6-Azauracil (6AU), plates with galactose as the only sugar (SC-GAL) and 1M urea. Pictures were taken starting at two days and a representative image of three independent experiments is presented.

Appendix 3.C

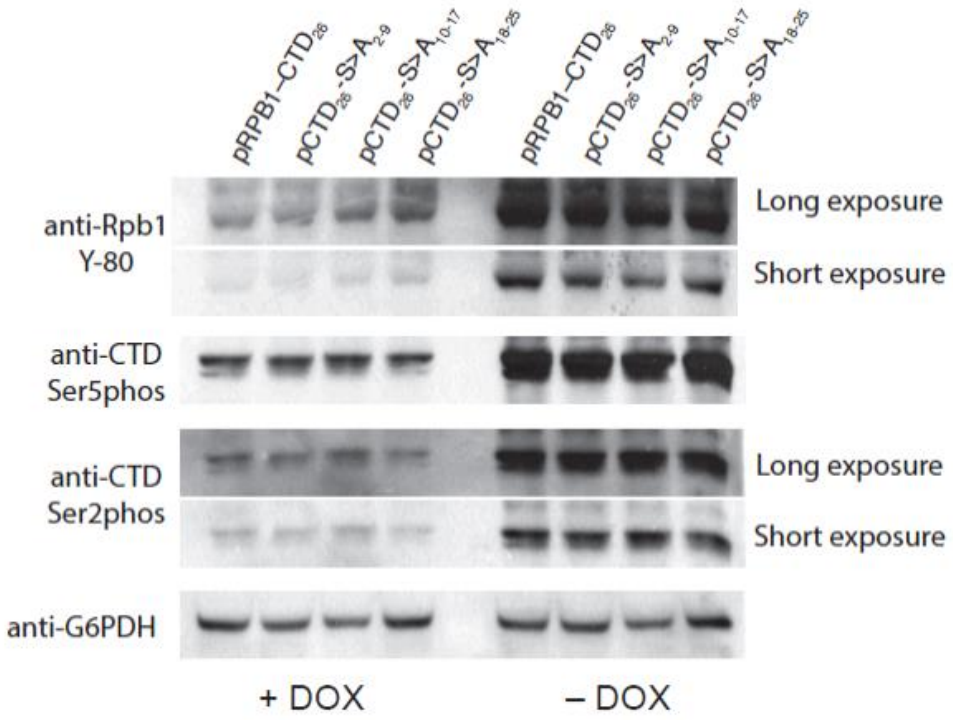


Figure 3.C-1. Expression and phosphorylation levels of position-specific mutants. Western blot of CTD constructs grown in the presence and absence of DOX. Total Rpb1p levels as well as Ser2 and Ser5 phosphorylation were assayed and compared to a G6PDH housekeeping gene loading control. A long exposure panel is provided for the Rpb1p and Ser2phos blots due to the faint signal observed for the +DOX samples.

Appendix 3.D

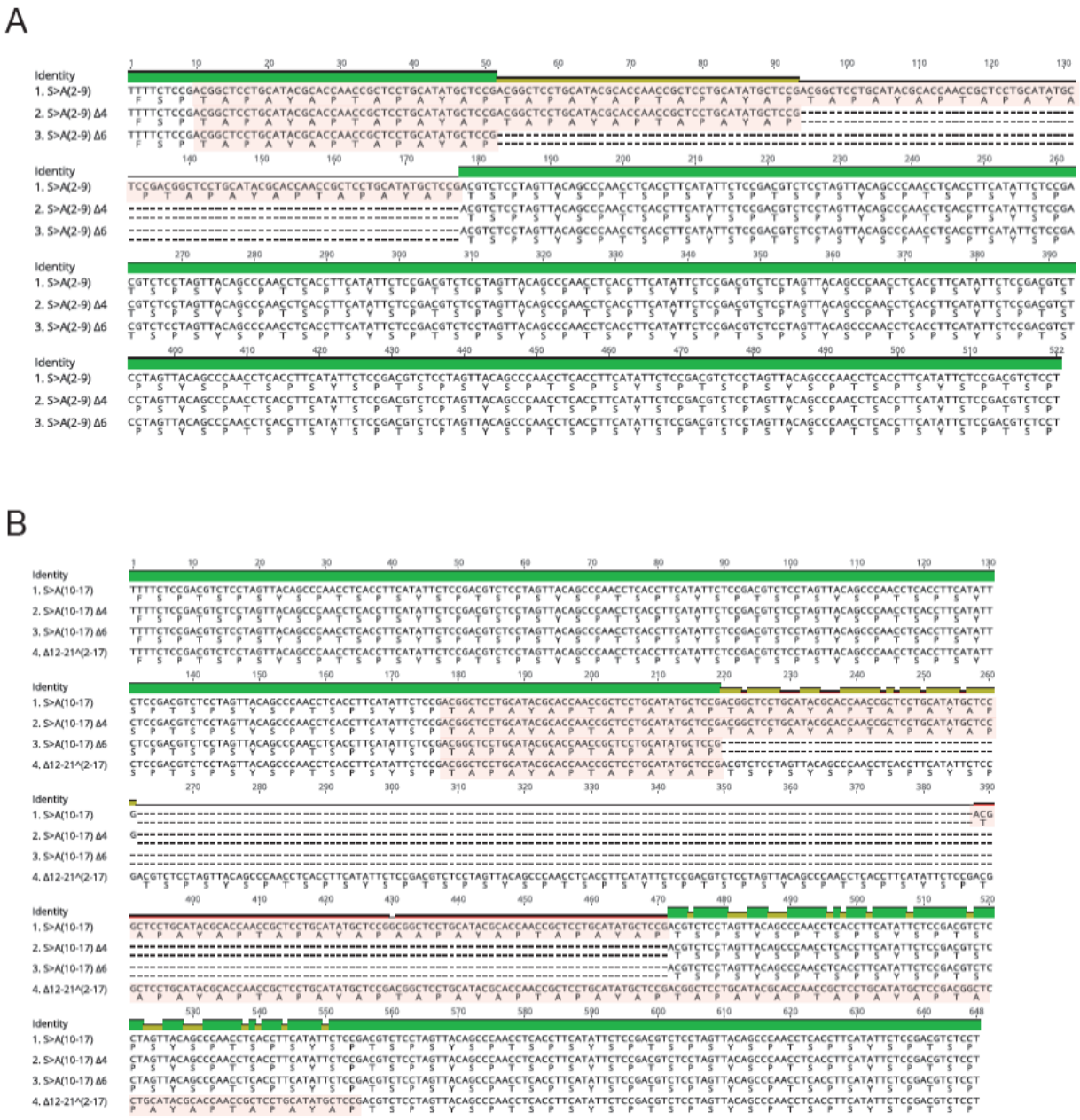


Figure 3.D-1. Sequence alignments of CTD coding region from $pCTD_{26}$–S>$A_{2\text{-}9}$ and $pCTD_{26}$–S>$A_{10\text{-}17}$ and corresponding suppressor mutants. Changes in CTD length that were observed by colony PCR were confirmed by Sanger sequencing. The position of mutated repeats is highlighted in pink. **A**) Alignment of suppressors with either four (Δ4) or six (Δ6) mutant repeats deleted from the $pCTD_{26}$–S>$A_{2\text{-}9}$ region-specific mutant. **B**) Alignment of suppressors with either four (Δ4) or six (Δ6) mutant repeats deleted or a more complex rearrangement (Δ12-21^(2-17)) from the $pCTD_{26}$–S>$A_{10\text{-}17}$ mutant.

Appendix 3.E

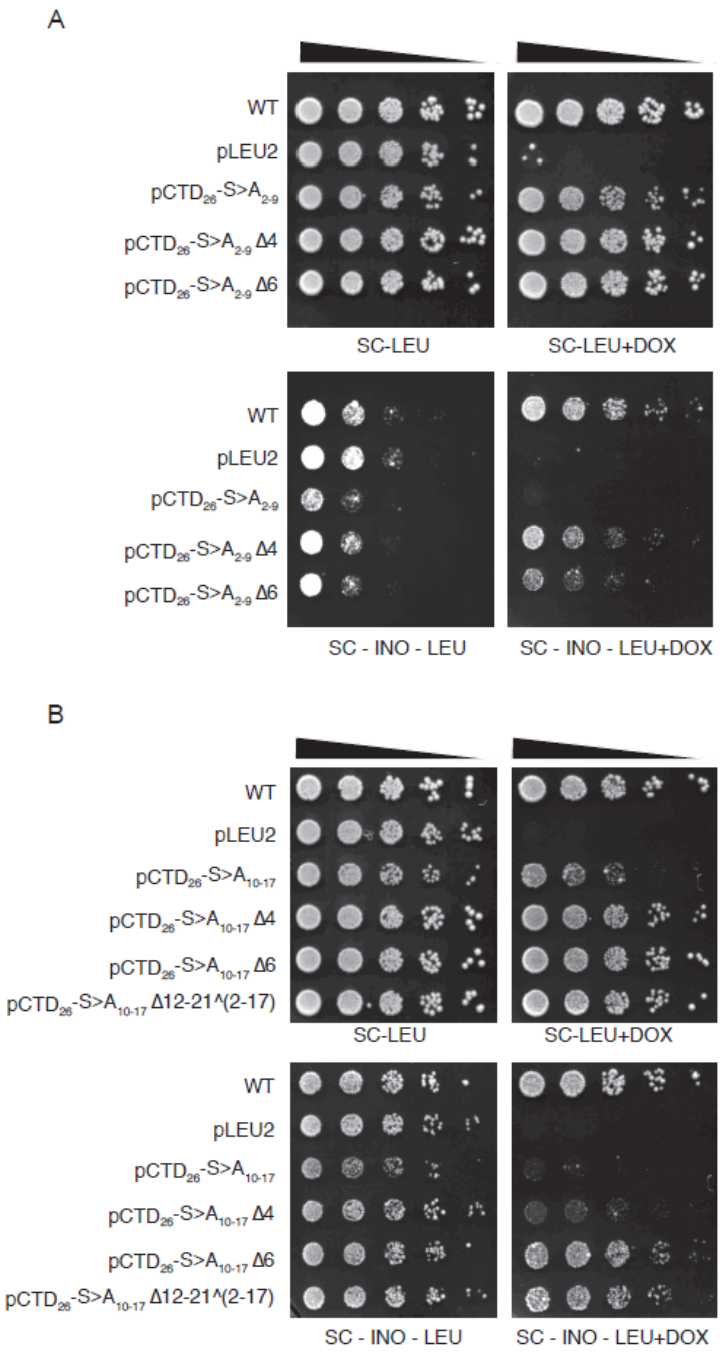


Figure 3.E-1. Improved growth of region-specific suppressors. Suppressor plasmids were retransformed into the tet-off strain GRY3019 and scored for growth on standard (+DOX) and inositol deficient (–INO+DOX) media. **A**) Spotting assay for suppressor mutants of the $pCTD_{26}$–$S>A_{2-9}$ region-specific mutant. **B**) Spotting assay for suppressor mutants of the $pCTD_{26}$–$S>A_{10-17}$ region-specific mutant.

Appendix 3.F

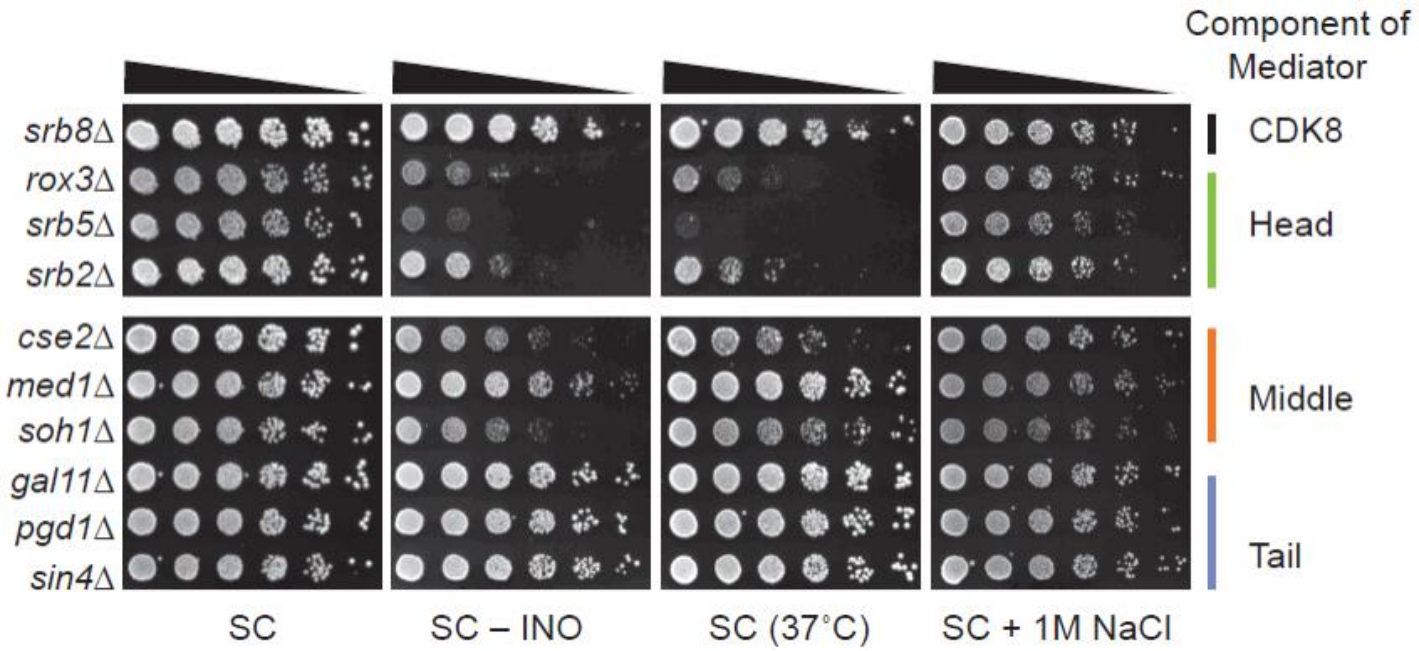


Figure 3.F-1. Phenotypes of Mediator subunit deletion strains. Yeast strains harboring deletions of the listed Mediator subunits were grown up and spotted on the indicated plate types. Strains were grouped based upon their predicted localization into the Cdk8, head, middle or tail subcomplexes (MALIK AND ROEDER 2010). Conditions tested are: inositol deficient (–INO+DOX) plates, 37° C and 1M NaCl. Pictures of plates were taken starting at two days and a representative image was selected to display.

Chapter 3 Literature Cited

Adams, A., D. E. Gottschling, C. A. Kaiser and T. Stearns, 1997 *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. Cold Spring Harbor Press, Plainview, NY.

Archambault, J., D. B. Jansma and J. D. Friesen, 1996 Underproduction of the largest subunit of RNA polymerase II causes temperature sensitivity, slow growth, and inositol auxotrophy in Saccharomyces cerevisiae. Genetics 142**:** 737-747.

Bartolomei, M. S., N. F. Halden, C. R. Cullen and J. L. Corden, 1988 Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. Mol Cell Biol 8**:** 330-339.

Corden, J. L., 2013 RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem Rev 113**:** 8423-8455.

Eick, D., and M. Geyer, 2013 The RNA polymerase II carboxy-terminal domain (CTD) code. Chem Rev 113**:** 8456-8490.

Fabrega, C., V. Shen, S. Shuman, and C. D. Lima, 2003 Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol. Cell 11: 1549–1561.

Feaver, W. J., O. Gileadi, Y. Li and R. D. Kornberg, 1991 CTD kinase associated with yeast RNA polymerase II initiation factor b. Cell 67**:** 1223-1230.

Fong, N., and D. L. Bentley, 2001 Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes Dev 15**:** 1783-1795.

Fuchs, S. M., R. N. Laribee and B. D. Strahl, 2009 Protein modifications in transcription elongation. Biochim Biophys Acta 1789**:** 26-36.

Greber, B. J., T. H. D. Nguyen, J. Fang, P. V. Afonine, P. D. Adams *et al.*, 2017 The cryo-electron microscopy structure of human transcription factor IIH. Nature 549**:** 414-417.

Hsin, J. P., A. Sheth and J. L. Manley, 2011 RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing. Science 334**:** 683-686.

Janke, C., M. M. Magiera, N. Rathfelder, C. Taxis, S. Reber *et al.*, 2004 A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast 21**:** 947-962.

Jeronimo, C., and F. Robert, 2014 Kin28 regulates the transient association of Mediator with core promoters. Nat. Struct. Mol. Biol. 21: 449–455.

Kizer, K. O., H. P. Phatnani, Y. Shibata, H. Hall, A. L. Greenleaf et al., 2005 A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Mol. Cell. Biol. 25: 3305–3316.

Kubicek, K., H. Cerna, P. Holub, J. Pasulka, D. Hrossova *et al.*, 2012 Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. Genes Dev 26**:** 1891-1896.

Lee, J. M., and A. L. Greenleaf, 1989 A protein kinase that phosphorylates the C-terminal repeat domain of the largest subunit of RNA polymerase II. Proc Natl Acad Sci U S A 86**:** 3624-3628.

Liu, P., A. L. Greenleaf and J. W. Stiller, 2008 The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. Mol Biol Evol 25**:** 719-727.

Malagon, F., M. L. Kireeva, B. K. Shafer, L. Lubkowska, M. Kashlev *et al.*, 2006 Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. Genetics 172**:** 2201-2209.

McDaniel, J. R., J. A. Mackay, F. G. Quiroz and A. Chilkoti, 2010 Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes. Biomacromolecules 11**:** 944-952.

Morrill, S. A., A. E. Exner, M. Babokhov, B. I. Reinfeld and S. M. Fuchs, 2016 DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. J Biol Chem 291**:** 11540-11550.

Nonet, M., D. Sweetser and R. A. Young, 1987 Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. Cell 50**:** 909-915.

Nonet, M. L., and R. A. Young, 1989 Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of Saccharomyces cerevisiae RNA polymerase II. Genetics 123**:** 715-724.

Nozawa, K., T. R. Schneider, and P. Cramer, 2017 Core Mediator structure at 3.4 A extends model of transcription initiation complex. Nature 545: 248–251.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt *et al.*, 2004 UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25**:** 1605-1612.

Phatnani, H. P., and A. L. Greenleaf, 2006 Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev 20**:** 2922-2936.

Powell, W., and D. Reines, 1996 Mutations in the second largest subunit of RNA polymerase II cause 6-azauracil sensitivity in yeast and increased transcriptional arrest in vitro. J Biol Chem 271**:** 6866-6873.

Robinson, P. J., D. A. Bushnell, M. J. Trnka, A. L. Burlingame and R. D. Kornberg, 2012 Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II. Proc Natl Acad Sci U S A 109**:** 17931-17935.

Robinson, P. J., M. J. Trnka, D. A. Bushnell, R. E. Davis, P. J. Mattei *et al.*, 2016 Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell 166**:** 1411-1422 e1416.

Schneider, S., Y. Pei, S. Shuman and B. Schwer, 2010 Separable functions of the fission yeast Spt5 carboxyl-terminal domain (CTD) in capping enzyme binding and transcription elongation overlap with those of the RNA polymerase II CTD. Mol Cell Biol 30**:** 2353-2364.

Schwer, B., and S. Shuman, 2011 Deciphering the RNA polymerase II CTD code in fission yeast. Mol Cell 43**:** 311-318.

Sennett, N. C., and D. J. Taatjes, 2014 Mediator redefines itself. Cell. Res. 24: 775–776.

Singh, H., A. M. Erkine, S. B. Kremer, H. M. Duttweiler, D. A. Davis et al., 2006 A functional module of yeast mediator that governs the dynamic range of heat-shock gene expression. Genetics 172: 2169–2184.

Spahr, H., G. Calero, D. A. Bushnell and R. D. Kornberg, 2009 Schizosacharomyces pombe RNA polymerase II at 3.6-A resolution. Proc Natl Acad Sci U S A 106**:** 9185-9190.

Stiller, J. W., and M. S. Cook, 2004 Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. Eukaryot Cell 3**:** 735-740.

Suh, H., S. B. Ficarro, U. B. Kang, Y. Chun, J. A. Marto *et al.*, 2016 Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Mol Cell 61**:** 297-304.

Valay, J. G., M. Simon, M. F. Dubois, O. Bensaude, C. Facca *et al.*, 1995 The KIN28 gene is required both for RNA polymerase II mediated transcription and phosphorylation of the Rpb1p CTD. J Mol Biol 249**:** 535-544.

Villa-Garcia, M. J., M. S. Choi, F. I. Hinz, M. L. Gaspar, S. A. Jesch *et al.*, 2011 Genome-wide screen for inositol auxotrophy in Saccharomyces cerevisiae implicates lipid metabolism in stress response signaling. Mol Genet Genomics 285**:** 125-149.

Wang, X., Q. Sun, Z. Ding, J. Ji, J. Wang et al., 2014 Redefining the modular organization of the core Mediator complex. Cell. Res. 24: 796–808.

Werner-Allen, J. W., C. J. Lee, P. Liu, N. I. Nicely, S. Wang *et al.*, 2011 cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. J Biol Chem 286**:** 5717-5726.

West, M. L., and J. L. Corden, 1995 Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics 140**:** 1223-1233.

Wilcox, C. B., A. Rossettini and S. D. Hanes, 2004 Genetic interactions with C-terminal domain (CTD) kinases and the CTD of RNA Pol II suggest a role for ESS1 in transcription initiation and elongation in Saccharomyces cerevisiae. Genetics 167**:** 93-105.

Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson et al., 1999 Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901–906.

Wong, K. H., Y. Jin and K. Struhl, 2014 TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape. Mol Cell 54**:** 601-612.

Yoh, S. M., J. S. Lucas and K. A. Jones, 2008 The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. Genes Dev 22**:** 3422-3434.

Young, B. P., J. J. Shin, R. Orij, J. T. Chao, S. C. Li et al., 2010 Phosphatidic acid is a pH biosensor that links membrane biogenesis to metabolism. Science 329: 1085–1088.

Zhang, M., G. N. Gill, and Y. Zhang, 2010 Bio-molecular architects: a scaffold provided by the C-terminal domain of eukaryotic RNA polymerase II. Nano Rev. 1.

Zhang, J., and J. L. Corden, 1991 Phosphorylation causes a conformational change in the carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J Biol Chem 266**:** 2297-2302.

Chapter 4

Molecular mechanisms of CTD repeat-specific activity

Abstract

The repeats of the C-terminal domain (CTD) of RNA polymerase II display a remarkable specialization of function despite their identical sequence in budding yeast. As outlined in chapter 3, the proximal set of repeats were required specifically for inducible gene expression, the middle set had a general viability requirement and the distal set of repeats appeared to be functionally redundant. However, the mechanisms and the protein factors behind this region-specific activity were not explored in depth. In this chapter I will present genetic and biochemical evidence for the interactions that underlie the region-specific mechanism of the CTD repeats. Genetic analyses of Mediator complex subunits indicate distinct interactions with certain regions of the CTD depending on the cellular pathway in question. These studies are complemented by immunoprecipitations of Mediator-CTD complexes that demonstrate differential recruitment of Mediator to the various regions of the CTD. Lastly, the middle set of repeats is specifically required for histone modification by Set2p. These preliminary results provide the first hints as to how identical repeats can have specific functions and marks the initial step towards comprehensively mapping protein-protein interactions on the CTD repeats of RNA polymerase II.

Introduction

The essential transcriptional activity of RNA polymerase II is dependent on the repetitive C-terminal domain (CTD) of the largest subunit Rpb1p. The CTD consists of heptad repeats of the sequence $Y_1S_2P_3T_4S_5P_6S_7$ whose copy number roughly correlates with the complexity of the organism. The CTD repeats promote transcription by recruiting various protein factors that produce the mature transcript and facilitate transcription through chromatin (Corden 2013). Presently, one of the main outstanding questions in the field is how the CTD is able to spatially and temporally organize the wide range of protein factors throughout the transcription cycle to enable RNA polymerase II activity. A commonly cited framework for understanding the function of the CTD is the “CTD code” model. This model states that certain combinations of post-translational modifications (PTMs) of the heptad repeats mediate protein factor binding at the appropriate step of the transcription cycle. The CTD code model has proved to be a powerful framework to explain the serine phosphorylation-mediated transitions through the main steps of the transcription cycle (Harlen and Churchman 2017). However, recent discoveries of functional redundancies in the sites of CTD modification have cast doubt on the exact code-like nature of the CTD (Suh et al. 2016, Schuller et al. 2016). As a consequence, there is now a need to examine additional mechanisms of CTD functional specialization during transcription.

Repeat degeneration by sequence changes in the heptad repeat is the best understood process that leads to functional specialization in the CTD. While almost all eukaryotes, with the exception of constitutive parasites, maintain the heptad register of the CTD,

divergence from the consensus $Y_1S_2P_3T_4S_5P_6S_7$ sequence is commonly observed (Yang and Stiller 2014). One particularly well-known example is the CTD of the fruit fly *Drosophila melanogaster* that contains residue substitutions in almost all of the heptads of its 45 repeat CTD. Mammalian heptads typically contain substitutions at the S_7 position of the distal half of their CTDs, and these substitutions are associated with PTMs that recruit mammalian-specific protein factors to the site of transcription (Eick and Geyer 2013). Region specialization was also hinted at in budding yeast (West and Corden 1995), although the near-perfect consensus repeats in yeast ruled out the kind of sequence-specific functions that are found in more complex organisms. Instead, specific function was tied to the position of particular repeats along the yeast CTD (Babokhov et al. 2018). Further dissection of this region specificity mechanism will require examination of the protein factors that could differentially bind to certain regions of the CTD.

Out of all of the protein factors that bind the CTD, the Mediator complex is by far the largest and most multifunctional interactor. Mediator is a large complex consisting of up to 25 subunits in yeast and up to 30 in humans. The complex is an important transcription factor and is especially required for activated (enhancer-driven) gene expression. Mediator can be divided up into four sub complexes or modules consisting of the head, the middle, the tail and Cdk8 modules (Poss et al. 2013). Each of the modules are separated by flexible linker domains that enable drastic conformational changes in the complex that are essential for function (Tsai et al. 2014). Mediator primarily acts by interacting with DNA-bound transcription factors, integrating and

transmitting the signals to RNA polymerase II. Mediator subunits are also required to promote the phosphorylation of the Serine-5 residue in the CTD repeats by TFIIH, which signals the start of transcription (Baidoobonso et al. 2007). The CTD has been shown to bind Mediator along the boundary between the head and middle modules and unmodified serine residues are especially important for this interaction (Robinson et al. 2012). Although general models of Mediator binding to RNA polymerase II exist, how the CTD binds Mediator and the effect of this binding on other protein factors is still unknown. Given the requirement of the CTD-Mediator interaction for the initiation of transcription, further elucidation of the binding arrangement is necessary to appreciate this important step of transcription.

In addition to the Mediator complex, there are a number of other important protein factors that bind to the CTD to enable transcription. Following the initiation of transcription, capping enzymes bind to Serine-5 phosphorylated repeats and add the 5' cap to the emerging transcript to protect it from degradation (Cho et al. 1997). During active transcription, elongation factors such as Spt4/5 and the Paf1 complex associate with the CTD to promote elongation of the transcript (Mayekar et al. 2013). Finally, transcription termination factors bind the Serine-2 phosphorylated CTD to finish transcription and add the poly-A tail to the end of the transcript to promote nuclear export (Dunn et al. 2005). The elongating RNA polymerase II complex must also pass through and regulate chromatin structure, which is aided by a variety of polymerase-interacting proteins. Once such enzyme is the methyltransferase Set2p, which methylates histone H3 at the Lysine-36 residue (H3K36me) to antagonize histone

acetylation, preventing cryptic transcription (Fuchs et al. 2012). The variety of protein factors that must bind to the CTD further highlights the need to understand how region specific determinants of the repeats coordinate this suite of functions.

In this chapter, I present evidence that builds upon the foundational work of chapter 3 and demonstrates how region specific functions track with known CTD binding proteins. Although many of the experiments are preliminary in nature, they help to establish the future directions for studying region specificity in the budding yeast CTD. Genetic analyses of three viable deletion mutants of Mediator subunits demonstrate subunit and stress-specific phenotypes when combined with the region specific mutants from chapter 3. These genetic studies are complemented with immunoprecipitations of the Mediator-CTD complex, indicating a requirement for the proximal CTD region to recruit Mediator. Lastly, western blot analysis of H3K36me3 levels in the region specific mutants implies a role for the middle eight CTD repeats in recruiting the methyltransferase Set2p. Taken together, these experiments begin to piece together the region specific requirements of protein factor binding and pave the way for future analysis of the biochemical organization of co-transcriptional processes on the CTD.

Methods

Strains and plasmids

A description of the strains used in this chapter is available in available in Appendix 4.A. Mediator subunit deletion strains with the RPB1 Tet-off system were created by crossing derivatives of GRY3019 (Malagon et al. 2006) with selected strains of the

yeast deletion collection (Winzeler et al. 1999) using standard methods. The resulting double mutants were then transformed with the region specific block mutants described in chapter 3 and listed in Appendix 3.A. Tagged Mediator subunits were created through homologous recombination using c-myc cassettes targeted to the C-terminus of the selected subunits and verified by PCR and western blotting (Janke et al. 2004). Dominant drug resistance markers KanMX6, HphNT1 and NatNT2 were selected for using 50 µg/mL of geneticin (G418), hygromycin B and nourseothricin (ClonNAT), respectively. Ammonium sulfate was replaced with 1 g/L of monosodium glutamate as a nitrogen source whenever these drugs were used for selection in liquid media or plates.

Spotting assays

The growth of Mediator double mutants was examined by spotting assays as described previously (Babokhov et al. 2018). The four conditions tested were standard synthetic complete (SC) media lacking leucine (–LEU) to select for CTD plasmids, SC–LEU at an elevated 37° C growth temperature, SC–LEU media lacking inositol (SC–LEU–INO) and SC–LEU with 1M NaCl added as an osmotic stress (SC–LEU+1M NaCl). Strains were spotted to the four conditions with and without doxycycline (+/– DOX), pictures were taken starting at two days and representative images were selected for display.

Western Blots

Western blotting was performed as described previously (Babokhov et al. 2018). Primary antibodies used for tagged Mediator strains were anti-c-myc (9E10, Invitrogen) and anti-HA (Gallus Immunotech). For the histone modification westerns the antibodies

used were: anti-H3 (ab1791, abcam), anti-H3K36me3 (ab9050, abcam), anti-Set2p (Fuchs et al. 2012) and anti-G6PDH (ab9485, abcam).

Mediator immunoprecipitation

Co-immunoprecipitation of myc-tagged Srb4p (Med17 under the standard nomenclature) Mediator subunits and HA-tagged Rpb3p was adapted from (Wittermeyer et al. 2004). Briefly, tagged yeast strains were grown up in SC–LEU+DOX media to mid log phase, harvested, washed twice with PBS, split into pellets corresponding to 100 mL of culture and stored at –80° C. Pellets were taken out and resuspended in lysis buffer (50 mM HEPES, pH 7.5; 250 mM potassium acetate, pH 7.5; 10% glycerol (v/v); 10 mM Na-EDTA; 0.5 mM DTT; and protease inhibitors) and glass beads were added for mechanical lysing. Cells were lysed by vortexing for 1 min at 4° C and cooled for 1 min at 4° C for 5 cycles. Cell lysate was then clarified by centrifugation and nucleic acids were precipitated out using 0.1% polyethylenimine in the presence of 400 mM potassium acetate. The clarified lysate was incubated with 10µL of anti-c-myc primary antibody for one hour and then with protein A/G beads (88802, Pierce) for two hours to bind complexes. The beads were washed three times with lysis buffer and complexes were eluted by boiling in the presence of SDS loading buffer. The eluted samples were loaded into a 10% SDS-PAGE gel and visualized by western blotting. A full protocol for the experiments reported in this chapter is available in Appendix 4.B.

Results

Mediator-CTD region double mutant rationale

My previous examination of region specific function of the CTD repeats found that three broad regions of the CTD (proximal, middle and distal) all had different properties under a range of tested phenotypic conditions (Babokhov et al. 2018). In order to begin to unravel the protein factors responsible for these different behaviors, I turned to viable deletion mutants of Mediator complex subunits with known defects under the inositol auxotrophy phenotype (Young et al. 2010). I found that these deletion mutants had different responses to the stresses (high temperature, osmotic, no inositol) that I commonly worked with and would be ideal to test in conjunction with my region specific CTD mutant strains. Genetic crosses between these two strains and transformation of CTD mutant plasmids yielded double mutants that had both a Mediator subunit deleted and a region specific mutant CTD. Comparing the response of the double mutant versus each of the single mutants to stress therefore permits an analysis of the genetic interaction between these two factors. Under classical yeast genetics, an additive effect is indicative of different pathways while no additive effect suggests the two factors are in the same pathway. These genetic interactions can then be used to infer potential region specific interactions between the CTD and Mediator.

Double mutants grow normally under standard conditions

To begin the genetic analysis, I started with three Mediator subunit deletion mutants, one in the head module (*SRB2/MED20*) and two in the middle module (*CSE2/MED9* and *SOH1/MED31*) (Figure 8A). All three mutants of these subunits demonstrated mild

phenotypes upon deletion that are amenable to genetic interaction analysis (Young et al. 2010). The single mutants were crossed to the *RPB1* TET-off strain and the resulting spores were transformed with the region specific CTD block mutants to obtain the final double mutant strains. Examining the double mutants by spotting assay on SC–LEU plates with and without DOX enabled the viability of the strains to be scored in the absence of stress (Figure 1). The double mutants grew comparably to both the TET-off (Figure 1A) and subunit deletion (Figure 1B) single mutants in the absence of DOX and this condition served as the loading control for the spotting assay experiments. Upon the addition of DOX, the *srb2Δ* double mutant grew comparably to the TET-off single mutants (Figure 1D) while the *ce2Δ* and *soh1Δ* double mutants showed slightly more robust growth (Figure 1C and 1E).

Subunit-specific effects at high temperature

The overall good health of the double mutants under standard conditions facilitates the study of these mutants under stress conditions. Increasing the incubation temperature of yeast from a permissive 30° C to a restrictive 37° C places a heat stress on the $pCTD_{26}$–$S>A_{10-17}$ (Figure 2A) and the *srb2Δ* and *cse2Δ* single mutants (Figure 2B). The double mutants responded much more drastically to the heat stress. The *srb2Δ* double mutant showed a complete loss of viability for all CTD constructs under both –DOX and +DOX conditions (Figure 2C). The *cse2Δ* double mutant was comparable to the single mutants at –DOX, and in the +DOX condition demonstrated an additive effect with the $pCTD_{26}$–$S>A_{2-9}$ plasmid while showing no interactions with the other CTD constructs (Figure 2D). The *soh1Δ* double mutant had a slight growth defect in –DOX while

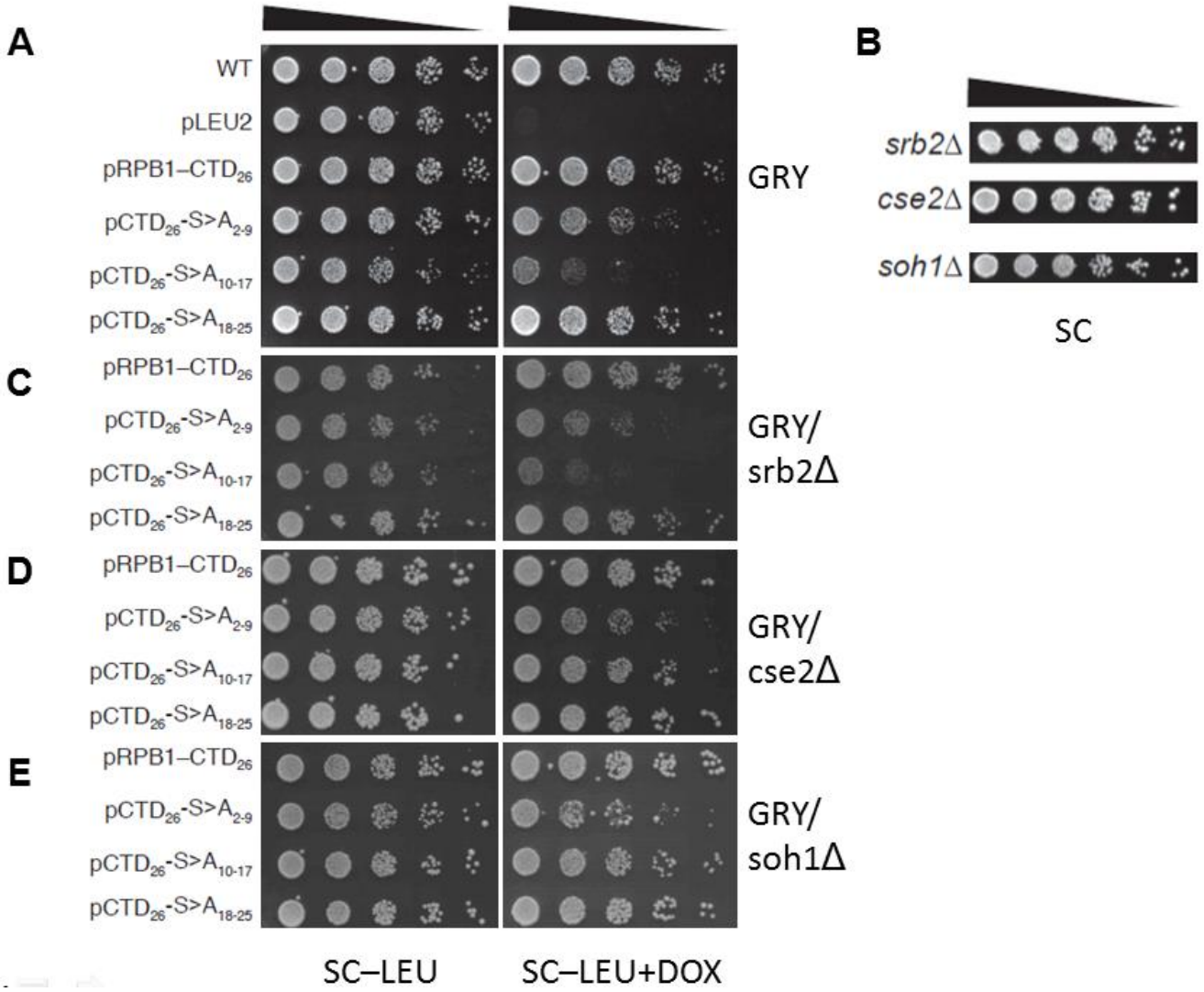


Figure 1. Phenotypes of double mutants under standard growth conditions. Spotting assays are displayed measuring growth of five-fold dilutions of the indicated strains. **A**) Growth of the TET-off strain (GRY) single mutants. Panel adapted from figure 2B of (Babokhov et al. 2018). **B**) Growth of Mediator subunit deletion single mutants. Panel adapted from figure S5 of (Babokhov et al. 2018). TET-off and Mediator subunit deletion double mutant *srb2Δ* (**C**) *cse2Δ* (**D**) and *soh1Δ* (**E**) growth were measured by spotting assay to determine genetic interactions.

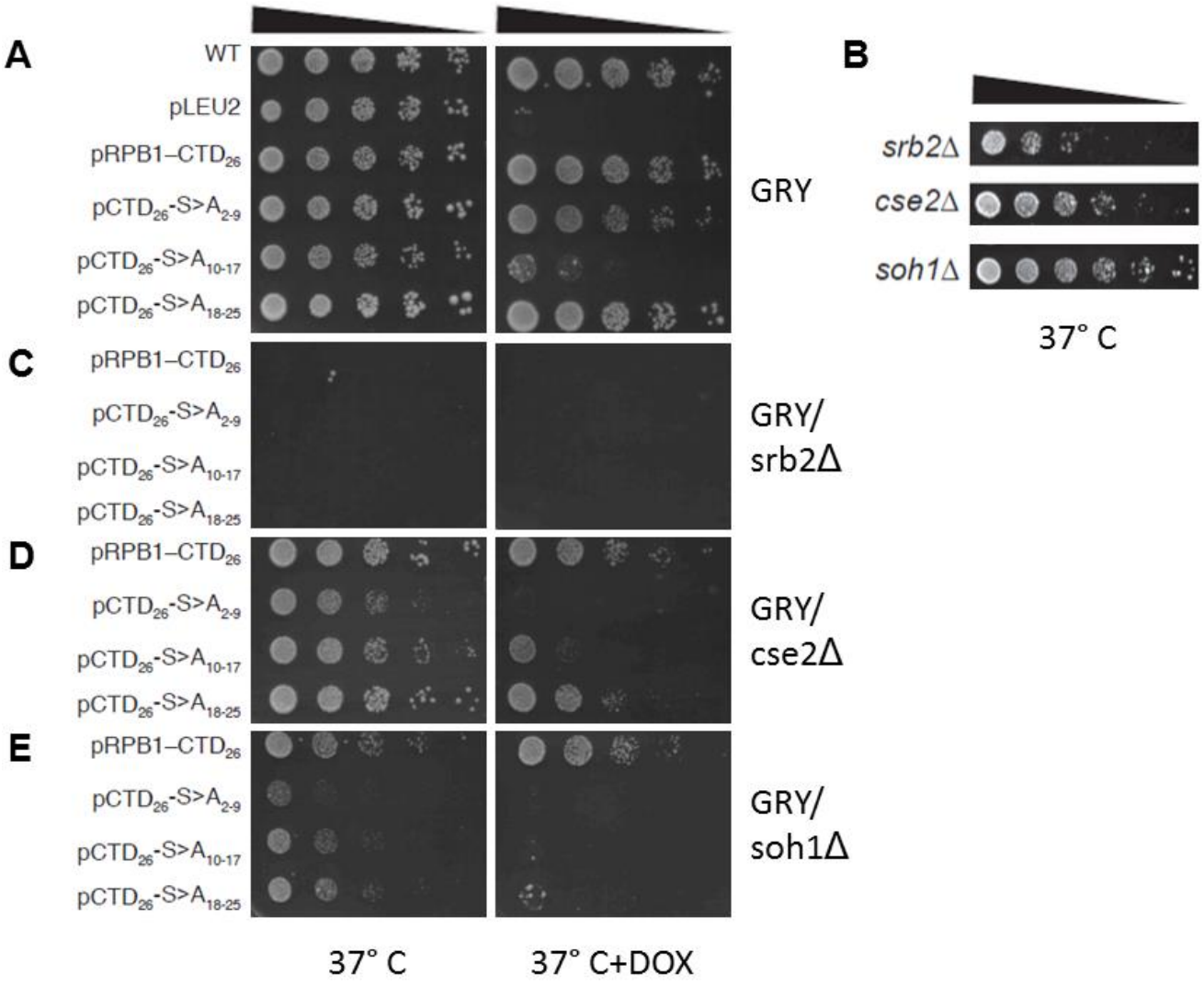


Figure 2. Phenotypes of double mutants under high temperature (37° C) stress. Spotting assays are displayed measuring growth of five-fold dilutions of the indicated strains. **A**) Growth of the TET-off strain (GRY) single mutants. **B**) Growth of Mediator subunit deletion single mutants. Panel adapted from figure S5 of (Babokhov et al. 2018). TET-off and Mediator subunit deletion double mutant *srb2Δ* (**C**) *cse2Δ* (**D**) and *soh1Δ* (**E**) growth were measured by spotting assay to determine genetic interactions.

showing a complete loss of viability for the three mutant CTD constructs in +DOX (Figure 2E). Overall, the three double mutants all displayed different behaviors to the original single mutants and to each other, indicating the complex genetic interactions between Mediator and the CTD.

Double mutants are inviable under inositol auxotrophy

Under the original study of CTD region specificity, only the inositol auxotrophy phenotype was linked to a specific defect in the first eight repeats of the CTD (Babokhov et al. 2018). Therefore, it was of considerable interest to see if there were any genetic interactions between the first eight repeats and the Mediator subunit deletion mutants. However, parsing out any interactions was made difficult by the complete loss of viability of the double mutants in media lacking inositol for both the –DOX and +DOX conditions (Figure 3C-D). While the $pCTD_{26}$–$S>A_{10\text{-}17}$ single mutant showed an almost complete loss of viability (Figure 3A), the subunit deletion single mutants showed only modest defects in media lacking inositol (Figure 3B). Consequently, the complete loss of viability in all of the double mutants in –INO was surprising. This result suggests that other factors aside from the specific interaction between Mediator and the CTD may be at play in determining function under inositol auxotrophy.

Subunit-specific response to osmotic stress

Osmatic stress due high levels of salt in the growth media triggers an osmotic and general stress response in yeast that is exacerbated by transcriptional mutants. In the

presence of 1 M NaCl, only the $pCTD_{26}$–$S>A_{10\text{-}17}$ single mutant is rendered inviable while the other block mutants are unaffected (Figure 4A). The subunit deletion single mutants are also unaffected by the presence of 1 M NaCl (Figure 4B). In contrast, the combined double mutants all showed different responses to the osmotic stress. The *srb2Δ* double mutant grew at a significantly reduced rate for all except the $pCTD_{26}$–$S>A_{10\text{-}17}$ plasmid, where it mimicked the single mutant for both the –DOX and +DOX conditions (Figure 4C). The outcome was much different for the *cse2Δ* double mutant, which had an additive effect for the $pCTD_{26}$–$S>A_{2\text{-}9}$ plasmid while rescuing the growth defect seen in the single mutant of the $pCTD_{26}$–$S>A_{10\text{-}17}$ plasmid (Figure 4D). Lastly, the *soh1Δ* double mutant fully copied the phenotypes of the single mutants for all plasmids (Figure 4E). These assays reveal that there are subunit-specific interactions underlying the Mediator-CTD stress response to high environmental salt levels.

Srb4-myc reveals genetic interactions with CTD regions

The genetic interaction analysis of Mediator and RNA polymerase II indicated that there were multiple pathway and subunit-dependent interactions between these two complexes (Figure 8B). In order to gain further insight into these interactions, I launched a complementary biochemical approach to look at the physical interaction between the Mediator complex and the various CTD region mutants. I introduced c-myc tags into the C-termini of Mediator subunits that were previously used for immunoprecipitations (Liu and Myers 2012). All but one (Med8-myc) of the c-myc tags were successfully integrated (Figure 6A) and Srb4-myc was selected due to its considerable difference in

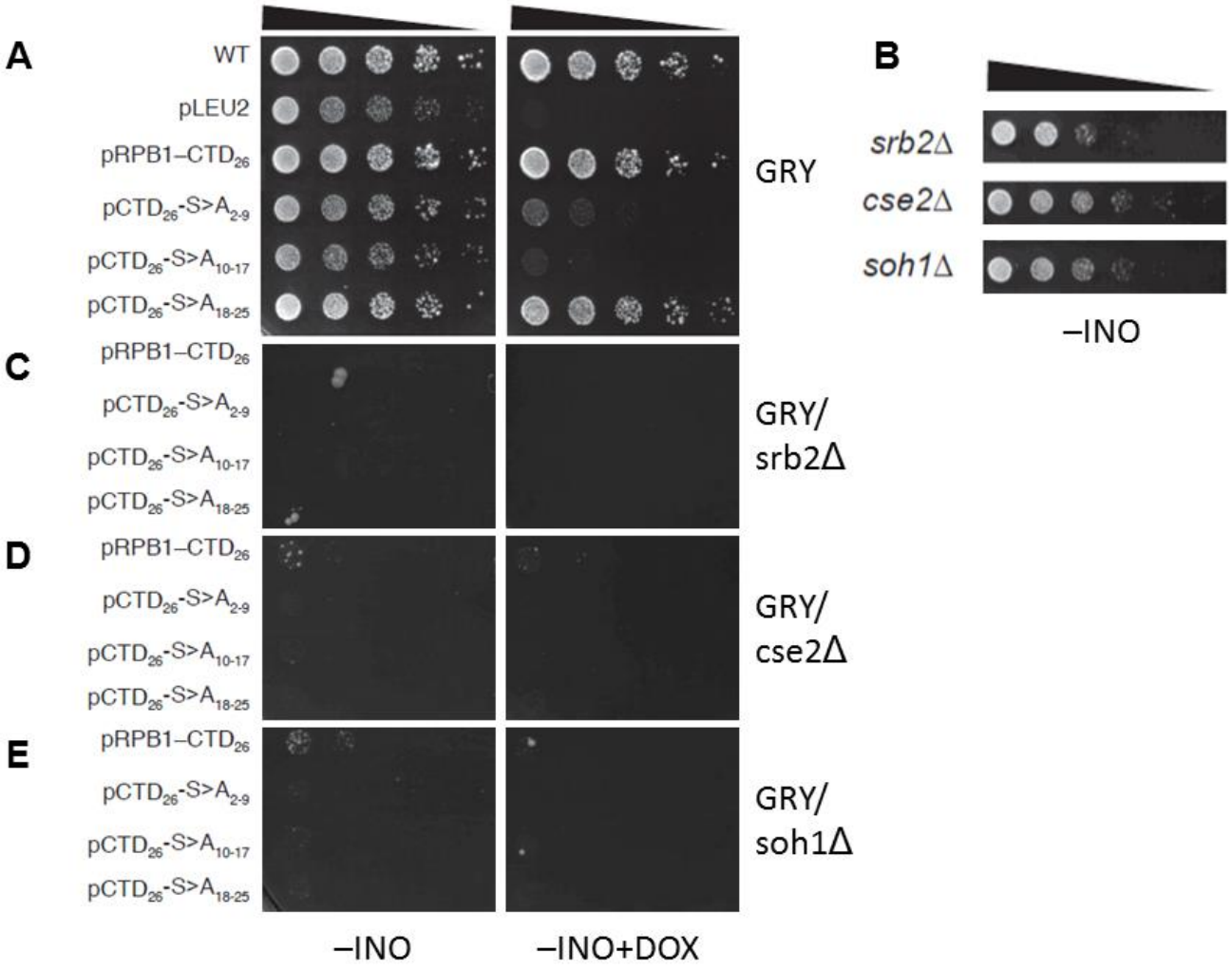


Figure 3. Phenotypes of double mutants under inositol auxotrophy. Spotting assays are displayed measuring growth of five-fold dilutions of the indicated strains. **A**) Growth of the TET-off strain (GRY) single mutants. Panel adapted from figure 2C of (Babokhov et al. 2018). **B**) Growth of Mediator subunit deletion single mutants. Panel adapted from figure S5 of (Babokhov et al. 2018). TET-off and Mediator subunit deletion double mutant *srb2Δ* (**C**) *cse2Δ* (**D**) and *soh1Δ* (**E**) growth were measured by spotting assay to determine genetic interactions.

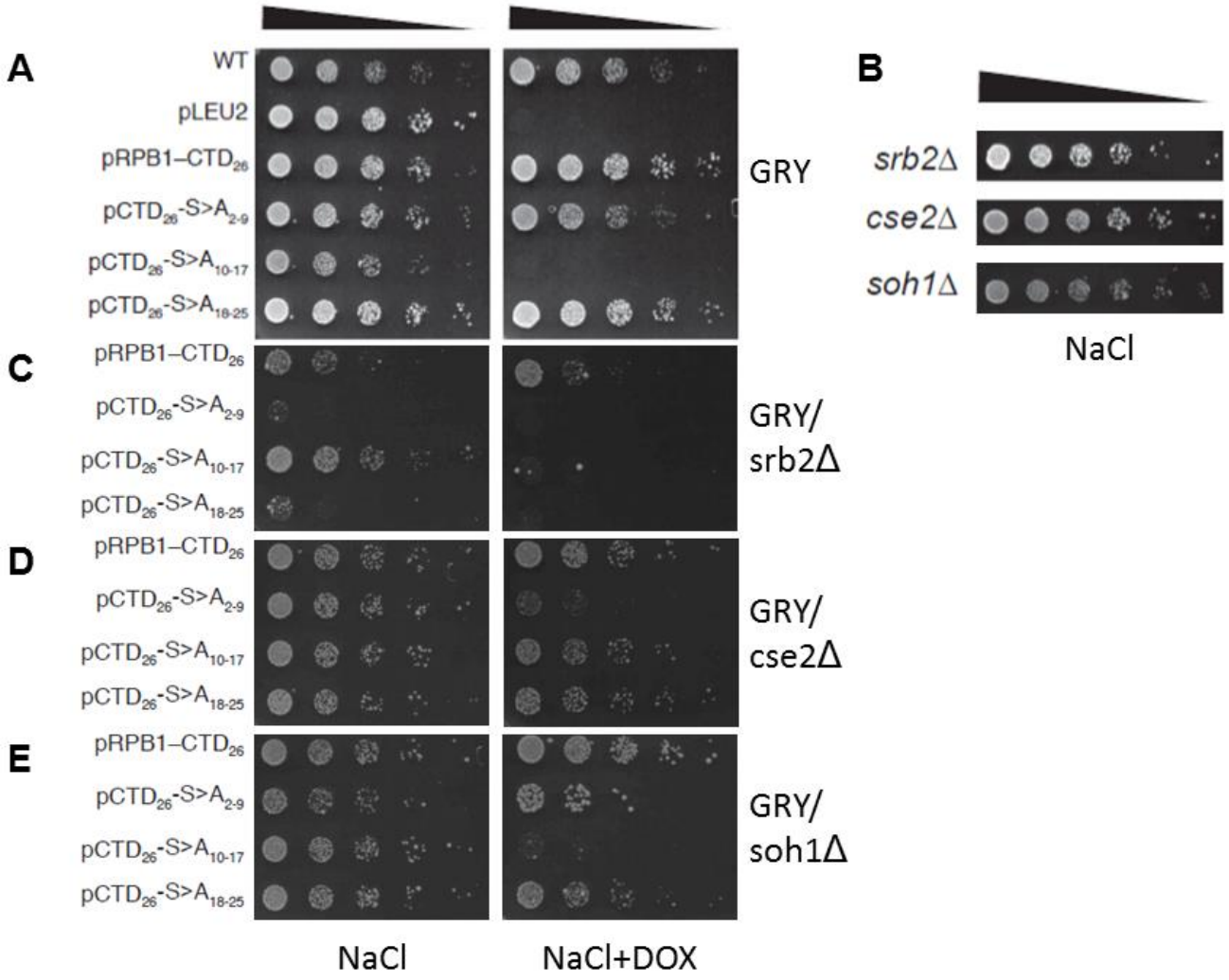


Figure 4. Phenotypes of double mutants under osmotic stress (1 M NaCl). Spotting assays are displayed measuring growth of five-fold dilutions of the indicated strains. **A**) Growth of the TET-off strain (GRY) single mutants. Panel adapted from figure 3 of (Babokhov et al. 2018). **B**) Growth of Mediator subunit deletion single mutants. Panel adapted from figure S5 of (Babokhov et al. 2018). TET-off and Mediator subunit deletion double mutant *srb2Δ* (**C**) *cse2Δ* (**D**) and *soh1Δ* (**E**) growth were measured by spotting assay to determine genetic interactions.

size from the Rpb3p subunit that would be used to check for pulldown of the RNA polymerase II complex.

During the initial culturing of yeast cells to produce protein extracts, I noticed that the tagged Mediator strain grew slightly slower than the typical TET-off strains, likely due to the nine integrated c-myc tags interfering slightly with the function of the complex. I reasoned that this was an opportunity to study the genetic interactions between *SRB4*, a known CTD interactor (Robinson et al. 2012), and the region specific mutants of the CTD, given that *SRB4* is an essential gene with no available deletion mutants. Spotting assays of the tagged Mediator strains for the standard conditions and the three stress phenotypes (high temperature, no inositol, osmotic) described above revealed different genetic interactions for the three CTD regions. Under standard conditions, the tagged strains behaved roughly the same as the untagged strains, with a slight additive effect in the $pCTD_{26}$–$S>A_{2\text{-}9}$ region (Figure 5A). At high temperature, the tagged strains showed an additive effect for the $pCTD_{26}$–$S>A_{2\text{-}9}$ region while the other two regions were unchanged (Figure 5B). The tagged strains were particularly sensitive to the inositol auxotrophy phenotype and were all inviable when combined with a mutant CTD plasmid (Figure 5C). Lastly, the 1 M NaCl condition was similar to the 37° C stress, with an additive effect for the $pCTD_{26}$–$S>A_{2\text{-}9}$ region while the other regions showed the same outcome (Figure 5D). As a result, the slight growth defect caused by the c-myc tag under stress conditions expands the analysis of Mediator-CTD genetic interactions to the Srb4p subunit.

Mediator preferentially binds proximal CTD repeats

Having established the phenotypes of the Srb4-myc strain under stress, I next moved on to immunoprecipitating RNA polymerase II complexes. The Srb4-myc strain grew poorly under stress conditions (Figure 5), so I was limited to growth in standard media to examine the physical interaction of Mediator with the region specific CTD. Other tagged subunits, including the known CTD binder Med6p (Figure 6A), may not have the same phenotype and could be used to examine interactions under stress conditions. As an initial preliminary trial, I immunoprecipitated Srb4-myc in +DOX conditions and examined the amount of RNA polymerase II (as measured by the structural Rpb3p subunit) that eluted in the various CTD mutants. Input controls all eluted at equal levels, indicating that equivalent amounts of protein were loaded onto the beads (Figure 6B). Looking at eluted Rpb3p, all three region specific mutants showed reduced levels compared to a wildtype CTD, with almost no eluted protein for the $pCTD_{26}$–$S>A_{2\text{-}9}$ mutant. There was also slightly different elution of the Srb4-myc protein, indicating uneven elution likely due to loss of beads during the boiling and recovery process (Figure 6B). Overall, these results suggest that Mediator binding to RNA polymerase II is affected by region specific mutants, especially by mutation of the proximal repeats.

The CTD middle region is required for H3K36 trimethylation

Many protein factors are known to bind to the CTD to ensure smooth progression through the transcription cycle. To look for further potential region specific binding events I turned to the histone methyltransferase Set2p, which binds to the CTD to methylate histone H3 at the lysine 36 residue (H3K36me). I probed protein extracts from the CTD region mutant strains using an antibody specific to the H3K36me3 mark.

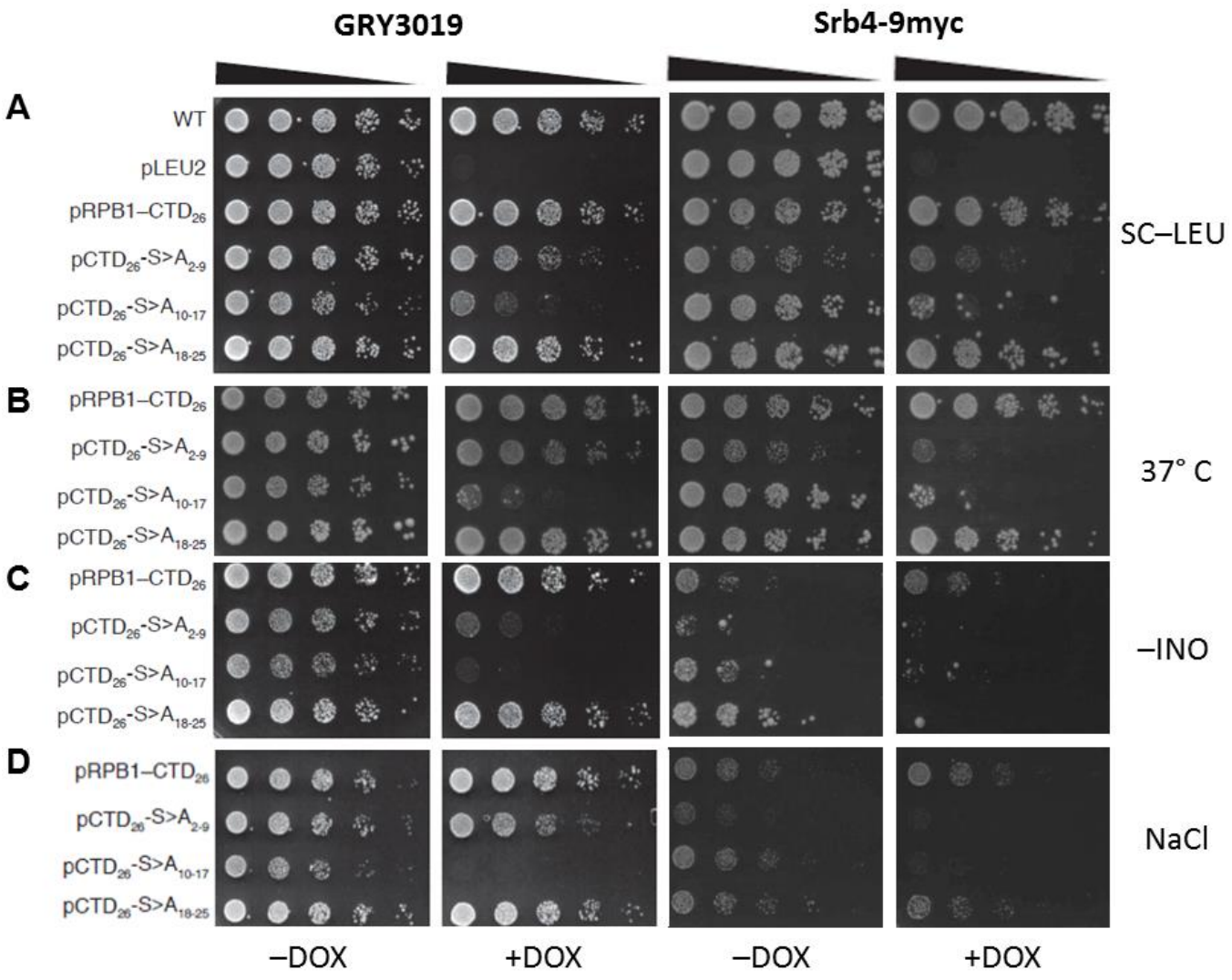


Figure 5. Phenotypes of CTD region mutant and Srb4-myc double mutants under a panel of stresses. Spotting assays are displayed measuring growth of five-fold dilutions of the indicated strains. Untagged GRY3019 spotting assay panels were adopted from (Babokhov et al. 2018) **A**) Growth of both the untagged GRY3019 strain and the tagged Srb4-myc strain with CTD region mutants under standard growth conditions. Stress conditions were tested at high temperature (**B**) inositol auxotrophy (**C**) and osmotic stress (**D**)

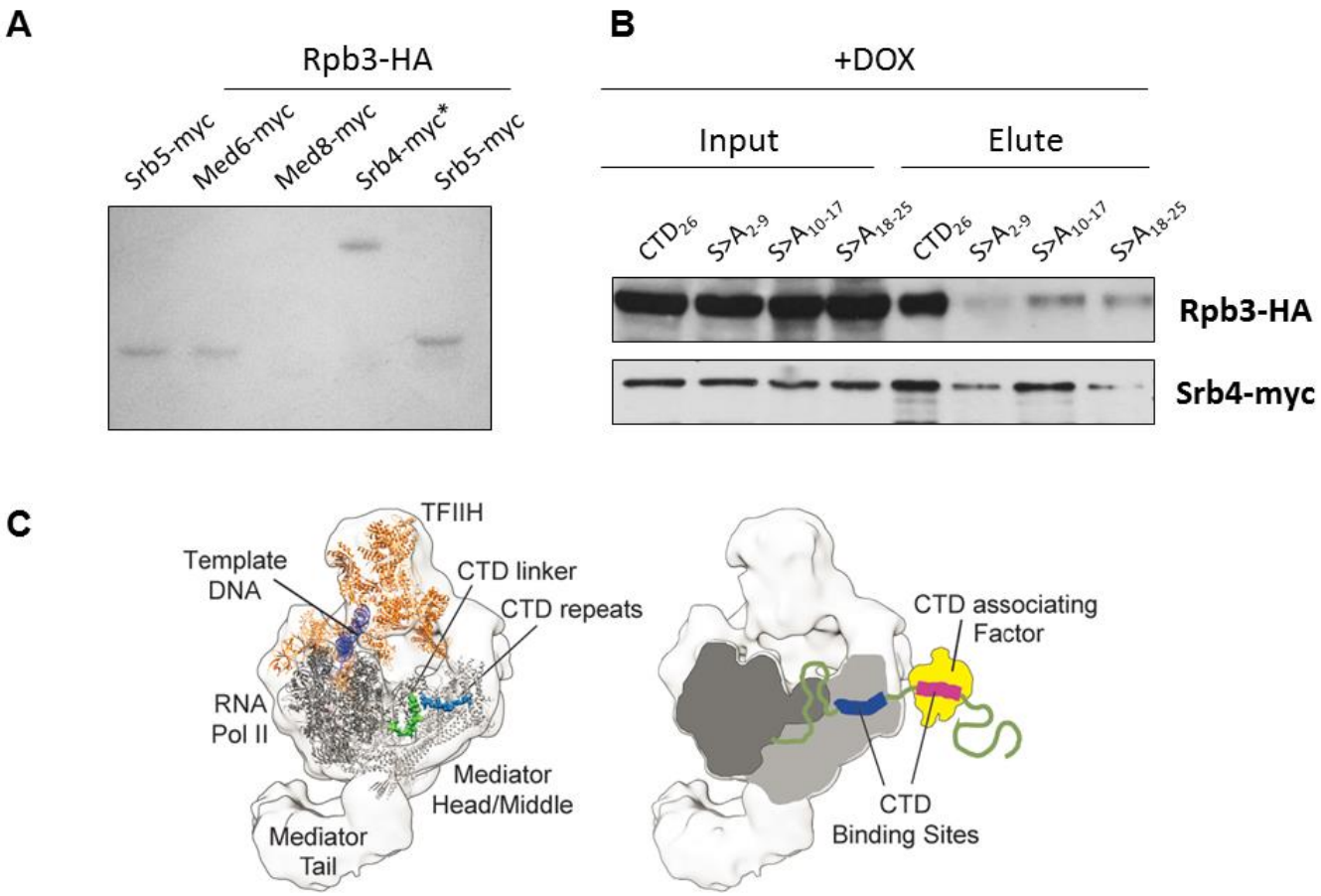


Figure 6. Srb4p preferentially interacts with the pCTD$_{26}$–S>A$_{2\text{-}9}$ region of the CTD. **A**) Anti-c-myc western blot of the indicated myc-tagged Mediator strains. The Srb4-myc/Rpb3-HA double tagged strain, indicated by the asterisk, was selected to perform immunoprecipitations. **B**) Immunoprecipitation of Mediator-CTD complexes grown in SC–LEU+DOX media. Srb4-myc was used as the bait and eluted RNA polymerase II complexes were measured by blotting for the tagged Rpb3-HA subunit. Eluted proteins are displayed together with 5% input controls. **C**) Structural model of Mediator-CTD binding adapted from figure 6C of (Babokhov et al. 2018).

Yeast grown in the absence of DOX showed equal levels of trimethylated H3, H3, Set2p itself and the loading control G6PDH (Figure 7A). In contrast, protein extracts from strains grown in the presence of DOX showed a modest but reproducible reduction of bulk H3K36me3 levels in the $pCTD_{26}$–$S>A_{10\text{-}17}$ mutant. Histone H3 levels themselves were unchanged, indicating that the decrease was due specifically to a loss of methylation at K36 possibly due to a loss of the Set2p enzyme itself (Figure 7). These results demonstrate that other proteins besides Mediator may be interacting with specific regions of the CTD to carry out their functions.

Discussion

The physical interaction of protein factors along the CTD is of great interest in the study of transcription, although the precise arrangement of these factors has remained a mystery. A significant goal for the field is moving away from individual CTD peptide-based studies of protein factor interactions and towards a more holistic view of binding on the CTD. In chapter 3 of this thesis I outlined a property of the CTD, region specific function, which could provide an explanatory framework for the arrangement of protein factors on the CTD (Babokhov et al. 2018). The region specific model posits that for at least a subset of protein factors there is a specific stretch of repeats that serves as a binding site for those factors. While broad phenotypes for several regions of the CTD repeats have been identified, the challenge of assigning particular protein factors to these regions remains. In this chapter, I have started to address this challenge by examining genetic and biochemical data of two factors, Mediator and Set2p, and assigning putative binding regions. The analysis of the data is provided below with

further discussion of the implications of these data for mechanisms of region specific binding on the CTD.

Srb2p is a subunit of the head module of the Mediator complex that has previously been associated with telomere maintenance (Peng and Zhou 2012). Out of the four subunits examined in this chapter Srb2p is the most sensitive to the phenotypes tested, completely losing viability at high temperature (Figure 2C) and no inositol (Figure 3C) while being very sick on osmotic stress (Figure 4C). Under standard conditions with DOX, the *srb2Δ* double mutant shows no additive effects with the CTD region mutants (Figure 1C) suggesting that the subunit in general is required for CTD-Mediator function. Srb2p itself sits on the outside edge of the head module of Mediator, facing away from the CTD binding site (Figure 8A) (Tsai et al. 2014) so it is unlikely to directly interact with the CTD repeat regions. Instead, Srb2p may play a role in stabilizing the head module to enable proper interaction with the CTD in general, regardless of region. Mediator complexes lacking the stabilizing Srb2p subunit would therefore be highly sensitive to protein folding stresses brought about by high temperature (Figure 2C) and would have difficulty responding to transcriptional programs associated with inositol production (Figure 3C) and stress response (Figure 4C). This model of Srb2p function is supported by findings that it is required for the assembly of parts of the head module (Shaikhibrahim et al. 2009), including Med8p which is a known CTD interactor (Robinson et al. 2012). Although the Srb2p data do not contain any indications of region specific activity, they still demonstrate an important function of Srb2p in stabilizing the CTD-Mediator interaction.

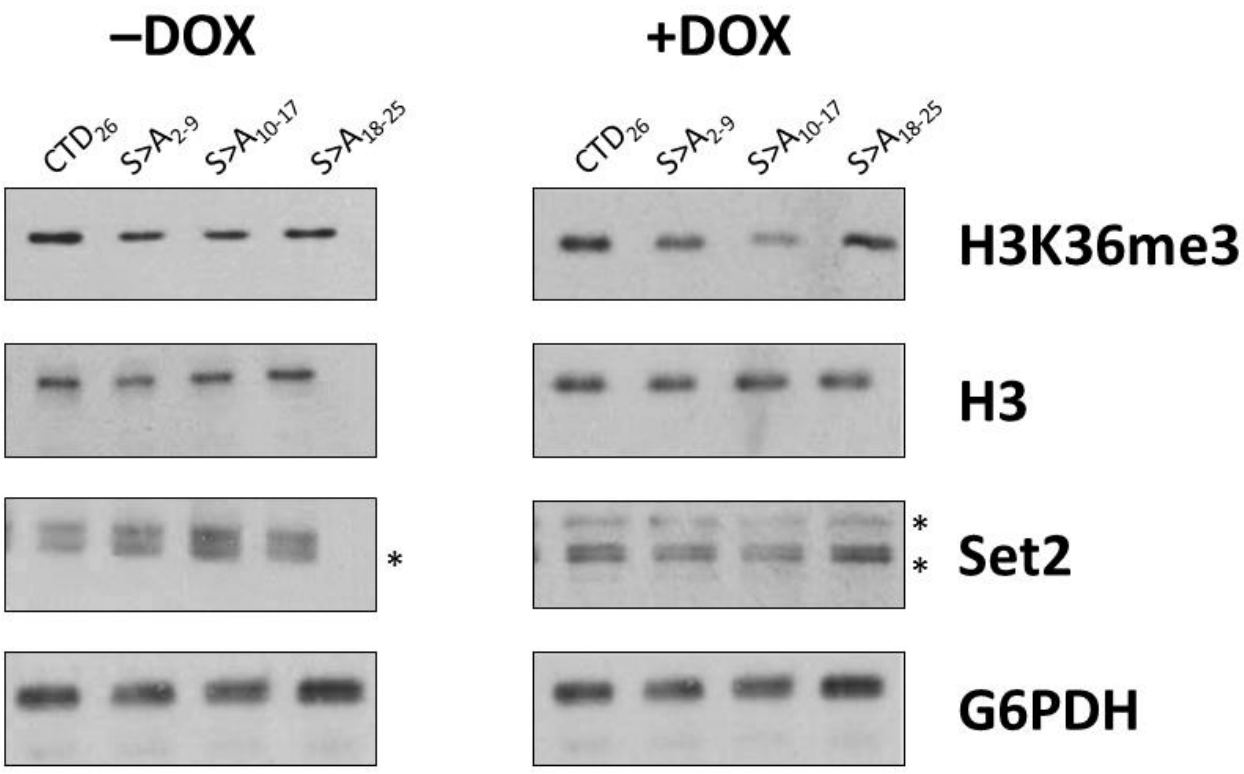


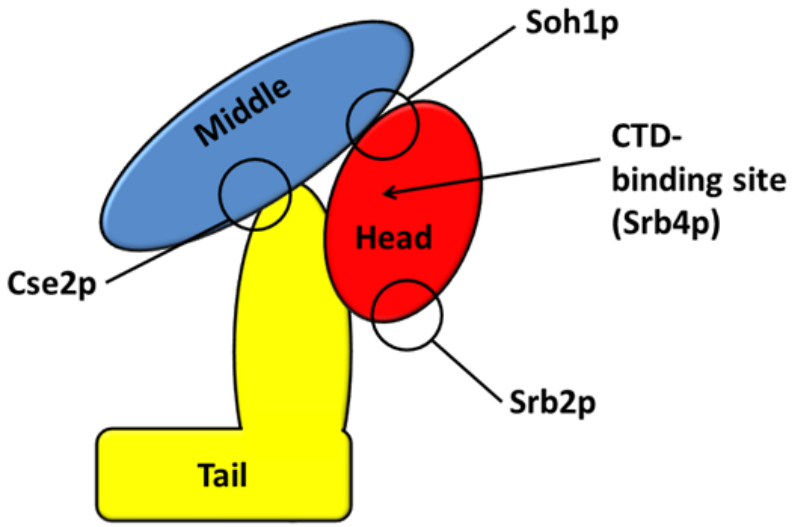
Figure 7. Repeats 10-17 are important for histone H3 methylation. Western blotting of yeast cell extract shows a decrease in histone H3 trimethylation at lysine 36 (H3K36me3) even though overall H3 levels are unchanged (third column) under the TET-off system (Doxycycline). Set2 methyltransferase that places the H3K36me3 mark appears unchanged (asterisk represents non-specific bands). G6PDH was blotted as a loading control. No changes in protein levels are seen when the wild type copy of Rpb1 is expressed (No Doxycycline).

One of the most interesting indications of region specific interactions comes from the Cse2p subunit of Mediator. Cse2p is a component of the middle module of Mediator and is located at the contact point between the middle and tail modules (Figure 8A). The *cse2Δ* double mutant demonstrated a strong additive genetic interaction with the $pCTD_{26}$–S>$A_{2\text{-}9}$ region under stress (Figure 2D, 4D), indicating that *CSE2* is operating in a different pathway than the first eight repeats of the CTD. In contrast, the *cse2Δ* double mutant rescues the growth defect of the $pCTD_{26}$–S>$A_{10\text{-}17}$ region under all conditions except inositol auxotrophy (Figure 1D, 2D, 4D), while the $pCTD_{26}$–S>$A_{18\text{-}25}$ region is unchanged. This suppression of the $pCTD_{26}$–S>$A_{10\text{-}17}$ growth defect is indicative of a typically repressive role of Cse2p at this region of the CTD. Cse2p has been previously associated with a repressive role in transcription (Han et al. 2001), perhaps by transferring transcription factor-induced conformational changes from the tail to the middle module. The genetic interaction data argue that Cse2p's repressive function is mediated specifically through the middle eight repeats of the CTD (Figure 8B). The Mediator complex is known to flexible in both its subunit composition and function (Anandhakumar et al. 2016), and complexes recruited for a repressive function could specifically utilize the middle eight repeats of the CTD while activating Mediator could use a different region. This model could be further tested by studying the interactions of other known repressive subunits, such as those found in the CKM subcomplex of Mediator (Tsai et al. 2013). The CKM subcomplex is particularly interesting, as mutations to its subunits rescue CTD truncations (Liao et al. 1995), similar to the rescue of the $pCTD_{26}$–S>$A_{10\text{-}17}$ region by the *cse2Δ* mutant.

Soh1p is a subunit of the middle module and is one of the most highly-conserved Mediator subunits among eukaryotes (Fan and Klein 1994). Genetic analysis of the *soh1Δ* double mutant reveals a phenotype-dependent interaction with the CTD. At 37° C, the *soh1Δ* double mutant has a strong additive effect with the CTD region specific mutants indicating that Soh1p acts in a separate pathway to the CTD to maintain transcription at high temperature (Figure 2E). Soh1p is located at the boundary of the head and middle modules (Figure 8A) and could play a stabilizing role similar to Srb2p during transcription. In contrast, the *soh1Δ* double mutant shows no additive effects under 1 M NaCl (Figure 4E), suggesting that it is working in the same pathway equally with all three regions of the CTD. Soh1p physically contacts the CTD-binding Med6p subunit (Tsai et al. 2014), and may exert its genetic influence through this physical interaction. Alternatively, cryo-EM studies have hinted that the CTD can bind along the middle module (Tsai et al. 2013) and Soh1p would be a candidate for this interaction during stress conditions such as high salt. The lack of any region specific interactions may mean that Soh1p can bind to any set of CTD repeats, perhaps acting to anchor the CTD under stress conditions and allow for the proper alignment of other binding factors along it.

Aside from the Mediator subunit deletion mutants, the C-terminal c-myc tag on Srb4p proved to be a useful tool in studying the genetic interactions of this subunit with the CTD. While useful for immunoprecipitation, the c-myc tag introduced a slight growth defect that may be caused by interfering with the structure of the head module of Mediator. However, this growth defect under stress allowed me to study the genetic

A



B

Repeat	*SRB4*	*SRB2*	*CSE2*	*SOH1*
CTD_{2-9}	+	+	+	+/-
CTD_{10-18}	-	+	Sup.	+/-
CTD_{19-25}	+	+	+	+/-

Figure 8. Summary of Mediator subunit genetic analysis. **A**) Cartoon of the three main modules of Mediator with the positions of the subunits investigated in this chapter labeled. Panel was based off of structural data from (Tsai et al. 2014). **B**) Summary of results of the genetic interaction studies between Mediator subunit deletion mutants and the CTD region mutants. A plus sign indicates an additive genetic interaction while a minus sign indicates no additive interactions. Sup. indicates a suppression effect of in the *cse2Δ* double mutant. Scoring is based on the overall trend of the double mutants, and certain genetic interactions may differ depending on the phenotype. The *soh1Δ* was evenly split among the various phenotypes between additive and non-additive interactions and is scored as +/-.

interactions of this subunit in the absence of a deletion mutant. The overall pattern that emerged from the spotting data in Figure 5 is that Srb4p-myc has an additive effect with the $pCTD_{26}$–$S>A_{2\text{-}9}$ region and no effect with the $pCTD_{26}$–$S>A_{10\text{-}17}$ region, arguing for a specific interaction with the middle eight CTD repeats (Figure 8B). While an interaction with the $pCTD_{26}$–$S>A_{18\text{-}25}$ region cannot be ruled out, the lack of any discernable phenotypes of this region in larger phenotypic screens (Babokhov et al. 2018) suggests that the last eight repeats have a redundant function at best. This interaction with the middle region of the CTD is in line with the *cse2Δ* double mutant and could represent a significant region specific binding activity of Mediator to the CTD.

In contrast to the genetic interaction data of *CSE2* and *SRB4*, the immunoprecipitation experiments show that Mediator most strongly prefers the first eight repeats of the CTD as a binding region (Figure 6B). These findings can be reconciled by considering Mediator binding as sampling different regions of the CTD depending on growth and stress conditions. The immunoprecipitation experiments were conducted under normal growth conditions, and the preference for the first eight repeats could represent the default binding arrangement of Mediator on the CTD. The additive genetic interaction between *SRB4* and the first eight repeats (Figure 5A) would in turn represent the contribution of binding to the other two regions of the CTD. Alternatively, the c-myc tag may be interfering with other Srb4p functions unrelated to CTD binding such as assembly of the subunit into the head module of Mediator (Robinson et al. 2012). The other regions of the CTD might also host Mediator, or different compositions of Mediator subunits, specifically under stress conditions. This model would explain the specific

genetic interactions of *CSE2* with the middle eight repeats of the CTD under stress (Figure 4D). Further immunoprecipitations of Mediator under osmotic stress would be expected show a shift of preferential binding from the first eight to the middle eight CTD repeats. Finding a tagged Mediator subunit that does not lose viability under osmotic stress will be especially important to obtain the results of such an experiment. Overall, the initial findings of Mediator binding to the CTD are in line with the prediction from previous genetic and structural work (Figure 6C).

The Mediator complex is not the only known CTD-binding protein and a true grasp of region specific binding on the CTD will require a screen of other CTD interactors. As a first step towards this process, I examined the function of the histone methyltransferase Set2p. H3K36me3 levels were specifically reduced in the $pCTD_{26}$–S>$A_{10\text{-}17}$ mutant in the presence of DOX (Figure 7B). Set2p binding to the CTD is required for its activity likely through the stabilization of the protein following binding (Fuchs et al. 2012). It was unclear from the blots of the anti-Set2 antibody if Set2p levels were decreased as expected from a loss of binding (Figure 7B). The presence of residual H3K36me3 signal indicates that some Set2p is still present, perhaps by binding semi-redundantly to other CTD repeats. Alternatively, Set2p region specific binding may be gene dependent, and the bulk reduced levels are only representative of a fraction of region specific binding. Tagging Set2p directly with a c-myc tag would allow more accurate determination of protein levels as well as characterizing the region specific binding of Set2p directly. Additionally, locus specific or genome-wide chromatin immunoprecipitation experiments could further refine the effect of CTD mutations on histone methylation levels. Further

exploration of the Set2p pathway by these methods could also be used to determine the effects of the region specific mutants on downstream histone acetylation levels (Lickwar et al. 2009). These Set2p data are therefore a useful launching point to analyze further region specific interactions along the CTD.

Although the data presented in this chapter provide some examples of region specific interactions, the larger question of what causes region specificity still remains to be solved. The middle eight repeats of the CTD yielded two potential region specific interactors: Mediator and Set2p. Mediator is bound during the initiation phase while Set2p is active during the elongation phase and both factors recognize differently modified CTD repeats. Enzymatic writers of CTD PTMs may therefore play a significant role in establishing region specific interactions throughout the transcription cycle, perhaps by altering an intrinsic property of the CTD such as its secondary structure. Although the CTD is disordered, residual structure has been detected in the repeats and the overall packing of the CTD is known to change with modification (Portz et al. 2017). Serine phosphorylation is distributed evenly throughout the CTD repeats (Suh et al. 2016, Schuller et al. 2016), so it is unlikely that the modifications themselves are establishing region specific information. CTD packaging or positioning relative to the body of RNA polymerase II or the Mediator complex could instead determine the initial positioning of protein factors. In this model, the first wave of factors bound to the CTD would then determine the arrangement of factors that come in later during the transcription cycle. Alternatively, region specific binding could be determined by protein factor binding affinity to the CTD, with high-priority factors receiving additional stabilizing

interactions with the RNA polymerase II holoenzyme. This mechanism may be involved in Srb4p binding to the CTD, where all three CTD region mutants reduce Srb4p binding, but the first eight repeat region shows the strongest defect (Figure 6B). Additional biochemical analysis of entire transcription complexes using a variety of region specific mutants will be essential to piece together the mechanism that determines repeat and region specific binding along the CTD.

In conclusion, this chapter provided initial data on two protein factors that demonstrate region specific binding to the CTD. Using genetic analysis of Mediator subunit deletion and region specific double mutants the Cse2p and Srb4p subunits were found to interact specifically with the middle region of the CTD. This finding indicates that this region is specifically required for at least a subset of Mediator interactions and could explain the general requirement of this region for viability in yeast. Immunoprecipitation of Mediator-CTD complexes complemented these genetic studies and revealed a specific requirement for the first eight CTD repeats in binding Mediator. Lastly, analysis of histone methylation levels in region specific mutant strains uncovered a specific role of the middle eight repeats in promoting H3K36 trimethylation. Combined, these studies reveal the importance of two broad regions of the CTD and lay the groundwork to pursue deeper mechanistic understanding of the discrete essential regions of the CTD.

Appendix 4.A

Table 4.A-1. List of strains used in chapter 4.

Name	Mating Type	Genotype
YMB001	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **rpb3-6HA**-natNT2
YMB002	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **med6-9myc**-HIS3MX6
YMB003	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1**, srb5-9myc**-HIS3MX6
YMB005	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **rpb3-6HA**-natNT2, **med6-9myc**-hphNT1
YMB007	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **rpb3-6HA**-natNT2, **med8-9myc**-hphNT1
YMB009	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **rpb3-6HA**-natNT2, **srb4-9myc**-hphNT1
YMB011	a	his3Δ, leu2Δ, lys2Δ, met15Δ, trp1Δ::hisG, URA::CMV-tTA, kanRPtetO7-TATA-RPB1, **rpb3-6HA**-natNT2, **srb5-9myc**-hphNT1
YMB012	?	his3?, leu2Δ, lys2?, met15?, trp1Δ::hisG?, URA::CMV-tTA, HPHNT2-RPtetO7-TATA-RPB1, **cse2Δ**::KanMX
YMB013	?	his3?, leu2Δ, lys2?, met15?, trp1Δ::hisG?, URA::CMV-tTA, HPHNT2-RPtetO7-TATA-RPB1, **srb2Δ**::KanMX
YMB014	?	his3?, leu2Δ, lys2?, met15?, trp1Δ::hisG?, URA::CMV-tTA, HPHNT2-RPtetO7-TATA-RPB1, **soh1Δ**::KanMX

Appendix 4.B

A detailed protocol of the immunoprecipitation of Srb4-myc/Rpb3-HA complexes is outlined below.

Original protocol: Wittmeyer J, Saha A, Cairns B. DNA translocation and nucleosome remodeling assays by the RSC chromatin remodeling complex. Methods Enzymol. 2004;377:322-43.

1) Grow up 500 mL cultures of YMB009 transformed with the four region specific CTD mutant plasmids to an OD of ~0.8 in SC–LEU+DOX media.
    - Typical growth times are six to seven hours.
2) Wash cells twice with PBS and resuspend the pellet in 500 µL of water in a 1.7 mL tube.
    - First add 400 µL to the pellet and resuspend. Then add water up to 500 µL to ensure equal amounts of cells in the aliquots.
3) Aliquot 100 µL to separate tubes, spin down the cells and decant the supernatant.
4) Freeze cell pellets at -80° C.
5) Use one cell pellet per IP experiment.
    - Typically one pellet for the experiment and one pellet for the no antibody control
    - One pellet corresponds to 100 mL of culture. This is in extreme excess for 50 µL of magnetic beads. Future experiments should determine the optimal number of cells to incubate with the beads.
6) Resuspend the pellet in 500 µL of lysis buffer with protease inhibitors in a 1.7 mL tube.
    - Buffer: 50 mM HEPES pH 7.5, 250 mM KOAc, 10 % glycerol, 10 mM EDTA, 0.5 mM DTT (store at 4° C)
    - 100X inhibitor cocktail: 0.03 mg/ml leupeptin, 0.14 mg/ml pepstatin, 0.02 mg/ml chymostatin, 8.5 mg/ml phenylmethanesulfonyl fluoride, 33 mg/ml benzamidine solubilized in ethanol (store at -20° C)
7) Add 0.5 mm diameter glass beads to an equal volume of the cell suspension.
    - After resuspending pellets in lysis buffer, indicate the liquid level with a marker. Then add beads up to the level of the mark.

8) Vortex the cell/bead suspension in the cold room for 1 min, then sit to cool on ice for 1 min. Repeat vortex/cool cycle 5 times for a total of 10 mins.
9) Flip the tubes upside down and let the suspension settle to the cap side of the tube. Poke a hole in the bottom of the tube using a red-hot needle. Transfer the poked tube to a second 1.7 mL tube and spin down to collect the supernatant.
    - Use the short spin button on the micro-centrifuge for about 5 sec (until the speed reaches 3000 rpm). Any faster or longer will dislodge the poked tube and spill beads everywhere.
    - Check that the poked tube only has dried-out beads remaining. If any liquid persists then spin again for 5 sec.
10) Clarify the lysate by spinning at 15,000 rpm for 10 min in the cold room.
11) Transfer the lysate (500 µL) to a fresh 1.7 mL tube.
12) Adjust the salt concentration to 400 mM KOAc by adding 42 µL of 3M KOAc stock solution.
13) Precipitate nucleic acids by adding 5.42 µL of 10 % polyethylenime to get a final concentration of 1 % PEI.
    - Add the volume slowly and vortex well until the solution turns milky white.
14) Clarify the lysate by spinning at 15,000 rpm for 45 min in the cold room.
15) Transfer the supernatant (500 µL) to a fresh 1.7 mL tube.
16) Take a 50 µL sample to use as an input control for the final western blot.
17) Add 10 µL of anti-c-myc 9E10 antibody to the lysate and incubate rotating for 1 hr in the cold room.
    - Add 10 µL of lysis buffer for the no antibody controls.
18) Take 50 µL of magnetic protein A/G beads per reaction and wash 3 times with 1 mL of lysis buffer.
19) Resuspend the beads in 50 µL of lysis buffer and add to the lysate-antibody mix. Incubate rotating for 2 hr in the cold room.
20) Decant the supernatant and keep it as the unbound fraction.
21) Wash the beads 3 times with 1 mL of lysis buffer. Keep the first wash as the wash fraction.
22) Resuspend the washed beads in 100 µL of lysis buffer and 20 µL of 5X SDS-PAGE loading buffer.
23) Boil beads for 10 min at 95° C in the heat block to elute bound proteins.
24) Load 10 µL of eluted sample on a 10 % SDS-PAGE gel and analyze by western blot.
    - The high KOAc salt concentration will precipitate out some of the SDS at room temperature, so load samples while they are still hot from the heat block.

Chapter 4 Literature Cited

Anandhakumar J, Moustafa YW, Chowdhary S, Kainth AS, Gross DS2. Evidence for Multiple Mediator Complexes in Yeast Independently Recruited by Activated Heat Shock Factor. Mol Cell Biol. 2016 Jun 29;36(14):1943-60.

Babokhov M, Mosaheb MM, Baker RW, Fuchs SM. Repeat-Specific Functions for the C-Terminal Domain of RNA Polymerase II in Budding Yeast. G3 (Bethesda). 2018 May 4;8(5):1593-1601.

Baidoobonso SM, Guidi BW, Myers LC. Med19(Rox3) regulates Intermodule interactions in the Saccharomyces cerevisiae mediator complex. J Biol Chem. 2007 Feb 23;282(8):5551-9.

Cho EJ, Takagi T, Moore CR, Buratowski S. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev. 1997 Dec 15;11(24):3319-26.

Corden, J. L., 2013 RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem Rev 113: 8423-8455.

Dunn EF, Hammell CM, Hodge CA, Cole CN. Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. Genes Dev. 2005 Jan 1;19(1):90-103.

Eick D, Geyer M. The RNA polymerase II carboxy-terminal domain (CTD) code. Chem Rev. 2013 Nov 13;113(11):8456-90.

Fan HY, Klein HL. Characterization of mutations that suppress the temperature-sensitive growth of the hpr1 delta mutant of Saccharomyces cerevisiae. Genetics. 1994 Aug;137(4):945-56.

Fuchs SM, Kizer KO, Braberg H, Krogan NJ, Strahl BD. RNA polymerase II carboxyl-terminal domain phosphorylation regulates protein stability of the Set2 methyltransferase and histone H3 di- and trimethylation at lysine 36. J Biol Chem. 2012 Jan 27;287(5):3249-56.

Han SJ, Lee JS, Kang JS, Kim YJ. Med9/Cse2 and Gal11 modules are required for transcriptional repression of distinct group of genes. J Biol Chem. 2001 Oct 5;276(40):37020-6.

Harlen KM, Churchman LS. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. Nat Rev Mol Cell Biol. 2017 Apr;18(4):263-273.

Janke C, Magiera MM, Rathfelder N, Taxis C, Reber S, Maekawa H, Moreno-Borchart A, Doenges G, Schwob E, Schiebel E, Knop M. A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast. 2004 Aug;21(11):947-62.

Liao SM, Zhang J, Jeffery DA, Koleske AJ, Thompson CM, Chao DM, Viljoen M, van Vuuren HJ, Young RA. A kinase-cyclin pair in the RNA polymerase II holoenzyme. Nature. 1995 Mar 9;374(6518):193-6.

Lickwar CR1, Rao B, Shabalin AA, Nobel AB, Strahl BD, Lieb JD. The Set2/Rpd3S pathway suppresses cryptic transcription without regard to gene length or transcription frequency. PLoS One. 2009;4(3):e4886.

Malagon F, Kireeva ML, Shafer BK, Lubkowska L, Kashlev M, Strathern JN. Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. Genetics. 2006 Apr;172(4):2201-9.

Mayekar MK, Gardner RG, Arndt KM. The recruitment of the Saccharomyces cerevisiae Paf1 complex to active genes requires a domain of Rtf1 that directly interacts with the Spt4-Spt5 complex. Mol Cell Biol. 2013 Aug;33(16):3259-73.

Peng J, Zhou JQ. The tail-module of yeast Mediator complex is required for telomere heterochromatin maintenance. Nucleic Acids Res. 2012 Jan;40(2):581-93.

Portz B, Lu F, Gibbs EB, Mayfield JE, Rachel Mehaffey M, Zhang YJ, Brodbelt JS, Showalter SA, Gilmour DS. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nat Commun. 2017 May 12;8:15231.

Poss ZC, Ebmeier CC, Taatjes DJ. The Mediator complex and transcription regulation. Crit Rev Biochem Mol Biol. 2013 Nov-Dec;48(6):575-608.

Robinson PJ, Bushnell DA, Trnka MJ, Burlingame AL, Kornberg RD. Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II. Proc Natl Acad Sci U S A. 2012 Oct 30;109(44):17931-5.

Schüller R, Forné I, Straub T, Schreieck A, Texier Y, Shah N, Decker TM, Cramer P, Imhof A, Eick D. Heptad-Specific Phosphorylation of RNA Polymerase II CTD. Mol Cell. 2016 Jan 21;61(2):305-14.

Shaikhibrahim Z, Rahaman H, Wittung-Stafshede P, Björklund S. Med8, Med18, and Med20 subunits of the Mediator head domain are interdependent upon each other for folding and complex formation. Proc Natl Acad Sci U S A. 2009 Dec 8;106(49):20728-33.

Suh H, Ficarro SB, Kang UB, Chun Y, Marto JA, Buratowski S. Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Mol Cell. 2016 Jan 21;61(2):297-304.

Tsai KL, Sato S, Tomomori-Sato C, Conaway RC, Conaway JW, Asturias FJ. A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. Nat Struct Mol Biol. 2013 May;20(5):611-9.

Tsai KL, Tomomori-Sato C, Sato S, Conaway RC, Conaway JW, Asturias FJ. Subunit architecture and functional modular rearrangements of the transcriptional mediator complex. Cell. 2014 Jun 5;157(6):1430-44.

West ML, Corden JL. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics. 1995 Aug;140(4):1223-33.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Véronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science. 1999 Aug 6;285(5429):901-6.

Wittmeyer J, Saha A, Cairns B. DNA translocation and nucleosome remodeling assays by the RSC chromatin remodeling complex. Methods Enzymol. 2004;377:322-43.

Yang C, Stiller JW. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A. 2014 Apr 22;111(16):5920-5.

Young BP, Shin JJ, Orij R, Chao JT, Li SC, Guan XL, Khong A, Jan E, Wenk MR, Prinz WA, Smits GJ, Loewen CJ. Phosphatidic acid is a pH biosensor that links membrane biogenesis to metabolism. Science. 2010 Aug 27;329(5995):1085-8.

Chapter 5

Future perspectives of tandem repeat structure and function

Abstract

Recent research progress in the unique properties of tandem repeats and disordered regions has pushed the boundaries of our understanding of protein function. Proteins are now known to function by mechanisms other than the conventional folded protein domain. However, presently our knowledge of disorder and tandem repeat function is restricted to a limited number of well-characterized proteins. Further progress in the field is therefore dependent on a comprehensive approach to identify overarching biological principles behind disordered and repetitive protein regions. In this final chapter I will present several promising future perspectives in the field, focusing specifically on tandem repeats. I frame these perspectives into three broad questions: how repeats emerge, how repeats change and how repeats specialize. Along the way I will refer to previous studies in addition to my own thesis findings to predict how addressing these three questions will further deepen our appreciation of the role tandem repeats play in influencing protein structure and function. I will also propose a number of approaches that would contribute to the resolution of the three broad questions. A comprehensive approach that identifies general principles in the study of disorder and repeats has an amazing potential to transform our understanding of how evolutionary and biochemical forces shape the biological function of proteins in the cell.

Contribution of thesis research to the study of disorder and repeats

One of the stunning aspects of the study of disorder and repeats is the explosive progress in the past two decades, largely driven by the parallel expansion of computational power. In chapter 1 of this thesis, I introduced the growth in knowledge of disordered regions and repeats from dismissal to appreciation of their biological function. My research as outlined in chapters 2, 3 and 4 then sought to further probe the unique properties of disordered regions and repeats using budding yeast as a model organism. Chapter 2 presented the striking overlap of disordered regions and repeats in the variable loci of 93 strains of budding yeast and provided evidence for their conserved function. Chapter 3 focused on one of these disordered and repetitive regions, the CTD of RNA polymerase II, and found that the identical repeats of this region had specific functions that were governed by their position in the repetitive array. In chapter 4, I expanded upon the genetic studies of chapter 3 and provided evidence that the region specific functions of the CTD were due to differential protein factor activity across the repeats. Taken together, this work revealed additional interactions between disorder and repeats and provided a new region specific mechanism for repeat function. My thesis work focused on broad approaches, examining the whole disordered proteome of budding yeast and the repeat functions of the entire CTD. Continuing to pursue such a big-picture approach will be vital for further advances in the field.

Future perspectives of tandem repeat function

Previous studies of disorder and tandem repeats have tended to focus on the properties of one or a set of proteins. While this work helped establish disorder and repeats as

significant modulators of protein function, databases of disordered and repetitive regions have grown to the point where more comprehensive approaches are possible. In this chapter, I will introduce three broad questions that I believe will shape the study of repetitive sequences in the coming years. While I have chosen to focus on repetitive protein sequences, many of these questions and the methodologies used to address them will also apply to disordered regions given the significant overlap between the two that was reported in chapter 2 of this thesis. The three questions are: 1) **How do repeats emerge?** 2) **How do repeats change?** 3) **How do repeats specialize?** Each of these questions are drawn both from observations in the field as well as data from my thesis research. Approaches that work towards answering these questions will enable the development of universal principles of tandem repeat structure and function that can be used to predict repeat behavior in whole proteomes.

How do repeats emerge?

In order to properly understand how repeats function, it is important to determine how repetitive sequences emerge and become fixed in the population. Much work has been done to establish how present repeats expand, but how tandem repeats emerge from a non-repetitive sequence context has yet to be systematically addressed. While the overall repetitive DNA content of a genome varies wildly between species, tandem repeats in the coding frames of proteins are more prevalent in eukaryotes than prokaryotes (Dunker et al. 2000). This observation suggests that tandem repeats were selected for in parallel with, or as a response to, increasing cellular complexity. What

therefore are the mechanisms both at the coding sequence level and at the protein context level that give rise to tandem repeats?

In chapter 2, I described the different types of tandem repeats that were copy number variable in the genomes of 93 strains of budding yeast. One interesting class of variable repeats was minisatellites that appeared to be a segmental duplication to make a novel repeat of two copies (Appendix 2.C). These new repeats typically emerge in only one or two of the 93 strains of yeast, suggesting that they are novel events and not a degradation of a previous minisatellite repeat. Further examination of these sequences could give us hints as to the mechanism behind the emergence of new repeats. Segmental duplications are known to arise from replication slippage events (Fan and Chu 2007) and studying this set of repeats may reveal commonalities that suggest the sequence elements that give rise to new repeats. For example, the repeat sequences could be examined for possible secondary structure formation or GC skew that may have caused their duplication. Any mechanisms discovered could be tested to arrive at a common pathway for the emergence of novel minisatellites that could then be selected for or against. Such an understanding would allow the field to grasp how the raw material for repeat evolution could be created from non-repetitive sequence.

Determining how repeats are created is only the first step. Another important aspect of how repeats emerge is whether or not the newly created repeats are fixed in the population. Not all of the repeats that could potentially be created by the mechanisms describe above are necessarily advantageous to the organism. Although the cost and

benefit of a particular tandem repeat sequence will ultimately depend of the protein in question, what are general forces that shape which new repeats are selected for and which are selected against? One hint comes from the observation that tandem repeats, and indeed disordered regions, are frequently found in proteins that serve as nodes in protein-protein interaction networks (Chavali et al. 2017). The emergence of new repeats can be studied and the effects quantified by comparing repeat emergence with the growth of protein-protein interaction networks. Under this model, new repeats are expected to be selected for in existing nodes and lead to additional connections in the network. New repeats would have less of an effect in proteins at the periphery of the network and these repeats would likely degenerate via genetic drift (Figure 1). Significantly, this approach is independent of the particular repeat sequence of cellular pathway being studied, making it generally applicable. The interest is not in the specific repeat function, but instead in overall effect of repeat emergence in a protein-protein interaction network. Going forward, additional genome and interactome data across a range of strains and species will be especially important to enable the comparative approaches described above.

The CTD of RNA polymerase II is an especially illustrative example of how a tandem repeat emerges and changes a protein-protein network. The CTD is a unique adaptation of RNA polymerase II that is not present in the prokaryotic RNA polymerase or in any other of the eukaryotic polymerases. The CTD was predicted to emerge from two independent motifs that combined to form the YSPTSPS heptad in the ancestral eukaryote. RNA polymerase was already an important node in the network that

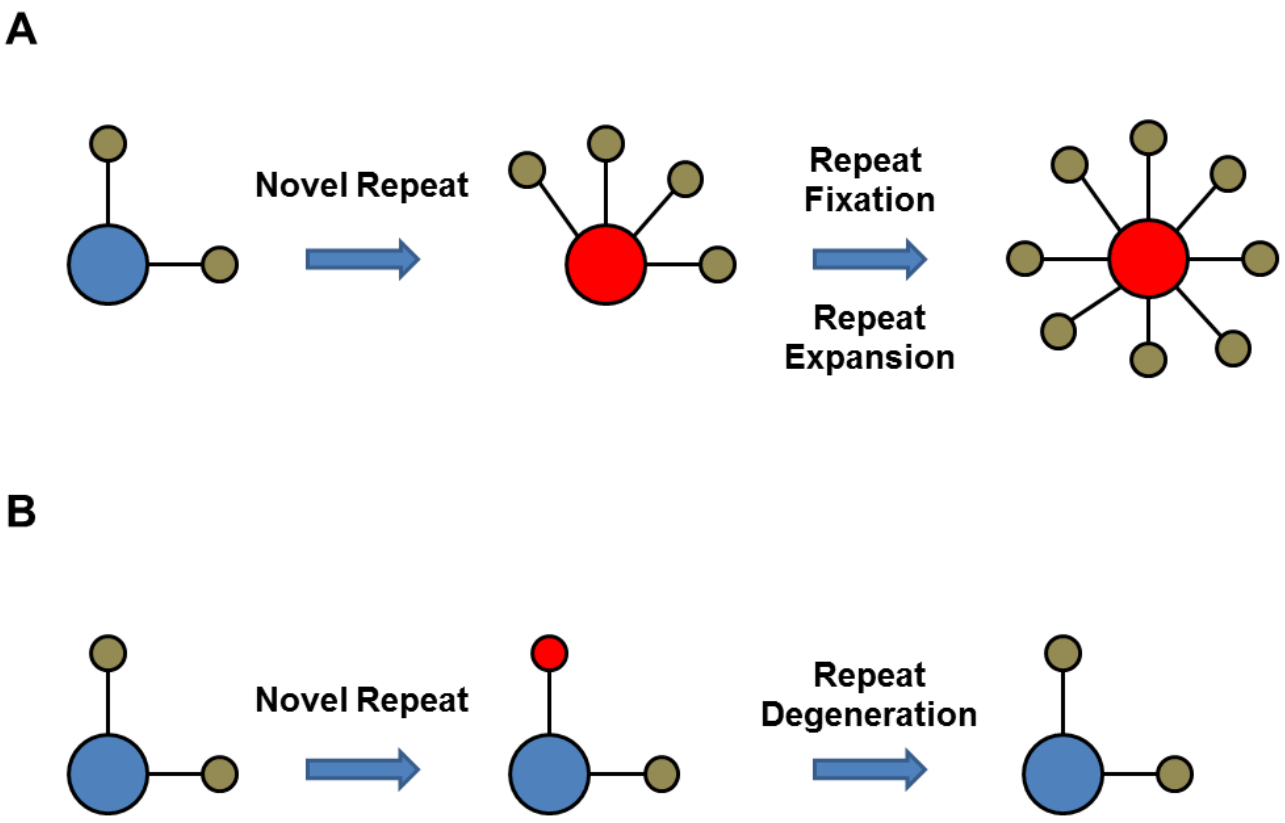


Figure 1. Model for tracking repeat formation through protein network growth. **A**) Novel repeats that form in a node protein (blue to red transition) could lead to increased interactions with other proteins in a network. Fixation and expansion of the repetitive sequence could further expand the connections to the node. **B**) Repeats that form in terminal proteins of a network would not increase the number of connections compared to a node protein. The repeat would eventually be selected against or lose its repetitive nature due to genetic drift.

coordinated transcription and the CTD greatly expanded this network by recruiting co-transcriptional proteins (Yang and Stiller 2014). The original function that was selected for in eukaryotes is currently hypothesized to be co-transcriptional splicing, as the ancestral eukaryote had to cope with the influx of mobile genetic elements from the incorporated prokaryote that would later become the mitochondria (Irimia and Roy 2014). Eventually, additional functions such as elongation factors took advantage of the repetitive sequence and lead to the complex network of interactions present on the CTD. The example of the CTD thus illustrates how repeats can emerge in the context of protein interaction networks and serves as a conceptual basis to understand the creation of novel repeats in other systems.

How do repeats change?

Tandem repeats are notable for their high rate of instability that leads to expansion or contractions of the repeat number (Brinkmann et al. 1998). The resulting variation can then lead to physiologically relevant changes to protein function (Verstrepen et al. 2005, Gemayel et al. 2017), supporting the need for a greater understanding of how repeats can change. Currently, there are a few described examples of how particular repeats expand, but the general principles governing repeat instability for all repeats or classes of repeats are still unknown. The most well understood example is repeat instability of polyQ repeats coded by the CAG trinucleotide, especially in instances of neurodegenerative diseases. This mechanism of instability occurs through DNA replication and repair coupled processes and explains well how CAG microsatellites expand and contract (La Spada and Taylor 2010). However, looking through the data of

variable polyQ repeats from chapter 2 of this thesis, it becomes clear that not all polyQ repeats consist of pure CAG repeats (Appendix 2.B). Additionally, alternative instability mechanisms have also been proposed for polyA repeats (Albrecht and Mundlos 2005), indicating that the question of how microsatellites change is still open for future investigation. A number of examples of different codon usage in microsatellites are presented in (Figure 2). A broader understanding of how microsatellite repeats change will require experimentation on the pathways involved in instability for all of the common microsatellites and bioinformatics approaches to determine the role of codon usage on repeat instability. Poly-serine repeats in particular will be interesting to examine, as they have six codons to choose from that could significantly impact the propensity for expansions or contractions. Given that as many as half of all tandem repeats in a proteome are microsatellites, working out the mechanisms of microsatellite instability will go a long way to answering the question of how repeats change.

Repeat instability is even less well understood for larger minisatellite repeats. Minisatellite instability is thought to be mediated through homologous recombination events, where misalignments of the repeat sequence lead to expansion or contraction of the copy number (Richard and Paques 2000). However, the mechanism behind minisatellite instability is more complex than previously appreciated, as a report from our lab showed that homologous recombination is necessary for the expansion of CTD repeats but competes with mechanisms causing contractions (Morrill et al. 2016). More studies are therefore required to work out the mechanisms that lead to both expansions and contractions of minisatellites, as both can act on a sequence to change repeat

number (Appendix 3.D-1). One promising mechanism that can act broadly to affect repeat change is GC skew. An imbalance of guanine vs. cytosine on the coding and template strands may lead to secondary structures like G-quadruplexes on one strand that can induce instability in the repeat (Paeschke et al. 2013). These secondary structures have been detected in microsatellites and our lab has also found G-quadruplex-like structures in the repeats of the CTD (Morrill et al. 2016). Future approaches could examine the GC skew of known variable minisatellites to establish the DNA signatures that characterize repeat instability. Any broad patterns that are discovered in this way could then be used to predict the tandem repeats that are most prone to changes in their repeat number.

Examining how repeats change also prompts the reverse question: why are some repeats not variable? While chapter 2 focused on the variable repeats in budding yeast, I also found close to 1000 tandem repeats that did not show any length variation. If repeat variation has the potential to increase and tune the functions of biologically-relevant proteins, why do some repeats maintain a constant repeat number? One hypothesis is that these repeats are acting in a structural capacity that requires them to keep a specific length, for example between two important domains in a protein. The disordered nature of the repetitive sequence may also be particularly important in this case, to prevent structure formation that would alter the required spacing. An alternative hypothesis is that the repeat copy number is kept invariable to preserve a bulk biophysical property of the repetitive sequence such as charge or propensity to aggregate. Expansion or contraction of the repeating units may upset the functionally

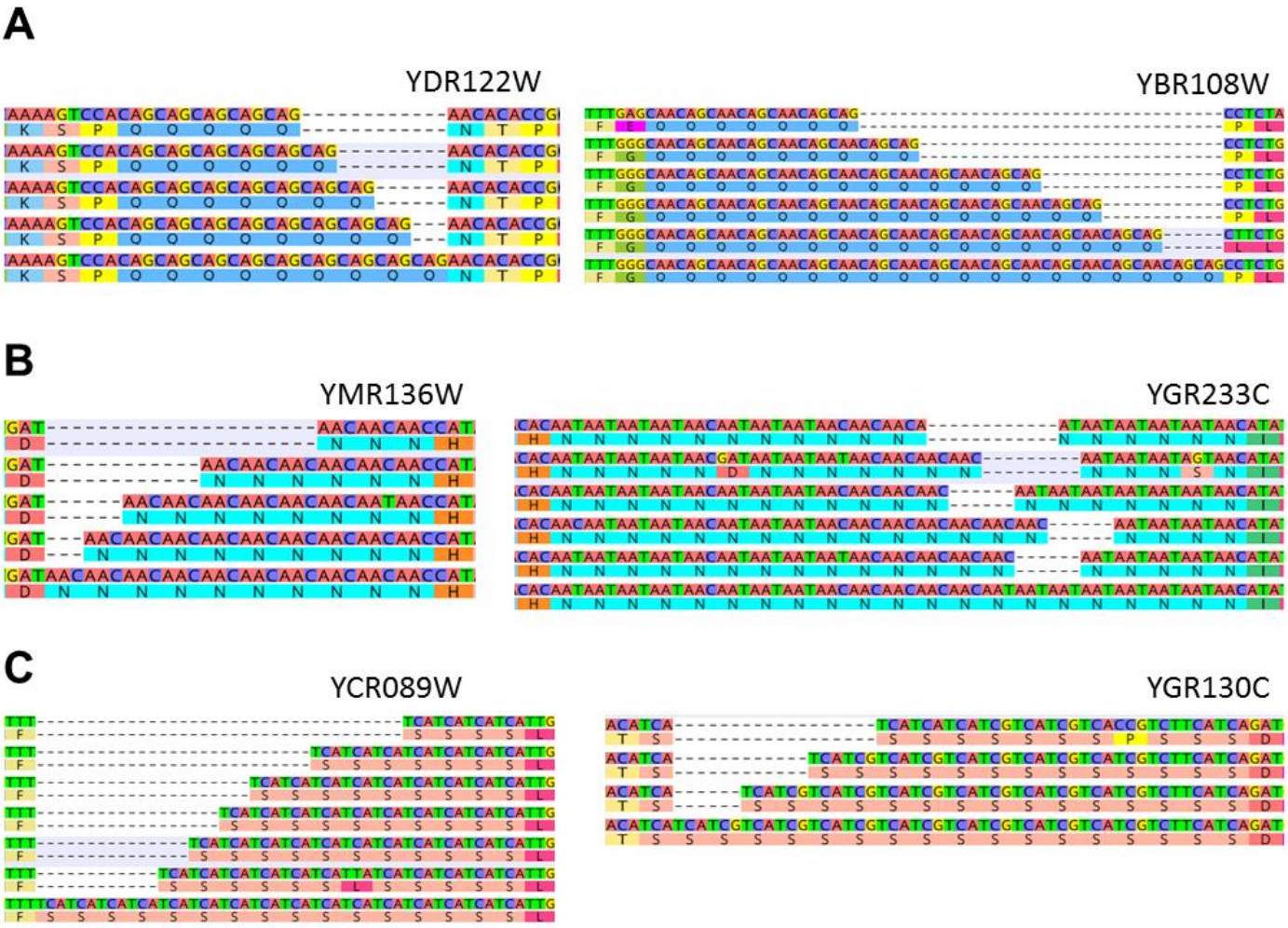


Figure 2. Examples of different codon usage among homopolymer repeats. **A**) Comparison of a pure CAG polyQ and a mixed CAG/CAA polyQ repeat. **B**) Comparison of a pure AAC polyN and a mixed AAT/AAC polyN repeat. The mixed repeat shows discreet partitioning of the AAT and AAC codons, with both sets demonstrating variability. **C**) Comparison of a pure TCA polyS with a mixed codon polyS.

relevant property and would be selected against in the population. These two hypotheses are not mutually exclusive, as length requirements could be important for mixed repeat sequences while bulk properties could act on pure or mixed microsatellites like polyE/D. Dissecting the mechanisms that keep some repeats constant would expand our knowledge of these important repeats while also providing an interesting contrast to variable repeats to further our understanding of the process that underline how repeats change.

How do repeats specialize?

While some tandem repeats are characterized by their ability to perform one function really well, many repeats mediate multiple biological functions (Gemayel et al. 2010). Consequently, understanding how repeats specialize and handle these multiple functions is an important future direction in the study of repeat function. Tandem repeat variation is not only characterized by changes in the repeat number, but also in changes to the consensus repeat sequence termed degeneration. While some of these substitutions are expected to be neutral changes, they can further specialize one or a block of repeat for a specific function. The serine-7 position in the CTD heptad repeat in mammals is a prime example, where substitutions of serine to lysine open up specific PTM-mediated interactions with co-transcriptional processes (Simonti et al. 2015). Degeneration from the consensus sequence is a promising metric to study how repeats specialize irrespective of any particular repeat sequence or function. The prediction from such a model is that increasing organismal complexity is correlated to increasing specialization of key repetitive sequences (Figure 3). The opposite case has been also

observed with the CTD, where simpler organisms have repeats that adhere closely to the consensus sequence and organisms that adapt to a parasitic life history lose their CTDs in general (Yang and Stiller 2015). Focusing on repeat sequence specialization would enable the clarification of further links between repeat structure and cellular function.

Specialization of repeat function can also be established through mechanisms outside of sequence divergence. In chapters 3 and 4 of this thesis, I laid out the region specific effects of the CTD repeats on transcriptional activity of RNA polymerase II (Babokhov et al. 2018). This finding raised the possibility that repeats can specialize even in the absence of sequence differences. One way to examine how this can occur is to again look at protein-protein interaction networks and examine how identical repeat structure correlates to the creation of additional interactions in node proteins. Here the prediction would be that additional interactions without changes in repeat sequence would suggest a region specific specialization similar to that of the CTD. However, a drawback of this approach is the additional functions could also be explained as redundant binding along any of the repeating units, especially if there are changes in repeat copy number. Looking for region specialization may be especially difficult to tackle with a systems biology approach and will likely require further molecular studies of a few representative repeats. One promising repeat is the C-terminal region (CTR) repeats of the transcription elongation factor Spt5p. Spt5p is currently the only known elongation factor that is conserved in all the domains of life (Zhou et al. 2009), and its CTR repeats behave similarly to the CTD repeats of RNA polymerase II (Ding et al. 2010). Studying

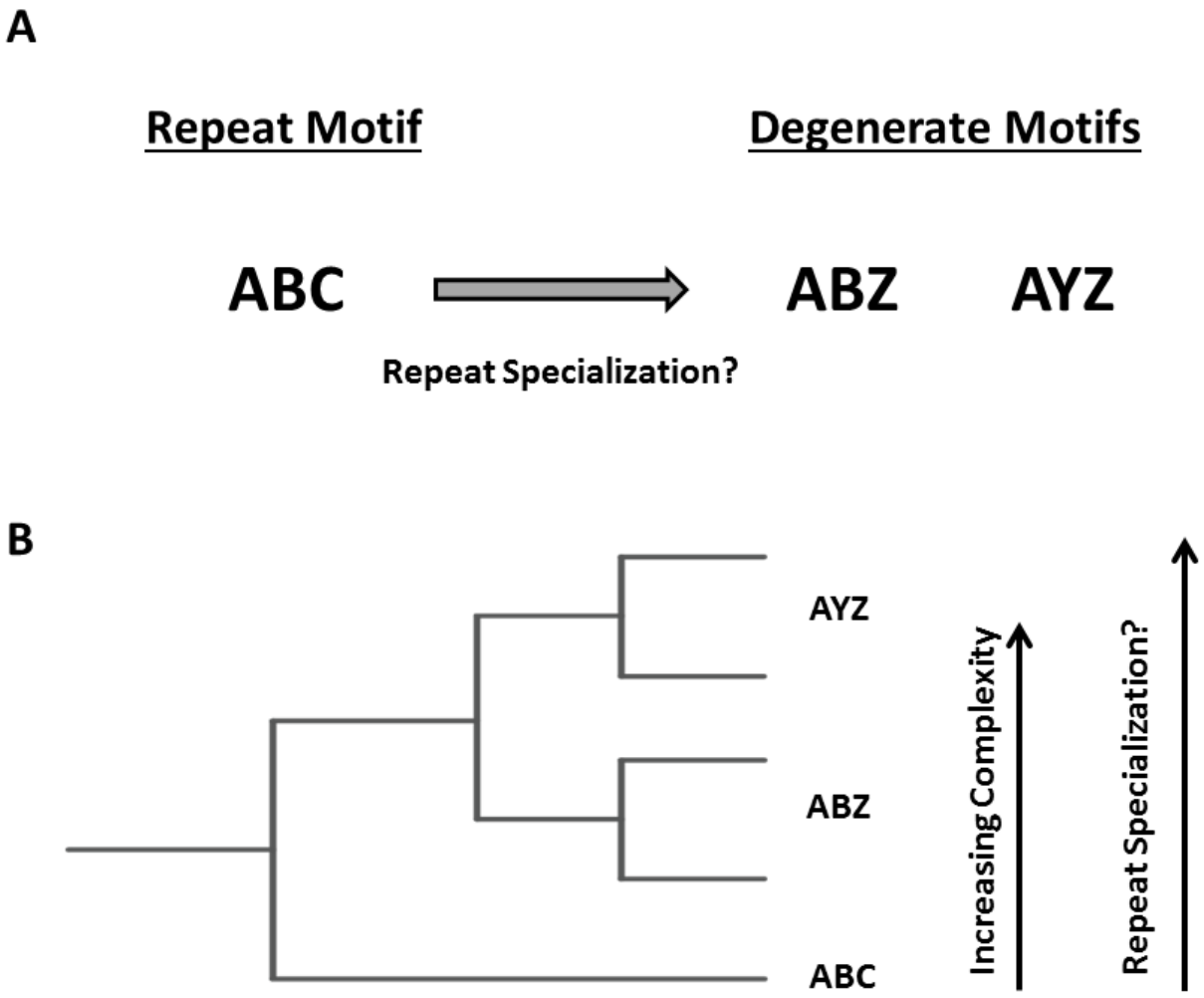


Figure 3. Model to track repeat specialization throughout evolutionary time. **A**) A hypothetical trimer repeat ABC can undergo substitutions to form degenerate repeats ABZ and AYZ. This model predicts that increasing conserved substitutions would be correlated to increasing organism complexity. **B**) Prediction of repeat ABC specialization through evolutionary time. With increasing complexity, there would be increasing amounts of residue substituted repeats that could be quantified as a function of evolutionary distance.

Spt5p would be a powerful comparative approach to begin to determine region specificity prevalence and its broader role in shaping repeat specialization.

So far the discussion of repeat specialization has been focused on specialized functions of CTD-like minisatellites. However, can simpler microsatellite repeats like polyQ or polyE also exhibit specialized functions? On the surface the answer appears to be no, as microsatellites are typically thought to act through bulk biophysical properties such as electrostatic interactions or aggregation (Gemayel et al. 2015). These simple repeats also frequently appear as mixed repeats (*e.g.* polyE/D or polyQ/N), further arguing that the overall biochemical properties are more relevant than the specific sequence of the repeat. However, in the course of examining variable repeats in chapter 2, I identified a number of repeats that have a partition of poly amino acid repeats (Figure 4). These types of microsatellite arrangements could represent a rudimentary form of repeat specialization among low complexity repeats. Additional analyses of these arrangements will be required to determine if they correlate to specific functions or added network interactions of the protein in question. This approach would expand the field to cover all of the common repeat types while addressing the question of repeat specialization.

Towards a comprehensive perspective of repeat function

Ultimately, answering the three questions of how repeats emerge, change and specialize will establish a framework that can be used to address repeat function as a whole and not on a case-by-case basis. While there will likely be different classes of

A

YML049C

B

YER002W

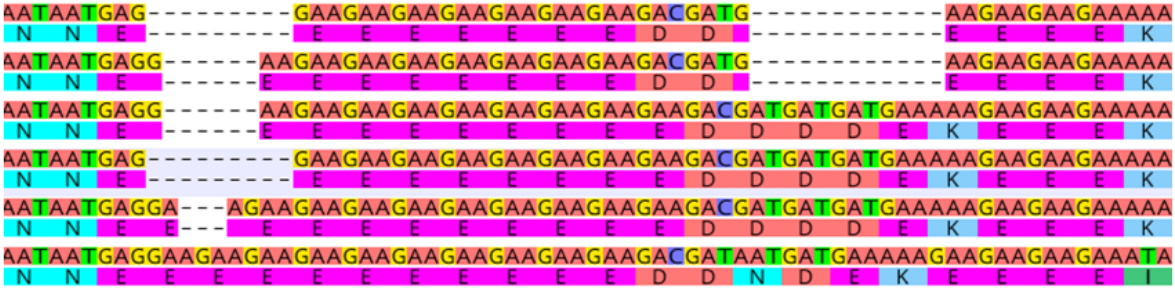


Figure 4. Mixed amino acid repeats with potential region specific functions. **A**) Most mixed repeats with similar amino acid content are similar to YML049C, containing either evenly spaced E/D or a more random mix. **B**) Some mixed repeats have discreet regions of polyE and polyD that can be variable, raising the potential of subtle region specific effects even in homopolymeric repeats.

repeats that share particular functional properties, such an approach to looking at repeats will enable scientists to comprehensively address difficult questions of repeat function. One such example is if repeat function is universal or strictly dependent on the protein context, *i.e.* can a repeat only provide its function in the protein it was discovered in? If a microsatellite was known to increase transcription factor binding in its natural protein, would it have this effect if it was engineered onto other transcription factors (Gemayel et al. 2015)? We already know that there is some wiggle room in repeat function, for instance the CTD can be moved to other RNA polymerase II subunits and still preserve its function provided it is near to the site of transcription (Suh et al. 2013). Understanding how repeats emerge, change and specialize would allow us to apply the mechanisms obtained from these questions to pose and address problems such as repeat universality.

In conclusion, our understanding of protein function has benefitted greatly from advances in the study of tandem repeats. The three questions that I covered in this chapter will guide further progress in the field, aided by advances in computational and systems biology approaches. Importantly, many of the insights gained from repeat function will also be applicable to the study of disordered regions and would help clarify the relationship between disorder and repetitiveness. Working out the general mechanisms behind how repeats emerge, change and specialize not only explains the properties of already identified repeats but can also predict the properties of novel repeats identified from genomic data. The massive expansion in available sequence data will enable the prediction of repeat emergence and expansion and the subsequent

effects on protein function. This large dataset of predicted and tested repeat interactions will allow the true extent of the effect of repeats on protein function to be determined to build a more complete model of the relationship between the structure of a protein and its biological function.

Chapter 5 Literature Cited

Albrecht A, Mundlos S. The other trinucleotide repeat: polyalanine expansion disorders. Curr Opin Genet Dev. 2005 Jun;15(3):285-93.

Babokhov M, Mosaheb MM, Baker RW, Fuchs SM. Repeat-Specific Functions for the C-Terminal Domain of RNA Polymerase II in Budding Yeast. G3 (Bethesda). 2018 May 4;8(5):1593-1601.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet. 1998 Jun;62(6):1408-15.

Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, Latysheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, Babu MM. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol. 2017 Sep;24(9):765-777.

Ding B, LeJeune D, Li S. The C-terminal repeat domain of Spt5 plays an important role in suppression of Rad26-independent transcription coupled repair. J Biol Chem. 2010 Feb 19;285(8):5317-26.

Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71.

Fan H, Chu JY. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics. 2007 Feb;5(1):7-14.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445-77.

Gemayel R, Chavali S, Pougach K, Legendre M, Zhu B, Boeynaems S, van der Zande E, Gevaert K, Rousseau F, Schymkowitz J, Babu MM, Verstrepen KJ. Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. Mol Cell. 2015 Aug 20;59(4):615-27.

Gemayel R, Yang Y, Dzialo MC, Kominek J, Vowinckel J, Saels V, Van Huffel L, van der Zande E, Ralser M, Steensels J, Voordeckers K, Verstrepen KJ. Variable repeats in the eukaryotic polyubiquitin gene ubi4 modulate proteostasis and stress survival. Nat Commun. 2017 Aug 30;8(1):397.

Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol. 2014 Jun 2;6(6).

La Spada AR, Taylor JP. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet. 2010 Apr;11(4):247-58.

Morrill SA, Exner AE, Babokhov M, Reinfeld BI, Fuchs SM. DNA Instability Maintains the Repeat Length of the Yeast RNA Polymerase II C-terminal Domain. J Biol Chem. 2016 May 27;291(22):11540-50.

Paeschke K, Bochman ML, Garcia PD, Cejka P, Friedman KL, Kowalczykowski SC, Zakian VA. Pif1 family helicases suppress genome instability at G-quadruplex motifs. Nature. 2013 May 23;497(7450):458-62.

Richard GF, Pâques F. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. 2000 Aug;1(2):122-6.

Simonti CN, Pollard KS, Schröder S, He D, Bruneau BG, Ott M, Capra JA. Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. BMC Evol Biol. 2015 Mar 10;15:35.

Suh H, Hazelbaker DZ, Soares LM, Buratowski S. The C-terminal domain of Rpb1 functions on other RNA polymerase II subunits. Mol Cell. 2013 Sep 26;51(6):850-8.

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nat Genet. 2005 Sep;37(9):986-90.

Yang C, Stiller JW. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A. 2014 Apr 22;111(16):5920-5.

Zhou K, Kuo WH, Fillingham J, Greenblatt JF. Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. Proc Natl Acad Sci U S A. 2009 Apr 28;106(17):6956-61.

Appendix

The following appendix compiles the strains and protocols that I used when training in the use of the fission yeast model system.

Table 1. List of fission yeast strains available in the Fuchs Lab.

From Dr. Hiroaki Kato, Department of Biochemistry Shimane University Medical School

Name	Mating Type	Genotype
HKM-475	-	leu1-32, tfa2-3HA::kanMX6
HKM-476	-	leu1-32, tfa2-13myc::kanMX6
HKM-483	-	srb4-13myc::kanMX6
HKM-972	-	
HKM-975	+	
HKM-1100	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+
HKM-1102	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L::ura4+, otr1R::ade6+
HKM-1219	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, ∆clr4::hphMX
HKM-1334	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, rpb2-m203
HKM-1685	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, rpb3-5FLAG-kanMX
HKM-1740	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, spt6-3HA-natMX
HKM-1747	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, spt6-3HA-natMX
HKM-1766	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, rpb3-5FLAG-natMX
HKM-1967	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, rik1-10myc-kanMX
HKM-2124	-	ade6-DN/N, ura4-DS/E, imr1L::ura4+, otr1R::ade6+, rpb1-GFP(S65T)::kanMX

From Dr. Takuya Kajitani, Department of Chemistry Hokkaido University		
TKY032	90	ura4-5BoxB-hph, tas3λN-kanMX, chp1-mycx6-his3, ade6-m210, rpb2+, his3-D1?, otr1R(SphI)::ade6, Δeri1::ble
TKY328	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, chp1-13myc-nat
TKY407	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, chp1-13myc-kan, Δclr4::nat
TKY410	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, chp1-13myc-kan, Δdcr1::nat
TKY546	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rpb1-CTD S2A-kan
TKY549	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6 S165A/L238R-3HA-kan
TKY555	-	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6ts1-3HA-kan
TKY557	-	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6ts2-3HA-kan
TKY573	?	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6-3HA-kan
TKY579	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rpb1-CTD S2A-kan
TKY582	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6-S165A-kan
TKY584	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, mcs6-S165A-kan
TKY842	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Δlsk1::kanMX
TKY843	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Δcsk1::kanMX
TKY846	?	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Nat-pago1-3FLAG-ago1::ago1
TKY856	?	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Nat-pago1-3FLAG-ago1::ago1, Δclr4::kan
TKY859	?	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Nat-pago1-3FLAG-ago1::ago1, Δdcr1::kan
TKY887	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rik1-13myc-nat
TKY893	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rik1-13myc-nat, Δclr4::kan
TKY896	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rik1-13myc-nat, Δdcr1::kan
TKY919	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rdp1-13myc-nat
TKY922	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rdp1-13myc-nat, Δclr4::kan
TKY925	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rdp1-13myc-nat, Δdcr1::kan

TKY990	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, cid12-13myc-nat
TKY993	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, cid12-13myc-nat, Δclr4::kan
TKY995	-	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, cid12-13myc-nat, Δdcr1::kan
TKY1513	90	ade6-DN/N, leu1-32, ura4-DS/E, otr1R(Sph1)::ade6+, his2, chp1+, his3-D1, kint2::ura4+
TKY1586	-	rpb1-(CTD-wt)11-MCE1-nat
TKY1595	-	spt5ΔCTR-kan
TKY1596	-	spt5ΔCTR-ura4
TKY1602	+	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, cdk9T212A-kan
TKY1603	+	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, cdk9T212E-kan
TKY1635	?	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rpb1-CTD-S7E-nat
TKY1652	?	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, rpb1-(CTD-S5A)11-MCE1-nat
TKY1674	+	ade6-DN/N, leu1-32, his2-, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, spt6-13myc-kan
TKY1685	+	ade6-DN/N, leu1-32, his2-, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, spt6-13myc-kan, Δclr4::nat
TKY1687	+	ade6-DN/N, leu1-32, his2-, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, spt6-5flag-kan, Δdcr1::nat
TKY1781	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, chp1+, his2, kint2::ura4+, otr1R(Sph1)::ade6+, tas3-13myc-nat
TKY1785	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, chp1+, his2, kint2::ura4+, otr1R(Sph1)::ade6+, stc1-13myc-nat
TKY1849	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Δiwr1::kan
TKY1908	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, chp1+, his2, kint2::ura4+, otr1R(Sph1)::ade6+, stc1-13myc-nat, Δclr4::kan
TKY1911	90	ade6-DN/N, leu1-32, his3-D1, ura4-DS/E, chp1+, his2, kint2::ura4+, otr1R(Sph1)::ade6+, stc1-13myc-nat, Δdcr1::kan
TKY1945	?	ade6-DN/N, leu1-32, ura4-DS/E, otr1R(SphI)::ade6+, imr1L(NcoI)::ura4+, mcs6-as2-hph, FLAG-rpb3
TKY1992	-	ade6-DN/N, leu1-32, ura4-DS/E, chp1+, his3-D1, otr1R(Sph1)::ade6+
TKY2002	+	ade6-DN/N, leu1-32, ura4-DS/E, chp1+, his3-D1, otr1R(Sph1)::ade6+
TKY3157	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, Δpin1::hph
TKY3351	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, kanMX6-Purg1-3flag-fcp1
TKY3355	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, kanMX6-Purg1-3flag-ssu72
TKY3359	+	ade6-DN/N, leu1-32, ura4-DS/E, imr1L(Nco)::ura4+, otr1R(Sph1)::ade6+, kanMX6-Purg1-3flag-rtr1

TKY4017	-	ade6-DN/N, leu1-32, ura4-DS/E, otr1R::ade6+, imr1L::ura4+, flag-rpb3, kan-Purg1-3ha-iwr1
TKY4021	-	ade6-DN/N, leu1-32, ura4-DS/E, otr1R::ade6+, imr1L::ura4+, flag-rpb3, hph-Purg1-3ha-iwr1

Pombe Lithium Acetate Transformation Protocol

From Dr. Hiroaki Kato, Shimane University Medical School Department of Biochemistry

Reagents:

0.1M LiAc (pH 4.9), autoclaved

50% PEG3350, autoclaved in glass bottle

MilliQ Water, autoclaved

1. Incubate cells in YES medium at 30°C overnight. The cells should be well-growing. Cell concentration should be 0.1 – 1.0 x 10^7 cells/mL. When it is over 1.0 x 10^7 cells/mL, dilute cells with fresh YES and incubate several hours. We use 50 – 200 mL YES in a flask.
2. Harvest cells in a 50 mL plastic tube at 3,000 rpm. 30 seconds at 3,000 rpm is enough to get solid cell pellets. You need 1 x 10^8 cells for one plasmid to be introduced. Suppose that you have two plasmid DNA. If the cell concentration at step 1 is 0.5 x 10^7 cells/mL, pour 40 mL of cell culture into a 50 mL tube. Similarly, use 20 mL of culture when the concentration is 1 x 10^7 cells/mL.
3. Discard the supernatant by decantation. Carefully look at the pellet. If the pellet is soft and it loses its shape, stop decantation and centrifuge the tube again. After decantation, remove remaining supernatant on the pellet with a pipet.
4. Suspend the ells with 40 mL of Milli Q water. After you add water, close the cap tightly, hold the tube horizontally and tap the tube hardly to suspend cells.

5. Harvest the cells at 3,000 rpm. You need 5 minutes to get solid cell pellets when the cells are suspended in water.
6. Discard the supernatant by decantation and then with a pipet.
7. Suspend the cells with 10 mL of 0.1M LiAc (pH 4.9)
8. Discard the supernatant by decantation and then with a pipet.
9. Suspend the cells with a suitable amount of 0.1 M LiAc. Cell concentration should be 1 x 10^9 cells/mL. If there are 2 x 10^8 cells in the tube, use 200 µL of LiAc to suspend them.
10. Store the cell suspension at 30°C for 1 hour.
11. Put 15 µL of DNA solution (1 ng in TE) in a 1.5 mL microcentrifuge tube.
12. Add 100 µL of cell suspension prepared at step 10.
13. Add 290 µL of 50% PEG3350 and mix well by vortexing.
14. Store the tube at 30°C for 1 hour.
15. Incubate the tube at 42°C for 15 minutes.
16. Centrifuge at 5,000 rpm for 2 minutes. Remove the supernatant with a pipet. Suspend the cells with 500 µL of Milli Q water.
17. Centrifuge at 5,000 rpm for 2 minutes. Remove the supernatant with a pipet. Suspend the cells with 100 µL of Milli Q water.
18. Transfer the cell suspension onto a selective plate. Spread the cells with a spreader unevenly.
19. Incubate the plate at 30°C for 3 – 5 days.

Pombe Colony PCR Protocol

From Dr. Takuya Kajitani, Hokkaido University Department of Chemistry

Stock solution of Zymolyase 20T: 10 mg/mL of zymolyase 20T in distilled water as 10x stock.

1) Prepare working solution of Zymolyase in 1.5 mL microtube:
    - 27 µL of 0.1x TE + 3 µL of 10 mg/mL zymolyase 20T per sample
2) Pick small colonies that have been freshly streaked (2 – 4 days after streaking)
3) Resuspend the colony in the solution from step 1 and incubate at 37°C for 3 – 4 hr
    - Can incubate overnight for better signal
4) Vortex tube and boil at 98°C for 10 min (do not boil longer than 20 min)
5) Vortex again and spin down 15,000 prm for 1 min
6) Use 1.5 µL of supernatant per 10 µL of PCR reaction for 38 cycles.

Additional Pombe resources

Genome database: https://www.pombase.org/

Repository of fission yeast resources (stains and plasmids): http://yeast.nig.ac.jp/yeast/

General fission yeast methods:

Sabatinos SA, Forsburg SL. Molecular genetics of Schizosaccharomyces pombe. Methods Enzymol. 2010;470:759-95.

Petersen J, Russell P. Growth and the Environment of Schizosaccharomyces pombe. Cold Spring Harb Protoc. 2016 Mar 1;2016(3):pdb.top079764.