STRUCTURAL HETEROGENEITY IN THE INTRINSICALLY DISORDERED
RNA POLYMERASE II C-TERMINAL DOMAIN

A Dissertation in
Biochemistry, Microbiology and Molecular Biology
by
Bede Portz

© 2017 Bede Portz

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2017
The dissertation of Bede Portz was reviewed and approved* by the following:

David S. Gilmour  
Professor of Molecular and Cell Biology  
Dissertation Advisor  
Chair of Committee

Joseph C. Reese  
Professor of Biochemistry and Molecular Biology

Katsuhiko Murakami  
Professor of Biochemistry and Molecular Biology

B. Franklin Pugh  
Evan Pugh University Professor  
Willaman Chair of Molecular Biology  
Professor of Biochemistry and Molecular Biology

Song Tan  
Professor of Biochemistry and Molecular Biology

Scott A. Showalter  
Associate Professor of Chemistry

Wendy Hanna-Rose  
Associate Professor and Head, Department of Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School
ABSTRACT

RNA polymerase II contains a repetitive and intrinsically disordered C-Terminal Domain (CTD) composed of heptad repeats of the consensus sequence YSPTSPS. The CTD can be heavily phosphorylated and serves as a scaffold, interacting with factors involved in transcription initiation, elongation, termination, RNA processing and chromatin modification. Despite its role as a nexus of eukaryotic gene regulation, the structure of the CTD and the structural implications of CTD phosphorylation, are poorly understood. Additionally, there is an increasing awareness of the importance of intrinsically disordered proteins (IDPs) that function without adopting a stably folded structure. Here I present a biophysical and biochemical interrogation of the structure of the full-length CTD of D. melanogaster, which I conclude is a compact random coil. I find that the repetitive CTD is structurally heterogeneous as evidenced by a discontinuous pattern of cutting in limited proteolysis assays. Small Angle X-Ray scattering (SAXS) is a method ideally suited for the structural interrogation of large IDPs and can be employed to measure the size of a protein and to monitor structural changes in response to post-translational modification. Using SAXS I determined that phosphorylation by the kinase P-TEFb caused an increase in CTD radius and stiffness. Limited proteolysis of the phosphorylated CTD showed these gross structural changes are accompanied by increased protease accessibility and an alteration in relative protease accessibility across the length of the CTD.
Additionally, we show that the human CTD is also structurally heterogeneous and able to substitute for the *Drosophila melanogaster* CTD in supporting the development of flies to adulthood. These finding implicate conserved structural organization, not a precise array of heptad motifs, as important to CTD function.

The CTD is attached to the catalytic core of Pol II via a linker. I show that this linker is more compact than the CTD repeats and serves as an independent structural unit. The phosphorylated linker-CTD remains flexible relative to the phosphorylated CTD alone. Together, these results support a mechanism by which phosphorylation reduces the conformational entropy of the CTD, generating a more binding competent dock for CTD:protein interactions, with the linker region maintaining the ability of CTD bound factors to sample the 3-dimensional space which may be required for RNA processing and histone modification.

The data described herein represent the most thorough structural characterization to date of the full length CTD on the global and local scales, examining both the overall size and local structural organization of the CTD. These studies establish the *Drosophila* CTD as an attractive model for the biophysical, biochemical and genetic interrogation of the structure and function of the CTD from a developmentally complex organism.
# TABLE OF CONTENTS

LIST OF FIGURES.............................................................................................................vii

LIST OF TABLES................................................................................................................xi

ACKNOWLEDGEMENTS....................................................................................................xii

Chapter 1 Introduction: The CTD as an IDP: What might lessons from IDP biology teach us about gene regulation via the CTD?...........................................1

1-1.1 The CTD is a nexus of eukaryotic gene regulation ............................................1
1-1.2 What are IDPs, and what do they do?.................................................................6
1-1.3 The evolution of the CTD ......................................................................................8
1-1.4 The CTD is a series of SLiMs .............................................................................9
1-1.5 SLiMs and multivalency enable evolutionary experimentation .....................13
1-1.6 CTD organization in metazoans ......................................................................15
1-1.7 Sequence vs. structure of CTD regions ............................................................18
1-1.8 IDP allostery and the CTD .................................................................................20
1-1.9 The structure of the CTD is poorly understood .................................................21
1-1.10 Questions to be addressed ..............................................................................23

Chapter 2 Biophysical characterization of the RNA polymerase II C-terminal domain.........................................................................................................................26

2-1 Introduction.............................................................................................................26
2-2 Results.....................................................................................................................31

2-2.1 The CTDs of yeast, fly and human are structural homologs .........................31
2-2.2 Phosphorylation by P-TEFb extends the CTD ..................................................36
2-2.3 Mass spectrometry characterization of P-TEFb phosphorylated CTD .........39
2-2.4 Substrate specificity of P-TEFb for the CTD ....................................................43
2-2.5 Small angle X-ray scattering studies to characterize the CTD structure ......50
2-2.6 SAXS sample preparation .................................................................................53
2-2.7 SAXS analysis ....................................................................................................57
2-2.8 The CTD is semi-compact ..................................................................................61
2-2.9 Ensemble optimization models of the CTD ......................................................65

2-3 Discussion.................................................................................................................67

Chapter 3 Local structural heterogeneity and ensemble allostery in the Pol II CTD .................................................................................................................................70

3-1 Introduction.............................................................................................................70
3-2 Results.....................................................................................................................76

3-2.1 The CTD of *D. melanogaster* is structurally heterogeneous .......................76
3-2.2 Structural heterogeneity persists after phosphorylation……87
3-2.3 CTD phosphorylation increases protein accessibility……89
3-2.4 The CTD stiffens as a function of phosphorylation……90
3-2.5 The human CTD is also structurally heterogeneous……94
3-2.6 The Human CTD can function in place of the fly CTD in vivo………………………………………………………………………95
3-3 Discussion…………………………………………………………………95

Chapter 4 The linker region is a distinct structural unit of the CTD……118
4-1 Introduction………………………………………………………………118
4-2 Results……………………………………………………………………124
  4-2.1 Defining the linker………………………………………………124
  4-2.2 The CTD with the linker is more compact than predicted…127
  4-2.3 The linker is a structurally independent unit of the CTD…..131
  4-2.4 Conformational changes in the CTD with the linker……135
4-3 Discussion………………………………………………………………149

Chapter 5 Structure enters the CTD code……………………………153

Chapter 6 Materials and Methods……………………………………172

Appendix A CG3773 is the Drosophila homolog of Gdown1…………188
  A-1 Introduction…………………………………………………………188
  A-2 Results………………………………………………………………191
    A-2.1 Generation and characterization of a FLAG tagged Rpb1 fly line……………………………………………………………………191
    A-2.2 Purification of Pol II and Pol II(G)…………………………194
  A-3 Discussion………………………………………………………………201
  A-4 Materials and Methods………………………………………………203

Appendix B Nucleosome organization in D. melanogaster is stable in response to heat shock………………………………………………………206
  B-1 Introduction……………………………………………………………206
  B-2 Results………………………………………………………………210
  B-4 Materials and Methods………………………………………………211

References………………………………………………………………213
LIST OF FIGURES

Figure 1-1 Specific CTD phosphoisoforms recruit distinct interactomes...........4
Figure 1-2 Energy landscapes of native and intrinsically disordered proteins......7
Figure 1-3 CTD structures in complex with globular domains..................12
Figure 1-4 A comparison of CTD sequences from S. cerevisae, D. melanogaster
and H. sapiens..................................................................................................................................................17
Figure 1-5 Existing models of CTD structure........................................23
Figure 2-1 Consensus motif comparison of H. sapiens and D. melanogaster CTD
heptads........................................................................................................................................................31
Figure 2-2 Native gel electrophoresis MBP-CTD fusion proteins..............33
Figure 2-3 Size exclusion chromatography of MBP-CTD fusion proteins......34
Figure 2-4 The CTD extends after P-TEFb phosphorylation as measured by size
exclusion chromatography.................................................................37
Figure 2-5 The apo and phospho MBP-D.melCTD fusions elute as sharp peaks
from a size exclusion column...............................................................38
Fig 2-6 MBP-D.melCTD sample preparation for mass spectrometry analysis....41
Figure 2-7. MALDI-TOF mass spectrometry of MBP-D.melCTD to quantify
phosphate incorporation........................................................................42
Figure 2-8. Tandem mass spectrometry to map phosphate incorporation by P-
TEFb........................................................................................................48
Figure 2-9. Small Angle X-Ray Scattering of proteins in solution.................52
Figure 2-10 MBP-D.melCTD sample preparation for small angle x-ray
scattering.................................................................................................54
Figure 2-11. Components of the P-TEFb reaction are resolved from the CTD via preparative size exclusion chromatography..................................................56

Figure 2-12 Averaged, buffer subtracted SAXS curves from three protein concentrations plotted as I(q) vs q.................................................................58

Figure 2-13 Guinier fits of SAXS data for the apo and phospho CTD..............59

Figure 2-14 The CTD is semi compact..........................................................63

Fig 2-15 Ensemble optimization method models of the CTD..........................66

Figure 3-1 The human and fly CTDs have different sequences but a similar consensus motif.................................................................73

Figure 3-2 Amino acid sequence alignment of 12 species of *Drosophila*........75

Figure 3-3 Limited Proteolysis Experimental design........................................77

Figure 3-4 Generation of CTD specific MW markers........................................78

Figure 3-5 Limited proteolysis with chymotrypsin.........................................79

Figure 3-6 Multiple proteases yield similar patterns of proteolysis across the CTD..................................................................................81

Figure 3-7. A GST-CTD fusion has the same proteolytic pattern as the MBP-CTD fusion.................................................................................82

Figure 3-8 A comparison of protease accessible sites........................................84

Figure 3-9 The structure of the CTD is sensitive to the denaturant SDS...........86

Figure 3-10 Local structural heterogeneity is maintained in the phospho CTD...88

Figure 3-11. Phosphorylation increases the proteolytic sensitivity of the CTD...89

Figure 3-12 Phosphorylation stiffens the CTD.................................................91

Figure 3-13 Validation of phosphorylation dependent stiffening....................93
Figure 3-14 The human CTD is structurally heterogeneous.................................95
Figure 3-15 Gal4 driven Rpb1 knockdown and expression system....................97
Figure 3-16 Knockdown of endogenous Rpb1 and coincident expression of
humanized Rpb1 in flies....................................................................................99
Figure 3-17 Results of RNAi rescue experiments with Rpb1 derivatives...........101
Figure 3-18 Quantification of mRNA levels for Rpb1 derivatives in adult flies from
RNAi rescue crosses.........................................................................................102
Figure 3-19 Immunofluorescence staining of polytene chromosomes detects
ectopically expressed Rpb1 on chromatin.....................................................104
Figure 3-20 Ectopically expressed Rpb1 derivatives co-localize with the Rpb3
subunit of Pol II across chromosomes..............................................................105
Figure 3-21 The central, protease resistant region of the Drosophila CTD is
conserved between Drosophila and human.....................................................116
Figure 4-1 A diagram of CTD domains..............................................................118
Figure 4-2 Defining the linker region in D. melanogaster.................................126
Figure 4-3 Native gel electrophoresis of the linker containing CTD compared to
CTDs lacking the linker....................................................................................128
Figure 4-4 The CTD with the linker region is more compact than predicted.....130
Figure 4-5 The linker is a structurally independent subdomain of the CTD......134
Figure 4-6 Phosphorylation leads to extension of the CTD containing the linker
region.................................................................................................................136
Figure 4-7 Expression and purification strategy for linker CTD proteins for small
angle X-ray scattering......................................................................................138
Figure 4-8 SAXS curves and Guinier fits of phospho MBP-$D.mel$/CTD w/linker
Figure 4-9 Pairwise distance distribution function of phospho MBP-$D.mel$/CTD w/linker SAXS data
Figure 4-10 Flexibility analysis of phospho MBP-$D.mel$/CTD w/linker SAXS data
Figure 4-11 Kratky plot of phospho MBP-$D.mel$/CTD w/linker SAXS data extrapolated to zero concentration
Figure 6-1 Representative purification of MPB-$D.mel$/CTD6HIS
Figure 6-2 Purification of $D. melanogaster$ P-TEFb from Sf9 cells
Figure 6-3 CTD ladder cleavage fragments
Figure A-1 Map of the FLAG-Rpb1 genomic insert
Figure A-2 FLAG-Rpb1 is expressed in the 10M.11M fly line
Figure A-3 Pol II purification strategy and activity assay results
Figure A-4 Silver stained SDS-PAGE gel of peak 1 and peak 2 Pol II
Figure A-5 Coomassie stained gel of TCA precipitated FLAG peak 1 and peak 2 Pol II for mass spectrometry
Figure A-6 anti-Rpb3 and anti-Gdown1 western blot of mono Q peak 1 and peak 2 pol II
Figure A-7 anti-Gdown1 western blot of embryo lysate and purified Pol II
Figure B-1 Immunofluorescence of polytene chromosomes before and after heat shock
Figure B-2 MNase ChIP-seq before and after heat shock
LIST OF TABLES

Table 2-1 Summary of Guinier analyses……………………………………………………………60
Table 5-1 Summary of Guinier analysis of phosho MBP-\textit{D.mel}CTD w/linker…141
ACKNOWLEDGEMENTS

Training to become a scientist shapes one’s worldview beyond experiments and data, changing the way you encounter the natural world. To have Dave as my mentor on this journey has been one of the great joys of my life. I cannot thank him enough for his teaching, wisdom, patience, and willingness to let me pursue my own questions.

Likewise, my thesis committee members have been helpful in many important ways; I appreciate their insights and suggestions. Beyond scientific advice, members of my committee have taken a genuine interest in my career and I thank them for their helpful advice regarding my goals.

Helpful collaborators have profoundly shaped my training and thinking. In particular, I thank Scott Showalter and Eric Gibbs for opening up a new world of biology to me through countless conversations and tireless explanations. I also thank Feiyue Lu for his help with the fly genetics and collaborators at the University of Texas at Austin for their help with mass spectrometry, Jessie Zhang, Josh Mayfield, Rachel Mehaffey and Jennifer Broadbelt.

I would like to thank Surinder Chopra for providing my first chance to do science. I am also sincerely thankful to Song Tan, Phillip Mohr, and Carl Sillman for their concern so many years ago. I am grateful to past and current members of the Gilmour lab for their help and friendship.

My family has been incredibly supportive of my pursuit of a Ph.D. and career in science. I thank my parents and my sister for their encouragement and
my mother in law for the many weekends she helped watch the twins so I could go to the lab. I also thank my friends for their support.

Finally, Anna, Katherine and Everett have supported me in countless ways and sacrificed much in my pursuit of a Ph.D. Throughout graduate school Anna has been unwavering in her love and support and has been a constant source of joy in my life. Her love and support and belief in me has been a blessing beyond words.
CHAPTER 1

THE CTD AS AN IDP: WHAT MIGHT LESSONS FROM IDP BIOLOGY TEACH US ABOUT GENE REGULATION VIA THE CTD?

1-1.1 The CTD is a nexus of eukaryotic gene regulation.

The study of transcription regulation and intrinsically disordered proteins (IDPs) converge on the C-terminal domain (CTD) of RNA polymerase II (Pol II). Gene transcription by Pol II is tightly regulated at each stage of the transcription cycle: recruitment, initiation, elongation, and termination(1). The RNA transcript itself must be correctly capped, spliced and processed at the 3’ end, and the chromatin environment characteristic of transcribed loci must be established and maintained. The regulatory nexus of Pol II transcription and co-transcriptional processes is a highly repetitive, intrinsically disordered C-terminal domain emanating from the largest Pol II subunit proximal to the RNA exit channel(2-4). The CTD is a low complexity domain (LC domain), comprised of repeating heptads of the consensus sequence \( Y_1S_2P_3T_4S_5P_6S_7 \)(5). Each non-proline residue of the consensus repeat has been shown to be phosphorylated, and the proline residues can exist in either the \textit{cis} or \textit{trans} conformation(2, 3). The CTD serves as a scaffold, tethering factors involved in every stage of the transcription cycle, the recycling of Pol II for subsequent rounds of transcription, RNA processing, and chromatin modification, to the polymerase. The CTD orchestrates regulation via association with an ever-growing list of more than 200 known interacting proteins(6). These CTD:factor interactions are controlled in
part by the dynamic application and removal of phosphorylation marks and the regulation of proline isomerization(3). Collectively, the phenomena of dynamic and reversible CTD modifications and the recognition of specific modifications by factors have been dubbed “the CTD code(7, 8).” Much of what is known about the CTD resulted from pioneering genetic studies by Corden and colleagues that identified the CTD, and specific positions within the heptad repeat, as essential for cell survival(9). More recently genomic studies found specific patterns of phosphorylation predominate at specific stages of the transcription cycle(10, 11) (Fig. 1a adopted from Geyer and Eick(3)). The distribution of Pol II phospoisoforms often correlates with the occupancy of various factors shown to bind those marks(3). The concept of tethering factors to particular CTD modifications at specific stages of transcription has been dubbed, “the CTD code(8).” For example capping enzyme, responsible for addition of the 7-methylguanosine cap on the 5’ end of nascent RNA, is recruited by S5 phosphorylation (pS5), which predominates at the 5’ end of genes(10, 12). Another well characterized CTD modification, S2 phosphorylation (pS2), predominates at the 3’ end of genes and recruits factors involved in transcription termination(10, 13). In the most thorough validation of the CTD code hypothesis to date, Harlan and colleagues found Pol II, purified with antibodies directed against specific phosphorylation marks, associates with distinct interactomes(6). In this study, Pol II elongation complexes were immunoprecipitated from yeast extract using an antibody against an epitope tag on the Rpb3 subunit of Pol II. These complexes were eluted and subsequently immunoprecipitated with
antibodies specific for the phosphorylated residues of the heptad repeat or with an antibody against mCherry which was used as a control. Co-immunoprecipitated proteins were identified via mass spectrometry and the enrichment of factors from the phospho-specific CTD antibodies was compared to the mCherry controls. The experiment was done in triplicate for each CTD antibody and the data were subject to principal component analysis (6) (Fig. 1-1b adopted from Harlen et al. (6)). This analysis reduces the dimensionality of the data and enables a comparison of the variance in the data between like triplicates or across different antibodies (14). The authors found that the individual experiments from each phospho-specific immunoprecipitation clustered together, and individual triplicates clustered more closely to one another than to the clusters arising from the other phospho-specific immunoprecipitations (Fig. 1b) (6). This observation suggests that as Pol II traverses a gene, a wholesale exchange of factors associated with specific Pol II phospho-isoforms occurs in concert with changes to the phosphorylation status of the CTD (Fig. 1-1).
Fig 1-1 Specific CTD phosphoisoforms recruit distinct interactomes.

(a) ChIP experiments in yeast and mammalian cells localize specific phosphorylation marks to particular stages of the transcription cycle, relative to gene features including the transcription state site (TSS), poly-adenylation site (pA) and transcription termination site (TTS)(3). Distribution of phosphorylation is largely similar in yeast and mammalian cells. Threonine 4 phosphorylation is shown as a dashed line in mammalian cells where the pattern differs from that of yeast. (b) Principal component analysis of factors enriched by immunoprecipitation of Pol II using antibodies against specific CTD phosphorylation marks and identified via mass spectrometry. Marks cluster separately, with close grouping among experimental replicates and larger variance between different phosphoisoforms, indicating Pol II enriched for each mark associates with a distinct interactome(6).

The repeated nature of the CTD, and the possibility of phosphorylation on five of the seven residues in the consensus heptad, give rise to a potentially vast array of phospho-species, which could in theory underlie an incredibly complex
Two recent studies used a series of CTD mutants and mass spectrometry to map post translational modifications across the CTD in yeast and human cells, and the data suggest a more simple CTD code than theoretically possible(17, 18).

In both model systems, the data show that most CTD heptaheptads are phosphorylated only once, suggesting the removal of early S$_5$ phosphorylation must precede or quickly follow the addition of S$_2$ phosphorylation(17, 18). These results, combined with the recent work from Harlen and colleagues, paint a picture of Pol II switching out largely non-overlapping sets of phosphorylation marks and corresponding interacting partners across the length of the CTD during gene transcription. The nature of this dynamic and precise regulation begs the question: what regulates the regulation? If the CTD is disordered, and present in the nucleus with hundreds of proteins that make up the readers and writers of the CTD code, what might protect against spurious modification and recruitment of factors at the inappropriate time and place?

Existing data pertaining to the pS$_5$ to pS$_2$ transition suggests the potential for an additional structural layer of CTD code modulation. Ssu72, the CTD phosphatase responsible for erasing pS$_5$, has higher enzymatic activity for those pS$_5$ heptads with proline 6 (P$_6$) of the repeat in the cis conformation(19, 20). Additionally, among the readers of the pS$_5$ code is the peptidylprolyl isomerase (PPIase) Pin1, which may stimulate the pS$_2$ to pS$_5$ transition catalyzed by Ssu72 via isomerization of the peptide bond at P$_6$(19) (21).
The suggestion that the pS$_5$ to pS$_2$ transition and interactome exchange may be modulated in part by cis-trans structural switches raises questions about CTD organization on other length scales, such as whether or not the CTD is compact or fully extended, and whether or not it is structurally homogenous across its length. More specifically if structure is “read” by enzymes at the level of cis-trans isomerization, might other structural organizing exist within the CTD, such as variation in the structures adopted by different regions of the CTD that also contributes to protein binding? These questions are amplified in the context of the more diverse CTD sequences found in higher eukaryotes, and by a growing appreciation for the structural complexity of IDPs. Here, I address questions about the structural organization of the CTD in the context of an emerging understanding of the structure, function and evolution of intrinsically disordered proteins.

1-1.2 What are IDPs, and what do they do?

A structure-dictates-function paradigm long dominated understanding of protein folding and function. In this model, the amino acid sequence of a protein encodes a fold that in turn enables function. The conformational space occupied by natively folded proteins can be conceptualized by strongly funneled energy landscapes with one or few energy minima (Fig 1-2 left panel). A protein conformation can be thought of as the structural states accessible under a given set of conditions (temperature, ionic strength, buffer composition, etc).
Figure 1-2 Energy landscapes of native and intrinsically disordered proteins. (a) A representative energy diagram of a globular protein that adopts a stable fold, present in the blue energy well. This state minimizes free energy and conformational entropy(22). (b) A representative energy diagram of an intrinsically disordered protein that lacks a deeply funneled energy landscape and instead has a rough topology(22). Different protein conformations that contribute to the overall ensemble structure of the IDP are depicted in shallow energy wells between which there are small barriers to interconversion.
More recently, understanding of protein structure and function has expanded to accommodate the large segment of the eukaryotic proteome which consists of proteins not thought to adopt a folded conformation, or that contain regions that fail to adopt a fold (23). In contrast to natively folded proteins, IDPs have much rougher energy topologies. Instead of an energy funnel giving rise to a narrow range of conformers, the conformational landscape of IDPs can be described by a rough energy topology, with many shallow energy minima. Thus, the conformation of IDPs is actually an ensemble of conformers with low energetic barriers to interconversion. As such, a single IDP may rapidly interconvert between conformations, or wells in the rough energy landscape, and a population of molecules will be comprised of subpopulations dominated by the most energetically favorable conformers. This conformational plasticity enables IDPs to evolve unique functional properties, such as the ability to bind to a diverse array of partners.

1-1.3 The evolution of the CTD

The CTD is thought to have evolved from tandemly repeated heptads, consistent with data suggesting the functional unit of the CTD spans adjacent repeats (24, 25). The overall array of repeats has diverged throughout evolution, resulting in CTDs differing in length and sequence complexity among phyla with a trend toward longer CTDs and a higher prevalence of non-consensus heptads as a function of developmental complexity (24, 26, 27). Recent genetic analysis
suggests that the CTD evolved in a common ancestor containing a di-heptapeptide repeat, and this sequence diverged and re-emerged throughout evolution, sometimes resulting in organisms with highly degenerate CTDs(24). While much of what is known about the function and essential units of the CTD comes from studies in yeast, it is important to note that yeast is unique in the extent to which its CTD heptads directly match the consensus motif. In contrast, the CTDs of higher eukaryotes are comprised predominantly of non-consensus repeats that resemble the consensus, but differ at one or more positions(24, 26, 27). The evolutionary plasticity of the CTD exhibited across phyla must be balanced by the continued need to recruit conserved and essential proteins to Pol II. This balance, and the multiple binding modalities that the CTD can achieve is perhaps best examined in the context of disordered protein sequences collectively known as Short Linear Motifs, or SLiMs.

1-1.4 The CTD is a series of SLiMs

SLiMs are short, 3-10 amino acid motifs found within IDPs that function as binding modules and sites of regulatory post-translational modifications(28). The concept of SLiM binding by globular proteins describes a process that differs categorically from interactions between globular proteins. Common examples of globular binding domains with SLiM ligands include SH3 domains which recognize PxxP motifs, and SH2 domains which bind phospho-YxxI/V motifs, where x denotes any amino acid(28). Generally, interactions between folded proteins typically involve 50-200 amino acids across discontinuous surfaces (i.e.
residues may be close in space but distant in sequence) and bury in excess of 1000Å² of surface area, which can result in stable interactions with dissociation constants ($K_d$) in the nanomolar to picomolar range. In contrast, the recognition of SLiMs by folded proteins typically involves short stretches of 3-10 amino acids, ~500Å² of buried surface area and dynamic binding with $K_d$ values in the micromolar range(28). Such $K_d$ values could be important in the dynamic exchange of factors on the CTD that occurs on the timescale of gene transcription, where distinct interactomes are exchanged as a function of the CTD modification state and the position of Pol II on the gene(6, 29, 30).

Owing to the increase in solvent accessible surface area inherent to IDPs vs. globular domains, the buried surface area at the interface between globular proteins or globular proteins and SLiMs does not scale linearly with the number of residues involved(28, 31, 32). Thus, SLiMs are evolutionarily useful recognition motifs as they provide recognition surfaces that are large relative to the number of amino acids in the motif and their short length means multiple SLiMs can be combined to achieve high valency without a large increase in protein mass, as is the case in the CTD. Thus, the CTD can be thought of as a repeated series of SLiMs, resulting from a linear array of heptad repeats, capable of being read in different registers and with different post-translational modification patterns(3, 33). These reading frames and phosphorylation patterns give rise to a diverse array of binding motifs, some preferentially recognized by particular factors(3). Additionally, different factors could induce different SLiM structures upon binding. Indeed, multiple different heptad conformations, biased
by phosphorylation, have been observed for CTD peptides bound to CTD interacting factors and this structural plasticity enables the same heptad sequence to recruit multiple factors (34-38).

I will briefly discuss two examples of different modes of interaction between SLiMs within the CTD and globular domains. X-ray crystallography, NMR and biochemical data have characterized the interaction between CTD peptides and the CTD interacting domain (CID) of the transcription termination factor Pcf11 (39, 40). Residues P6S7Y1pS2pT4pS5P6S7 are ordered and visible in Pcf11 CID:CTD structure with a β-turn motif adopted by the central pS2pT4pS5 residues (39). Subsequent NMR studies revealed that additional contacts between the Pcf11 CID and CTD occur, but the overall binding affinity remains in the millimolar range (40). It is also notable that while phosphorylation of S2 stabilizes the interaction, the phosphate is not directly recognized by any residues of the CID and may contribute to binding by enabling an additional intrastrand hydrogen bond stabilizing the β-turn (39, 40). The interaction between the Pcf11 CID with a CTD peptide contrasts with the WW domain of the proline isomerase Pin1 and the CTD. In the case of Pin1 binding with micromolar affinity is achieved with only a short doubly-phosphorylated peptide encompassing Y1pS2pT4pS5 (34). As was the case for the Pcf11 CID, the phosphate on S2 contributes to increased binding affinity but is not directly recognized by the WW domain of Pin1, however, the phosphate on S5 is (34). The WW structures a smaller, 7 residue SLiM that adopts an extended structure that differs from the tight turn motif central to the larger nine residue SLiM bound by Pcf11 (34, 39).
The differences in the structures adopted by the CTD when bound to the Pcf11 CID and the Pin1 WW domain highlight the structural plasticity of the CTD and exemplifies how the CTD repeats can be recognized by multiple factors (Fig. 1-3).

The various structures consensus heptads adopt upon binding raises questions about how non-consensus CTD heptads could bias these induced structures. Additionally, degenerate heptads could alter the array of structures sampled and presented to factors for binding and post-translational modification relative to consensus heptads in ways that could serve to localize interactions or modifications to specific regions of the CTD.

**Figure 1-3 CTD structures in complex with globular domains.**
Cartoon model of the backbone of CTD peptides as they are structured in complex with globular domains. Residues in the structure are listed and color coded below each structure, phosphoserines are depicted in red, serines in green. (a) The structure of a CTD peptide in complex with the CTD interacting domain of Pcf11. S2-S5 forms a β-turn. (b) The CTD adopts a more extended structure without a β-turn when bound to the WW domain of Pin1. Image generated from PDB files 1SZA and 1F8A using PyMol (34, 39, 41).
1-1.5 SLiMs and multivalency enable evolutionary experimentation

Surprisingly, despite the short length of SLiMs, often only particular residues at specific positions within the SLiM are essential for binding by a partner(28). This enables the evolution of degenerate SLiM variants containing permitted substitutions to continue to support core functionality while also enabling the evolution of additional functions(28, 42, 43). The multivalency of the CTD may thus support evolvability provided sufficient motifs remain to support essential functions. These concepts are supported by classical experiments from Corden and colleagues that identified Y₁, and SP motifs at positions 3-4 and 5-6 within the heptad as essential, but mutations at essential positions could be tolerated in a limited number of heptads(9). Additionally, the CTD is required for survival, but can be truncated to 8 heptads in S. cerevisae and S. pombe and to 29 heptads in mammals(9, 44, 45). Additionally, the strict sequence requirements at some but not all SLiM positions for a given interaction could enable permissive mutations to arise that are both capable of retaining existing binding functionally while also enabling the evolution of new interactions or post-translational modifications. Thus, the multivalent CTD could be a hotspot for the evolution of new functionality provided a minimum number of contiguous repeats remain to support core functionality(46). In theory, co-evolution of CTD interacting factors along with degenerate heptads could recruit new functions to the CTD or result in the targeting or blocking of binding to specific heptads allowing for the spatial orientation of multi-protein complexes (43). Consistent with this idea are the observations that non-consensus repeats do not support full CTD function. The
S. cerevisae CTD, comprised nearly entirely of consensus motifs, can be replaced with the mammalian CTD, the first half of which closely matches that of yeast and possesses 17 consensus heptads, exceeding the number required for yeast viability (47) (Fig. 1-4). Conversely, the more degenerate Drosophila CTD, which maintains a heptad repeat architecture but containing only 2 of 42 repeats exactly matching the consensus, does not support viability in yeast, potentially due to an insufficient number of heptads capable of being recognized by yeast CTD interacting factors (9, 47) (Fig. 1-4). Further, experiments in human cells show that fully consensus CTDs support cell growth, but CTD mutants lacking consensus repeats do not (48), again suggesting unique roles for consensus motifs. Despite significant differences between more distantly related species, the non consensus motifs are themselves highly conserved among related species in a phylum, suggesting selective pressure to maintain the non-consensus motifs (24, 49, 50). The extent to which the CTD may target factors to specific non-consensus motifs, and possible determinants of such specificity are not understood.
1-1.6 CTD organization in metazoans

Multiple observations suggest there may be functional differences encoded in different regions of the CTD in higher eukaryotes, including mammals and Drosophila, whose CTDs are composed predominantly of non-consensus heptads (Fig. 1-4). Both share an acidic tip appended to the final repeat, which in mammals has been shown to be important for the stability of the CTD in vivo and is constitutively phosphorylated by casein kinase II (48, 51) (Fig 1-4). Curiously, the sequence of the acidic tips differs between mammals and flies, yet the fly sequence can function to stabilize the mammalian Rpb1 in vivo, suggesting charge and not sequence of the tip provides function (48). Like the final repeat, the repeats at the very beginning of the mammalian CTD may also provide unique functionality as deletion of repeats 1-3 in human cells also leads to degradation of Rpb1 and this degradation phenotype is synergistic with deletion of the acidic tip (48). In another similarity to the acidic C-terminal tip, the proximal repeats can be replaced with a variety of amino acid sequences that impart stability to Rpb1, suggesting an important feature of this region of the CTD cannot be reduced to sequence alone. Instead, these proximal repeats may serve a structural function (48).

In contrast to the ends of the CTD, the remaining internal heptads have been mutated and truncated extensively in a variety of studies with limited phenotypes or effects on cell viability (18, 44, 48). However, in the more degenerate CTD of Drosophila there exists evidence for region specific functions within the central region of the CTD. First, the central region of the Drosophila
CTD, which contains the only two consensus repeats, preferentially binds the transcription termination factor Pcf11(52). Additionally, recent genetic studies used RNAi and ectopic expression of Pol II mutants to knock down the endogenous pool of Rpb1 to lethal level and replace it with an RNAi resistant Rpb1 transgene harboring mutations in Pol II(53). In this study, 4 internal deletions in the CTD were tested, only one of which failed to rescue the lethality caused by RNAi. The lethal mutant harbored a central deletion that removed the two consensus repeats preferentially bound by Pcf11 in vitro (52, 53). Together, these results argue that the central region of the CTD is functionally distinct. One explanation is that some consensus heptads are essential to support viability, perhaps because they specifically recruit an essential activity to the CTD. Support for this notion comes from data in mammalian cells that showed a fully-consensus CTD was functional provided it contained the acidic tip, but a fully-non consensus CTD had growth phenotypes despite being equivalent in length to a mutant containing consensus heptads(48). One hypothesis is that consensus heptad sequences have strong affinity for one or more essential CTD interactions, and that impairing or abrogating those interactions results in growth phenotypes or lethality.
Figure 1-4 A comparison of CTD sequences from *S. cerevisae*, *D. melanogaster* and *H. sapiens*. The consensus heptads are depicted in boldface. In contrast to the yeast CTD, the fly and human CTDs are comprised primarily of non-consensus heptads. The CTD tip is highlighted in yellow. The tip is acidic in flies and mammals, but not yeast.
1-1.7 Sequence vs. structure of CTD regions

Might there be an alternative explanation for the contribution of non-consensus repeats to cell growth and proliferation beyond sequence recognition by proteins? The CTD is an IDP. As such, it is reasonable to consider observations, models and methodologies from the IDP field pertaining to the nature of IDP evolution, structural ensembles, and allosteric regulation as potentially applicable to our efforts to understand how a highly repetitive and disordered protein imposes order on co-transcriptional processes. The growing IDP literature provides many examples of proteins that harbor local structural elements or variation in compaction across their length despite being disordered on the global scale(54, 55). Some IDPs that are also subject to multiple post translational modifications respond differently at the local level despite relatively modest global changes in structure or compaction(56). In the following sections, I will briefly discuss three examples of structural variation and differential responses to post-translational modifications within IDPs using the proteins NS5A, α-synuclein, and Ash1 as examples. Together, these studies on IDPs provide a theoretical framework to view the global and local structure of the CTD. The CTD could be sampling a multitude of structures, each described by some gross structural measurement, such as radius of gyration ($R_g$). The $R_g$ of the ensemble is thus the population weighted average of the $R_g$ of each individual conformer. The rough free energy landscape in figure 1-2 provides a visual conceptualization of this phenomenon, with each shallow well representing a conformer. Nature can exploit this structural plasticity in a number of ways, such
as enabling induced fit with interacting factors that bind most avidly to a particular conformer. Evolving non-specific heptads that respond uniquely to binding or post-translational modification could bias interactions to particular regions of the CTD. Alternatively, and not mutually exclusively, certain heptads in isolation or combination could give rise to conformers with local structural variation that render some regions more binding competent than others for a particular interaction. In this case, some structure encoded in the CTD exists prior to a factor interaction, preferentially exposing particular regions of the CTD to factors, or biasing interactions evolved to recognize a pre-formed, albeit transient, structure. In this model, the structure of the CTD can be conceptualized as a flexible ensemble, interconverting between conformers, but with particular conformers over-represented in the ensemble relative to others, depicted as deeper, more blue wells in the otherwise rough energy landscape (Fig. 1-2). Here, individual wells of the CTD energy landscape could contain conformers with variation in local structure, or compaction, across the length of the CTD.

One recently characterized example of an IDP harboring pre-formed structural elements is the hepatitis C viral protein NS5A. Like the CTD, NS5A is a phosphoprotein that utilizes phosphorylation to modulate and expand its functionality in essential ways(57, 58). The disordered regions of NS5A bind multiple globular proteins and are substrates for multiple kinases. Despite being intrinsically disordered, evidence of preformed regions of structure and compaction are observed in NS5A, and these regions are recognized by a globular SH3 domain and by Casein Kinase II (CK2)(59). The regions of
preformed structure are allosterically modulated by phosphorylation, suggesting transient preformed substructures within an IDP can be exploited for recognition\(^{(55, 59)}\). The IDP domain of NS5A is an example of intrinsic disorder existing on a structural continuum involving local differences in compaction and extension within an IDP\(^{(54, 55)}\). Examples like NS5A raise interesting possibilities for the CTDs in metazoans where non-consensus heptad motifs could impart local structural heterogeneity in a largely homogenous and disordered array of repeats.

### 1-1.8 IDP allostery and the CTD

The textbook view of allosteric regulation involves structural changes to a protein fold by virtue of binding to a ligand, or resulting from post translational modification\(^{(60)}\). How might proteins with structures defined by an ensemble rather than a stable fold be allosterically modulated? Recent examples show that just as protein structure exists on a continuum between order and disorder, so too does allostery. One example is the protein \(\alpha\)-synuclein, which undergoes nitration of tyrosine residues in the disordered C-terminal domain that results in structural changes that alter lipid binding by the structured N-terminal domain\(^{(61)}\). Alterations in binding are thought to occur via changes in the ensemble of structures \(\alpha\)-synuclein adopts after nitration, specifically a de-population of conformers in the ensemble that are binding competent\(^{(61)}\). Here, \(\alpha\)-synuclein provides an example of structural perturbations in an IDP ensemble manifested in functional ways at the opposite termini of the protein\(^{(61)}\). Allostery
can also result in local changes that alter protein structure in equally opposing ways resulting in little to no change in the global dimensions of a protein. An interesting recently described example is the IDP domain of the yeast transcriptional regulator Ash1, which like the CTD is multiply phosphorylated on serine and threonine residues with this phosphorylation serving regulatory functions that in the case of Ash1 leads to binding by the ubiquitin ligase Cdc4 and subsequent degradation by the proteosome(62). In contrast to predictions based on sequence composition and charge, Ash1 does not undergo significant extension in response to being phosphorylated at 10 sites across the IDP(56). Based on molecular modeling, the authors propose that proline residues in the SLiM phosphosites buffer charge dependent extension. Moreover local extension and compaction at different places on the IDP are compensatory, resulting in little change in global dimensions observed experimentally using small angle X-ray scattering (SAXS)(56). Ash1 provides an example of allostery on the local level, with differing local responses to phosphorylation. In this way, post-translational modification could bias the ensemble by enriching or depleting it of conformers competent for a particular interaction. As these examples illustrate, in much the same way folded proteins can be allosterically modulated, so to can intrinsically disordered proteins(63).

1-1-9 The structure of the CTD is poorly understood

Almost nothing is known about the structure of the CTD. The yeast CTD, if fully extended, would be 650Å in length(64) (Fig 1-5). By comparison, a loose β-
spiral model was proposed based on NMR data collected on CTD peptides that would give rise to a CTD 280Å in length (Fig 1-5)(65). However, these turns were rare in solution, so the likelihood of the CTD adopting these structures across its length is exceedingly slim(65). A more compact β-turn model is based on applying turn structures found in CTD peptides when bound to factors to the length of the CTD(13) (Fig 1-5). Compact random coil structures are based on the space occupied by the CTD in crystallographic studies(4, 66) (Fig 1-5). These models are each very different from one another, and would thus present the CTD repeats to binding factors very differently.

The aforementioned examples of α-synuclein, NS5 and Ash1 provide distinct examples of structural variation within an IDP. Understanding the structure of the Pol II CTD, and thus how it presents itself to other factors, could contribute to our understanding of transcriptional regulation and provide a framework to generate hypotheses about how the CTD organizes a complex interactome. The study of SLiM evolution in IDPs hints at ways novel, lineage specific functionality could emerge in multivalent sequences. Structural models of the CTD vary considerably, and are assembled from data collected on short CTD fragments, or under conditions that could potentially bias the structure of the CTD in unnatural ways (Fig 1-5). To date, these models have failed to consider the implications of non-conserved heptads that diverge from the YSPTSPS consensus. Such heptads comprise the bulk of metazoan CTDs, and recent comparative genetic analysis suggest yeast is unique in the homogeneity of the CTD(24, 67). In this thesis, I explore the global and local structure of the CTD of
Drosophila melanogaster, employing tools and concepts from the IDP field to address both new and longstanding hypotheses in the transcription field.

From Meinhart and Cramer, Genes and Dev. 2005 19(12):1401-1415

**Figure 1-5 Existing models of CTD structure.** Various models of CTD structure have been proposed; here they are applied to the yeast CTD. The CTD interacting domain (CID) of the transcription termination factor Pcf11, is shown to scale for comparison. Each model would present the CTD to factors in different ways.

**1-1.10 Questions to be addressed**

There are five unanswered questions about the CTD I seek to address in this thesis. First, what effect do non-consensus heptads have on the global structure of the CTD? This question is biologically important because both CTD length and sequence complexity generally scale with evolutionary complexity. Understanding if these evolutionary changes resulted in structural changes at the global level is important, especially if I seek to relate findings generated on
metazoan CTDs to the more simple but better understood CTD of *S. cerevisiae*.

To best leverage the unique elements of *Drosophila* as a model system for structure-function analysis it is important to understand how the *Drosophila* CTD structure relates to other systems where the bulk of genetic and genomic analyses have taken place. Second, how might the CTD organize factors across its length? The work described herein uncovers previously unappreciated structural organization present in the *Drosophila* CTD that I propose could serve to organize CTD:factor interactions. I also describe aspects of this local structural organization that are shared with the human CTD. Third, given the fact the CTD is thought to be an extended structure and that it exists in the nucleoplasm with hundreds of interacting factors, how does the CTD protect itself from interactions and post translational modifications occurring at the inappropriate time and place? I describe structural changes to the CTD and propose models of ensemble allostery as a means of regulating CTD:factor interactions that transcend the phospho-CTD code as it has been extensively described. Fourth, CTD sequences are conserved within phyla, but differ across them, and so it has been suggested that lineage specific CTD sequences have co-evolved to interact preferentially with specific factors. We directly test this hypothesis, and determine that a precise array of heptad motifs is dispensable for CTD function in *Drosophila*. I also show that there are structural similarities between the *Drosophila* and human CTD at the global and local levels, enabling the hypothesis that conserved structural organization, not a precise array of heptad motifs, is important for CTD function. Finally, the CTD heptad repeats are
connected to the catalytic core of Pol II via a linker domain that is important for Pol II function but has been largely ignored in the literature. No data exist pertaining to the impact the linker region has on the overall structure of the CTD. Here, I compare the structure of the linker to that of the CTD, including the response to phosphorylation, and determine that the linker is a distinct structural element of the CTD that may serve as a hinge to maximize the radius of access for CTD bound enzymes.
CHAPTER 2

BIOPHYSICAL CHARACTERIZATION OF THE RNA POLYMERASE II C-TERMINAL DOMAIN

The data in this chapter are published in Portz, B. et al. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nature. Communications. 2017 8, 15231

2-1.1 Introduction

Despite its role as a regulatory nexus, little is known about the structure of the CTD, and most studies have focused on the structure of short CTD peptides. Such peptides have been shown to adopt β-turn structures when bound to CTD interacting proteins, and β-spiral models of the CTD have been proposed that extrapolate β-turns across the entire length of the CTD(39). However, nuclear magnetic resonance (NMR) and circular dichroism studies of CTD peptides suggest such turns are rare among the ensemble of CTD structures, and only populated to a high degree in turn-promoting solvents at low pH or in the context of circularized peptides(65, 68-70). Phosphorylation enables formation of additional hydrogen bonds that have been shown to contribute to turn structures in CTD peptides bound to factors, but phosphorylation fails to increase the β-turn propensity of unbound CTD peptides in solution(69, 71). Together, these data argue against models iterating turns across the full CTD in either the apo or phospho state. Such models are entropically unlikely, as they assume the CTD simultaneously adopts multiple low-probability turns. Additionally, β-spiral models
fail to account for local structural variation potentially imparted by those heptads deviating from the consensus sequence, and such repeats comprise the majority of the CTD in developmentally complex organisms (24).

Existing structural information describing the full-length CTD is limited. The CTD, likely due to its flexibility, is absent from X-ray crystal structures of Pol II, but one study showed it may share space with the CTDs of adjacent Pol II molecules in the crystal, a space too small to accommodate extended structures (4). This lead to a model that was sufficiently compact to fit into the space provided in the crystal, yet lacked the order associated with folded, globular structures (4). The CTD is also absent from cryoEM derived models of elongating Pol II, again likely due to its flexibility (72). Additional evidence for a compact CTD was obtained from cryoEM models of a GST-CTD fusion protein bound to the middle module of the multi-subunit transcriptional regulator known as the Mediator complex (73). However, it is unknown to what extent binding to Mediator and possible CTD:CTD interactions resulting from GST-CTD dimers contribute to the observed compaction (73). Earlier studies of negatively stained 2-D crystals of Pol II and Pol II mutants lacking the CTD revealed density differences attributed to a flexible CTD and linker occupying a space measuring ~100Å in diameter (66). The extent to which fixation on a solid surface alters the CTD ensemble in the 2D crystal studies is unknown, and like other X-ray crystallography studies, the proximity of CTDs from adjacent Pol II molecules in the 2D crystals could potentially alter the CTD ensemble through CTD:CTD interactions. One study by Corden and Zhang examined the structure of the
mammalian CTD in solution and found the hydrodynamic radius of the CTD extends as a function of phosphorylation (74). Consistent with extension was the observation by Laybourn and Dahmus that the CTD is more proteolytically labile in the phosphorylated state, suggesting conformations more accessible to CTD:protein interaction (75). However, it is unclear from this work if the CTD was rendered more accessible by CTD phosphorylation alone, or in concert with possible interactions with proteins present in HeLa extract, which was a component of the cleavage reactions of the phosphorylated Pol II, but not added to the unphosphorylated Pol II cleavage reactions (75). Nevertheless, evidence supports an unphosphorylated CTD existing as a compact structural ensemble, capable of extending as a function of phosphorylation. Such an ensemble is consistent with emerging structural understanding of intrinsically disordered proteins (IDPs), which can be more compact than chemically denatured proteins of the same size, and can be structurally heterogeneous, possessing transient structural features (54, 55, 77-79).

The work described in this chapter was aimed at understanding the global structure of the monomeric CTD in solution, the relationship between this structure and the prevalence of non-consensus repeats, and the effects of characterized patterns of phosphorylation on the global structure of the CTD. I find that, despite sequence differences and variations in the amount of non-consensus heptads, the yeast, fly and human CTDs exist on a structural continuum, adopting similar structures that differ as a function of their molecular weights. This relationship establishes the fly as an ideal model for studies on the
CTD as it is developmentally complex, genetically tractable, and has a native sequence compatible with phosphorylation site mapping and structural studies using NMR(53). This motivated us to characterize CTD phosphorylation by a CTD kinase known as the Positive Transcription Elongation Factor (P-TEFb), which I can then relate to changes in the overall structure of the CTD. In doing so, I clarify existing discrepancies about the specificity of P-TEFb for the full-length CTD.

Prior work mapping the sites of phosphorylation on the CTD in yeast and human cells required extensive mutagenesis of the CTD(17, 18). The sequence repetition of all or a portion of the CTDs of these organisms mandated mutagenesis in order to distinguish fragments from different regions of the CTD from one another(17, 18). The extent to which this mutations alter the substrate recognition by CTD modifying enzymes is unknown. The Drosophila CTD contains only two identical repeats in the YSPTSPS register such that digestion with the protease chymotrypsin can generate unique fragments compatible with identifying the sites of phosphate incorporation using mass spectrometry(80).

Recently, Mayfield and colleagues exploited this feature of the Drosophila CTD to map phosphate incorporation on the fly CTD after incubation with recombinant mouse p42 MAP kinase(80). Interestingly, several rules emerged about the specificity of this kinase for the varied heptads present in the CTD. First, phosphorylation occurred on heptads irrespective of what amino acid was present in the 4th and 7th position of the repeat(80). This is interesting, as these positions are varied in the fly CTD and non-essential in the yeast CTD(9).
such these positions may represent permissive positions in the heptad SLiM for evolutionary experimentation where factor binding or structural parameters of the CTD could be tuned without abrogating recognition by some kinases. Second, phosphorylation occurred both at positions analogous to $S_2$ and $S_5$, with a preference for $S_5$ phosphorylation provided that an SP motif occurs in this position in a given heptad(80). Finally, consistent with in vivo observations from both yeast and human cells, no heptads were phosphorylated more than once(17, 18, 80). This work established the ability to localize the sites of phosphate incorporation to specific heptads across much of CTD of the important model organism D. melanogaster, without mutagenesis. However, the impact of these results is limited by the use of a non-native kinase:substrate pair (mouse p42 MAP kinase:fly CTD)(81). Furthermore, phosphorylation by p42 MAP kinase was very incomplete, with ~50% of the GST-CTD proteins used in these studies remaining unphosphorylated(81). In collaboration with the authors of this work, we sought to generate highly purified and more completely phosphorylated Drosophila CTD phosphorylated by a physiologically relevant Drosophila kinase, P-TEFb. In addition to exploring the rules for substrate preference by this enzyme, this system would enable us to relate structural differences observed between the apo and phospho CTD in biophysical and biochemical experiments to a characterized pattern of phosphate incorporation.
2-2 Results

2-2.1 The CTDs of yeast, fly and human are structural homologs

CTD length and sequence complexity differ among organisms, with non-consensus heptads predominating in the longer CTDs of metazoans (24, 26). Nevertheless, aligning blocks of seven amino acids beginning with YSP for the CTDs of *D. melanogaster* and *H. sapiens* using Weblogo returns a similar consensus motif with most deviation occurring in the seventh position (82) (Fig. 2-1).

![Weblogo comparison of CTD heptads](image)

**Figure 2-1 Consensus motif comparison of H. sapiens and D. melanogaster CTD heptads.** A comparison of the consensus motifs from the *H. sapiens* and *D. melanogaster* CTDs generated using WebLogo (82). Despite only 2 heptads from the Drosophila CTD precisely matching the YSPTSPS motif compared with 26 such repeats in human, both motifs are similar, and both are most degenerate at the 7th position, shown in yeast to be non-essential (9, 26).
I sought to determine if the presence of non-consensus repeats impacts the global structure of the CTD. Because the CTD naturally exists as a terminal extension emanating from the globular and acidic Pol II, and to aid in purification and solubility, I expressed and purified the CTDs of *S. cerevisae*, *D. melanogaster*, and *H. sapiens* as maltose binding protein (MBP) fusions and compared their mobility using native-PAGE. The relationship between CTD length and mobility for the three species suggests gross structural similarity despite differences in the relative proportion of non-consensus heptads (Fig. 2-2). This similarity places the fly CTD on a structural continuum with yeast, the source of much of our knowledge of the CTD, and humans, which shares a longer and more complex CTD with similarities to the fly.
Figure 2-2 Native gel electrophoresis MBP-CTD fusion proteins. 6% native page gel comparing the mobility of the *S. cerevisiae*, *D. melanogaster* and *H. sapiens* MBP-CTD fusions, stained with coomassie. The acidic MBP imparts a negative charge, such that each fusion has a similar isoelectric point (pI). Mobility scales with molecular weight (MW), suggesting gross structural homology between each CTD despite sequence differences. pI and MW calculated using ExPASy ProtParam(83).
To quantify the relationship between CTD length and overall structure more precisely and independently of pl, I compared the three CTDs using size exclusion chromatography (SEC) determining the Stokes radii ($R_s$) using a standard curve derived from natively folded proteins (Fig. 2-3a), and found a linear relationship between $R_s$ and molecular weight (MW) for the CTD fusions (84, 85)(Fig 2-3b).

**Figure 2-3 Size exclusion chromatography of MBP-CTD fusion proteins.** (a) Analytical gel filtration standard curve used to determine CTD Stokes Radii ($R_s$). Standards used, in order of increasing $R_s$, were RNAseA, ovalbumin, aldolase, apoferritin, and thyroglobulin. Y-axis values were the same for three experimental replicates to two significant digits. (b) Comparison of MW vs $R_s$ for the protein standards (squares) and MBP-CTD fusions (triangles), including MBP alone. The curves converge near MBP (leftmost triangle), but have very different slopes, with smaller changes in MW yielding large changes $R_s$ for the CTD fusions relative to the globular standards, suggesting an extended CTD emanating from the globular MBP. MBP fusion values are the average of two replicates which are plotted as individual values in Figure 2-4 to show variance.
Just as the migration of globular proteins on the SEC column are linearly correlated, so too are the various MBP-CTD fusions (Fig 2-3b). This suggests that the CTDs are structurally similar to one another, despite sequence differences. A direct comparison of the slopes of the globular protein standard curve, and the MBP-CTD curve, yields insight into the structure of the CTDs (Fig. 2-3b). The slope of the CTD curve is much flatter, indicating that small increases in CTD length (MW) cause more significant increases in size ($R_s$) compared to natively folded proteins (Fig. 2-3b b). This difference suggests the CTD is much more extended than a natively folded protein of similar MW. In fact, the MBP-CTD fusions have molecular weights in the range of 58-82kD, but have a size more comparable to the natively folded standard protein, aldolase, which has a MW of 158kD (Fig. 2-3b).

At some minimum CTD length, the MBP portion of the fusion could dominate the analysis, and it could be expected that this linear relationship would break down, confounding interpretation of the results. Thus, I included in this analysis a truncated version of the fly CTD, dubbed W81, shown to have functional defects in vivo(86, 87). Despite removing ~50% of the CTD, generating a CTD shorter than that of yeast, this fusion also remained in the linear range of the CTD curve (Fig. 2-3b). The relationship between $R_s$ and MW for all of the CTD fusions suggests that they are structural homologs despite sequence differences, sharing an extended conformation compared to natively folded proteins of similar MW (Fig. 2-3b). The correlation between $R_s$ and MW
for the CTDs creates a CTD standard curve, with the *Drosophila* CTD at the center, which can be used to interrogate structural changes resulting from phosphorylation.

**2-2.2 Phosphorylation by P-TEFb extends the CTD**

Phosphorylation of the *D.mel* CTD by the CTD kinase P-TEFb lead to an apparent ~26% increase in the Stokes radius of the CTD, causing the phospho-CTD to fall off the CTD standard curve (Fig 2-4). This observation is notable for three reasons. First, it relates the *Drosophila* CTD to its mammalian homolog in the phosphorylated state, as the mammalian CTD was also observed to extend in a similar analysis, leading the authors to later propose that stiffening of the CTD could contribute to the observed extension(74, 88). That the mammalian and fly CTDs are structural homologs globally, and both respond similarly to phosphorylation, validates the fly CTD as a useful structural model for understanding the more complex CTD in metazoans. This relationship allows us to leverage the developmental complexity and genetic tractability of *Drosophila* to explore the CTD structure:function relationship. Moreover, the native *Drosophila* CTD sequence is sufficiently complex to enable NMR and mass spectrometry analyses that are inaccessible to the yeast and mammalian CTDs without extensive mutagenesis and its associated caveats(17, 18). Lastly, that the *Drosophila* CTD extends upon phosphorylation provides further insight into its structure. Simply, extension suggests a structure that must have been somewhat compact in the first place. Protein folding results from the polypeptide chain's
intra-chain attractions supplanting those between the protein and solvent. Unfolded random coil structures occur in so called “good” solvents where protein:solvent interactions dominate (89, 90). SEC results clearly suggest the CTD is more extended than natively folded globular proteins of the same mass, but phosphorylation driven extension provides evidence that the physiologically relevant buffer conditions employed in my experiments (150mM NaCl, 50mM HEPES pH 7.5,10% glycerol) are insufficient to generate a random coil structure, allowing for the possibility of some structural organization within the CTD.

Figure 2-4 The CTD extends after P-TEFb phosphorylation as measured by size exclusion chromatography. $R_s$ vs. molecular weight (MW) derived from size exclusion chromatography analysis of MBP-CTD fusions. Two replicates of each protein are plotted (several points appear as one because of overlapping values). $R_s$ is linearly related to MW, suggesting gross structural homology. Phosphorylation of the D. melanogaster CTD by P-TEFb increases $R_s$, causing the CTD to deviate from the line (red triangles). Two replicates for each protein are plotted, some points appear as one due to overlapping values.
Notably, the MBP-\textit{D. mel}CTD elutes from the size exclusion chromatography column as a distinct peak in a volume (peak width) similar to those of the globular proteins used as standards for determining $R_s$ (\textbf{Fig. 2-5}). This chromatographic behavior persists after phosphorylation (\textbf{Fig. 2-5}). If the CTD ensemble included structures ranging from very extended to compact, the peaks would be expected to be broader than those of globular proteins.

\textbf{Figure 2-5} The apo and phospho MBP-\textit{D. mel}CTD fusions elute as sharp peaks from a size exclusion column. (a) The globular standard proteins elute as sharp peaks from the Superose 6 column. (b and c) The apo and phospho MBP-\textit{D. mel}CTD proteins also elute as sharp peaks similar in volume to the
globular standard proteins. The arrows point at the protein beaks. The large peak after the MBP-CTD in (c) is ATP from the kinase reaction.

2-2.3 Mass Spectrometry characterization of P-TEFb phosphorylated CTD

I sought to characterize the extent and location of phosphate incorporation by P-TEFb under my in vitro conditions in order to relate the extension I observed to a characterized pattern of phosphorylation. To this end, I employed a similar expression and purification strategy to that used previously for SEC experiments. In brief, the MBP-\textit{D.mel}CTD fusion was expressed overnight at 11°C and purified using affinity resins selecting for opposite termini of the protein (Fig 2-6 a). This scheme selects for full-length MBP-\textit{D.mel}CTD (Fig 2-6b lane 1). The CTD was then subject to phosphorylation by recombinant P-TEFb purified from baculovirus infected Sf9 cells (see materials and methods). Motivated by a desire to achieve more complete phosphorylation, my reaction conditions differed from those of a previously published study using p42 MAP kinase in that I used a longer reaction time (1hr vs. ~2days, respectively) and buffer conditions that better replicate cellular ionic strength(80). The kinase reaction or mock kinase reaction lacking P-TEFb, was then purified over a size exclusion chromatography column and concentrated to 1mg/ml for mass spectrometry analysis. This strategy resulted homogenous preparations of apo and phospho species with discrete mobility in SDS-PAGE and with no evidence of unphosphorylated protein remaining in the phosphoCTD sample (Fig. 2-6b lanes 2 vs. 3).

Using MALDI-TOF mass spectrometry, we determined that phosphorylation by P-TEFb under our conditions lead to an increase in MW of
our MBP-*D. melanogaster* CTD fusions of ~1.9kDa, equivalent to the mass of approximately 25 phosphates (Fig. 2-6). This density of phosphorylation, less than one phosphate per heptad with no heptads phosphorylated more than once (Fig 2-7), is in good agreement with *in vivo* observations, suggesting our biophysical analyses focus on a biologically relevant CTD phospho-isoform(17, 18). Thus, in addition to deciphering rules of substrate preference for much of the full-length CTD using an intra-species CTD:kinase pair, I can also relate structural changes observed in biophysical and biochemical experiments to characterized patterns of phosphorylation.
Figure 2-6 MBP-\textit{D.mel}CTD sample preparation for mass spectrometry analysis. (a) A schematic of the purification strategy of apo and phospho \textit{D. melanogaster} CTD used for mass spectrometry. Low temperature expression, and successive purification steps select for opposing termini facilitate enrichment of full-length protein. (b) Coomassie stained SDS-PAGE gel of \textit{D. melanogaster} CTD preparations subsequent to amylose purification (lane 1), or after kinase/mock kinase reaction and gel filtration chromatography using a Superose 6 10/30 column (lanes 2 and 3). 0.5ul of each preparation at 1mg/ml were loaded. These preparations were used for mass spectrometry. No visible unphosphorylated protein remains in the phospho CTD sample.
Figure 2-7. MALDI-TOF mass spectrometry of MBP-\textit{D.mel}CTD to quantify phosphate incorporation. (a) MALDI-TOF mass spectra of apo and phospho MBP-\textit{D.mel}CTD fusions indicates an increase in mass of $\sim$1.9 kDa. (b) Zoomed-in section of the spectra from panel a shows a distribution of phosphorylated species for the phospho MBP-\textit{D.mel}CTD with the average change in mass consistent with the addition of 25 phosphates. (Protein expression, purification and sample preparation by Bede Portz, mass spectrometry by Josh Mayfield)
2-2.4 Substrate specificity of P-TEFb for the CTD

Some controversy exists about the substrate specificity of P-TEFb for the CTD, and as of this writing, no high-resolution structural data exists for P-TEFb in complex with the CTD that could inform us about determinants of specificity(91). P-TEFb is commonly considered a $S_2$ kinase in the context of the CTD, but evidence for this preference is indirect and conflicting evidence for the specificity of P-TEFb also exists in the literature. P-TEFb was initially described as a $S_2$ kinase because Pol II elongation complexes incubated in nuclear extracts depleted of P-TEFb exhibited reduced accumulation of $S_2$ phosphorylation, suggesting P-TEFb phosphorylates $S_2$(92). However, P-TEFb exists as a component of other complexes and is modulated by the activity of other factors, including Brd4 which is involved in recruiting P-TEFb to promoters and activating P-TEFb kinase activity(93, 94). Importantly, Brd4 is itself a CTD $S_2$ kinase, so immunodepletion of P-TEFb may have resulted in the loss of other $S_2$ kinase activities, including Brd4(95). To address similar possibilities, the authors inhibited P-TEFb activity with DRB, which resulted in a loss of phosphorylation, but the extent to which other kinases are also inhibited by the nucleotide analog DRB is unknown, though it has been shown to inhibit another CTD kinase TFIIH,(96) itself shown to have $S_2$ kinase activity in vitro(81). Additionally, the antibody specific for $S_2$ phosphorylation used to establish P-TEFb as a $S_2$ kinase has been shown to preferentially recognize $S_5$-$S_2$ doubly phosphorylated repeats, making it difficult to determine the exact specificity from immunoblots(97).
Together, the nature of these experiments and observations makes it difficult to conclusively label P-TEFb as a S₂ kinase.

More recent and direct evidence suggests P-TEFb may actually phosphorylate S₅(98). These experiments were conducted using recombinant P-TEFb on GST-CTD constructs containing varying numbers of heptad repeats of either the consensus motif, YSTPSPK, or the native mammalian sequence, that were characterized by using mutagenesis, mass spectrometry, and western blot(98). The findings of this paper challenged the precedent that P-TEFb is a S₂ kinase, and found a clear preference for S₅ phosphorylation of short CTD fragments(98). On short constructs, a density of 1 phosphate per repeat was achieved, a higher density than has been observed for Pol II purified from yeast or human cells(17, 18, 98). Interestingly, evidence emerged from this paper that the length of iterated sequences can alter P-TEFb activity. Constructs with 3 repeats of the non-consensus heptad YSPTSPK were phosphorylated with similar kinetics to a construct with 3 consensus repeats, but constructs of 9 or 13 repeats with K₇ showed slower kinetics relative to the shorter constructs, suggesting the length of the substrate could alter P-TEFb activity(98). Together, this study employing various CTD constructs to map P-TEFb phospho sites determined that in the context of short peptides, P-TEFb shows a clear preference for S₅ phosphorylation. However, this study does not map phospho sites on the native CTD sequence with anything other than a western blot, and also provides evidence for a length dependent alteration of P-TEFb activity(98),
so the rules governing P-TEFb specificity in the context of a full-length native CTD sequence remain unexplored.

Unlike the more repetitive CTDs of other model organisms, the sequence of the *Drosophila* CTD makes it an ideal substrate to test the specificity of kinases and phosphatases using both tandem mass spectrometry (tandem MS/MS) and nuclear magnetic resonance spectroscopy (NMR) (53, 80). Tandem mass spectrometry requires initial fragmentation of the protein, typically in the form of enzymatic digestion by trypsin to generate precursor ions (99). After an initial mass spectrometry scan, these precursor ions are themselves fragmented into smaller peptides enabling localization of the post translational modification via a second mass spectrometry scan (99).

Precise mapping of post-translational modifications (PTMs) requires the generation of precursor ions and peptides that can be uniquely identified by their mass. Tandem MS/MS faces limitations on repetitive sequences, including the CTD of *S. cerevisiae* and mammals, because different regions of the CTD are indistinguishable from one another on the peptide level. For example, the chymotryptic fragment YSPTSPS, and any associated PTMs, could arise from nearly anywhere in the yeast CTD, or anywhere on the proximal half of the mammalian CTD. This challenge is analogous to mapping very short DNA sequencing reads to large genomes, where the identical sequence could map to multiple genomic locations. Indeed, mapping the location of phosphates to the CTDs of *S. cerevisiae* and humans purified from cells required extensive mutagenesis to generate sufficient trypsin digestion sites and peptides of unique
mass\cite{17, 18}. Such mutagenesis strategies are hamstrung by the fact they fail to present native sequences, and possibly structures, to their cognate modifying enzymes. This limitation could be particularly acute for an IDP like the CTD, which lacks a stable structure to compensate for the local effects of mutated residues, and relies on short stretches of amino acids and small amounts of buried surface area for recognition\cite{28}. Additionally, IDP structure is related to charge distribution, and the mutations required to make the yeast and human CTDs suitable for mass spectrometry involved introducing charged lysine residues into regions of the CTD that are typically neutral in the unmodified state\cite{17, 18, 56, 79}. In addition to these potential structural changes, evidence exists for the evolutionary tuning of IDP:ligand interactions via charge, with protein ligands evolved to recognize the charge state of the IDP \cite{100}. One characterized example is the phosphoprotein Sic1, which is a multivalent IDP that is multiply phosphorylated on repeated motifs recognized in a phosphorylation dependent manner by the cell cycle regulator Cdc4\cite{100}. In this example, only after a threshold amount of phosphorylation is achieved on Sic1 does robust binding by Cdc4 occur\cite{100}. If the CTD were recognized in a charge dependent manner by CTD modifying enzymes, the addition of basic amino acids for the purpose of mass spectrometry could conceivably alter the phosphorylation of the CTD. Notably, the Cdk9 subunit of P-TEFb has an unusually high pi of \(~9, and is stimulated by the preexisting S\textsubscript{7} phosphorylation, suggesting the CTD code could be tuned by polyelectrostatic effects, and thus potentially altered by mutations that introduce positive charges\cite{91}. 
Unlike the yeast and human CTDs, the native sequence of the *Drosophila* CTD is sufficiently degenerate to enable phosphosite identification using tandem MS/MS across the majority of the CTD without the aforementioned caveats of mutagenesis. Here, *Drosophila* presents us with the ability to interrogate a full length CTD:kinase pair from the same organism while also allowing us to relate discrete, biologically relevant patterns of phosphorylation to changes in measurable biophysical and molecular parameters like $R_g$, $D_{max}$, protein accessibility, and flexibility.

Two different tandem mass spectrometry methods utilizing different secondary fragmentation strategies were employed in an effort to localize and cross-validate sites of phosphate incorporation by P-TEFb. We mapped phosphorylation to primarily the 5th position, unless a heptad lacks a SP motif at positions 5-6 (Fig 2-8), consistent with the *in vitro* preference for S$_5$ for P-TEFb observed previously(98). Notably, those heptads deviating significantly from the consensus can be phosphorylated on the equivalent of the S$_2$ position (Fig 2-8) similar to observations made previously using p42 MAP kinase(80).
### a. HCD Mass Spectrometry

<table>
<thead>
<tr>
<th>Heptad Number</th>
<th>Peptide Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YSPTS PN</td>
</tr>
<tr>
<td>2</td>
<td>YTASSPGGASPN</td>
</tr>
<tr>
<td>3</td>
<td>YSPSSPN</td>
</tr>
<tr>
<td>4</td>
<td>YSPTSPLYASPR</td>
</tr>
<tr>
<td>5</td>
<td>YASTTPNFNPQRSTG</td>
</tr>
<tr>
<td>6</td>
<td>YSPSSSG</td>
</tr>
<tr>
<td>7</td>
<td>YTPSTPV</td>
</tr>
<tr>
<td>8</td>
<td>YSPTVQFQ</td>
</tr>
<tr>
<td>9</td>
<td>SPSFAGSGSNI</td>
</tr>
<tr>
<td>10</td>
<td>YSPGNA</td>
</tr>
<tr>
<td>11</td>
<td>YSPSSSN</td>
</tr>
<tr>
<td>12</td>
<td>YSPNSPS</td>
</tr>
<tr>
<td>13</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>14</td>
<td>YSPSSPS</td>
</tr>
<tr>
<td>15</td>
<td>YSPSTPC</td>
</tr>
<tr>
<td>16</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>17</td>
<td>YSPSTPN</td>
</tr>
<tr>
<td>18</td>
<td>YTPVTPS</td>
</tr>
<tr>
<td>19</td>
<td>YSPSTPN</td>
</tr>
<tr>
<td>20</td>
<td>YASAPQ</td>
</tr>
<tr>
<td>21</td>
<td>YSPASPA</td>
</tr>
<tr>
<td>22</td>
<td>YSQTGVK</td>
</tr>
<tr>
<td>23</td>
<td>YSPSTPT</td>
</tr>
<tr>
<td>24</td>
<td>YSPPSPS</td>
</tr>
<tr>
<td>25</td>
<td>YDGSFGSPQ</td>
</tr>
<tr>
<td>26</td>
<td>YTPGSPQ</td>
</tr>
<tr>
<td>27</td>
<td>YSPASPK</td>
</tr>
<tr>
<td>28</td>
<td>YSPSTPL</td>
</tr>
<tr>
<td>29</td>
<td>YSPSSPQHSPSNQ</td>
</tr>
<tr>
<td>30</td>
<td>YSPSTGST</td>
</tr>
<tr>
<td>31</td>
<td>YSATSPR</td>
</tr>
<tr>
<td>32</td>
<td>YSPNMSI</td>
</tr>
<tr>
<td>33</td>
<td>YSPSSTK</td>
</tr>
<tr>
<td>34</td>
<td>YSPSTPT</td>
</tr>
<tr>
<td>35</td>
<td>YPTAPSH</td>
</tr>
<tr>
<td>36</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>37</td>
<td>YSPSPTPA</td>
</tr>
<tr>
<td>38</td>
<td>YSPSPSFTFESEED</td>
</tr>
</tbody>
</table>

### b. UVPD Mass Spectrometry

<table>
<thead>
<tr>
<th>Heptad Number</th>
<th>Peptide Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YSPTS PN</td>
</tr>
<tr>
<td>2</td>
<td>YTASSPGGASPN</td>
</tr>
<tr>
<td>3</td>
<td>YSPSSPN</td>
</tr>
<tr>
<td>4</td>
<td>YSPTSPLYASPR</td>
</tr>
<tr>
<td>5</td>
<td>YASTTPNFNPQRSTG</td>
</tr>
<tr>
<td>6</td>
<td>YSPSSSG</td>
</tr>
<tr>
<td>7</td>
<td>YSPSTPV</td>
</tr>
<tr>
<td>8</td>
<td>YSPTVQFQ</td>
</tr>
<tr>
<td>9</td>
<td>SPSFAGSGSNI</td>
</tr>
<tr>
<td>10</td>
<td>YSPGNA</td>
</tr>
<tr>
<td>11</td>
<td>YSPSSSN</td>
</tr>
<tr>
<td>12</td>
<td>YSPNSPS</td>
</tr>
<tr>
<td>13</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>14</td>
<td>YSPSSPS</td>
</tr>
<tr>
<td>15</td>
<td>YSPSTPC</td>
</tr>
<tr>
<td>16</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>17</td>
<td>YSPSTPN</td>
</tr>
<tr>
<td>18</td>
<td>YTPVTPS</td>
</tr>
<tr>
<td>19</td>
<td>YSPSTPN</td>
</tr>
<tr>
<td>20</td>
<td>YASAPQ</td>
</tr>
<tr>
<td>21</td>
<td>YSPASPA</td>
</tr>
<tr>
<td>22</td>
<td>YSQTGVK</td>
</tr>
<tr>
<td>23</td>
<td>YSPSTPT</td>
</tr>
<tr>
<td>24</td>
<td>YSPPSPS</td>
</tr>
<tr>
<td>25</td>
<td>YDGSFGSPQ</td>
</tr>
<tr>
<td>26</td>
<td>YTPGSPQ</td>
</tr>
<tr>
<td>27</td>
<td>YSPASPK</td>
</tr>
<tr>
<td>28</td>
<td>YSPSTPL</td>
</tr>
<tr>
<td>29</td>
<td>YSPSSPQHSPSNQ</td>
</tr>
<tr>
<td>30</td>
<td>YSPSTGST</td>
</tr>
<tr>
<td>31</td>
<td>YSATSPR</td>
</tr>
<tr>
<td>32</td>
<td>YSPNMSI</td>
</tr>
<tr>
<td>33</td>
<td>YSPSSTK</td>
</tr>
<tr>
<td>34</td>
<td>YSPSTPT</td>
</tr>
<tr>
<td>35</td>
<td>YPTAPSH</td>
</tr>
<tr>
<td>36</td>
<td>YSPSTPS</td>
</tr>
<tr>
<td>37</td>
<td>YSPSPTPA</td>
</tr>
<tr>
<td>38</td>
<td>YSPSPSFTFESEED</td>
</tr>
</tbody>
</table>

**Legend:**
- Dark Red: Localized phosphate
- Dark Blue: Ambiguous phosphate
- Dark Grey: Non-phosphorylated residue
- Light Grey: Not detected by LC-MS
Figure 2-8. Tandem mass spectrometry to map phosphate incorporation by P-TEFb. Tandem mass spectrometry defines P-TEFb specify on the full length CTD. (a) Map of phosphosites determined using higher energy collisional dissociation (HCD) to fragment precursor ions. Precisely mapped phosphosites are depicted in red. Ambiguous phosphosites that could be at one of two nearby residues are depicted in blue. Regions of the CTD lacking mass spectrometry coverage are shaded in grey. SP motifs are phosphorylated at the S$_5$ position, unless absent in a given heptad, leading to SP motifs being phosphorylated at the S$_2$ position. (b) Phospho-site mapping by ultraviolet photodissociation (UVPD) mass spectrometry is consistent with the HCD data, but with slightly reduced coverage. (Protein expression, purification and sample preparation by Bede Portz, mass spectrometry by Rachel Mehaffy)

Our mass spectrometry data is consistent with prior in vitro observations obtained on shorter CTD fragments with repeated heptads and shows a preference for S$_5$ phosphorylation when an SP motif is present at positions 5-6 of the repeat(98). However, regions of S$_2$ phosphorylation occur in closely spaced heptads lacking an SP motif in positions 5-6, such as those found in repeats nine through eleven.

Coverage was lacking in the central region of the CTD, an issue with previous efforts to map phosphosites within the Drosophila CTD that required generating a smaller CTD fragment corresponding to the central region(80). Fortunately, the central region missing from our mass spectrometry data largely overlaps with the central region of the CTD studied by Eric Gibbs and colleagues(53). In this study both NMR and mass spectrometry were used to map and cross validate phospho-sites, with P-TEFb showing a clear preference
for S$_5$ phosphorylation. Together, these results indicate that, at least in vitro, P-TEFb is a S$_5$ kinase and that modification of heptad sequences can give rise to regions with S$_2$ phosphorylation.

2-2.5 Small angle X-ray scattering studies to characterize CTD structure

Both native gel electrophoresis and analytical gel filtration chromatography indicate that the *D. melanogaster* CTD is structurally similar to its yeast and human homologs, globally (Figs. 2-2 and 2-3). Like the human CTD, the *Drosophila* CTD extends upon phosphorylation (Figs. 2-3 and 2-4)(74). These characteristics make the *Drosophila* CTD an ideal model to interrogate CTD structure, with the added benefit of being able to relate structural changes to amounts and patterns of phosphorylation determined by mass spectrometry.

To further interrogate differences in the structures of the apo and phospho CTD, I employed small angle X-ray scattering (SAXS), a high-precision technique well suited to the interrogation of IDPs in solution(101). SAXS lacks the resolution of NMR, X-ray crystallography, or cryo-electron microscopy, but presents advantages for disordered proteins that are inappropriately sized or too flexible to be studied using the aforementioned techniques. SAXS experiments provide information about the size and shape of biomolecules in solution and can also report on conformational changes resulting from phosphorylation. In a SAXS experiment, a pure solution of protein is exposed to X-ray radiation, which is scattered by both the protein and buffer components of the solution and imaged by a detector (Fig 2-9a, modified from Skout et al)(102). A buffer sample is
similarly analyzed, and the buffer data are subtracted from the data from the protein scattering sample, revealing the scattering contribution of the protein alone, averaged over all conformations and orientations (Fig 2-9 b and c). This buffer subtracted scattering data can be analyzed to reveal information about the size, shape and flexibility of the protein under study, and changes in these parameters resulting from ligand binding or post translational modifications(102-104). SAXS data can also be used to constrain structural models to visually represent the size of the CTD(105, 106).
Figure 2-9. Small Angle X-Ray Scattering of proteins in solution. (a) schematic of a SAXS experiment. The protein solution is illuminated with X-ray radiation, which is scattered by protein and buffer. $q$ is a scattering vector, which is defined by the equation $q=(4\pi \sin \theta)/\lambda$, where $2\theta$ is the scattering angle and $\lambda$ is the wavelength of the x-ray radiation. (b) Beamline software integrates the scattering pattern into one dimensional scattering curves where the scattering intensity $I(q)$ is plotted against the $q$, which has units of inverse angstroms ($102$, $107$). Plotted are data from a protein solution (red) and buffer (black). (c) Buffer scattering curves are subtracted from protein solution scattering curves leaving behind a scattering curve arising from the protein, which can be used in further analysis.
2-2.6 SAXS sample preparation

SAXS experiments are sensitive to sample quality. Scattering intensity is a function of molecular weight, and thus small amounts of protein aggregation can bias the SAXS data\(^{(102)}\). Also, the scattering curve contains data about all proteins in the sample, so impurities and breakdown products could also bias interpretation of the data. Preparation of the apo and phospho MBP-*D. mel*/CTD fusions for SAXS was designed to enrich for highly pure, full-length protein and mirrors the strategy used for tandem MS/MS sample preparation (**Fig 2-10a**). The purified protein was further purified and buffer exchanged over a 24ml Superose 6 size exclusion column, with the exact preparation of buffer used for blanks for buffer subtraction during data acquisition. This scheme yields homogenous preparations of protein, with no unphosphorylated protein detected in the phosphorylated sample, consistent with the mass spectrometry sample preparation (**Fig 2-10b and 2-6b**).
Figure 2-10 MBP-\textit{D. mel}CTD sample preparation for small angle x-ray scattering. (a) the purification scheme employed for SAXS sample preparation. Selection for opposing termini and subsequent gel filtration generates highly purified sample. (b) Coomassie blue stained gel of the protein samples used for SAXS, subsequent to data acquisition. The freeze thaw cycle inherent to transporting the protein for SAXS has not caused a loss of sample integrity.
The gel filtration column used in SAXS sample preparation was the same column, run under similar conditions and in the same buffer, as was used for analytical gel filtration (Figs. 2-3 and 2-4). This column clearly resolves the apo and phospho forms of the CTD, helping to minimize the likelihood of any protein that remains hypophosphorylated after the extended P-TEFb reaction contributing to the phospho CTD SAXS data (Fig. 2-11). Additionally, a kinase reaction lacking the CTD substrate was also run on the SEC column. This was done to ensure that the P-TEFb enzyme, although only present in enzymatic quantities in the CTD kinase reactions, does not co-elute with the monomeric phospho CTD peak that was used for SAXS analysis. No species present in sufficient quantity to be detected with our in-line UV detector co-elutes with the CTD samples (Fig. 2-11).
Figure 2-11. Components of the P-TEFb reaction are resolved from the CTD via preparative size exclusion chromatography. (a) A chromatograph of a P-TEFb reaction lacking the CTD substrate. The fractions where the phospho CTD sample eluted are shown in red, and apo in blue, with their corresponding elution volumes \( (E_v) \). (b) The same chromatograph, scaled to enable visualization of trace amounts of protein that could co-migrate with the CTD samples. No UV absorbance above background is present in the fractions where the CTD samples elute.
2.2.7 SAXS analysis

SAXS data are represented as curves relating scattering intensity \( I(q) \) to scattering angle \( q \), in reciprocal space, with signal intensity decaying at the high \( q \) regions of the curve. An initial approximation of data quality is ascertained by comparing the shape of the scattering curves collected across the concentration series. SAXS data collected on both the apo and phospho MBP-\textit{D.mel}CTD show no concentrations effects, with the curves having the same shape, differing only by intensity, across concentrations (Fig. 2-12). An additional quality control is Guinier analysis of SAXS curves, which attempts to fit the low \( q \) data to an equation, from which one can estimate \( R_g \). Barring concentration related interparticle effects, such as aggregation or repulsion, the values of these \( R_g \) estimates should be within ~10% of one another across the concentration series(102). Similarly, the MW of the protein can be estimated by comparing the scattering intensity at \( I(0) \) obtained from the Guinier approximation, to that of a protein standard of known MW and concentration(108). Guinier estimates of \( R_g \) and MW obtained from my scattering data showed no concentration effects across the dilution series in measured MW or radius of gyration (\( R_g \)). Phosphorylation lead to a ~15% increase in \( R_g \) (Fig 2-13 and table 2-1)(102, 107) and an increase in MW of ~1.9kDa, in excellent agreement with the MALDI-TOF data (Fig 2-7).
Figure 2-12 Averaged, buffer subtracted SAXS curves from three protein concentrations plotted as \( I(q) \) vs \( q \). (a) the apoMBP-\textit{D.mel}CTD and (b) phosphoMBP-\textit{D.mel}CTD. Scattering curves correspond to those used in the Guinier analysis (Fig. 2-1).
Figure 2-13 Guinier fits of SAXS data for the apo and phospho CTD. Curves plotted as Ln(I) vs q² (left panels) and residuals (right panels) of SAXS data collected at three protein concentrations for the apo (blue data) and phospho (red data) MBP-\textit{D.mel}/CTD fusion proteins. Protein concentrations are noted on the right.
Table 2-1 Summary of Guinier analyses. Calculated MW or $R_g$ are independent of protein concentration as expected for monodisperse samples. Calculated MW is close to that of the apo MBP-\textit{D.melICTD} fusion, 74.70kD. The $R_g$ of the CTD increases as a result of phosphorylation. The increase in mass of ~1.9kD is in excellent agreement with the increase in mass measured by MALDI-TOF (Fig 2-6).

<table>
<thead>
<tr>
<th>Protein</th>
<th>(mg/ml)</th>
<th>MW(kD)</th>
<th>$R_g$(Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apoCTD</td>
<td>1.56</td>
<td>79.73</td>
<td>50.69</td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>72.74</td>
<td>46.22</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>74.63</td>
<td>47.64</td>
</tr>
<tr>
<td>Avg and Std Dev.</td>
<td></td>
<td>75.7 +/- 3.6</td>
<td>48.2 +/- 2.3</td>
</tr>
<tr>
<td>phosphoCTD</td>
<td>1.9</td>
<td>78.98</td>
<td>56.02</td>
</tr>
<tr>
<td></td>
<td>0.97</td>
<td>77.04</td>
<td>54.22</td>
</tr>
<tr>
<td></td>
<td>0.45</td>
<td>81.67</td>
<td>54.70</td>
</tr>
<tr>
<td>Avg and Std. Dev.</td>
<td></td>
<td>77.6 +/- 1.6</td>
<td>55.0 +/- 0.9</td>
</tr>
</tbody>
</table>
These initial SAXS data analysis steps serve primarily as quality controls. An additional algorithm can transform the averaged scattering curves into real space, and graphically represent the distance between all points in the molecule as a histogram (102, 103, 109). Unlike the Guinier analysis, this operation reveals shape information about the molecule and utilizes a much wider range of the data, which can result in more accurate size calculations for IDPs (103). Such pairwise distance distribution functions for the averaged scattering curves from the apo and phospho MBP-D.melCTD fusions were generated using AUTOGNOM (Fig 2-14)(110). The $R_g$ values obtained via the $P_r$ vs $r$ analysis are slightly larger than obtained via Guinier analysis, which is not uncommon for IDPs, and the overall extension calculated in this analysis is in good agreement with that obtained from the Guinier fit (Table 2-1)(103). These analyses provide independent validation of the phosphorylation dependent extension I observe using SEC (Fig 2-4).

2-2.8 The CTD is semi-compact

Protein compaction is a function of the interactions between the polypeptide chain and the solvent, and intra-chain interactions between the polypeptide and itself (89). Folded proteins are compact under physiological conditions because intra-chain contacts dominant over interactions with the solvent, so called “poor” solvent conditions (89). Some IDPs can adopt extended conformations, adopting random coil dimensions similar to those of proteins chemically denatured by “good” solvents in which intra-chain contacts are broken.
and interactions between the protein and solvent dominate. The Flory chain model applied to disordered polypeptides relates protein length to $R_g$ via a power law, and has been used as a benchmark against which to compare the disorder, or unfoldedness, of a protein. Applied only to the CTD portion of the fusion protein and omitting MBP, the Flory equation using parameters for denatured proteins predicts an $R_g$ of $\sim 60\AA$ (Fig. 2-14 b and c). Notably, this is larger than the measured $R_g$ of 51.8Å for the entire MBP-\textit{D.mel}/CTD fusion, which includes MBP, a globular protein with a mass exceeding that of the CTD and an estimated $R_g$ of $\sim 22\AA$ (Fig. 2-14a). This comparison suggests that the CTD is considerably more compact than a chemically denatured coil. Together, our measurements support a semi-compact CTD, more extended than a globular protein of the same molecular weight, but more compact than a denatured coil.
a) $P(r)$ vs $r$

- apoCTD: 51.8Å/221.5Å
- phosphoCTD: 60.6Å/255.7Å
- MBP: 22.1Å/72.5Å

b) $R_g = R_0 N^\nu$

c) MBP

CTD only Flory chain Prediction $R_g \approx \sim 60\text{Å}$

MBP-CTD measured $R_g \approx \sim 52\text{Å}$
**Figure 2-14 The CTD is semi compact.** (a) Pairwise distance distribution function of the apo (blue) and phospo (red) MBP* D.mel* CTD SAXS data generated using AUTOGNOM\(^{110}\). A theoretical scattering curve was generated for MBP using CRYSOL\(^{111, 112}\). In contrast to the globular MBP, the CTD fusions have a long, rod-like extension emanating from the globular MBP. Phosphorylation extends the CTD, evidenced by increases in \(R_g\) and \(D_{\max}\). (b) The Flory equation predicts the \(R_g\) of random coil proteins as a function of polypeptide length. Random coil proteins behave like denatured proteins, and thus the model is a benchmark of unfoldedness \((90, 103, 111)\). \(N\) is the number of amino acids in the polypeptide, 318, \(R_0\) is a constant with a value 1.927, and \(\nu\) is an exponential scaling factor dependent on conditions. \(\nu\) of 0.598 was derived from chemically denatured proteins and used in my estimates\((90)\). (c) The Flory model predicts a \(R_g\) of the CTD alone of \(\sim 60\text{Å}\), but the experimentally measured \(R_g\) of the entire MBP-CTD fusion is \(\sim 52\text{Å}\), suggesting the CTD is considerably more compact than a random coil.
2-2.9 Ensemble optimization models of the CTD

To better conceptualize the CTD structure on the global scale, I generated ensemble models from the SAXS curves collected on apo and phospho MBP-\textit{D.mel}/CTDs using the Ensemble Optimization Method (EOM) 2.1, accounting for the contribution of the globular MBP\textsuperscript{106, 113}. This method generates 10,000 models for the CTD that are parsed using an algorithm to yield an ensemble whose average scattering curve agrees with the experimentally measured curve for our apo and phospho MBP-\textit{D.mel}/CTD fusions. These models yielded ensembles for the apo and phospho MBP-\textit{D.mel}/CTD with \( R_g \) values of 50.91Å and 60.98Å, which are in good agreement with \( R_g \) values derived from the \( P(r) \) vs \( r \) analysis which enables these models to be used to visually conceptualize the CTD including the extension resulting from phosphorylation (Fig 2-15 and Fig 2-14). The resulting ensembles are depicted oriented by MBP and displayed with the 12-subunit RNA Pol II elongation complex model to allow for size comparison (Fig. 2-15)(113, 114). Notably, both compact and extended structures are found in the ensembles describing both the apo and phospho CTD, with a bias toward extended structures in the phospho state consistent with the extension observed in SEC and SAXS. SAXS data lack the resolution to interpret the models on the amino acid or heptad scale, yet do provide a visual representation of the scale of the monomeric CTD in solution that is constrained by and in agreement with experimental data.
Fig 2-15 Ensemble optimization method models of the CTD. Averaged apo and phospho MBP-\textit{D. mel}/CTD scattering data were modeled using EOM 2.1 from the ATSAS suite\textsuperscript{(110)}. Individual models in the apo and phospho ensembles are shown in shades of blue and red, respectively, oriented by MBP and positioned adjacent to the 12 subunit RNA Pol II elongation complex (1Y1W.pdb) model (green) for scale\textsuperscript{(104, 114)}. 
2-3 Discussion

Despite significant sequence differences, the yeast, fly and human CTDs are structurally similar globally, as evidenced by the linear relationship between their mobilities in native gel electrophoresis and size exclusion chromatography experiments (Fig. 2-2 and 2-3). This relationship suggests that increases in CTD length and sequence complexity lead to an increase in CTD size but not a wholesale disruption of the global structure. The slope of molecular weight vs Stokes radius for the CTDs is more flat than that of native folded proteins, with the size of the CTDs growing more rapidly as a function of MW standards. This suggests the CTDs adopt more extended structures (Fig. 2-3). As has been observed for the mammalian CTD, the fly CTD extends in response to phosphorylation, arguing for an unmodified CTD that must be somewhat compact (Fig. 2-4)(74). Indeed, the entirety of the MBP D.mel/CTD fusion is more compact than random coil models predict the CTD to be in isolation (Fig. 2-14). The similarities between the fly CTD and its yeast and human counterpart, the combination of developmental complexity and genetic tractability, as well as the ability to and map phosphorylation to the native sequence establish Drosophila as an ideal model to study CTD structure.

Existing models of CTD structure were generated by iterating structural motifs across the length of the CTD. These models leveraged the best available structural data from CTD peptides, but evidence for turn motifs in unbound CTD peptides in physiologically relevant buffer conditions was lacking(65, 68-71).
Evidence for a compact but flexible CTD was more compelling, but was potentially biased by CTD:CTD or CTD:factor interactions or fixation conditions required by the various experiments (4, 66, 73). The biophysical data presented here on the monomeric full-length CTD obtained with both size exclusion chromatography and small angle X-ray scattering supports a model for CTD structure that I describe as a compact random coil: more extended than predicted for a natively folded protein of the same mass, but significantly more compact than a denatured coil.

The primary role of the CTD is thought to be that of a scaffold protein capable of recruiting proteins to Pol II involved in transcription initiation, elongation, termination, and co-transcriptional processes such as mRNA capping, splicing and polyadenylation; further, these interactions occur in a temporospatially regulated manner (2, 3). Compactness may protect the CTD from imprecise post-translational modification and factor recruitment out of phase with the transcription cycle.

A semi-compact structure capable of extension could confer organization to CTD interactions. Beyond that, it hints at the possibility of further structural organization in the CTD. IDPs exist on a structural continuum of order and compaction, and this continuum also applies within individual IDPs, which can be structurally heterogeneous, non-uniformly compact, and differentially responsive to post translational modifications across their length (54-56, 76).

Work from Mayfield and colleagues, Gibbs and colleagues, and the work presented here firmly establishes the Drosophila CTD as a biophysical tractable
native sequence enabling global and local structural changes to be related to characterized patterns of phosphate incorporation (53, 80). The methodologies brought to bear in these studies, namely tandem mass spectrometry, NMR, size exclusion chromatography and SAXS could be applied to the fly CTD modified by various kinases, alone or in isolation. Further, the CTD has been shown to be subject to other post-translational modifications, including acetylation, methylation, and glycosylation (115-119). Relating structural changes to discrete patterns of post translational modifications resulting from different enzymes could be a fruitful avenue of future study aimed at elucidating structural changes associated with subsets of the CTD code. In the following chapter, I examine the structure of the CTD on the local level, and relate structural changes in the CTD to phosphorylation by P-TEFb.
CHAPTER 3
LOCAL STRUCTURAL HETEROGENEITY AND ENSEMBLE ALLOSTERY IN THE POL II CTD

The data in this chapter are published in Portz, B. et al. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nature. Communications. 2017 8, 15231

3-1 Introduction

It is becoming increasingly appreciated that intrinsically disordered proteins exist on a structural continuum, endowed with the ability to adopt multiple conformers. Rather than being a completely random ensemble of structures, some IDPs can also populate particular conformers to a greater extent than others (76). Also, the order to disorder continuum applies within IDPs which can possess different amounts of compaction or propensity for order in some regions, and less in others (54, 55). If the CTD was structurally heterogeneous, with local variation in order and protein accessibility, this order could potentially be exploited by evolution to target post-translational modifications or factor interaction to specific regions of the CTD.

My observations, via SEC and SAXS, are consistent with a semi-compact CTD capable of undergoing extension. This compaction suggests there may be some structural organization to the CTD, and prompted me to examine the CTD on a more local level. SEC and SAXS provide information about the global
structure of the CTD, but do not provide information about local structural variation contained therein. Higher resolution structural methods like X-ray crystallography and CryoEM are poorly suited to interrogating flexible systems and have failed to provide local structural information about the CTD(4, 66, 72, 73). X-ray crystallography has provided high resolution data on a range of structures adopted by the CTD in complex with other folded domains(13, 33, 64). However, these studies fail to provide information about how the different regions of the CTD may be structured differently when the CTD is free in solution, and to what extent different regions of the CTD may in turn be differentially accessible to protein interaction.

Structural footprinting methods are ideal to characterize local variation in protein structure in solution under physiologically relevant conditions and are not limited by size, sequence repetition, or flexibility. Limited proteolysis experiments have been used to interrogate IDP structure, with regions of protease sensitivity and protection reflecting more unstructured and structured regions, respectively(120-124). In a striking recent example, this technique has been applied to entire proteomes and determined that many putative IDPs actually possess detectable structure(124). Because the CTD is thought to function primarily as a scaffold domain, recruiting other proteins to the polymerase, proteases can be used as protein scale probes to determine if local variation in structure and protein accessibility exist across the CTD. In contrast, chemical cleavage experiments would probe the disordered CTD with molecules orders of magnitude smaller than the proteins the CTD has evolved to recruit, analogous to
reading the newspaper through binoculars. Finally, limited proteolysis is factor agnostic, and examines local structure and accessibility independent of the evolutionary constraints that could target a particular CTD interacting factor to a particular region. Given these advantages, I used limited proteolysis to probe for variation in CTD structure.

The experiments described in this chapter discover previously unknown structural organization in the *Drosophila* CTD. It is evident that like globular proteins, IDPs can also undergo allosteric modulation by ligand binding and post translational modifications(63). My discovery of structural heterogeneity within the CTD lead me to question if the structure of the CTD could be allosterically modulated by phosphorylation. In this chapter, I describe structural changes to the CTD in response to phosphorylation that in combination with structural heterogeneity could aid in organizing CTD:factor interactions in time and space.

The notion that IDPs can adopt semi-compact, heterogeneous ensembles is important in the context of CTD evolution. Much of our knowledge comes from studies on the CTD of *S. cerevisiae* that is unique in its sequence homogeneity, with nearly every heptad exactly matching the YSPTSPS consensus(24). The CTD in mammals contains a consensus region proximal to the catalytic core of Pol II, and a non-consensus region at the distal end of the CTD composed of heptads differing primarily at the seventh position(24, 26) (Fig. 3-1a). The CTD of *D. melanogaster* is distinct from its human counterpart in that only two heptads exactly match the consensus, though a similar consensus motif is identified(26, 82)(Fig. 3-1b)
### Mammals

| 1 | YSPtSPA |
| 2 | YEPRSPPG |
| 3 | YTPQSPPS |
| 4 | YSPtSPS |
| 5 | YSPtSPS |
| 6 | YSPtSPN |
| 7 | YSPtSPS |
| 8 | YSPtSPS |
| 9 | YSPtSPS |
| 10 | YSPGNA |
| 11 | YSPSSSN |
| 12 | YSPNSPS |
| 13 | YSPtSPS |
| 14 | YSPtSPS |
| 15 | YSPtSPS |
| 16 | YSPtSPS |
| 17 | YSPtSPN |
| 18 | YTPVTPS |
| 19 | YSPtSPS |
| 20 | YSAEPA |
| 21 | YSPtSPN |
| 22 | YSPtSPN |
| 23 | YSPtSPT |
| 24 | YSPtSPT |
| 25 | YSPtSPT |
| 26 | YSPtSPT |
| 27 | YSPtSPT |
| 28 | YSPtSPL |
| 29 | YSPtSPPQHSPSNQ |
| 30 | YSPtSPT |
| 31 | YSPtSPT |
| 32 | YSPtSPT |
| 33 | YSPtSPT |
| 34 | YSPtSPT |
| 35 | YSPtSPT |
| 36 | YSPtSPT |
| 37 | YSPtSPT |
| 38 | YSPtSPT |
| 39 | YSPtSPT |
| 40 | YSPtSPT |
| 41 | YSPtSPT |
| 42 | YSPtSPT |
| 43 | YSPtSPT |
| 44 | YSPtSPT |
| 45 | YSPtSPT |
| 46 | YSPtSPT |
| 47 | YSPtSPT |
| 48 | YSPtSPT |
| 49 | YSPtSPT |
| 50 | YSPtSPT |
| 51 | YSPtSPT |
| 52 | YSLTSPAISPDDSEEN |

### D. melanogaster

| 1 | YSPtSPN |
| 2 | YTAESSPGGASPN |
| 3 | YSPSSPN |
| 4 | YSPtSPLYASPR |
| 5 | YASTTPNFPSQSTG |
| 6 | YSPSSSG |
| 7 | YSPtSPV |
| 8 | YSPtVQFQ |
| 9 | SSPSFAGSGSNI |
| 10 | YSPGNA |
| 11 | YSPSSSN |
| 12 | YSPNSPS |
| 13 | YSPtSPS |
| 14 | YSPSSPS |
| 15 | YSPtSPC |
| 16 | YSPtSPS |
| 17 | YSPtSPN |
| 18 | YTPVTPS |
| 19 | YSPtSPN |
| 20 | YSAEPA |
| 21 | YSPASPA |
| 22 | YSQTGVK |
| 23 | YSPtSPT |
| 24 | YSPtSPT |
| 25 | YDGSPGSPQ |
| 26 | YTPGSPQ |
| 27 | YSPASPK |
| 28 | YSPtSPL |
| 29 | YSPSSPSQHSPSNQ |
| 30 | YSPtGST |
| 31 | YSATSPR |
| 32 | YSPNMSI |
| 33 | YPSHSTK |
| 34 | YSPtSPT |
| 35 | YTPTARN |
| 36 | YSPtSPM |
| 37 | YSPtSPH |
| 38 | YSPtSPA |
| 39 | YSPSSPTFEESED |
| 40 | YSPtSPK |
| 41 | YSPtSPT |
| 42 | YSPtSPK |
| 43 | YSPtSPT |
| 44 | YSPtSPK |
| 45 | YTPtSPK |
| 46 | YSPtSPT |
| 47 | YSPtSPK |
| 48 | YSPtSPT |
| 49 | YSPtSPK |
| 50 | YSPtSPG |
| 51 | YSPtSPT |
| 52 | YSLTSPAISPDDSEEN |
Figure 3-1 The human and fly CTDs have different sequences but a similar consensus motif. a. The *H. sapiens* and *D. melanogaster* CTDs differ in length and sequence complexity. The *D. melanogaster* CTD contains only two consensus motifs (bold type), compared to 22 in *H. sapiens*. b. Despite these sequence differences, a similar consensus motif is present in both CTDs.

Evolutionary differences in CTD composition raise the possibility that different motifs may serve to organize CTD:protein interactions via the co-evolution of factors that favorably recognize particular CTD motifs(24). An alternative and not mutually exclusive role for non-consensus motifs could be the formation of structural heterogeneity, rendering particular regions of the CTD more accessible to protein interaction.

The structural similarity but sequence differences between the fly and human CTDs, coupled with the genetic tractability and developmental complexity of *Drosophila*, emboldened me to test the hypothesis that the evolution of non-consensus CTD repeats imparted important lineage specific functionality(24, 26). The high degree of conservation of the CTD sequence among the twelve species of *Drosophila* (Fig. 3-2) and separately among mammals contrasts with the sequence divergence between the two phyla (Fig. 3-1)(49). This suggests that within various lineages, specific CTD sequences arose and were maintained under selective pressure, perhaps co-evolving with CTD interacting factors in a way that could recruit new activities to RNA Pol II or organize multiple CTD:factor interactions across the CTD.
Figure 3-2 Amino acid sequence alignment of 12 species of *Drosophila*.

Amino acids identical in all 12 species are shown in black. Amino acids differing in one or more species are shown in white. The amino acid sequence of the CTD is nearly identical among the 12 species.
3-2 Results

3-2.1 The CTD of D. melanogaster is structurally heterogeneous

The repetitive nature of the CTD is ideally suited to limited proteolysis, as predicted cleavage sites are evenly distributed across the length of the CTD. If the CTD is either a completely disordered chain, or if it is structurally repetitive as β-spiral models predict, proteases would be predicted to cleave the CTD with equal probability along its length. Alternatively, local structural heterogeneity would manifest as regions of protease sensitivity or protection (Fig. 3-3). Due to the aberrant mobility of the CTD in SDS-PAGE gels, I generated a CTD specific MW ladder in order to relate proteolytic fragments from limited proteolysis assays to CTD fragments of known MW. To do so, I generated 4 separate MBP-D. melanogaster CTD fusion proteins with TEV protease sites at specific locations. These TEV site containing proteins are then mixed together, end-labeled, cleaved with TEV and resolved on SDS-PAGE gels alongside the limited proteolysis experiments. Comparison of the limited proteolysis fragments to the CTD ladder aids in localizing cleavage to regions of the CTD (Fig. 3-4).
**Figure 3-3 Limited Proteolysis Experimental design.** End-labeled MBP-D.mel/CTD is subject to limited proteolysis, SDS-PAGE, and autoradiography. The blue line represents the CTD, the red star the $^{32}$P end label. A completely unstructured, or alternatively, a structurally repetitive CTD is predicted to generate a uniform pattern of proteolytic fragments. A structurally heterogeneous CTD is predicted to give rise to a non-uniform pattern.
Figure 3-4 Generation of CTD specific MW markers. (a) Location of TEV protease recognition sites inserted into four separate MBP-\textit{D. mel}/CTD fusion proteins. Numbered red bars correspond to fusion proteins displayed in panel c. (b) Schematic of radiolabeled TEV product sizes. (c) TEV cleavage generates a CTD ladder ranging in size from the full CTD (35kD) to a 10kD distal fragment.
Chymotrypsin cleaves the peptide bond C-terminal to tyrosine and is predicted to cut the *Drosophila* CTD 43 times across its length (83). Surprisingly, chymotrypsin yields a pattern of proteolysis characterized by a hypersensitive site towards the distal part of the CTD, and sensitivity located across the proximal region, with a protease insensitive region spanning the central part of the CTD that is cut in only one location (Fig. 3-5).

**Figure 3-5 Limited proteolysis with chymotrypsin.** The protease chymotrypsin generates a non-uniform pattern of CTD fragments, with a hypersensitive site in the distal CTD, a sensitive region in the proximal CTD, and a largely protease insensitive region in the central CTD that is cleaved at only one site. The right most lane contains radiolabelled CTD fusion proteins with the molecular weights designated on the right of the pane.
One interpretation of this result is that the CTD is not uniformly accessible to protein-sized factors due to local structural variation. Alternatively, this result could be due to sequence variation in the Drosophila CTD giving rise to non-uniform recognition of potential cleavage sites. To distinguish between these hypotheses, I subjected the MBP-D.mel.CTD to limited proteolysis with subtilisin, a protease possessing similar specificity to chymotrypsin but that is structurally unrelated, and proteinase K, predicted to cut the CTD at 100 sites(83). Strikingly, the proteolytic pattern generated by the additional enzymes recapitulates that of chymotrypsin (Figs. 3-5 and 3-6).

In the case of each protease, the cleavage products arising from the closely spaced sensitive sites proximal to MBP do not appear to be as responsive to increased protease concentrations as the distal hypersensitive site. These nearly full-length products are formed at low protease concentrations and persist, only disappearing at the highest protease concentrations (Figs. 3-5 and 3-6). Further, these products do not accumulate linearly as a function of protease concentration as evenly as the product resulting from cleavage at the distal hypersensitive site (Figs. 3-5 and 3-6). It is worthwhile to consider that limited proteolysis is a bulk experiment that samples the structures of the entire ensemble. One possibility is the longer CTD fragments arising from the proximal sensitive sites represent cleavage of rare conformers in the ensemble, and the resulting fragments are resistant to further proteolysis on the timescale probed in my experiment. If these particular conformers were sufficiently rare, the resulting cleavage products may be generated to completion at even the lowest protease
concentrations used in my experiments. This would explain why the longer CTD fragments do not accumulate linearly as a function of protease concentration. If these fragments promptly aggregated or otherwise represent conformers protected from further proteolysis, this would explain their persistence until the highest protease concentrations are reached and nearly all the full length CTD is digested.

Figure 3-6 Multiple proteases yield similar patterns of proteolysis across the CTD. (a) Limited proteolysis with subtilisin. (b) Limited proteolysis with
proteinase K reveals similar sites of sensitivity and protection. Proteinase K is predicted to cleave the CTD at 100 sites, but recapitulates a similar pattern of proteolysis to that of subtilisin.

To determine if the MBP may be altering the structure of the CTD, I repeated the limited proteolysis of the CTD fused to glutathione S-transferase (GST), observing the same protease sensitivity when compared to my MBP-CTD experiments, arguing against structural influence on the CTD by the fusion protein (Fig 3-7).

**Figure 3-7.** A GST-CTD fusion has the same proteolytic pattern as the MBP-CTD fusion. Limited proteolysis of GST-\textit{D. mel}CTD compared to the CTD ladder and MBP-\textit{D. mel}CTD. The discontinuous pattern of proteolysis generated from
the GST fusion matches that of the MBP fusion, suggesting the local structure of the CTD is not affected by the fusion protein in either case.

Comparing directly the digested products from all three proteases at like titration points reveals a similar but not identical pattern of proteolysis (Fig. 3-8). For each protease the distal hypersensitive site of the CTD is similar, and resulted in bands that migrated near the 10kD CTD ladder band. This 10kD band was generated by introducing a TEV protease site upstream of the amino acid corresponding to residue 1818 of the *D. melanogaster* Rpb1 protein (Fig. 3-4, see also Materials and Methods Fig. 6-3).
Figure 3-8 A comparison of protease accessible sites. A direct comparison of proteolytic fragments generated by three proteases shows a similar by not identical pattern of proteolysis. This suggests similar regions of the CTD are more accessible to protein interaction, with the cleavage site varying as a function of the specificity of each enzyme for the underlying sequence.

I also performed limited proteolysis in the presence of the denaturant sodium dodecyl sulfate (SDS) in concentrations tolerated by proteinase K(125). SDS alters the pattern of proteolysis, altering the distribution of sensitivity relative to the no-SDS control, consistent with disruption of local structure in the CTD.
substrate(125) (Fig. 3-9). This interpretation is more likely than an alternative hypothesis that the structure of the protease itself is altered by SDS in such a way that stimulates overall activity. SDS actually lowers proteinase K activity for short, unstructured peptides, but stimulates apparent activity for folded proteins. This dichotomy suggests the increase in apparent activity is due to disruption of substrate, not enzyme, structure(125). In the presence of SDS, protease sensitivity is more evenly distributed among multiple protease sensitive sites on the CTD, compared to one hypersensitive site in the absence of SDS. Thus, I favor the hypothesis that the CTD is structurally heterogeneous and not uniformly accessible to protein interaction across its length and that this organization is an intrinsic feature of the CTD.
Figure 3-9 The structure of the CTD is sensitive to the denaturant SDS. Sodium dodecyl sulfate (SDS) alters the relative proteolytic sensitivity of the CTD to proteinase K, enhancing the sensitivity of the CTD at sites that generate bands near 35kD and below 19kD relative to the CTD ladder (0.1% SDS lane). 0.5% SDS renders the globular MBP portion of the fusion protein more susceptible to proteolysis, evidenced by the proteolytic fragment above the 35kD CTD ladder band but below the intact fusion.
3.2.2 Structural heterogeneity persists after phosphorylation

In light of the findings that the CTD undergoes extension as a function of phosphorylation and is structurally heterogeneous in the apo state, I sought to determine if the CTD is locally reorganized in the phospho state. To address this question, I performed limited proteolysis on the phospho CTD using proteinase K (Fig 3-10). The phosphorylated CTD retains a non-uniform pattern of cleavage with the resulting proteolytic fragments having shifted mobility due to phosphorylation, indicating local structural heterogeneity is maintained (Fig. 3-10). However, the sensitivity of the various regions changes relative to each other. For example, at digestion points with minimal amounts of protease, products in the lower region of the gel are more abundant than products near the central region for the apo CTD (Fig. 3-10) whereas they are nearly equivalent in the case of the phosphorylated CTD (Fig 3-10). Overall, phosphorylation of the CTD results in more even level of cutting at all of the labile sites.
Figure 3-10 Local structural heterogeneity is maintained in the phospho CTD.

Limited proteolysis of the phospho CTD shows an altered pattern of proteolysis. The hypersensitivity of the distal site is greatly reduced, and the pattern of proteolysis is more evenly distributed among the sensitive sites. The proximal sensitive region, and central protected region are preserved after phosphorylation, with proteolytic fragments shifted relative to the apo fragments due to phosphorylation across the CTD.
3-2.3 CTD phosphorylation increases protein accessibility

The more even cleavage of protease accessible sites across the phospho CTD prompted me to test whether phosphorylation increases CTD accessibility. I subjected a mixture of apo and phospho CTD to limited proteolysis and compared the level to which each isoform was digested as a function of increasing protease concentration (Fig. 3-11a). I quantified 3 experimental replicates, comparing the amount of intact CTD remaining at each protease concentration. This revealed a moderate increase in the level of proteolysis of the phospho CTD, which is digested past the point of single hit kinetics (> 50% cleavage(126) at a lower protease concentration, consistent with a moderate increase in accessibility as a function of phosphorylation (Fig. 3-11b).

**Figure 3-11. Phosphorylation increases the proteolytic sensitivity of the CTD.** (a) Limited proteolysis of a mixture of apo and phospho CTD to compare protease accessibility. The two phospho isoforms run as separate bands. Results are from three representative experiments, from two separate phosphorylation reactions. (b) Average of three replicates of the apo (blue bars) and phospho (red bars) mixing experiments quantified as the percentage of intact CTD remaining at
each protease concentration. 100 to 50% intact CTD is the single hit kinetics range of the experiment(126). Error bars depict standard error of the mean.

3-2.5 The CTD stiffens as a function of phosphorylation

Intrinsically disordered proteins possess a high degree of conformational entropy due to the ensemble of structures they adopt. Binding of globular proteins to IDPs comes at the entropic cost of ordering a structurally dynamic sequence(127). Recently, theoretical modeling suggests changes in stiffness due to post translational modifications could be exploited by IDPs to tightly regulate binding(127). SAXS data can be used to make qualitative comparisons of the flexibility of proteins and to monitor flexibility changes after post-translational modification(104). The flexibility analyses, including Kratky, Porod-Debye and Kratky-Debye plots exploit differences in the way rigid and flexible proteins scatter X-rays to enable qualitative comparison of protein flexibility. The contrast between protein scattering and buffer scattering is more discrete for rigid proteins, and more continuous in flexible proteins where each amino acid has more freedom of movement(104, 128). These plots transform the data mathematically to reveal differences in flexibility that manifest themselves in different regions of the scattering curves. Kratky plots indicate that the CTD becomes less flexible when phosphorylated as evidenced by a reduction in the signal in the higher q range of the data(104) (Fig. 3-12a). To further validate this result I generated Porod-Debye and Kratky-Debye plots (Fig. 3-12 b and c) from the apo and phospho CTD scattering curves. These analyses truncate the high q
region of the data and thus are less prone to potential artifacts of buffer subtraction and generate plots with characteristic shapes indicative of flexible or more rigid proteins(104). The asymptotic rise in the Porod-Debye plot exhibited by the apo CTD curve is characteristic of flexible proteins and is diminished for the phospho CTD (Fig. 3-12b)(104). Further, the plateau observed for the apo CTD in the Kratky-Debye plot is lost for the phospho CTD (Fig. 3-12c). The divergent behavior of the apo and phospho CTD between the Kratky-Debye and Porod-Debye plots and the differences within each plot for the apo and phospho CTD are indicative of a phosphorylation-dependent decrease in flexibility.

**Figure 3-12 Phosphorylation stiffens the CTD.** (a) Kratky Plot ($q^2 \times I(q)$ vs $q$) of averaged apo (blue) and phospho (red) MBP-CTD scattering curves. The more gradual rise in the high $q$ region of the plot for the phospho-CTD than the apo-CTD indicates that the phospho-CTD is less flexible than the apo CTD. (b and c) Porod-Debye ($q^4 \times I(q)$ vs $q^4$) and Kratky Debye ($q^2 \times I(q)$ vs $q^2$) plots of the apo (blue) and phospho (red) scattering curves. The increased rise in the Porod-Debye plot for the apo-CTD compared to phospho-CTD indicates a phosphorylation dependent decrease in flexibility, as does the loss of the plateau in the Kratky-Debye plot for the phospho-CTD.
A potential caveat to Kratky plots is that they multiply scattering intensity (I) by the square of the scattering angle, q and in doing so can amplify incorrect buffer subtraction because the high q region of the Kratky plot is where flexibility changes are manifest, and the high q region of SAXS curves is the region with lower signal:noise(104). These concerns are minimized by Kratky-Debye and Porod-Debye plots, which operate on data truncated to lower scattering angles. Nevertheless, I sought further validation of phosphorylation induced stiffening that could serve to ensure my observation that the CTD stiffens holds across the entire q range. To this end, I generated scattering data extrapolated to zero concentration and repeated the Kratky analysis. This procedure takes scattering data collected across a range of concentrations that has not yet been buffer subtracted, and computes the protein and buffer contributions to the scattering curve(110). This algorithm provides a buffer subtracted scattering curve, as well as curve with the estimated scattering contribution of the buffer(109). The estimated buffer scattering curve from this extrapolation closely matches that of actual buffer scattering curves used for buffer subtraction for the apo and phospho CTD data (Fig. 3-13a). My observation that the CTD stiffens upon phosphorylation is supported through comparison of Kratky plots of buffer subtracted scattering curves and extrapolated scattering curves, which overlay across the q-range (Fig 3-13b). This suggests I accurately accounted for the scattering contribution of the buffer in the flexibility analyses and argues against inaccuracies in buffer subtraction giving rise to the appearance of stiffening.
Figure 3-13 Validation of phosphorylation dependent stiffening. (a) $I(q)$ vs. $q$ plot of averaged buffer signal used in buffer subtraction of apo and phospho scattering data plotted with buffer scattering curves calculated from the zero concentration extrapolation performed using SAMBUF(110). The measured and extrapolated buffer scattering curves overlay indicating that $I$ correctly accounted for and subtract the scattering contribution of the buffer in my analyses. (b) Kratky plots with error bars for averaged scattering curves and for scattering curves derived from zero concentration extrapolation using SAMBUF. Buffer subtracted averaged curves and buffer independent zero concentration extrapolation yields the same result. The agreement validates that the loss of flexibility resulting from phosphorylation is not an artifact of buffer subtraction.
3-2.5 The human CTD is also structurally heterogeneous

Despite a large disparity in the proportion of the fly and human CTD comprised of consensus heptads, 5% vs 41% respectively, my native gel and size exclusion chromatography results indicated gross structural similarity. This led me to consider if the human CTD may also share the local structural heterogeneity I identify in the fly CTD. I subjected the human CTD to limited proteolysis with proteinase K, predicted to cleave the human CTD at 122 locations (83). Surprisingly, I observed a similar pattern of protease hypersensitivity between fly and human, with the human CTD also exhibiting a hypersensitive site in the distal region, in addition to a hypersensitive site in the proximal region near where the collection of sensitive sites reside in the Drosophila CTD, suggesting that conservation in CTD structure between fly and human exists despite length and sequence differences (Fig. 3-14).
**Figure 3-14 The human CTD is structurally heterogeneous.**

The human CTD is radiolabeled on the final acidic repeat by casein kinase II (CK2) (uncut lane). TEV cleavage to separate the human CTD from the MBP fusion demarcates the point below which protease sensitivity occurs in the CTD portion of the protein (TEV lane). Limited proteolysis with proteinase K reveals a distal hypersensitive site reminiscent of that observed in the *Drosophila* CTD (compare to *D.mel*CTD ProK lane), a protease hypersensitive proximal site near MBP, where the collection of sensitive sites resides in the *Drosophila* CTD, and a central region that is largely protease insensitive. The

3-2.6 The Human CTD can function in place of the fly CTD in vivo.

The importance of the lineage specific non-consensus heptads that predominate in the CTDs of metazoans is a longstanding and open question in the field(5, 24, 26, 27, 129, 130). To directly test of the importance the conserved and lineage specific CTD sequence in a higher organism, we sought to
functionally replace Rpb1 in *Drosophila* with a derivative harboring the human CTD sequence, in every tissue and throughout development. To this end, we used RNAi to knock down endogenous Rpb1 to lethal levels and tested whether the coincident expression of RNAi-resistant versions of wild-type or humanized Rpb1 could rescue the lethality caused by ubiquitously expressing Rpb1 RNAi in the entire animal and throughout development (*Fig. 3-15*). These experiments are facilitated by dual transgenes under the control of the yeast transcriptional regulator Gal4. Flies lack Gal4, but Gal4 will induce the expression of downstream genes when the cognate motif is present in the promoter region. Expression of the *Gal4* gene under the control of the *Actin* gene promoter ensures organism-wide expression of Gal4 throughout development. The Gal4 protein in turn binds motifs upstream of two separate transgenes, one expressing Rpb1 RNAi, the other expressing an epitope tagged Rpb1 derivative rendered RNAi resistant using synonymous codon mutations. This system leads to the coincident expression of Rpb1 RNAi, and either the wild-type of humanized Rpb1 derivatives. Thus, if lethal levels of RNA are achieved, we can test whether Rpb1 harboring the human CTD can functionally replace the endogenous CTD sequence in a developmentally complex system.
Figure 3-15 Gal4 driven Rpb1 knockdown and expression system.

(a) The Gal4 gene is expressed under the control of the actin gene promoter, leading to expression of the Gal4 protein. (b) Promoters for two separate transgenes, one encoding Rpb1 RNAi, the other an epitope tagged Rpb1 derivative, both containing binding sites for the Gal4 protein. Flies harboring the transgene depicted in panel a are mated with those harboring the transgenes in panel b, leading to ubiquitous Rpb1 RNAi and ectopic Rpb1 expression.
This coupled knockdown and expression scheme is achieved in flies as follows: flies containing Gal4 maintained over a balancer chromosome containing a gene for curly wings are mated to flies harboring both RNAi against Rpb1 and an Rpb1 derivative. The wing phenotype is used to distinguish progeny expressing Gal4, and thus expressing Rpb1 RNAi and an Rpb1 derivative, from those lacking GAL4, and thus not expressing Rpb1 RNAi or an Rpb1 derivative. Flies lacking Gal4 have curly wings while progeny expressing GAL4, and thus expressing Rpb1 RNAi and an RNAi resistant derivative, have straight wings.

This scheme is depicted in figure 4-4. (Fig. 3-16). Lethal levels of RNAi not rescued by the ectopic expression of a wild type or humanized Rpb1 derivative is indicated by the absence of straight winged offspring. Full rescue of lethal levels of Rpb1 RNAi is indicated 50% of the offspring having straight wings.
Figure 3-16. Knockdown of endogenous Rpb1 and coincident expression of humanized Rpb1 in flies. Experimental design of the human CTD rescue experiment. UAS-Rpb1i is a GAL4 inducible transgene that produces an RNAi against endogenous Rpb1. UAS-Rpb1 is a GAL4 inducible transgene that encodes a FLAG-tagged, RNAi-resistant derivative of Rpb1. Fly images were created with Genotype Builder(131). The Actin-GAL4 gene is maintained over a balancer chromosome marked by curly wings (CyO). Offspring from this cross that have Actin-Gal4 will not have the CyO balancer, and will have straight wings. Straight winged flies express Gal4, which in turn drives expression of Rpb1 RNAi and an Rpb1 derivative, as depicted in figure 4-3.
As expected, approximately 50% of the progeny had straight wings when the Actin-GAL4 driver was crossed to yw control flies lacking Rpb1 RNAi (Fig. 3-17yw). In contrast, none of the progeny had straight wings when the Actin-GAL4 driver was combined with Rpb1 RNAi alone, indicating lethal levels of Rpb1 knockdown are achieved using this system (Fig. 3-17 Rpb1i). In addition, no straight-winged progeny were present when an RNAi-sensitive version of Drosophila Rpb1 was over-expressed along with the Rpb1 RNAi (Fig. 3-17 Rpb1[wt, sens], Rpb1i) indicating that the level of RNAi knockdown is robust enough to kill the flies even when Rpb1 is ectopically overexpressed in addition to endogenous levels. In stark contrast, approximately 50% of the progeny have straight wings when an RNAi-resistant form of Drosophila Rpb1 is present, indicating rescue of Rpb1 RNAi by the ectopically expressed derivative (Fig 3-17 Rpb1[wt, res], Rpb1i). Remarkably, the RNAi-resistant Rpb1 harboring the human CTD also rescues the lethality caused by the Rpb1 RNAi (Fig. 3-17 Rpb1[hu, Rpb1i]).
Figure 3-17 Results of RNAi rescue experiments with Rpb1 derivatives. Rpb1 harboring the human CTD supports fly development when wild-type Rpb1 is knocked down to lethal levels. Each bar on the histogram indicates the percentage of progeny with straight wings. yw: ActGAL4/CyO x yw (n=137 progeny); Rpb1i: ActGAL4/CyO x UAS-Rpb1i (n=86 progeny); Rpb1\(^{wt}\)[sens],Rpb1i: ActGAL4/CyO x UAS-Rpb1\(^{wt}\),UAS-Rpb1\(_{sen}\) (n=102 progeny); Rpb1\(^{wt}\)[res],Rpb1i: ActGAL4/CyO x UAS-Rpb1\(^{wt}\),UAS-Rpb1\(_{res}\) (n=91 progeny); Rpb1\(^{hu}\),Rpb1i: ActGAL4/CyO x UAS-Rpb1\(^{hu}\), UAS-Rpb1\(^{hu}\) (n=78 progeny). (Data generated by Feiyue Lu and David Gilmour)
The ectopically expressed wild-type and humanized Rpb1 derivatives are rendered RNAi resistant via a short string of synonymous codon mutations. This enables one to differentiate between endogenous and ectopically expressed mRNA using a primer complementary to either wild type sequence or the RNAi resistant sequence, respectively. To quantify the relative amounts of the Rpb1 variants, I repeated the RNAi rescue crosses and isolated RNA from mixtures of male and female straight-winged adult flies in triplicate and performed rtPCR with yw flies functioning as a control (Fig 3-18). My qPCR analysis of the straight winged progeny shows that the RNAi reduces the endogenous Rpb1 mRNA by about 20 fold relative to control flies (Fig. 3-18a) Moreover, comparable levels of each ectopically expressed FLAG-tagged Rpb1 are being expressed (Fig. 3-18b).

**Figure 3-18** Quantification of mRNA levels for Rpb1 derivatives in adult flies from RNAi rescue crosses. a. qPCR analysis of RNA isolated from straight winged adult progeny from the Rpb1 RNAi rescue experiments using primers specific for endogenous Rpb1 RNA. Genotypes of the flies are as follows: yw is
ActGAL4/+. Rpb1i,Rpb1\textsuperscript{wt} is ActGAL4/+;Rpb1i,Rpb1\textsuperscript{wt}/+. Rpb1i,Rpb1\textsuperscript{hu} is ActGAL4/+; Rpb1i,Rpb1\textsuperscript{hu}/+. b. qPCR as in panel a, using primers specific for the ectopically expressed, RNAi resistant Rpb1 RNA. The sequence of the ectopically expressed Rpb1 RNA differs from the endogenous version by a 21 nucleotide region that contains the synonymous mutations which render it resistant to the RNAi.

I next used immunostaining of polytene chromosomes from third instar larvae to further validate that the humanized Rpb1 is functioning as expected. The ectopically expressed RNAi resistant Rpb1 derivatives carry a C-terminal FLAG tag. This epitope tag allows me to specifically probe for the ectopically expressed Rpb1 variants and ensure the protein is localized to chromatin. The larval stage precedes wing development and the wing phenotype is what allowed me to identify rescue individuals, so immunostaining serves as a blinded control that the wing phenotype is reporting functional humanized Rpb1. Because only ~50% of offspring from our RNAi rescue crosses are expressing FLAG-tagged Rpb1 derivatives and RNAi, immunostaining chromosomes obtained from multiple larvae generates instances where chromosomes staining positively for FLAG-Rpb1 were located next to chromosomes from individuals not expressing the FLAG-Rpb1 and Rpb1 RNAi. The latter fail to stain with FLAG antibody but are visible using DNA stain (Fig. 3-19). The positive FLAG staining shows that the ectopically expressed Rpb1 derivatives harboring either the wt or the human CTD associate with chromosomes (Fig. 3-19). The lack of FLAG staining on chromosomes visible with DNA stain arise from those larvae destined to be curly winged adults, lacking Actin-Gal4.
**Figure 3-19 Immunofluorescence staining of polytene chromosomes detects ectopically expressed Rpb1 on chromosomes.** Immunofluorescence of polytene chromosomes from salivary glands of third instar larvae. Larvae were derived from mating ActGAL4/CyO and UAS-Rpb1i, UAS-Rpb1 parents, so half of the larvae ectopically expressed *Drosophila* or humanized Rpb1. Left and right panels depict identical chromosomes for the given genotype. Slides were stained with Hoescht dye (DNA stain, left panels) and with antibodies against the FLAG tag (right panels). Arrows point to chromosomes that fail to stain with FLAG antibody.
Higher magnification images of chromosomes staining positively for FLAG-Rpb1 show that the ectopically expressed wild-type and humanized Rpb1 co-localize on chromosomes with the Rpb3 subunit of Pol II (Fig. 3-20) Together, these data indicate that that ectopically expressed Rpb1 with the human CTD is expressed, co-localizes on chromosomes with another subunit of Pol II, and is able to support growth and development when endogenous Rpb1 is knocked down by ~95% to lethal levels. This conclusion was further validated (by D.S.G.) by an orthogonal, RNAi independent, cross that showed humanized Rpb1 was able to complement an early-embryonic lethal allele of Rpb1(87).

**Figure 3-20** Ectopically expressed Rpb1 derivatives co-localize with the Rpb3 subunit of Pol II across chromosomes. High magnification images of polytene chromosomes double stained with Rpb3 (green) and FLAG (red) antibodies. The Rpb3 and FLAG-Rpb1 subunits of RNA Pol II co-localized across chromosomes for both the wild type (left panel) and humanized (right panel) Rpb1 derivatives.
3-3 Discussion

Biophysical analysis of the monomeric CTD, free in solution, yielded measurements consistent with a semi-compact protein, capable of undergoing extension as a function of phosphorylation. I then sought to determine if the CTD was iteratively structured, as β-turn models of the CTD predict, or if the CTD possessed a non-uniform organization across its length. In the latter case, a semi-compact CTD could result from compact regions combined with more extended regions. A non-uniform structure could have profound implications on the CTD code, preferentially rendering regions of the CTD more accessible to protein binding or post-translational modifications.

Using a series of proteases as protein scale probes, I interrogated the structure of the CTD and found a pattern of sensitivity and protection I interpret as evidence of a heterogeneous CTD structure containing regions with varying degrees of disorder. I favor this interpretation over the alternative hypothesis that variation in the amino acid sequence of the CTD gives rise to discontinuous proteolysis. First, three separate proteases with variations in structure and specificity all recapitulate a similar but not identical pattern of proteolysis, and these shared patterns result from proteolysis at only a subset of the total cleavage sites predicted by the specificity of each enzyme (Fig. 3-8). This suggests that each protease first interacts with the accessible regions of the CTD, before cleaving at subtly different sites therein as a function of their individual specificities. This observation also argues against the caveat that contaminants in each commercial preparation of protease give rise to the pattern
of proteolysis as the contaminants would need to be distinct yet share similar specificities. Second, the denaturant SDS, known to stimulate proteolysis through the disruption of protein structure, alters the relative pattern of proteolysis of the CTD by proteinase K (Fig. 3-9)(125). That some elements of the proteolytic sensitivity are altered by a denaturant suggests I am detecting a bona fide structural phenomenon. Finally, despite differences in primary amino acid sequence, the human CTD, like the fly CTD, also displays a discontinuous pattern of proteolysis with a hypersensitive site in the distal region of the CTD and a proximal protease sensitive region, suggesting structural heterogeneity is a feature common to both organisms (Fig. 3-14). This finding demands a new way of thinking about the CTD.

The CTD has been modeled as a repetitious spiral and is often depicted in the literature as a random, structure-less, noodle(27, 64). The consequence of either conceptualization is that the CTD structure is reduced to and explained predominantly by primary structure; a linear array of sequence to be “read” by interacting factors. Such a linear code provides little support for the spatial organization of multiple complexes with respect to the CTD. Any such organization would likely require the co-evolution of both the CTD and each individual factor benefitting from organization relative to its neighbor. In contrast, structural heterogeneity supports the hypothesis that different regions of the CTD are differentially structured and these differences could target factors to specific regions of the CTD through multiple mechanisms. The first is accessibility. A factor capable of binding to multiple CTD motifs with equivalent avidity will
preferentially bind those most accessible. Another mechanism could be structure preformation. Here, certain CTD structures that exist upon binding may be more readily accommodated, or even pre-selected among the ensemble, in certain regions of the CTD compared with other regions (54, 55, 59). The discovery of structural heterogeneity in the CTD of both *Drosophila* and human establishes for the first time the possibility of structure mediated CTD:factor interactions as a possible means of organizing the CTD interactome.

Recently, limited proteolysis was extended cell-wide to large segments of the yeast proteome (123). These experiments were done in lysates across a range of temperatures, resulting in thermal unfolding causing increased protease accessibility at elevated temperatures for globular proteins. Surprisingly, nearly half the proteins computationally predicted to be intrinsically disordered displayed increased proteolysis at elevated temperature, suggesting they contain some structure (123). Still others became less protease sensitive at high temperatures, which the authors interpret as resulting from aggregation after thermal unfolding (123). Where the CTDs of various metazoans falls on the order:disorder spectrum, and how their CTD structures relate to their optimum growth, is another interesting area of future work that could shed light on whether or not CTD structural heterogeneity is a common and adaptive feature in metazoans.

I characterized the global and local impacts of phosphorylation on CTD structure. Phosphorylation leads to extension of the CTD that is modest as a percentage of its initial size (Figs. 2-4 and 2-13), which results in modestly increased protein accessibility as measured by a comparison of proteolytic
sensitivity between the apo and phospho protein (Figs. 3-10 and 3-11). Meanwhile, local structural heterogeneity is maintained (Fig. 3-10). This suggests that the CTD does not simply “unfold” to become uniformly accessible to protein interactions in response to the biologically relevant phosphate density achieved in my studies. It is notable that the degree of extension observed after S5 phosphorylation is small relative to both the size of globular proteins with which the CTD interacts and also the proteinase K used in my studies which has an $R_g$ of ~17Å (112, 133). Nevertheless, this modest extension resulted in increased protease accessibility. $R_g$ approximates the protein under study as a sphere rotating about its center of mass. Considering the CTD as a sphere, modest changes in $R_g$ lead to larger change in volume and surface area via the equations $V=\frac{4}{3}\pi r^3$ and $A=4\pi r^2$, respectively. To what extent further phosphorylation, or different patterns of phosphorylation arising from different enzymes, would alter the CTD structure would be an exciting avenue for future experiments.

Analysis of the SAXS data indicates that the phosphorylated CTD is less flexible than the unphosphorylated CTD (Figs. 3-12 and 3-13) This observation is interesting in the context of those CID:CTD co-crystal structures of factors that favor binding phosphorylated CTD peptides without directly recognizing the phosphate via hydrogen bonding (39). Collectively, my data on the phospho CTD support a model for CTD binding in which multiple effects of phosphorylation may in concert facilitate factor binding. First, the CTD transitions from a polyampholyte to a polyanion, altering its affinity for factors, potentially as a
function of phosphate density (100). Extension de-protects the CTD, making it more uniformly accessible to binding, without disrupting local structural heterogeneity or preformation. A less flexible CTD with reduced conformational entropy may provide a more binding-competent dock by depleting the CTD ensemble of conformers incompatible with binding, thereby reducing the entropic penalty for binding (127). A compelling paradigm for phosphoCTD recognition may be the phospho-protein Sic1, which like the CTD is compact, intrinsically disordered, comprised of a repetitive and low complexity sequence, and is subject to phosphorylation at multiple repeats giving rise to multivalent binding (100, 127). Cdc4 can recognize phospho Sic1 on any one of multiple phosphorylated repeats, but only after a certain threshold of phosphorylation is reached that achieves adequate charge density (100) and possibly reduced conformational entropy does binding occur. This switch-like binding phenomenon is referred to as binding “ultrasensitivity”, and is characterized by extremely high Hill coefficients (127). These changes can result in “switch-like” binding after the phosphorylation threshold, and could be a mechanism by which an extended CTD is able to protect itself from premature interactions with particular factors. This model is compelling in part because it exploits the energy release from ATP transesterification to “pre-pay” the entropic penalty for binding. In this model, the high entropic cost associated with imposing local order on a dynamic structure like the CTD is sufficiently high to prevent binding until a net threshold of stiffness, a reduction in conformational entropy, is reached. It is notable that I observed stiffening of the CTD in response to S5 phosphorylation that
approximates the physiological density observed in yeast and mammalian cells in vivo (17, 18). Because S₅ phosphorylation predominates early in elongation, this stiffening could help prepare the CTD to bind the bevy of factors that associate with polymerase in a phosphorylation dependent manner throughout transcription (2, 6). The flexible CTD, hypophosphorylated, prior to initiation, would be predicted to be recalcitrant to the binding of multiple factors (75, 127). If different patterns of phosphorylation (i.e. S₂) also decrease flexibility, the CTD could maintain a binding competent status throughout elongation, with interactomes dictated by stereospecific recognition of particular patterns of phosphorylation. Thus, charge and stiffness changes in the CTD could tip the balance in favor of binding and explain how a CTD dynamically sampling extended structures avoids unregulated modification and premature binding in the presence of hundreds of interacting proteins in the nuclear compartment. Future SAXs experiments, or orthogonal experiments aimed at qualitatively comparing flexibility among CTDs phosphorylated by different CTD kinases, would be a compelling avenue of further study. In addition to SAXS, FRET may be a bench-top approach to interrogating changes in CTD stiffness. The MBP-\textit{D.mel}ICTD protein employed in my studies contains only one cysteine residue. This residue could be mutated, and new cysteines introduced at specific sites along the CTD could be used to conjugate thiol reactive fluorophores to generate FRET pairs. A loss of FRET efficiency as a function of phosphorylation would be predicted to result from reduced flexibility, as fluorophores would be less likely to enter FRET distances on an increasingly rigid molecule. Caution needs to be
taken in interpreting data from these experiments, as extension could also contribute to reduced FRET efficiency.

The comparative evolution of the CTD has been an active area of study for nearly 30 years (5, 24, 26, 27, 129, 132, 133). These studies were aimed in part at determining the functional units of the CTD. Extensive genetics in yeast, flies, mammalian cell culture, and mice, also sought to determine the function of positions within the heptad repeat as well as the minimum CTD length required to support viability (9, 44, 48, 86). Two central hypotheses emerge from these studies. First, some minimum CTD length is required to support essential cellular functions, with additional length required to survive stress such as heat shock (9). Second, lineage specific modifications arose and were selected to organize CTD:factor interactions, presumably through the co-evolution of CTD binding factors (24, 134).

The finding that fly and human CTD exhibited gross structural homology and similar patterns of proteolysis lead us to test the second of these long standing hypotheses: that the co-evolution of non-consensus heptads and CTD binding proteins imparts lineage specific functionality, via added specificity of CTD binding interactions (24, 26). Such specificity could in theory serve to organize multiple proteins on a single CTD or modulate binding specificities among competing factors.

Our ability to replace the D. melanogaster CTD sequence, which is highly conserved among Drosophila species, with the human CTD sequence bearing different heptads but a similar structure, suggests the CTD sequence is not
organizing binding interactions in an essential way, even in a developmentally complex context (Fig. 3-1). Instead, I posit that structural organization may serve to preferentially expose particular regions of the CTD that in turn could control the placement of post-translational modifications and organize binding interactions. In this model, different amino acid substitutions to consensus CTD SLiMs could yield similar adaptive structural variation to the CTD while continuing to permit recognition by factors. Different factors may be able to “read” heptads of varying sequences, but preferential binding to some heptads is mediated by variations in accessibility or pre-formed structural features across the CTD.

CTD:factor interactions may lie on a continuum between preferred and accessible sequence motifs, with factors “fitting” into accessible sub-regions of the CTD, then “reading” their preferred motif and register, therein. My limited proteolysis results exemplify this sequence-follows-structure recognition. All three protease recognize similar regions of the CTD, but cleave subtly different sites based on their intrinsic sequence specificities (Fig. 3-8). Presumably, different factors could evolve to interact with the CTD differently, with some exhibiting an overriding preference for a specific motif, such as a consensus heptad, while other factors may prefer structural preformation inherent to particular CTD regions. Still other factors may be capable of altering the structure of the CTD by binding to less accessible CTD regions which could in turn modulate downstream binding events. Indeed, more work needs to be done to characterize the determinants of binding for discrete CTD:factor interactions. By uncovering conserved structural heterogeneity, and by illustrating that complex
developmental programs withstand wholesale alteration of a conserved CTD sequence, this work provides a novel framework to conceptualize the nature of CTD:factor interactions on a sequence:structure continuum.

It warrants mention that the human CTD swap experiments also change the length of the CTD, as well as the sequence composition. It is possible that the human CTD contains a sub-optimal array of SLiMs for the purpose of recruiting Drosophila factors, but that this reduction in binding affinity is compensated by additional valency. If so, one implication would be that factors can function when bound to different locations along the CTD and would argue against models depicting multiple CTD interacting complexes binding the CTD in an organized way with respect to one another, with binding locations driven by variation in local CTD sequence. The differences between the fly and human CTD argue against this level of sequence driven organization.

It is important to consider the effect of increased valency in the context of my observation that the CTD stiffens upon phosphorylation. Swain and Lentz modeled the cumulative effects of local stiffening on the conformational entropy of a disordered protein upon phosphorylation. The model suggests binding affinity of a globular protein could be affected in a highly non-linear manner as the valency of a repetitive IDP increases. If stiffening upon phosphorylation does in fact modulate CTD:factor interactions as predicted by the model, the implications for the human CTD swap experiments could be significant. The human CTD could in fact present heptads that are vastly inferior binding sites for fly proteins, but this loss of binding affinity could be compensated by the
increased length of the human CTD and the non-linear reduction in conformational entropy upon phosphorylation this additional valency provides. Teasing apart the precise contribution of length, sequence and structure is difficult, because the human CTD contains many more consensus repeats that may be better binding sites. Mutants of varying lengths that contain the same number of consensus repeats as the native fly sequence, but otherwise contain non-native heptads would be best suited to dissecting length and sequence contributions to CTD function.

In contrast to sequence encoded binding organization, the structure of the CTD could have evolved to drive binding to specific parts of the CTD, thus organizing complex binding with respect to the polymerase itself. This model is perhaps better supported by the similarity between the limited proteolysis results from the fly and human CTD, and the observation that the region of the CTD most protease insensitive in both cases, is also the most well conserved between fly and human (Fig. 3-21).
Figure 3-21 The central, protease resistant region of the Drosophila CTD is conserved between Drosophila and human. a. Alignment of the D. melanogaster and H. sapiens CTDs. Black vertical bars indicate identical amino acids, white vertical bars indicate non-identical amino acids, and black horizontal lines in the Drosophila diagram indicate gaps in the alignment. (Adapted from Gibbs et al 2017)(53) b. Scale diagram of the Drosophila CTD indicating the location of TEV protease sites and the resulting MW of TEV cleavage products used for comparison of the limited proteolysis fragments. c. The central region of the Drosophila CTD, between TEV fragments with masses of 27kD and 10kD is largely protease insensitive, and corresponds to a region of the CTD that is well conserved between fly and human.

Mutants that alter the sequence composition and structural organization of the CTD could be used to understand the relative importance of conserved sequence elements and their positioning along the CTD. In future work, CTD mutants that result in lethality or phenotypic variation in vivo could be examined biochemically with limited proteolysis, and compared to mutants that rescue RNAi lethality or lethal alleles.
It also remains possible that a precise array of heptads and structural organization is crucial to CTD function under certain stress conditions not recapitulated in laboratory growth conditions. Notably, dozens of low complexity proteins, the majority of which function in transcription regulation and/or via nucleic acid binding, have recently been shown to adopt rare conformations that elicit distinct, heritable transcriptional outcomes in yeast (134). These cryptic structural variants remained “hidden in plain sight” in the unstructured domains of the best studied model of eukaryotic transcription, largely due to their specialized roles which are selected against in laboratory growth conditions (134). Thus, testing the ability of CTD mutants that alter structural organization to respond to heat shock may be an interesting avenue for future experiments relating the structure and sequence of the CTD.
CHAPTER 4

THE LINKER IS A DISTINCT STRUCTURAL UNIT OF THE CTD

4-1 Introduction

The heptad repeats of the CTD are connected to Pol II via a domain known as the linker domain, which is presumed to be flexible (2). A comprehensive review on the Pol II CTD written by Jeffrey Corden in 2013 cites over 350 references pertaining to the evolution, structure, and function of the CTD in transcription regulation (2). In a sense, the study of the CTD has become a sub-field of gene regulation unto itself. However, to the best of my knowledge, the number of references that address the linker region connecting the CTD to the rest of Rpb1, can be counted on one hand. Nevertheless, the work that has been done on the linker domain suggests an important role for this region in supporting Pol II function and cell viability.

Figure 4-1 A diagram of CTD domains. The CTD is affixed to the core of Poll II via the linker region, which emanates from the Rpb1 subunit near where the
Rpb1 4/7 stalk meet the core of Pol II, which is also close to the RNA exit channel(114). The linker domain does not contain the heptad repeats, but is enriched for similar amino acids. The distal end of the CTD contains a tip that deviates from the other heptads. The linker region is 120 amino acids in length, the heptad repeats and acidic tip total 309 amino acids.

Lis and colleagues showed the linker region is important for transcription activation in yeast(135). The linker can be subdivided into a region proximal to the core of Pol II beginning in the distal part of a region known as homology box H. This region of Box H is highly conserved among yeast, flies and mammals and is followed by a less conserved region that connects to the heptad repeats(135). In the work from Lis and colleagues, the authors noted the linker region in the yeast CTD is enriched for both acidic and aromatic amino acids, a characteristic shared with the acidic activation domains of some transcription factors(135). Interestingly, the acidic region of the linker domain activated transcription of a reporter gene when fused to the GAL4 DNA binding domain. Deletion of the acidic linker region from Pol II lead to slow growth and temperature sensitivity, and this phenotype could be rescued by replacing the linker region of Rpb1 with the acidic activation domain of the transcription factor VP16, but not with other amino acid sequences that fail to function as transcriptional activators when fused to the GAL4 DNA binding domain(135). The requirement that specific sequence features be present in the linker region to support normal cell growth illustrates the linker serves a role in addition to simply being a spacer that orients the heptad repeats relative to the core of Pol II(135). In further support of this
conclusion is the observation mutations in the acidic region of the linker could suppress or enhance phenotypes associated with CTD truncations(135). The authors used the Gal4-linker domain fusions to drive expression of a reporter gene with this experiment serving as a platform to screen for regions of the linker that are potentially important in the context of Rpb1. Removal of a portion of the acidic domain from the linker reduced transcription of the reporter. When this mutant was placed in the context of Rpb1 a slow growth phenotype was observed(135). Notably a CTD of 35 heptad repeats could complement this slow-growth linker mutant and restore normal growth(135). This is notable, as 35 heptads is ~1/3 longer than the length of the wild-type yeast CTD. That additional CTD length can compensate for a deleterious linker mutant hints at some functional redundancy between the linker and the heptad repeats. Moreover, a point mutant that enhanced transcription in the reporter gene assay was shown to support normal growth in the context of Rpb1 harboring a CTD of less than 20 heptads, whereas a point mutant that reduced transcription in the reporter gene assay required a longer CTD to support normal growth.

The region of the yeast linker that functions most potently as a transcriptional activator includes residues downstream of box H that are conserved between *Drosophila* and human but not yeast (Fig. 4-2a). This is interesting, as the region is more acidic in yeast than in fly and human, and this acidic bias contributes to function. Linker sequences that functioned as more potent transcriptional activators required fewer heptad repeats to support normal growth. This lead the authors to speculate that the absence of the extended
acidic region in the linker of fly and human necessitates their longer CTDs relative to yeast (135). The nature of the functional relationship between the linker region and the heptad repeats, and whether or not there is a structural interaction between the two regions, remains unexplored.

A second, more recent, paper affirms a functional role for the linker domain. In this work, Buratowski and colleagues append the CTD repeats to other Pol II subunits and found that the CTD can function on Rpb4 or Rpb6, which positions the CTD close to its normal location, but not Rpb9, which positions the CTD on the opposite side of Pol II (136). This lead the authors to speculate that the CTD must be near its normal location to support polymerase function (136). This observation is consistent with a semi-compact CTD with a radius of access that is limited enough to require the CTD to be positioned on the appropriate face of Pol II (136). Unlike the heptad repeats, the linker region was required to be on Rpb1 to support Pol II function (17). The linker and CTD together could not function on other subunits if the linker was not also present on Rpb1 (17). Together, the data support a requirement for the linker to be included in Rpb1, and for the heptad repeats to be proximal to their normal location, and thus also the linker. Thus, there are spatial constraints on both the CTD and linker that may or may not be separable, and may relate to the radius of access for the CTD. While the CTD itself is not folded, the many factors it recruits are. As such, depending on the location of the CTD, interacting factors may be sterically occluded from accessing chromatin or the nascent RNA on which they operate, either by the polymerase itself or other components of the elongation
machinery. Additionally, the functional interaction between the linker and heptad repeats could be related to a structural interaction. For example, it is conceivable that the linker, which bears resemblance to transcriptional activation domains, has some affinity for the heptad repeats, protecting them from modification or factor recruitment until this interaction is replaced by interactions between the linker and initiation machinery. Such interplay could serve to protect the CTD until the polymerase is in the appropriate context. Alternatively, the linker region could be structurally distinct and independent from the CTD, and the functional interaction could result from the bimodal and simultaneous recognition of both the linker and heptad repeats by an essential factor. Finally, the functional interaction could be completely independent of structure, with the heptad repeats sharing some redundant function with the linker, such that deleterious linker mutants are complemented by additional heptad repeats.

The functional importance of the linker in supporting normal cell growth motivated me to interrogate the impacts of the linker region on CTD structure. Using the array of biophysical and biochemical techniques previously brought to bear on the CTD; native gel electrophoresis, size exclusion chromatography, small angle X-ray scattering, and limited proteolysis, I characterized the impact of the linker on CTD structure. I found that the linker containing CTD is more compact than would be predicted if the linker were structured similarly to the CTD. I show that the linker is structurally independent from the heptad repeats, which suggests the increased compaction of the linker containing CTD stems from the compact nature of the linker itself. Finally, the linker buffers the overall
linker containing CTD from stiffening. Together, these data support a model for a structurally independent linker region, that serves as a “hinge” between the core of Pol II and a stiffened phosphorylated CTD. This hinge in turn could maintain a radius of access, allowing factors bound to a stiffened phospho-CTD to access a large three-dimensional space to capture their RNA and chromatin substrates as Pol II traverses a gene. These data represent the first structural interrogation of the monomeric CTD in solution that contains the functionally important linker domain that is conserved in fly and human.
4-2 Results

4-2.1 Defining the linker

In order to determine the structural impacts of the linker region on the CTD, I first needed to define the linker region in *Drosophila*. To this end, I used MUSCLE to align the amino acid sequences of Rbp1 from *S. cerevisae*, *D. melanogaster*, and *H. sapiens* (products of *RPO21*, *RpII215* and *POLR2A*, respectively)(137). In earlier X-ray crystallography studies on the yeast Pol II, the linker was defined as beginning with amino acid 1436, which is upstream of the last structured amino acid in the final structural model, residue 1450(4). In the aforementioned MUSCLE alignment, residue 1436 of *S. cerevisae* Rpb1 corresponds to residue 1459 of the *Drosophila melanogaster* protein. Thus, I defined the *Drosophila* linker as beginning with residue 1459, and generated an MBP-CTD fusion protein that fuses all residues C-terminal to 1458 to MBP (Fig 4-2a).

Comparison of the linker sequences from the three species, and examination of the *Drosophila* linker sequence composition leads to a few notable observations. First, the acidic homology box H is well conserved among all three species, as has been noted previously(135). However, the region enriched in acidic residues is shorter in fly and human compared to yeast. Where the acidic region continues in yeast, an interruption in the three-way sequence comparison is created with the human and fly linkers sharing more homology with one another than with yeast, and containing fewer acidic residues (Fig 4-2a). Examination of the *Drosophila* linker sequence reveals an enrichment in amino acids in the linker region that are also enriched in the heptad repeat
region. The consensus heptad repeat contains amino acids Y,S,P and T. A comparison of the percentage of each of these residues in the sequence composition of the heptad repeat region (defined as C-terminal to the end of the linker, omitting the acidic tip region) and the linker region show similar amounts of S, P and T exist in both the linker and repeat regions. The linker is depleted of tyrosine residues compared to the heptad repeats, but contains similar amounts proline which is underrepresented in helices, particularly downstream of homology box H, and is thus predicted to be disordered (Fig 4-2b). Overall, the linker region of fly and human are longer than yeast, more conserved with one another than with yeast, and share enrichment in S,P and T residues with the heptad repeat region (Fig. 4-2). The homology between the fly and human linker sequences led me to apply the biophysical and biochemical techniques to the CTD including the linker that I previously applied to the CTD(87).
Figure 4-2 Defining the linker region in *D. melanogaster*. a. Section of the MUSCLE alignment of the amino acid sequences of the largest subunit of yeast, fly and human Pol II(137). The start of the linker is defined by sequence homology to start of the linker in yeast, previously defined by Cramer and colleagues(4). Numbers at the beginning and end of each sequence correspond to the amino acid position for each species included in the region of the alignment shown. The linker ends upstream of the heptad repeat region. Gaps in the alignment occur where greater homology exists between fly and human than between those species and yeast. b. Comparison of the amino acid composition of the linker and heptad repeat regions as the percentage of each region corresponding to those residues, Y, S, P, and T, which make up the consensus CTD heptad. The linker and heptads are similarly enriched for S, P and T residues.
4-2.2 The linker CTD is more compact than predicted

Fusing the CTD to the globular and acidic MBP aids in migration in a native-PAGE gel. Previously, I characterized the relationship between the migration of the S. cerevisae, D. melanogaster, and H. sapiens MPB-CTD fusions in a native gel. These three CTD variants migrate in a native gel in a manner that scales with their molecular weight and suggests gross structural homology between the three species' CTDs. The addition of the linker region to the fly CTD generates a MBP-CTD fusion protein that is larger in MW than that of the human CTD fusion. The similar sequence composition of the linker and heptad repeat regions predicts that the MBP-D.melCTD fusion containing the linker (MBP-D.melCTD w/linker) should migrate in the native gel similarly to those CTD fusions lacking the linker. As such, the MBP-D.melCTD w/Linker would be predicted to run above the MBP-H.sapCTD fusion in a native gel. Surprisingly, this was not observed. The MBP-D.melCTD w/Linker migrates as a band only slightly larger in size than the fly CTD fusion lacking the linker, and runs as though it were smaller in size than the human CTD fusion (Fig. 4-3). One possible explanation for this behavior is the addition of the linker region, which contains the acidic portion at the distal end of homology box H, causes the overall MBP-D.melCTDw/linker to migrate faster due to the additional electronegative charge. However, this is unlikely, as the pl of MBP-D.melCTD w/linker is 5.77, which is only slightly less than to that of MBP-D.melCTD and MBP-H.sapCTD, 5.83 and 5.93, respectively (Fig. 4-3). The addition of more heptad repeat region as the CTD grows in length from yeast, to fly , to human
causes a similar increase in size and decreased mobility on a native gel (Fig. 4-3). Conversely, the addition of the linker region breaks this trend, despite a similar amino acid composition to the repeat region, particularly downstream of box H. This data suggests the MBP-\textit{D.mel}/CTD w/linker is relatively more compact than would be predicted if the linker was structured similarly to the heptad repeat region.

\textbf{MBP-CTD Fusion} \textit{S. cer.} \textit{D. mel.} \textit{H. sap.} \textit{D. mel. w/Linker}

<table>
<thead>
<tr>
<th>M.W. (kD)</th>
<th>62.8</th>
<th>74.7</th>
<th>81.8</th>
<th>86.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>pI</td>
<td>5.57</td>
<td>5.83</td>
<td>5.93</td>
<td>5.77</td>
</tr>
</tbody>
</table>

6\% Native PAGE Coomassie

\textbf{Figure 4-3} Native gel electrophoresis of the linker containing CTD compared to CTDs lacking the linker. The MBP-\textit{D.mel}/CTD w/linker protein is larger that the MBP-\textit{H.sap}/CTD protein yet migrates in the native-PAGE gel as a smaller protein, comparable in size to MBP-\textit{D.mel}/CTD without the linker.
Clear interpretation of the native gel electrophoresis result is stymied by the, albeit slight, additional negative charge imparted by the linker relative to the heptad repeat region. This lead me to directly compare the size of the MBP-\textit{D.mel}/CTD w/linker with the other MBP-CTD fusions using size exclusion chromatography, a technique that resolves proteins in solution by shape that is agnostic to the charge of the protein under study. Previously, I compared the size ($R_s$) of the yeast, fly and human CTDs as MBP fusions using size exclusion chromatography, which revealed that a linear relationship exists between the various CTDs. In doing so, I created a CTD standard curve relating $R_s$ to MW (Fig. 2-3). This curve predicts that a new CTD fusion containing additional mass structured similarly to the heptad repeat region would have a relationship between size and shape that is co-linear with the CTD fusions. The addition of amino acids that were more extended than the heptad repeats would be plotted above the CTD standard curve. Conversely, the addition of amino acids adopting a structure more compact than the heptad repeats would fall below the CTD standard curve for $R_s$ vs. MW. Thus, the CTD standard curve can be used to compare the relative compaction of the linker region vs. the heptad repeats. As predicted by the native-PAGE results, the MBP-\textit{D.mel}/CTD w/linker falls below the CTD standard curve in the SEC experiments (Fig. 4-4). This suggests the linker region is more compact than the heptad repeats, despite similar amino acid composition, and provides an independent validation of the native-PAGE result.
Figure 4-4 The CTD with the linker region is more compact than predicted. The black line is the linear fit of the CTD standard curve (y=0.5903x + 6.3568). If the additional MW contributed by the linker region was similarly compact to the heptad repeats, the predicted $R_s$ of the MBP-\textit{D. mel}CTD w/linker would be 57.4Å, larger than the actual measures $R_s$ of 50.8Å.
4-2.3 The linker is a structurally independent unit of the CTD

The behavior of the MBP-\textit{D. mel}CTDw/linker compared the MBP-CTD fusions lacking the linker in both native-PAGE and size exclusion chromatography experiments suggests the linker region adopts a structure more compact than would be predicted by the addition of heptad repeats of the same MW. However, alternative hypotheses exist that would give rise to similar observations in these experiments. The linker region could be physically interacting with the heptad repeat region in such a way that drives additional compaction to some or all of the CTD. Similarly, the addition of the linker could alter the CTD ensemble and bias it toward more compact conformations. In these models, the addition of the linker would alter the structure of the CTD heptad repeat region. Alternatively, the linker itself could be more compact than the CTD heptad repeats, behaving as a structurally independent unit that does not alter the conformation of the heptad repeats.

Previously, I employed limited proteolysis to interrogate the structure of the CTD(87). This experiment uses protease accessibility as a proxy for local compaction. By using a protein as a structural probe, limited proteolysis experiments are ideally suited to interrogate the structure of IDPs that function as protein scaffolds. I sought to use limited proteolysis of MBP-\textit{D. mel}CTD w/linker to biochemically distinguish between the two models for CTD compaction that are congruent with my biophysical observations. If the linker were driving the compaction of the entire CTD, or biasing the ensemble adopted by the heptad
repeats, the limited proteolysis pattern in the repeat region would be predicted to differ from the pattern created by cleaving the CTD lacking the linker.

Alternatively, if the linker were a more compact, independent structural unit of the CTD, the pattern of proteolysis in the heptad repeat region would be unchanged by the addition of the linker. Furthermore, if the linker were a more compact structural unit than the CTD, the linker region would be predicted to be relatively protease insensitive compared with the more extended heptad repeat region.

Proteinase K is predicted to cleave the *Drosophila* heptad repeat region at 100 sites distributed nearly equally across 309 amino acids, roughly one potential cleavage site every three amino acids (83). Likewise, the 120 amino acid linker region contains 39 predicted cleavage sites, roughly one potential cleavage site for every three amino acids (83) ([Fig. 4-5a](#)). Thus, proteinase K can be used in limited proteolysis experiments to differentiate between competing models explaining the influence of the linker on CTD structure.

The MBP-*D. mel*CTD w/linker was subject to limited proteolysis with Proteinase K and the products were resolved in a SDS-PAGE gel. Comparison of the mobility of the cleavage products to that of CTD ladder fragments generated using the site specific TEV protease, and by comparison to the full-length uncleaved MBP-*D. mel*CTD w/linker, enable mapping sites of protease sensitive to the linker region or heptad repeat region. Proteolytic fragments running below the intact MBP-*D. mel*CTD w/linker, but above the top band of the CTD ladder which contains the full heptad repeat region, indicate cleavage within the linker region.
Conversely, cleavage below the top band of the CTD ladder indicates cleavage within the heptad repeats.

When subjected to limited proteolysis by Proteinase K, MBP-\textit{D.mel}CTD w/linker generates a pattern on bands strikingly similar to that of MBP-\textit{D.mel}CTD lacking the linker. Preserved in the context of the linker are the hypersensitive site running near the 10kD CTD ladder band, the largely protease resistant central portion of the CTD between the 10 and 27kD ladder bands that is cut at only one location, and the pattern of protease sensitivity on the proximal side of the heptad repeats (\textbf{Fig. 4-5b}). Two new proteinase K sensitive sites exist in the MBP-\textit{D.mel}CTD w/linker that map at or near the junction between the linker and heptad repeat region, as evidenced by their mobility close to the full-length 35kD CTD ladder band. Interestingly, the region between these new protease sites and the intact MBP-\textit{D.mel}CTD w/linker is devoid of proteolytic fragments, suggesting the linker is less protease accessible than the CTD (\textbf{Fig. 4-5b}). This result suggests the linker is more compact than the heptad repeat region. One possible interpretation is that the linker region is simply sterically occluded from proteolysis by proximity to MBP. I argue this is unlikely as the proximal heptads of the Drosophila CTD are proteolytically labile in the absence of the linker, and the proximal region of the human CTD is protease hypersensitive, despite their proximity to MBP. Proteolysis occurs in the linker region at or near the junction with the heptad repeats, where the linker sequence bears the most similarity to the heptad repeats in terms of amino acid composition, specifically the enrichment for S and P residues and SP motifs. The lack of sequence repetition
in the linker region made proteinase K the most appropriate choice of enzyme for limited proteolysis, but limited my ability to use other enzymes as a control for proteinase K specificity. However, the general lack of cleavage in the linker is consistent with the linker being compact, as was suggested by the native page and SAXS results. Together these data support a model for a linker region that is more compact than the heptad repeat region, but that does not bias the conformation of the heptad repeat region.

Figure 4-5 The linker is a structurally independent subdomain of the CTD.

a. Proteinase K cleavage sites are distributed throughout the linker region(83). b. limited proteolysis of the MBP-\textit{D. melanogaster} CTD w/linker. Cleavage products resulting from cutting in the linker would be predicted to migrate between the 35kD ladder band, which is composed of the full heptad repeat region, and the intact MBP-
D. melCTD w/linker. Protease sensitive sites map near the 35kD CTD ladder band, otherwise, the linker is uncut even at high concentrations of protease. This recalcitrance to proteolysis indicates compaction.

4-2.4 Conformational changes in the CTD with the linker as a function of phosphorylation

Previously, I used the CTD standard curve generated by size exclusion chromatography to interrogate structural changes as a function of phosphorylation by the CTD kinase P-TEFb. I noted that phosphorylation lead to an increase in $R_s$. I subject the MBP-D. melCTD w/linker to P-TEFb phosphorylation and repeated the size exclusion chromatography experiments under conditions identical to those used previously. The MBP-D. melCTD w/linker also extends after phosphorylation, with the magnitude of this extension being somewhat greater than that observed for the CTD lacking the linker (Fig. 4-6). This could be due to additional phosphorylation of the SP motifs appearing in the distal part of the linker providing additional sites of P-TEFb phosphorylation.
Figure 4-6 Phosphorylation leads to extension of the CTD containing the linker region. The MBP-D.melCTD w/linker extends as a function of P-TEFb phosphorylation (brown triangles) relative to the unphosphorylated form (blue squares). Two replicates of each protein are plotted.
I next sought to measure the dimensions of the MBP-\textit{D. mel}\textit{CTD w/ linker} using SAXS. Expression, phosphorylation, sample preparation, and data acquisition for the MBP-\textit{D. mel}\textit{CTD w/ linker} were conducted using the same conditions as the MBP-\textit{D. mel}\textit{CTD} described previously. Despite obtaining high purity MBP-\textit{D. mel}\textit{CTD w/ linker} and selecting for monodisperse protein using size exclusion chromatography (Fig. 4-7) SAXS experiments on the apo MBP-\textit{D. mel}\textit{CTD w/ linker} were confounded by aggregation likely resulting from freeze/thaw of the sample required for transportation to the synchrotron. Thus, I focused my analysis on the phosphorylated version of MBP-\textit{D. mel}\textit{CTD w/ linker}.
Figure 4-7 Expression and purification strategy for linker CTD proteins for small angle X-ray scattering. (a) Low temperature expression followed by purification selecting for tags at opposing termini select for full-length protein, which is then subject to a kinase or mock kinase reaction lacking P-TEFb, and further purified over a 24ml Superose 6 column. (b) Protein samples used for SAXS analysis. Despite the purity of the apo MBP-*D. melanogaster* CTD w/linker sample, it was not monodisperse and the data were not usable for further analysis.
Scattering curves were collected on MBP-\textit{D.mel/CTD} w/linker at three concentrations. These curves were superimposable, showing no interparticle effects or concretion effects in the shape of the curves or in R\textsubscript{g} estimates from Guinier fitting (\textbf{Fig. 4-8}). The measured R\textsubscript{g} of MBP-\textit{D.mel/CTD} w/linker calculated from Guinier fits is 56.82 +/− 1.97 (\textbf{Table 4-1}). This is larger than the phospho MBP-\textit{D.mel/CTD} lacking the linker, consistent with the size exclusion chromatography results (\textbf{Fig. 4-6}). The scattering data collected at three concentrations were averaged for further analysis, including the pairwise distance distribution plot and flexibility analyses. The MW and R\textsubscript{g} estimates are well within 10% across the concentration series (\textbf{Table 4-1}).

I note that, unlike the apo and phospho MBP-\textit{D.mel/CTD} MW estimates, the MW estimate of 69kD for the MBP-\textit{D.mel/CTD} w/linker is lower than the 86kD MW of the protein (\textbf{Table 4-1}). MW measurements from SAXS data can be made to within ~10% accuracy provided the protein concentration is accurately measured\cite{108}. That my MW estimates for phospho MBP-\textit{D.mel/CTD} w/linker underestimate the weight of the protein by a margin exceeding 10% points to inaccuracies in measured protein concentration. Underestimation of the protein concentration by less than 20% is enough to explain the discrepancy between estimated and actual MW (For equation, see Materials and Methods). However, these MW estimates do not show concentration effects and are still a proxy for SAXS data quality. Further, MW estimates are independent of R\textsubscript{g} estimates made using the Guinier approximation, with the accuracy of R\textsubscript{g} estimates being
independent of precise protein quantification. Thus, while there is a discrepancy between the estimated and actual molecular weight, the Rg estimates are valid.

Figure 4-8 SAXS curves and Guinier fits of phospho MBP-\textit{D.mel}CTD w/linker a. Scattering curves from MBP-D.melCTD w/linker collected at three

\textbf{Figure 4-8} SAXS curves and Guinier fits of phospho MBP-\textit{D.mel}CTD w/linker a. Scattering curves from MBP-D.melCTD w/linker collected at three
concentrations. b. Guinier fit (top panel) and residuals of the fit (bottom panel) of the curves from panel a. Protein concentrations are indicated.

<table>
<thead>
<tr>
<th>Concentration (mg/ml)</th>
<th>MW(kD)</th>
<th>Rg(Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.27</td>
<td>66.69</td>
<td>55.61</td>
</tr>
<tr>
<td>0.50</td>
<td>69.59</td>
<td>55.75</td>
</tr>
<tr>
<td>1.09</td>
<td>71.01</td>
<td>58.09</td>
</tr>
<tr>
<td>Avg. + Std. Dev.</td>
<td>69.10 +/- 2.20</td>
<td>56.82 +/- 1.97</td>
</tr>
</tbody>
</table>

Table 4-1 Summary of Guinier analysis of phosho MBP-\textit{D.mel}CTD w/linker. Estimated MW and R\textsubscript{g} from the fits of the low angle scattering data from Fig. 5-8 MW and R\textsubscript{g} estimates are within 10% across the dilution series.
The pairwise distance distribution function for the phosphorylated MBP-

_D.mel_CTD w/linker shows an elongated shape consistent with that of the

phospho MBP-_D.mel_CTD (Fig. 4-9)(138). The _R_g_ derived from the Pr vs r plot is

67.25Å, a marginal increase over the phospho MBP-_D.mel_CTD (Fig 4-9). Thus,

consistent with the native gel electrophoresis and size exclusion chromatography

results, SAXS comparison of phospho MBP-_D.mel_CTD w/linker and phospho

MBP-_D.mel_CTD suggest that despite adding considerable MW, the linker

contributes only marginally towards increased in size. By comparison, the Flory

chain model can be used to predict the _R_g_ of denatured proteins and thus serves

as a benchmark for compaction or unfoldedness(90, 103, 139). The Flory model

predicts that the CTD portion of the MBP-_D.mel_CTD w/linker, including the linker

region, would have an _R_g_ of ~73Å. This estimate exceeds the measured _R_g_ of

the MBP-_D.mel_CTD w/linker, which includes the globular MBP, after

phosphorylation-induced extension.
**Figure 4-9 Pairwise distance distribution function of phospho MBP-D.melCTD w/linker SAXS data.** The phospho MBP-D.melCTD w/linker is only marginally larger than the construct lacking the linker and remains considerably smaller than the Flory chain estimate for the linker and CTD alone, lacking MBP, which is 73Å.
Using SAXS, I showed previously that after phosphorylation by P-TEFb the CTD becomes stiffer\(^{(17, 18)}\). I posit that this stiffness could give rise to binding ultrasensititivity, via a reduction in conformational entropy that yields a more binding-component protein scaffold. This hypothesis is consistent with theoretical models for other multivalent proteins that are phosphorylated multiple times, characteristics shared with the CTD\(^{(127)}\). One potential drawback to this entropic binding model with respect to the CTD is that a reduction in conformational entropy could potentially come at the cost of a reduction in radius of access. The CTD serves as a scaffold for multi-subunit enzymatic complexes that act on the nascent transcript in capping, splicing, and polyadenylation, and on chromatin through the modification of histone tails\(^{(2)}\). The RNA and protein substrates of CTD-associated factors are themselves flexible and capable of being in multiple orientations relative to Pol II as it traverses a gene. Thus, I sought to understand how the linker affects the flexibility of the overall CTD after phosphorylation.

To this end, I subjected the averaged SAXS scattering curve from the phosphorylated MBP-\(D.\text{mel}\)/CTD w/linker to flexibility analyses capable of qualitatively comparing the degree of flexibility in protein systems, and plotted this data with the averaged scattering curves from MBP-\(D.\text{mel}\)/CTD and phosphorylated MBP-\(D.\text{mel}\)/CTD (Fig 4-10)\(^{(104)}\). As previously described, Kratky plots of flexible systems exhibit a characteristic asymptotic rise in the higher q region of the plot\(^{(104)}\). The Kratky plot for MBP-\(D.\text{mel}\)/CTD exhibits such a rise, which is diminished upon phosphorylation (Fig. 4-10a) In contrast; this rise is
largely maintained after phosphorylation when the linker is present, suggesting
the linker CTD maintains system flexibility. As noted in chapter 4, Kratky plots are
potentially biased by improper buffer subtraction(104). Additional analyses on
truncated data reduce the likelihood of improper buffer subtraction biasing the
results(104). One such analysis is the Porod-Debye plot, which exhibits an
asymptotic rise for flexible systems(104) (Fig. 4-10b). As was observed in the
Kratky plot, both the MBP-D.melCTD and phosphorylated MBP-D.MelCTD
w/linker, display behavior in Porod-Debye analysis consistent with flexibility,
which is reduced in the case of the phosphorylated CTD construct lacking the
linker(104) (Fig. 4-10b). Finally, the Kratky –Debye plot reveals a plateau for
flexible systems, shared amongst the unphosphorylated CTD fusion, and the
phosphorylated CTD with the linker, but lost upon phosphorylation in the absence
of the linker (Fig. 4-10c) Collectively, these data show that the linker region
functions to maintain the flexibility of the CTD as a system, even after
phosphorylation.

As a fourth validation that flexibility is maintained in the presence of the
linker after phosphorylation, I generated scattering data extrapolated to so-called
zero concentration and repeated the Kratky analysis. Here, an algorithm takes
scattering data from multiple protein concentrations that has not been buffer
subtracted, and extrapolates the protein contribution to the scattering curve in an
infinitely dilute solution. This algorithm returns the protein scattering curve after
the buffer contribution has been removed, as well as a buffer scattering curve
with the estimated scattering contribution of the buffer(109). The estimated buffer
scattering curve from the extrapolation closely matches that of actual buffer
scattering curves used for buffer subtraction in prior analyses (Fig 4-11a). This
suggests I accurately accounted for the scattering contribution of the buffer in my
prior analyses. The Kratky plots derived from the extrapolated scattering curve,
and the buffer-subtracted scattering curve used previously, overlay across the q-
range, again suggesting there were no issues in buffer subtraction (Fig 4-11b)
These controls further validate the Kratky analysis conducted on the buffer
subtracted scattering data, as well as the Porod-Debye and Kratky-Debye plots,
and validate the observation that system flexibility is maintained in the CTD after
phosphorylation when the linker is present.
Figure 4-10 Flexibility analysis of phospho MBP-\textit{D.mel}CTD w/linker SAXS data. (a) Kratky Plot ($q^2 \times I(q)$ vs $q$) of averaged apo MBP-\textit{D.mel}CTD (blue) and phospho MBP-\textit{D.mel}CTD (red) and phospho MBP-\textit{D.mel}CTD w/linker (orange) MBP-CTD scattering curves. The more gradual rise in the high $q$ region of the plot for the phospho-CTD than the apo-CTD and phospho MBP-\textit{D.mel}CTD w/linker indicates that the phospho-CTD is less flexible than the apo CTD and phospho MBP-\textit{D.mel}CTD w/linker. (b and c) Porod-Debye ($q^4 \times I(q)$ vs $q^4$) and Kratky Debye ($q^2 \times I(q)$ vs $q^2$) plots of the data from panel a, colored accordingly. The increased rise in the Porod-Debye plots for the apo-CTD and phospho-linker CTD compared to phospho-CTD indicates the phosphorylated CTD without the linker is less flexible, as does the loss of the plateau in the Kratky-Debye plot for the phospho-CTD. The behavior of the apo-CTD and phospho MBP-\textit{D.mel}CTD w/linker are congruent, and deviate from the phospho MBP-\textit{D.mel}CTD in a manner that suggests flexibility is lost upon phosphorylation for the CTD without the linker while the linker enabled the system to remain flexible after phosphorylation.
Figure 4-11 Kratky plot of phospho MBP-*D. mel*CTD w/linker SAXS data extrapolated to zero concentration. (a) $I(q)$ vs $q$ plot of averaged buffer signal used in buffer subtraction of phospho MBP-*D. mel*CTD w/linker and scattering data plotted with buffer scattering curves calculated from the zero concentration extrapolation performed using SAMBUF(109). The measured and extrapolated buffer scattering curves overlay indicating I correctly account for and subtract the scattering contribution of the buffer in our analyses. (b) Kratky plots with error bars for averaged scattering curves and for scattering curves derived from zero concentration extrapolation using SAMBUF. Buffer subtracted averaged curves and buffer independent zero concentration extrapolation yield the same result.
4-3 Discussion

The linker region has been shown to be important for proper Pol II function but has been largely ignored in studies by the CTD field(135, 136). Two studies that establish a requirement for the linker region in supporting normal cell growth provide evidence of a functional relationship between the linker and the heptad repeats. The earlier work shows there is some functional redundancy between these two regions as linker mutants are suppressed by additional heptad repeats, and CTD truncations are suppressed by linker sequences that function as more robust transcriptional activators(135). A more recent study moved the CTD to other Pol II subunits. This work observed a requirement that the CTD repeats be located proximal to its natural location to function, and that the linker region must remain on Rpb1(136). A corollary is that the heptad repeats must be located near to the linker region in order to function. If true, this suggests a spatial, or structural, constraint between the linker and repeat region that could explain the genetic interactions observed in the earlier study(135). The necessity of the linker for proper CTD function, the evidence that the heptad repeats must be located near the linker, and my observation that the heptad repeats themselves are structurally heterogeneous prompted me to interrogate the structural implications of the linker on the CTD as a whole.

Using native gel electrophoresis and size exclusion chromatography, I observe that the linker is more compact than would be predicted if it adopts the same degree of compaction as the heptad repeats (Figs 4-3 and 4-4), despite an
enrichment for similar amino acids in the two regions (Fig 4-2). This prompted me to interrogate the structure of the linker on a more local level, using limited proteolysis, which uses protease accessibility as a proxy for protein compaction. My result corroborates the biophysical observations; the linker is more compact than the heptad repeat region. Finally, flexibility analyses of SAXS data show that the compact linker buffers the CTD against stiffening in response to phosphorylation.

Together, these results support a model by which the heptad repeat region is phosphorylated resulting in a reduction in conformational entropy and a more binding competent protein scaffold. The linker region buffers against the additional stiffness of the overall CTD, in essence functioning as a hinge between the globular catalytic Pol II core, and the heptad repeats. In this model, the CTD itself can stiffen locally to become a more binding competent scaffold, but the radius of access is maintained by a flexible, hinge-like linker, allowing CTD associated proteins to sample a large three-dimensional space and capture their RNA and chromatin substrates.

The linker region likely serves as a flexible hinge between the core of Pol II and the heptad repeat regions. Flexibility analyses of SAXS data provide a means of interrogating the ability of different sequences to serve in this capacity. Linker mutants designed to introduce or remove flexibility could be made and evaluated in vivo using the genetic methodologies employed to test human CTD mutants. Those mutations giving rise to phenotypes could be tested via SAXS as MBP fusions to correlate degrees of flexibility to function.
The linker region contains numerous SP motifs making it a potential substrate for P-TEFb and other kinases. Like phosphorylation of the heptad repeats, phosphorylation of the linker could be functionally important. Unfortunately, the CTD with the linker region is likely too large for MALDI-TOF mass spectrometry to be used as a means to quantify phosphate incorporation. However, a TEV protease site could be introduced to liberate the linker and heptad repeat region from MBP. This phosphorylated fragment would be well below the size of the MBP-CTD that was used successfully to measure phosphate incorporation with MALDI-TOF. Finally, the mass spectrometry methods brought to bear on the heptad repeats to map the sites of phosphate incorporation could be applied to fragments with the linker. It would be interesting to know if the more distal region of the linker, enriched for SP motifs, is phosphorylated, or if patterns of heptad repeat phosphorylation are altered by the presence of the linker.

Two major hurdles in my biochemical and biophysical studies of the linker were insolubility and poor expression. Using expression and purification strategies that yield highly pure and soluble MPB-CTD fusions worked poorly for the constructs containing the linker. For perspective, dozens of liters of culture were required to yield ~1mg of MPB-\textit{D.mel}ICTD w/Linker, which is 10 - 20x less protein than was recovered for \textit{Drosophila} CTD fusions lacking the linker. Further, the resulting linker containing fusion proteins were labile to freeze thaw. Optimization of the constructs, possibly by the addition of a sequence N-terminal
to the linker as I defined it based on structure and sequence information, could be a potential strategy to increase solubility and enable additional experiments.
CHAPTER 5
STRUCTURE ENTERS THE CTD CODE

At the outset of this work, almost nothing was known about the structure of the CTD, particularly the full-length CTD as a monomer in solution. This work combined biophysical, biochemical, and genetic techniques and leveraged the unique combination of the Drosophila CTD sequence, developmental complexity, and genetic tractability to yield four key insights pertaining to CTD structure, with implications for CTD function and transcription regulation.

The yeast, fly and human CTDs are structurally similar globally despite sequence differences

The CTD can be thought of as a series of SLiMs; the consensus YSPTSPS heptad. Permissive mutations have arisen in this short motif, which are typically well conserved among members of the same phylum (24). Despite significant differences in adherence to the consensus CTD SLiM, native gel electrophoresis and size exclusion chromatography show that the CTDs of yeast, fly and mammals are structural homologs in terms of gross dimensions. This observation suggests that permissive mutations may be under selective pressure to maintain some overall balance between compaction and extension.

Additionally, this finding places the Drosophila CTD on a size continuum with yeast, which has given rise to much of the existing knowledge about CTD function, and human, which has a longer and more complex CTD typical among more developmentally complex organisms. This relationship establishes
Drosophila as an ideal model to interrogate the structure of the CTD, due to the experimental tractability of the native sequence, developmental complexity, and genetic tractability.

**The Drosophila CTD is structurally heterogeneous**

I next asked how an extended CTD might organize factors across its length. It has long been proposed that sequence variation, well conserved among members of a phyla but different among different phyla, could give rise to CTD:factor organization\(^{(24, 26, 130, 133)}\). My biophysical characterization establishes a model for global CTD structure I describe as a compact random coil; more extended than globular proteins of the same size, but significantly more compact than predicted for denatured proteins. This description of the CTD is compatible with a growing body of literature characterizing local structure, and structural heterogeneity in proteins predicted to be intrinsically disordered\(^{(54, 55, 103, 123)}\). Together these prompted me to interrogate the structure of the CTD on a more local level, using limited proteolysis. This approach is ideally suited to interrogating the structure of scaffold proteins, as it employs protein-scale probes to measure variation in local structure. The repeated nature of the CTD sequence means properly selected proteases should cleave the CTD across its length, so variation in proteolytic sensitivity can be interpreted as variation in CTD structure. These experiments find previously undiscovered structural organization within the CTD, with significant variation in protein accessibility across the CTD. Similar patterns of proteolysis arise from three proteases with distinct structures and specificities, and the observed pattern is perturbed by a denaturant. Collectively,
these observations are consistent with *bona fide* structural organization within the CTD. My observations contradict common depictions of the CTD in the transcription literature, which often conceptualize the CTD as an extended “noodle” where all heptads are structurally equivalent. This notion likely arises from models that iterate turn structures across the CTD(13, 64). My experiments are the first to interrogate global and local CTD structure using full-length, monomeric CTD in physiologically relevant buffers. The result is a model for a semi-compact, structurally heterogeneous CTD with implications for CTD:factor interaction that involve not only sequence variation but structural variation and non-uniform motif accessibility.

The presence of proteolytic fragments arising from a closely spaced series of sensitive sites in the proximal region of the CTD could be an interesting avenue for further study. One possibility is that these sites are generated from a sufficiently small subset of labile conformers that undergo complete cleavage of the sensitive sites at the lowest protease concentration used in my experiments, and the resulting fragments are recalcitrant to further proteolysis. One interesting potential source of such conformers could be proline isomerization. The *cis* conformation of proline occurs 5-6% of the time in protein structures and interconversion can be slow, occurring on the timescale of minutes(140). Meanwhile, my limited proteolysis experiments are 30 seconds in duration. The proximal sensitive sites could thus arise from a small subset of conformers locked in the *cis* conformation on the timescale probed in my experiments.

Repeating the limited proteolysis experiments in the presence of the *Drosophila*
Pin1 homolog Dodo could be informative. If the pattern of proteolysis changes, it could suggest some CTD conformers are trapped in a particular isomeric state on biologically relevant timescales in the absence of Pin1.

**The CTD extends and stiffens upon phosphorylation**

Textbook explanations of allosteric modulation of proteins involve conformational changes in folded proteins resulting from ligand binding or post-translational modification, with the differing folds being functionally distinct. However, allosteric modulation extends to proteins that are not folded and more accurately described as disordered ensembles (63, 141). The CTD exists in the nucleus in the presence of “writers” and “readers” of the CTD code, hundreds of proteins in total whose interaction with the CTD is under tight temporospatial control (6, 29, 30). This raises the question about how an extended CTD occupying a shared compartment with a vast interactome protects itself from spurious interactions and modifications out of phase with the transcription cycle?

Using size exclusion chromatography and small angle x-ray scattering, I show the CTD extends in response to physiologically relevant phosphorylation. Comparison of the limited proteolysis pattern of the apo and phospho CTDs reveals an altered distribution of cleavage sites after phosphorylation. The phospho CTD remains structurally heterogeneous, but is more uniformly accessible to protein interaction. Overall protease accessibility increases moderately upon phosphorylation, consistent with moderate increases in size as measured by SEC and SAXS. Together these results describe a structural ensemble allosterically altered by phosphorylation.
While increased protein accessibility could potentially explain how the CTD protects itself from improper interactions and modifications, the degree of these changes was modest relative to the overall size and accessibility of the unphosphorylated CTD. Additionally, ensemble models of the CTD based on SAXS data are consistent with extended structures existing in the CTD ensemble prior to phosphorylation. This lead me to consider other means by which the CTD could protect itself from spurious interactions.

SAXS data can be used to qualitatively compare the flexibility of proteins(104). Applying these analyses to SAXS data collected on the apo and phospho CTD reveals the CTD stiffens in response to phosphorylation. This finding is intriguing given the flexible nature of the CTD. Conceivably, a high entropic cost must be paid to capture the dynamic CTD, and the little buried surface area characteristic of globular protein:SLiM interactions provide less currency to pay this penalty than is common among globular protein interactions(28). I posit that increased CTD stiffness could reduce the entropic penalty for binding by reducing the conformational entropy of the CTD. ATP is a common cellular currency driving reactions with kinetics unfavorable to biological timescales. In a model where reduced conformational entropy enhances CTD:factor interactions, the energy released from ATP by CTD kinases pre-pays the entropic penalty for binding. Such a payment-precedes-binding model could explain how the CTD protects itself from interactions out of phase with the needs of the cell and organizes the wholesale exchange of phospho-specific interactomes on the transcriptional timescale(6).
Binding mediated by reduced CTD conformational entropy is a model that is wholly consistent with theoretical work that posits that modifications to multivalent disordered proteins can regulate binding by local decreases in conformational entropy in a manner that results in binding ultrasensitivity\(^{(127)}\). In this way, a CTD phosphorylated \(x\) times may be largely unbound by a particular factor, whereas a CTD phosphorylated \(x+1\) times would be largely bound. This switch-like binding would be an ideal way to tune CTD:factor interactions to the transcription cycle and provides protection without requiring dramatic structural reorganization of the CTD. Further, longer CTDs would have a non-linear increase in their capacity to regulate binding in this manner, which may provide a rationale for the longer CTDs of metazoans, where genes are longer, enabling the CTD to serve as an entropic clock. A final intriguing element of this entropic binding model is it may explain the increased affinity exhibited by some CTD interacting factors for phosphorylated CTD substrates, even when those phosphates are not involved in direct recognition of the CTD, such as Pcf11\(^{(13)}\).

I posit a reduction in conformational entropy could be a form of ensemble allostery exploited by the CTD to regulate a complex and dynamic interactome.

**The human CTD functions in place of the fly CTD *in vivo***

The human CTD had previously been observed to extend upon phosphorylation\(^{(74)}\). I observed gross structural similarities between the human and fly CTD using native-PAGE and SEC. The human CTD is similar to that of the fly in length, and both deviate from the consensus motif at the majority of heptads. These similarities lead me to test whether or not these metazoan CTDs
were structurally similar locally using limited proteolysis. I found that both the fly and human CTD are characterized by a largely protease insensitive central region flanked by regions of protease sensitivity. This structural similarity on the local level exists despite significant sequence differences that are highly conserved within each lineage. Collectively, these structural similarities and sequence differences lead me to test a longstanding hypothesis in the field; that lineage specific variation in the CTD is functionally important. Leveraging the genetic tractability of *Drosophila*, we humanized the fly *Rbp1* by replacing the endogenous CTD with the human counterpart in every tissue and throughout development using a dual *GAL4* RNAi/ectopic expression system. In contrast to predictions based on lineage specific sequence differences, the human CTD functions in the fly, supporting development of flies to adulthood. This is unlikely to arise from residual activity of endogenous Rpb1 harboring the fly CTD for a number of reasons. First, over-expressing the wild-type Rpb1 in the context of Rbp1 RNAi is lethal. This means that the RNAi is robust enough to knock down Rpb1 to lethal levels even when Rpb1 is being overexpressed ectopically. Second, I show using rtPCR primers specific for the ectopic and endogenous Rpb1 variants, that endogenous Rpb1 mRNA is knocked down to ~5% of normal levels in cells robustly expressing humanized Rpb1. These data suggest that the precise array of heptads conserved amongst *Drosophila* species is not essential for Pol II function, even in the complex developmental context of the fly. Instead our data support a hypothesis that the structure and biophysical properties of the CTD support function, not a specific CTD sequence.
The linker is an independent structural unit of the CTD

Finally, some of the biochemical and biophysical techniques brought to bear on the CTD were applied to the CTD containing the functionally important but oft-ignored linker region. These experiments show that the linker is more compact than would be predicted by a heptad repeat region of similar length, despite similar amino acid composition. Further, the linker is a structurally independent sub-domain of the CTD that is compact and does not alter the conformation of the heptad repeats. Finally, the linker buffers the CTD from overall stiffening in response to phosphorylation. Together, this suggests a model in which the linker functions as a hinge, separating the core of Pol II from the heptad repeats, allowing the repeats to stiffen while maintaining an overall radius of access compatible with CTD interacting factors operating on RNA and chromatin substrates.

Testing the entropic binding model

The model enabled by this work states that some CTD interactions may be controlled entropically as a function of phosphorylation. Further, the linker region serves as a hinge to maintain a radius of access for factors bound to a stiffened CTD. Two predictions emerge from the entropic binding model put forth by Swain and Lenz that may be testable in the context of the CTD(127). First, the longer the disordered region, the more cooperative binding will become as a function of phosphorylation(127). Second, removing some disordered repeats will favor binding(127). This can be rationalized by considering that longer repetitive IDPs will have greater conformational entropy than shorter ones, and thus can
theoretically undergo larger decreases in conformational entropy as a function of phosphorylation. However longer IDPs would require more phosphorylation to achieve this reduction.

It follows that this hypothesis could be tested with binding experiments from which a $K_D$ can be measured, such as pull down experiments or surface plasmon resonance. The prediction is that shorter CTDs would be more completely bound than longer CTDs assuming an equivalent number of total phosphates, but longer CTDs would be more completely bound assuming similar phosphate density. If binding is entropically controlled, non-linear changes in binding should result from linear changes in repeat number or phosphate density.

One caveat to changing the length of the fly CTD is that invariably the sequence composition also changes. Recently, Feiyue Liu in the Gilmour laboratory has shown that CTDs of varying lengths composed only of consensus repeats and the final acidic tip, support viability in flies. Thus, fully consensus CTD sequences could be used in binding assays in vitro to test the entropic binding model against recombinant proteins or nuclear extracts, allowing for changes in valency independent of changes in sequence. Feiyue’s finding also provides a series of mutant flies that could be used to test the entropic binding hypothesis in vivo. One potential experiment would be to perform ChIP-exo using antibodies against CTD associated factors in fly lines with varying length consensus CTDs. The signal would be normalized to the signal from total Pol II. Factors whose binding is entropically controlled should show non-linear changes in ChIP signal as a function of linear changes in CTD length. Three major
caveats to these in vivo studies exist. First, changing the CTD length could alter CTD phosphorylation in unpredictable ways, and interpreting the data in the context of the entropic binding model requires similar phosphate density be achieved among CTD mutants. Measuring this phosphate density across a sequence that has been rendered fully repetitive would be difficult. The best approximation would be MALDI-TOF mass spectrometry to measure total phosphate incorporation on Pol II purified from the various lines while assuming the phosphates are distributed evenly. Incorporating a protease site to liberate the CTD from purified Pol II could yield a CTD suitable in size for such analysis. A second caveat is that certain factors may interact with the CTD independently of entropic control. Sic1 interacts with Cdc4 in a phospho-specific manner that responds non-linearly to phosphorylation, consistent with entropic control(100). However, this interaction was shown experimentally to be sensitive to charge density in a non-linear manner(100). De-convolving electrostatic vs. entropic contributions would be difficult. Finally, approaches that rely on candidate factors may not be informative. Hundreds of factors interact with the CTD, not all of which may be entropically regulated. Quantitative, label-free mass spectrometry approaches, such as those employed by Churchman and colleagues to identify phosphorylation dependent CTD associated factors from immunopurified (IP) polymerase, could be employed as a more unbiased way to compare factor recruitment among CTD mutants(6). The caveat to this and other IP based approaches is dilution of CTD interacting factors occurs upon cell or nuclear lysis. Swain and Lenz predict the highly cooperative binding that could be
achieved through entropic control also leads to high sensitivity to ligand concentration (127). Experiments that rely on IP’ing native Pol II invariably result in dilution which may result in the loss of entropically regulated interactions.

Protein:protein crosslinking with agents compatible with the CTD sequence could potentially circumvent this issue by covalently linking CTD:factor interactions in vivo prior to cell lysis and accompanying dilution.

This work also supports the hypothesis the linker region of the CTD functions as a flexible hinge. Generating a series of CTD mutants with linker sequences predicted to be structured, i.e. helical, could test the importance of linker flexibility. If these CTDs undergo an overall loss of flexibility upon phosphorylation as measured by SAXS flexibility analyses, they could be incorporated into flies to test whether or not they generate a phenotype. The caveat to analyzing the linker is the importance of sequence is poorly understood. One approach could be to generate mutants with similar biophysical properties, such as charge distribution, but differences in predicted structure. If mutants were achieved that generate a gradient of flexibility as measured by SAXS flexibility analyses, they could be compared to see if a gradient of phenotypic severity results in flies.

Lastly, it is worth noting that Swain and Lenz modeled the local effect of phosphorylation in a disordered, multivalent protein as a kink in the chain (127). Eric Gibbs recently discovered context dependent structural switches within the CTD in response to phosphorylation (53). Eric found that heptads of the sequence Y₁S₂P₃T₄S₅P₆N₇ populated the cis conformation at P₆ to a much greater extent.
than heptads of other sequences (53). This suggests that specific regions of the CTD could respond locally to the distributive application of phosphates, including via a local reduction in conformational entropy. It had been found previously that the CTD phosphatase Ssu72 preferentially recognizes heptads containing cis $P_6$ (67). Eric's finding provides a mechanism of preferentially localizing this phosphatase activity to specific regions of the CTD, which could serve to nucleate the $S_5P$ to $S_2P$ transition therein (53). It remains to be seen if this cis-trans structural switch is of biological importance, or is overridden by the activities of PPIases like Pin1 (34, 142). Nevertheless, an overarching theme of both Gibbs et al. and Portz et al. is that the CTD is not a uniform polymer, but rather varies in its local structural organization, response to phosphorylation, and interactions with proteins (50, 53).

**The future of the field of transcription regulation is liquid-liquid phase separation.**

During my time in graduate school a series of papers emerged that I believe ushered in a new epoch for the study of how eukaryotic cells are organized via the phenomenon of liquid-liquid phase separation (143, 144). Liquid-liquid phase separation is a process by which components of a liquid demix, generating liquid droplets enriched for particular components relative to the bulk solution (143-146). Examples of assemblies of RNA and protein in cells thought to result from liquid-liquid phase separation are stress granules (147, 148), germ line granules in *C. elegans* (149-151), the nucleolus (152-154),
heterochromatin domains, and the centrosome. Additionally, evidence suggests that components of the cell cytoskeleton could grow from monomers highly concentrated by phase separation. More transient structures are also likely phase separations, including complexes of repair machinery formed on sites of DNA damage. That these various structures, many co-existing in the same cell, do not simply coalesce together is illustrative of the fact that phase separation is a specific process resulting in phases with distinct compositions. It is likely that many more biological phase separations exist, evading detection because they are below the diffraction limit of light microscopy.

Phase separation is exquisitely and non-linearly tied to the concentration of the protein or RNA undergoing the phase transition. Below a critical threshold concentration, $C_{\text{sat}}$, the protein or RNA is freely mixed with the solvent. Above $C_{\text{sat}}$, nearly all of the protein or RNA will de-mix, creating a liquid droplet within the larger solvent.

An important aspect of these phase separations are their dynamics, they are not solid aggregates. Phase separated droplets can wet surfaces, shear, and fuse with other droplets. Fluorescence recovery after photo bleaching experiments show components of the phase separated droplets interchange rapidly with the bulk solution, on the order of seconds. However, the morphology of the droplets themselves can be stable over much longer timescales, providing a means of organizing RNA and protein in what have been called “membraneless organelles.”
Phase separated liquid droplets are distinct from macromolecular protein and RNA complexes in that they are not stereospecific(145). The protein components of diverse liquid droplets share common features including intrinsic disorder, low sequence complexity, multi-valency and nucleic acid binding domains(144). These features enable proteins to transiently interact with multiple partners and generate liquid phases that in contrast to macromolecular complexes lack defined stoichiometry and stereospecific interaction surfaces. Conceptually, instead of thinking about stoichiometric binding between mated stereospecific surfaces, think about molecules with flexible regions “slithering” with one another, held together in dynamic droplets by transient interactions(143, 144). The nature of the forces that hold liquid phases together remains poorly understood.

**Clients and scaffolds**

Not every RNA or disordered protein partitions into every type of phase separation. A striking example of this is the nucleolus, which is thought to be at least two separate concentric phases, each enriched for different protein components and with different viscosities and dynamics(154). Some proteins and RNAs are likely responsible for nucleating liquid droplets at cellular concentrations and have been dubbed “scaffolds,” whereas other proteins and RNA specifically recruited to the phase are referred to as “clients(144).” This distinction is important. An inherent feature of the energetics of phase separations is that liquid droplets will increase in size, ultimately coalescing into one large droplet via fusion of small droplets and/or diffusion of molecules out of
smaller droplets and into larger ones (143). Thus, any protein that needs to be localized to many sites benefits from being a client with a partition function for the droplet phase that allows for diffusion into and out of the droplet.

**The Pol II CTD as a client**

RNA polymerase II shares characteristics common among proteins that phase separate at physiological concentrations; it binds nucleic acids and has a long, multivalent and low-complexity CTD. However, Pol II needs to be distributed at thousands of genes in the cell at any given time, a process that would be antagonized by Pol II droplets ripening into one large nuclear droplet as would be predicted if Pol II was a scaffold. Instead, Pol II is likely a transient client of phase separations that I posit are nucleated by locally high concentrations of disordered protein domains associated with chromatin bound transcription factors.

The CTD, fused to MBP or fluorescent proteins has been shown to behave as a client, partitioning into liquid and gel-like phase separations of the low complexity domains of FUS, TAF15, and EWS (161-164). This is notable, as these low complexity domains are drivers of cancer when fused to certain DNA binding domains as a result of chromosomal translocations (162). This suggests a mechanism by which low complexity domains, which are highly enriched among transcription factors, can recruit Pol II via the CTD to activate transcription (162). Curiously, not every region of the human CTD behaves equivalently as a client, the distal region is most avidly bound by gel phases of FUS, TAF15, and EWS (162). This suggests that the sequence composition of the distal part of the
human CTD, which is comprised of non-consensus repeats, could be important for regulating phase behavior. This is interesting in light of our finding that the human CTD functions in place of the fly CTD, and both CTDs share distal protease hypersensitivity. Perhaps evolution has arrived at different sequence solutions to the same biophysical problem, the regulation of Pol II phase behavior.

Towards a model system for phase separation in transcription regulation

Phase separation as a biological phenomenon has thus far been explored in the context of very large assemblies of protein and RNA that exist in cells above the diffraction limit of light microscopy. The large size of these membraneless organelles has provided an experimentally tractable system to interrogate phase dynamics and compare the results to mathematical predictions of phase behavior. Testing whether or not transcription is regulated via phase separation of transcription machinery that recruits Pol II would benefit from a readily inducible phase that is larger than the diffraction limit of light microscopy. I posit that heat shock puffs of *Drosophila* polytene chromosomes are a liquid phase separation and suggest them as the ideal model for interrogating this phenomenon in the context of transcription.

The Lis laboratory had a series of papers that established an essential role for poly-ADP ribosylase (PARP-1) in heat shock gene activation(165-167). This enzyme generates chains of ADP ribose (PAR) that are chemically similar to RNA. A finding from the Lis laboratory is that PAR chains are formed around heat shock loci(165). Further, the Lis laboratory showed that the recruitment of
fluorescently labeled Pol II to heat shock puffs continues well after the DNA is saturated with Pol II as measured by ChIP(167). This lead the authors to conclude that “cages” of PAR chains are formed around heat shock loci and these chains recruit Pol II due to their similarity to nucleic acids(167). The authors posited that these chains keep transcription machinery localized to the heat shock genes. However, any cage capable of keeping factors in, would also in theory keep factors out, so newly transcribed heat shock RNA could be trapped in the nucleus and additional recruitment of Pol II would not aid in transcription. Phase separation provides an alternate hypothesis.

RNA accelerates phase separation of many proteins, lowering the $C_{\text{sat}}$. The Pol II CTD enters FUS phase separations at lower concentrations in the presence of RNA(163, 164). I posit that PAR chains mimic RNA and aid in the nucleation of a phase separated state at heat shock loci that in turn recruits transcription machinery, including Pol II. This hypothesis is strengthened by a recent report that shows PARP generates PAR chains to nucleate a phase separation at sites of DNA damage(159).

This model of PAR dependent phase separation also provides a self-regulating mechanism by which the heat shock genes could be shut off. If proteins essential to the continued expression of the heat shock genes bind RNA more avidly than PAR, the accumulation of heat shock RNA and diffusion away from the heat shock locus, or alternatively up-regulated transcripts from elsewhere in the genome, could ultimately compete away these essential factors from the heat shock locus. In this way, the heat shock response could be self-
limiting. While RNA is thought to lower the $C_{\text{sat}}$ for many proteins, it has also
been shown to dissolve phases at higher concentrations\((168)\), another way
robust local RNA production could lead to self-limiting transcriptional bursting if
scaffold proteins involved in gene activation also behave in this manner.

**Testing the phase separation model for heat shock induction**

Phase separation differs from canonical binding due to the switch-like
dependence on protein concentration or valency. Expressing transcription
factors above the $C_{\text{sat}}$ could induce a phase separation. Alternately, knockdown
of key scaffold components below $C_{\text{sat}}$ would prevent phase separation. If the
CTD is indeed a client, mutations in the CTD that reduce the number of heptads,
i.e. reduce valency, would be predicted to have a non-linear change in the
recruitment of Pol II, which could be monitored with microscopy at heat shock
puffs, and in living cells with fluorescently tagged Pol II harboring CTD mutations.
Likewise, drugging PARP-1 with a titration of inhibitor to reduce but not abolish
activity would be predicted to have a non-linear effect on heat shock induction.

I believe liquid-liquid phase separation will provide a means to understand
many elements of gene regulation that are incompletely understood, from the
regulation of transcriptional bursting, genome organization, the function some
non-coding RNAs, and even the evolution of gene regulatory networks via
competition for limiting pools of factors involved in chromatin associated droplet
formation. The CTD has been shown to be a client of droplets \textit{in vitro}, and super
resolution microscopy experiments show Pol II recruitment to genes with dwell
times similar to the FRAP recovery times for many liquid phases\((169)\). As new
rules elucidating the structural and biophysical properties that facilitate phase separation emerge, they can inform mutations to the *Drosophila* CTD aimed at disrupting phase behavior, with the heat shock loci serving as the ideal experimental system.
CHAPTER 6
MATERIALS AND METHODS

Sequence Motif Generation

The amino acid sequences of the Pol II CTDs of *D. melanogaster* and *H. sapiens* were entered into WebLogo as chains of 7 amino acids, beginning with YSP. For compatibility with the tool, which requires each line of text contain the same number of letters, repeats longer than 7 amino acids were trimmed, those shorter were extended with “X.”

Cloning MBP-CTD plasmids

PCR amplicons encoding the *S. cerevisiae, D. melanogaster,* and *H. sapiens* CTD with either 6HIS or FLAG-STREP (DYKDDDDK-WSHPQFEK) tags were inserted into the XbaI digested pMalX vector using infusion cloning (Clontech) and transformed into Stellar competent *E. coli* (Clontech). CTD sequences were verified by sequencing from the plasmid across the insert using primers pMALXSeqE 5' GCG TAC TGC GGT GAT CAA CG 3' and M13/PUC 5' CGC CAG GGT TTT CCC AGT CAC GAC 3'.

Expression of MBP-CTD proteins

pMalX vectors containing the CTD fragments with either C-terminal 6HIS tags or C-terminal FLAG-STREP (DYKDDDDK-WSHPQFEK) tags were transformed into BL21(DE3) *E. coli* cells and grown in LB media with 0.2%
dextrose and 100µg/ml ampicillin at 37°C shaking at 250RPM to an optical density at 600nm of ~0.5. Protein expression was induced with 300µM IPTG, and proteins were expressed overnight at 11°C with shaking at 250RPM. Cells were harvested by centrifugation for 10 minutes at 3,000 x g, flash frozen, and stored at -80°C. Thawed pellets were lysed by sonication, or by three passes through a microfluidizer, in 20ml/1L culture volume of Binding Buffer (50mM HEPES pH 7.5, 500mM NaCl, 10% glycerol, 1% NP-40, 2.5mM imidazole, 2.5mM β-mercaptoethanol) plus protease inhibitors (0.1mM PMSF, 1.6µg/ml benzamidine HCL, 1µg/ml aprotinin, 1µg/ml pepstatin A, 1µg/ml leupeptin).

**Purification of MBP-CTD proteins**

Cell lysates were centrifuged at 100,000xg for 30 minutes and the clarified lysates were bound to TALON superflow resin (Clontech) for 6HIS tagged proteins or Streptactin superflow resin for FLAG-STREP (iba) in batch for 30 minutes at 4°C. Resin was centrifuged at 500xg for 5 minutes and washed with 20-40 column volumes (CV) of Wash Buffer A (50mM HEPES pH 7.5, 500mM NaCl, 10% glycerol, 1% NP-40, 10mM imidazole, 2.5mM β-mercaptoethanol) in batch. Resin was centrifuged as above, resuspended in 6 CV of wash buffer A and poured into a column and washed with an additional 3 CV of wash buffer. Protein was eluted in 8 CV of Elution Buffer (50mM HEPES pH 7.5, 500mM NaCl, 10% glycerol, 1% NP-40, 200mM imidazole, 2.5mM β-mercaptoethanol; for FLAG-STREP tagged proteins, this buffer also contained 2.5mM desthiobiotin). Eluted protein was passed twice over an amylose resin column (0.5ml column
resin/L cell culture). The resin was rinsed twice with 10CV of wash buffer B (50mM HEPES pH7.5, 150mM NaCl, 10% glycerol, 5mM DTT, 1mM EDTA) and eluted in CTD buffer (50mM HEPES pH7.5, 150mM NaCl, 10% glycerol, 5mM DTT, 10mM maltose, 1mM EDTA) (Fig. 6-1).

**Figure 6-1** Representative purification of MPB-\textit{D.mel}/CTD6HIS. Whole cell extract refers to lysate after sonication, clarified supernatant refers to the lysate after ultracentrifugation. This purification resulted from 10L of BL21(DE3) cell culture.
Expression and Purification of *D. melanogaster* P-TEF\(\text{b}\)

600ml of Sf9 cells were grown in suspension in baffled flasks at 27°C in Sf-900 II SFM media (Thermo Fisher Scientific), diluted to 1.5 million cells/ml, and infected with 1/10 culture volume of P-TEF\(\text{b}\) virus (a gift from J.T. Lis) and incubated at 27°C with shaking at 75rpm. 72hrs after infection, the cells were harvested at 500 x g for 5 minutes and resuspended in P-TEF\(\text{b}\) binding buffer (50mM HEPES pH 7.5, 500mM NaCl, 10% glycerol, 1% IGEPAL, 2.5mM Imidazole and 2.5mM \(\beta\)-mercaptoethanol and protease inhibitors as defined above) and lysed in a dounce homogenizer. Lysate was centrifuged at 100,000 x g for 30 minutes at 4°C. Clarified supernatant was bound in batch to 1.0ml bed volume of TALON resin for 30 minutes at 4°C, resin was pelleted at 250 x g for 5 minutes, washed with wash buffer, pelleted as before, washed and decanted into a gravity column and protein was eluted in 8ml 500mM NaCl, 50mM HEPES pH 7.5, 10% glycerol, 1% NP-40, 2.5mM \(\beta\)-mercaptoethanol and 200mM imidazole (Fig. 6-2).
Figure 6-2 Purification of *D. melanogaster* P-TEFb from Sf9 cells.
Coomassie stained SDS-PAGE gel depicting representative P-TEFb purification. Whole cell extract refers to cell lysate following homogenization in a glass dounce. Clarified supernatant refers to the lysate after ultracentrifugation. Two predominate bands migrate near the expected sizes of Cyclin T (118kD) and Cdk9 (47kD).

Phosphorylation and purification of MBP-CTD

MBP-*D. meli*/CTD6HIS was purified as described above, concentrated with a Vivaspin 30,000MWCO concentrator (GE Life Sciences) to ~10mg/ml and kinased in CTD buffer supplemented with 10 mM MgCl$_2$, 10mM ATP and P-TEFb at 24°C for ~48hrs with additions of 1/10$^{th}$ reaction volume 100mM ATP and 10mM MgCl$_2$ at ~8-12hr intervals for the duration of the reaction to restore
ATP concentrations to ≥10mM, while maintaining MgCl$_2$ concentration at 10mM. The kinase reactions were loaded on a Superose6 10/300 column (GE Life Sciences) in CTD buffer and the monomeric peak fractions were collected and concentrated to ~1-2mg/ml using a Vivaspin 30,000MWCO concentrator previously rinsed with CTD buffer. For SAXS, exact concentrations were obtained via UV absorbance at 280nm and the molar extinction coefficient of the fusion protein (124460 M$^{-1}$cm$^{-1}$). Dilutions for SAXS were done in the same preparation of CTD buffer used for the size exclusion chromatography step of the sample preparation.

**Size Exclusion Chromatography**

A 100ul protein standard mix containing thyroglobulin, apoferritin, aldolase, ovalbumin, and RNAseA, with stokes radii of 85Å, 67.1Å, 48.1Å, 30.5Å and 16.4Å, respectively or MBP or MBP-CTD6HIS fusion proteins, were loaded onto a Superose 6 10/300 column at 0.2ml/min in CTD buffer at 4°C. Elution was monitored by UV absorbance and the identity of the protein in each peak was verified by SDS-PAGE. Standards were run before and after MBP-CTD fusions were analyzed. The standard curve was generated by plotting $\sqrt{-\log K_{av}}$ vs Rs for the average of three standard runs, where $K_{av} = V_e-V_o/V_c-V_o$ ($V_e$ is elution volume, $V_o$ is void volume as determined by the elution of blue dextran and $V_c$ is the total column volume).
Small Angle X-Ray Scattering

SAXS data (I(q) vs q, where q=4\pi \sin(\lambda/\lambda) were collected at the G1 beamline at the Cornell High Energy Synchrotron Source with an X-ray beam energy of 9.944keV using dual Pilatus 100K-S detectors. Samples were exposed for 20 x 1s frames. For each sample and at every concentration, buffer blanks were collected immediately before and after sample data acquisition. Data acquisition, buffer subtraction, data reduction, Guinier analysis to a qR_g<1.1, and molecular weight determination by extrapolation to I(0) and comparison to a glucose isomerase control were carried out using BioXtas RAW(107) and the equation, MW_p = I(0)/c_p [MW_std/(I(0)_std/c_std)](108) where MW_p is the sample protein mass, MW_std is the standard protein mass, c_p is the sample protein concentration, c_std is the standard protein concentration, I(0)p is the sample I(0) from the Guinier approximation and I(0)std is the standard protein I(0) from the Guinier approximation. Guinier fits were plotted in Matlab or RAW. Buffer subtracted averaged scattering curves collected at three concentrations were averaged using PRIMUS, and distance distribution functions, R_g and D_{max} were calculated using AUTOGNOM(82, 109). Theoretical scattering data for MBP was generated using CRYSOl from 1ANF.pdb and plotted using AUTOGNOM. The Flory equation used to predict the R_g of the denatured CTD is, R_g=R_0 N^v \text{, where } R_0=1.927, v=0.598 and N=318, the number of amino acids in the D.melCTD portion of the MBP-D.melCTD fusion, including the C-terminal hexahistidine tag(90). Ensemble optimization models were generated using EOM 2.1, with averaged scattering curves and 1ANF.pdb as inputs(106, 113). Best agreement
of $R_g$ for EOM derived ensembles and the experimentally obtained data were obtained using a compact chain and native-like chain options for the apo and phospho CTD, respectively. Final ensembles were visualized using PyMol, oriented by MBP, and displayed adjacent to the structure of RNA polymerase II from 1Y1W.pdb for comparison(114). Flexibility analysis was carried out using ScÅtter version 2.3 and plotted in Excel or RAW.

**MALDI-TOF intact mass analysis**

Proteins at an initial concentration of 1 mg/mL were diluted four-fold using an aqueous solution of trifluoroacetic acid (TFA). The final concentration of TFA was maintained at 0.1% and the pH was verified to be <4 using pHydrion pH test paper (Sigma). Samples were desalted and concentrated using ZipTips with 0.6µL C₄ resin (Millipore) according to manufacturer instructions and eluted in acetonitrile/water/TFA (60/40/0.1 v:v:v). 0.5µL of desalted sample was spotted and overlaid with 0.5µL sinapinic acid matrix solution (saturated in acetonitrile/water 50/50 v:v) onto a metal sample plate. Spots were allowed to air dry at room temperature until crystals formed. MALDI-TOF spectra were acquired on a Voyager-DE™ PRO (Applied Biosystems) instrument using the preloaded myoglobin_linear parameters in negative mode. Laser intensity and number of flashes were manually adjusted to provide the greatest signal over noise. Spectra for the apoMBP-\textit{D.mel}CTD6HIS and phosphoMBP-\textit{D.mel}CTD6HIS were acquired independently from neighboring sample spots to reduce spectra variation. Spectra were analyzed and post-processed in Data
Explorer (Applied Biosystems). Spectra were baseline corrected and the Noise Removal utility was used to remove noise within two standard deviations. Masses were determined via single point calibration using the theoretical mass of unmodified apoMBP-\textit{D.mel}CTD6HIS (74,704.13 Da) as the reference standard for the highest intensity peak of the post-processed apoMBP-\textit{D.mel}CTD6HIS spectra. The resultant calibration was applied to the post-processed phosphoMBP-\textit{D.mel}CTD6HIS spectra to determine molecular weight.

**Mass Spectrometry for phospho-site identification**

Reduction of MBP-\textit{D.mel}CTD samples using 5 mM dithiothreitol (60 min at 55 °C) was followed by alkylation of reduced cysteines with 15 mM iodoacetamide (30 min at room temperature in the dark). Samples were then diluted into 100 mM Tris-Cl (pH 8) containing 10 mM CaCl\textsubscript{2} and digested overnight at room temperature with chymotrypsin (1:50 enzyme to substrate ratio). Digests were desalted on C18 spin columns and resuspended at 1 µM with 0.1% formic acid for LC-MS analysis.

Chromatographic separations were performed using a Dionex Ultimate 3000 RPLC nanoLC system configured for preconcentration. Integrafrit trap columns (30 mm × 0.1 mm) and picofrit analytical columns (20 cm × 0.075 cm) were packed in-house using 3.5 µM Waters Xbridge BEH C18 (Milford, MA). Peptides were loaded onto the trap column in aqueous solvent containing 2% acetonitrile and 0.1% formic acid for 5 min at a flow rate of 5 μL/min. Water (mobile phase A) and acetonitrile (mobile phase B), each containing 0.1% formic
acid, were used with a linear gradient of 2-40% B over 60 min at a flow rate of 300 nL/min. All spectra were acquired in the Orbitrap analyzer of an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Higher-energy collisional dissociation (HCD) was performed using a normalized collisional energy (NCD) of 35% in a 3 msec top speed data-dependent method. Dynamic exclusion was enabled with an exclusion duration of 8.00 seconds. MS1 (from m/z 400-2000) and MS2 spectra were collected at resolving powers of 60K and 15K (at m/z 200), respectively.

Proteome Discoverer 2.0 with Sequest HT and ptmRS site localization software was used to search all data against a forward and reverse FASTA database containing the entire Drosophila melanogaster proteome. All searches included phosphorylation of serine, threonine, and tyrosine as a variable modification and carbamidomethylation of cysteine as a fixed modification. Mass tolerances of 10 ppm and 0.02 Da were used for precursor and fragment ions, respectively. Product ions typically observed during collisional activation (a, b, and y-type) were considered for spectrum matching. A fixed value PSM validator filtered matches based on a maximum delta C_n of 0.05. For both PSMs and peptides, strict and relaxed target FDR settings were 0.01 and 0.05, respectively. PtmsRS operating in PhosphoRS mode was used for phosphorylation site localization in which identified sites with isoform confidence probability less than 99% were further inspected manually.
**CTD Ladder Generation**

TEV sites were cloned into the pMalX MBP-\textit{D.mel}/CTD FlagStrep plasmid using in fusion cloning (Clontech) and expression and purification was done as described previously. **Figure 6-3** shows the products of the TEV cleavage reaction after an overnight digest. Note the aberrant migration of CTD fragments and their weak staining with Coomassie blue.

**Figure 6-3 CTD ladder cleavage fragments.** Depiction of the location of TEV sites in the CTD fusions (top panel) and a Coomassie stained SDS-PAGE gel of overnight TEV digestions (bottom). Samples were centrifuged for 5 minutes at
13,000xG and the pellet was resuspended in SDS-PAGE loading buffer. Equal amounts of pellet and supernatant (denoted P and S, respectively, were loaded) CTD fragments are annotated with a star and migrate aberrantly relative to their MWs. For example the 35kD fragment from CTD #4 runs above the maltose binding protein band, which is 42kD. The CTD fragments stain poorly and the smallest fragment generated by CTD #1 destains completely and is not visible in this gel.

Radio labeled CTD ladder was generated by mixing equal volumes of each of the four TEV constructs and labeling the mixture with γP$^{32}$ATP using Caesin Kinase II (NEB), Caesin Kinase II buffer, and 0.5mM cold ATP and incubating at 30°C for 30 minutes followed by buffer exchange into CTD buffer using a micro bio spin column (BioRad). This mixture was digested with TEV protease (a gift from Song Tan) in CTD buffer overnight at 34°C.

**Limited Proteolysis**

MBP-*D. meli*CTDFLAG-STREP or MBP-*H. sap*CTD6HIS were end labeled on the terminal CTD repeat with casein kinase II (New England Biolabs) in protein kinase buffer (50mM Tris-HCl pH 7.5, 10mM MgCl$_2$, 0.1mM EDTA, 2mM DTT, 0.01% Brij 35), 0.5mM ATP and $^{32}$PγATP(MP Biomedical) for 30 minutes at 30°C, and then buffer exchanged into CTD buffer using a BioRad BioSpin6 column. To reduce background signal from auto-labeled CKII in the MBP-*H. sap*CTD6HIS experiments, CKII was first pre-incubated in kinase buffer with 4mM ATP, in the absence of $^{32}$PγATP, for three hours at 30°C. This pre-incubation reaction was then added to a MBP-*H. sap*CTD6HIS labeling reaction
resulting in final reaction conditions as described above for labeling MBP-
*D.mel*CTDFLAG-STREP. End labeled CTD was digested in CTD buffer for 30
seconds at room temperature by adding protease concentrations ranging from
2ng to 6000ng per 12µl reaction. Reactions were quenched by the addition of
PMSF to a final concentration of 16mM and flash frozen on liquid nitrogen.
Reactions were thawed in SDS-PAGE sample buffer and resolved in 15% SDS-
PAGE gels and exposed to a phosphoimager. For comparison of digestion of apo
and phospho CTD, labeling and buffer exchange were carried out as described
above, and aliquots of equal volume were phosphorylated by P-TEFb overnight
at 24°C in CTD buffer containing 1mM ATP and 1/10th reaction volume of Protein
Kinase Buffer (NEB) or subject to a mock kinase reaction of identical composition
but lacking P-TEFb. The P-TEFb and mock reactions were mixed and
immediately subject to limited proteolysis as described above, and fragments
were resolved on a 12% SDS-PAGE gel. Signal intensity of the full length
proteins were quantified using ImageJ and data from three experimental
replicates, representing two separate labeling and kinase reactions, were plotted
as the percentage of remaining undigested signal relative to the no protease
control at each protein concentration.

**Flexibility Analysis.**

Plots were generated in Excel. q range for Kratky-Debye and Porod-Debye plots
was based on the flexibility analysis tools in Scatter version 3 and represent the q
range were plateaus in the data first become visible(104). Extrapolation to zero
concentration was done using SAMBUF in Primus(109) and the data were plotted in RAW(107).

**Drosophila CTD sequence Alignments**

Amino acid sequences of the Pol II CTD for twelve *Drosophila* species, beginning with the sequence YSPTSPNYTAS were aligned using MUSCLE(137) improved manually, and visualized using BOXSHADE.

**Fly strains and lethality test**

*UAS-Rpb1i* and yw; Act-GAL4/CyO and C5-Gal4 fly lines were obtained from the Bloomington Stock Center (BDSC 36830, 4414, and 30839, respectively). Sequences encoding the Rpb1\(^{WT}\), Rpb1\(^{W81}\) or Rpb1\(^{hu}\) with double FLAG-tags at the C-termini were subcloned into the pUASt-attB vector, followed by transformation into the *PhiC31 attP 86Fb* strain(170). Rpb1i-resistance of the ectopically expressed Rpb1 variants was achieved by changing the part of the coding sequence of Rpb1 that is targeted by the 21 nucleotide shRNA (sense strand: AACGGTGAAACTGTCGAACAA) to AACCGTCAAGTTGAGCAACAA. The UAS-Rpb1i, UAS-Rpb1 lines were generated by routine matings and meiotic recombination. Rescue tests were done by mating virgin female yw; Act-GAL4/CyO (for ubiquitous expression) or C5-GAL4 (for wing expression) to male *UAS-Rpb1i, UAS-Rpb1\(^{WT}\) or mut*. Animals were raised at 24°C for wing experiments or 21°C for adult fly rescue experiments. For wing experiments, at least 50 individual offspring were examined and photographs were taken of
representative individuals. For adult rescue experiments, rescue was assessed by the emergence among the progeny of straight-wing adults (Act-GAL4/+; UAS-Rpb1i, UAS-Rpb1WT or mut/+ or UAS-Rpb1i, UAS-Rpb1WT or mut). 

Reverse transcription and quantitative PCR

Extent of endogenous Rpb1 knockdown was assessed by measuring endogenous Rpb1 or ectopically expressed Rpb1 using qPCR. At least three groups of 4-10 rescue animals or offspring from yw control crosses were collected and RNA was isolated using 100ul of Trizol reagent (Invitrogen). cDNA was prepared using MMLV reverse transcriptase (Promega) with random hexamers and oligo dT. cDNA was used to template qPCR reactions with a forward primer hybridizing to the 21nt region of wild-type or ectopically expressed Rpb1 mutated to generate RNAi resistance, and a reverse primer GCCTCCAGTTCTGGATG. Levels of endogenous Rpb1 were normalized to Actin expression levels (primers TCAGTCGGTTTATTCCAGTCATTCC and CCAGAGCAGCACTTCTTGATCA) and displayed as percentage of endogenous Rpb1 expression for the yw control crosses. Expression of ectopic Rpb1 was assessed by normalizing expression levels to Rp49 (primers TACAGGCCAAGATCGTGAA and ACGTTGTGCACCAGGAACTT) and displayed as a percentage of expression from UAS-Rpb1i, UAS-Rpb1WT rescue animals.
**Immunofluorescence**

Salivary glands from third instar larvae from the same cross as described in the RNAi lethality test were dissected and squashed as previously described (171) and incubated first with anti-FLAG M2 monoclonal antibody (Sigma Aldrich) and anti-Rpb3 antisera overnight at 4°C, and then with goat anti-mouse Alexa Fluor 568 conjugated IgG and goat anti-rabbit Alexa Fluor 488 conjugated IgG secondary antibodies for 3hr at room temperature. DNA was stained with Hoescht dye. Images from multiple channels were overlaid in Photoshop.
APPENDIX A

CG3773 IS THE DROSOPHILA ORTHOLOG OF GDOWN1

A-1 Introduction

In 2006, a 13\textsuperscript{th}, substoichiometric subunit of RNA Pol II was characterized by the Gnatt and Roeder laboratories\cite{172}. This 13\textsuperscript{th} subunit is the product of the \textit{POLR2M} gene in mammals and was named Gdown1. Gdown1 was found to be associated with \textasciitilde30-50\% of Pol II in mammalian cells\cite{172}. The Pol II form associated with Gdown1, (PolII(G)), can be separated from Pol II lacking Gdown1 chromatographically, with each Pol II variant eluting as a separate peak off an anion exchange column\cite{172}. Pol II(G) was found to be competent for elongation in a promoter independent \textit{in vitro} transcription assay, but required mediator for activation in a promoter driven transcription assay that included purified general transcription machinery, mediator complex, and transcriptional activators\cite{172}.

Subsequently, David Price and colleagues characterized the function of Gdown1 using a combination of biochemical and genomic approaches and concluded Gdown1 has a role in stabilizing paused RNA polymerase II\cite{173}. As such, Gdown1 serves to connect transcription initiation via mediator, to the early elongation control phenomena of promoter proximal pausing. However, the role of Gdown1 in pausing is not completely clear from the data presented in this work, and some data argue against a role in stabilizing paused Pol II\cite{173}.
Promoter proximal pausing is mediated by two complexes, negative elongation factor (NELF) and DRB sensitivity inducing factor (DSIF), which bind Pol II early in elongation and stop its forward translocation ~50nt downstream of the transcription start site (TSS) in metazoans (174-176). Phosphorylation of DSIF and NELF by the kinase P-TEFb is thought to relieve the pause, enabling Pol II to enter productive elongation (177, 178). Curiously, Gdown1 appears to antagonize the negative elongation activities of DSIF and NELF in in vitro transcription reactions where those pausing factors were added to Pol II(G) compared to Pol II, suggesting Gdown1 renders Pol II resistant to pausing (173). Furthermore, the authors note an anti-correlation between the amount of Pol II in the promoter region of genes, and the amount of Gdown1 (179).

The Gilmour lab has developed an assay that recapitulates Pol II pausing that is promoter driven and dependent on DSIF and NELF (180, 181). The reaction is carried out in Drosophila nuclear extract and recapitulates the location of the pause relative to the TSS, is dependent on core promoter elements and is sensitive to depletion of NELF and DSIF from the extract (180). By incorporating a Pol II immunodepletion step prior to this assay, or purifying Pol II and Pol II(G) from tagged, α-amanatin resistant mutant flies, this assay could be a powerful platform to dissect the role of Gdown1 in early elongation control.

To facilitate this and other biochemical analysis, I sought to identify and purify the Gdown1 ortholog in D. melanogaster. To this end, I generated a fly line ectopically expressing Rpb1 under the control of the Hsp83 gene promoter with a 2X FLAG tag on the C-terminus. Using a combination of ion exchange and
affinity chromatography, I purified two chromatographic forms of Pol II. Using mass spectrometry, I identified the product of the CG3773 gene as the putative ortholog of POLR2M in *Drosophila*. Western blot analysis identifies two bands that react to anti-Gdown1 serum in crude extract, but only one form from purified Pol II(G). This observation is consistent with recent work identifying two phospho-isoforms of human Gdown1, only one of which associates with Pol II(182). Together, these observations identify CG3773 as the putative POLR2M homolog in *Drosophila*, and provide a means for the specific enrichment of Pol II and Pol II(G) from *Drosophila* embryos that could be used in future biochemical studies.
A-2 Results

A-2.1 Generation and characterization of a 2X FLAG-tagged Rpb1 fly line

There was no cDNA available encoding the full Rpb1 gene. In order to generate an Rpb1 transgene with C-terminal FLAG tags, I amplified a portion of Rpb1 from the LD43558 cDNA (Drosophila Genome Resource Center) and another portion of the gene from gDNA. The resulting fragments were cloned into pCaSpeRhs83 to be inserted into the genome of yw flies using p-element insertion (Rainbow Transgenic Flies Inc.) to be expressed constitutively under the control of the Hsp83 gene promoter (Fig A-1). Expression of epitope tagged Rpb1 protein was validated by western blot from 0-12hr embryos lysed in SDS-PAGE sample buffer, and via immunofluorescence of polytene chromosomes from third instar salivary glands (Fig. A-2). One transgenic fly line, dubbed 10M.11M was positive for anti-FLAG staining in both the western blot and immunofluorescence experiments and was also capable of complementing an early embryonic lethal Pol II allele in experiments carried out by David Gilmour. This line was chosen for purification of Pol II.
**Figure A-1 Map of the FLAG-Rpb1 genomic insert.** FLAG-Rpb1 contains a double flag tag of the sequence DDYKDDDDKLDYKDDDDK on the C-terminus. This tag can be used to both aid in purification as well as distinguish wild-type Pol II from CTD mutants. The transgene was inserted into the genome using P-element insertion. FLAG-Rpb1 is under the control of the Hsp83 promoter, which is constitutively expressed but also heat shock inducible. This approach has been employed successfully in the expression and purification of FLAG-NELF-D from *Drosophila* embryos(176).
**Figure A-2** FLAG-Rpb1 is expressed in the 10M.11M fly line. (a). Western blot of 0-12hr embryo lysates from transgenic fly lines probed with anti-FLAG M2 antibody (Sigma Aldrich). (b) Immunofluorescence staining of polytene chromosomes from salivary glands dissected from third instar larvae from the 10M.11M transgenic line. DNA is stained in blue, FLAG-Rpb1 in green and Rpb3 in red. FLAG-Rpb1 and Rpb3 co-localize across the chromosomes suggesting FLAG-Rpb1 is incorporated into Pol II complexes.
A-2.2 Purification of Pol II and Pol II(G)

It was observed previously that mammalian Pol II activity could be separated into two chromatographic peaks by eluting the bound Pol II from an anion exchange column with a salt gradient (172). The peak eluting at the lower salt concentration was found to be tightly associated with Gdown1. I sought to isolate the two species of FLAG Pol II and subsequently immuno-purify each population using anti-FLAG resin. The anti-FLAG column should remove any proteins that co-elute with Pol II from the ion exchange column, leaving behind only those proteins associated with the Pol II complex.

In brief, I collected 0-14hr old 10M.11M embryos, and generated an embryo lysate that was passed over a Poros heparin column (GE Life Sciences). Pol II activity was assayed with an in vitro transcription assay and the peak fractions were pooled and passed over a Mono Q column (GE Life Sciences). This strong anion exchange resin resolved the single peak of activity obtained from the Poros heparin column into two peaks of activity (Fig. A-3).
Figure A-3 Pol II purification strategy and activity assay results. (a) Pol II was successively purified over Poros heparin and Mono Q columns and eluted with gradients. The peaks of activity from the Mono Q column were separately immuno-purified with anti-FLAG resin. (b and c). Activity assay results from fractions eluted off the Poros heparin (a) and Mono Q columns (b). The Mono Q column separates Pol II into two peaks of activity.

Fractions from each peak were pooled, and aliquots were purified over an anti-FLAG column and eluted with FLAG peptide. The resulting fractions, before and after immuno-purification, were analyzed on an SDS-PAGE gel and a number of bands specific for peak 1 were identified (Fig. A-4). An aliquot of immuno-purified peak 1 and peak 2 Pol II was TCA precipitated and resolved on a SDS-PAGE gel (Fig. A-5). A band migrating above 40kD was specific for peak 1. This band was excised from the gel along with putative Rpb1 and Rpb2 bands.
as controls and was identified via mass spectrometry to be the protein product of CG3773, a protein with homology to mammalian Gdown1 (PSU Mass Proteomics Core Facility).

**Figure A-4 Silver stained SDS-PAGE gel of peak 1 and peak 2 Pol II.** Equal volumes of peak 1 and peak 2 Pol II from the Mono Q column, before and after immuno-purification with anti-FLAG resin are resolved on a 10% SDS-PAGE gel and silver stained. A band later identified as the protein product of CG3773 is highlighted. As yet identified bands specific for peak 1 are highlighted.
Figure A-5 Coomassie stained gel of TCA precipitated FLAG peak 1 and peak 2 Pol II for mass spectrometry. Peak 1 and peak 2 FLAG immunopurified samples were TCA precipitated and resolved on a 12% SDS PAGE gel. The labeled bands were excised for mass spectrometry analysis and labeled according to their protein identification. A gel slice from peak 2 corresponding to the MW of the Gdown1 band was used as a negative control.
The unknown band was identified by mass spectrometry to be the gene product of CG3773, the putative Drosophila homolog of Gdown1. Roeder lab had previously generated antiserum against recombinant CG3773 protein, which they provided. This antiserum specifically recognized a band that migrated at the same position as Drosophila Gdown1, and this band is specific for Mono Q peak 1 (Fig. A-6). Thus, I conclude that peak 1 Pol II is Drosophila Pol II(G).

<table>
<thead>
<tr>
<th>Mono Q Peak 1</th>
<th>Mono Q Peak 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction #</td>
<td>11</td>
</tr>
</tbody>
</table>

![Image of western blot](image)

Figure A-6 anti-Rpb3 and anti-Gdown1 western blot of mono Q peak 1 and peak 2 pol II. Gdown1 is specific for peak 1. Gdown1 levels scale with Rpb3 levels.
I next sought to determine if the Gdown1 antiserum specifically recognized Gdown1 from crude embryo lysates. A crude lysate from dechorionated 0-12hr embryos was subject to western blot analysis alongside Pol II and Pol II(G) fractions. Notably, the Gdown1 antiserum reacts with two closely migrating bands in crude extract. Only the faster migrating of these bands associates with Pol II (Fig. A-7). This lead me to speculate that a post-translational modification of Gdown1, possibly phosphorylation, creates a slower migrating species of Gdown1 that fails to associate with Pol II, a hypothesis later corroborated by Price and colleagues(182). Notably, my efforts to purify the slower migrating “free” form of Gdown1, failed. Efforts to purify Gdown1 free from Pol II have also failed in the Roeder lab (Miki Jishage, private communication).
Figure A-7 anti-Gdown1 western blot of embryo lysate and purified Pol II.
Two closely migrating bands react with Gdown1 antiserum from crude embryo lysates. Only the faster migrating species associates with Pol II(G).
A-3 Discussion

I have determined that the protein product of CG3773 is the Drosophila homolog of Gdown1. Like the mammalian Gdown1, Drosophila Gdown1 associates with a distinct chromatographic species separable from Pol II using anion exchange. I have presented a purification strategy that can enrich for Pol II(G). It has been speculated that Pol II(G) is involved in promoter proximal pausing, but the data are unclear. Isolation of Pol II(G) could enable comparison of the association of NELF and DISF with Pol II and Pol II(G) in electrophoretic mobility shift assays. Further, if Pol II can be effectively immunodepleted from nuclear extracts, pausing activity of Pol II and Pol II(G) could be compared using in vitro transcription reactions. Finally, other bands that co-immunopurify with Pol II(G) were visible in a silver stained gel. Identification of these bands by mass spectrometry could reveal other Pol II(G) associated factors. The use of affinity tagged Pol II to identify these bands reduces the likelihood they are simply factors that co-elute with Pol II(G) rather than bound to the Pol II complex itself.

The 10M.11M fly line used in these studies no longer exists. However, this line served as a proof of concept for the many tagged Pol II lines currently in the lab. Defining CG3773 as the likely Gdown1 homolog in flies could motivate efforts to make a tagged Gdown1 fly line with epitope tags distinct from those currently in use on Rpb1. In this way, a fly line with both tagged Rpb1 and tagged Gdown1 could be generated. This would enable affinity purification of total Pol II, and subsequent enrichment for Pol II(G).
The Gdown1 antibody Roeder lab provided failed to stain polytene chromosomes or identify any locations of specific enrichment of Gdown1 in ChIP-EXO experiments. The antibody may only recognize Gdown1 when denatured and/or may recognize an epitope buried upon association with Pol II. Epitope tagged Gdown1 could mitigate this issue and better Gdown1 genome-wide data could do much to clarify the role of Gdown1 in promoter proximal pausing. My efforts to characterize Gdown1 genome-wide using ChIP-EXO failed for Gdown1, but included successful experiments for GAGA factor, Rpb3, NELF-E and Spt5.

Thus, Gdown1 ChIP-EXO data derived from a tagged Gdown1 cell line could be compared with NELF and DSIF localization to shed light on the role of Gdown1 in pausing. Finally, I suggest CRISPR as a means to generate epitope tagged Gdown1 in cells or flies. Gdown1 is a sub-stoichiometric subunit of Pol II. As such, altering Gdown1 levels though ectopic expression could skew the ratio of Pol II to Pol II(G) and confuse the interpretation of in vivo experiments.
A-4 Materials and methods

Cloning double FLAG tagged Rpb1

The 5’ portion of Rpb1 was amplified from gDNA purified from S2R+ cells using primers rpb1_redo_F and rpb1_redo_R to generate a 5’ Rpb1 fragment containing the first two introns. This fragment was subcloned into pCaSpeRhs83 that contained the 3’ region of Rbp1 derived from the LD43558 cDNA clone from the Drosophila Genome Resource Center as an Hpal/Nhe1 fragment. The resulting clone contained a single C-terminal flag tag and was verified by sequencing using primers 8821, CaSpeR L, 3’rpb1_seq1, 3’rpb1_seq2, 3’rpb1_seq3 and 3’rpb1_seq4. A second FLAG tag was added to the C-terminus of this construct by PCR amplifying the 3′ region of Rpb1 with Rpb1_DoubleFLAG_F and Rbp1_DoubleFLAG_R primers and subcloning this amplicon into FLAG Rbp1 pCaSpeRhs83 that was digested with Kpn1 and Not1. The resulting 2xFLAG Rpb1 pCaSpeRhs83 plasmid was verified by sequenced using rpb1_Seq1, CaSpeRhs83L, 8821, 3’rpb1_Seq1, 3’rpb1_Seq2, rpb1_upstream, 3’rpb1_seq4, rpb1_seq3, rpb1_seq5, rpb1_seq6, rpb1_seq7, rpb1_Cterm1, rpb1_seq8, CaSpeRhs83R primers. The sequences of these primers are available in the Gilmour Lab primer inventory. The plasmid was sent to Rainbow Transgenic Fly for microinjection.

Immunofluorescence of Polytene Chromosomes

Polytene chromosomes squashes were carried out as described previously in Chapter 7 materials and methods.
Purification of Pol II

0-12hr embryos were collected from 10M.11M transgenic flies and dechorionated in 50% bleach, washed with 0.7% NaCl and 0.04% Triton X-100, rinsed with diH<sub>2</sub>O and frozen at -80°C. Embryos were homogenized in 40ml of 300mM HGAE (25mM HEPES pH7.6, 15% glycerol, 300mM ammonium sulfate, 0.1mM EDTA 1mM PMSF, 16µg/ml benzamidine HCL, 10µg/ml aprotinin, 10µg/ml peptstatin 10µg/ml leupeptin, 5µg/ml soybean trypsin inhibitor and 1mM DTT) with 10 strokes each of the loose and tight fitting pestle in a dounce. The lysate was passed through miracloth and centrifuged 1hr at 100,000xG at 4°C. The clarified supernatant was diluted in 0mM HGAE (same recipe as above but lacking ammonium sulfate) and loaded onto a 20ml Poros heparin column pre-equilibrated with 200mM HGAE (same as above but 200mM ammonium sulfate).

The column was washed with 5 column volumes of 200mM HGAE and Pol II was eluted with a 6 column volume gradient from 200mM to 600mM HGAE and 10ml fractions were collected and assayed for activity. Fractions 6-8 had peak activity and were pooled and adjusted to 200mM salt by adding 37.5ml 0mM HGAE. The sample was loaded onto a MonoQ 10/10 column pre-equilibrated with 200mM HGKE (same as HGAE, but with 200mM KCl instead of ammonium sulfate). The column was washed with 3 column volumes of 350mM HGKE. The sample was eluted from the Mono Q column with a 8 column volume gradient from 350mM HGKE to 600mM HGKE and collected in 2ml fractions which were assayed for activity.
Fractions 12-14 were pooled as “peak 1” and 22-24 were pooled as “peak 2.” The salt concentration of peak 1 was determined to be 420mM and peak 2 was 450mM, based on conductivity relative to a standard curve generated from an HGKE dilution series. The two samples were loaded onto separate 1ml anti-FLAG M2 (Sigma Aldrich) columns previously equilibrated with 6 column volumes of 400mM HGKE. The flow through was passed back over the column a second time. The columns were washed with 10 column volumes of 400mM HGKE and the sample was eluted with 5 x 500µl elutions of 400mM HGKE plus 100µg/ml 3x FLAG peptide (Sigma Aldrich).

**Mass spectrometry sample preparation**

300µl of FLAG elution 3 and 500ml of FLAG elution 2 were pooled, TCA precipitated and resuspended in 25ul of 1x SDS PAGE sample buffer containing unbuffered Tris. The samples were boiled and loaded onto a 12% SDS-PAGE gel, coomassie stained, and the bands of interests were excised for mass spectrometry by the PSU proteomics core facility.

**Western Blots**

Western blots were probed with anti-Gdown1 antisera (Roeder Lab) at 1:2500 and anti-Rpb3 antisera at 1:2500 diluted in 2% milk in TBS-T pH 8.0 and probed with anti-rabbit alexa-fluor conjugated anti-rabbit IgG (GE Life Sciences) diluted 1:3000 in 2% milk TBS-T pH 8.0.
APPENDIX B

NUCLEOSOME ORGANIZATION IN  D. MELANOGASTER IS STABLE IN RESPONSE TO HEAT SHOCK

B-1 Introduction

In 2008 the Pugh lab succeeded in mapping nucleosome positioning across the genome of Drosophila embryos(183). This work discovered an ordered array of nucleosomes across many genes beginning with a +1 nucleosome centered 135-145bp downstream from the transcription start site(183). This finding is notable, as it positions the +1 nucleosome immediately downstream of the leading edge of paused Pol II, and thus implicates the nucleosome in contributing to promoter proximal pausing, perhaps by acting as a physical barrier into which the polymerase collides(183) (Fig. B-1).

The NELF complex is a key component in Pol II pausing(174). Work from Karen Adelman and colleagues employed RNAi against NELF to reduce levels of paused Pol II in Drosophila tissue culture cells and found that nucleosome organization was also disrupted(184). This work found that genes with high levels of paused Pol II that also exhibited the greatest loss of paused Pol II upon NELF knockdown are actually depleted of nucleosomes in the promoter proximal region(184). This finding argues against a role for nucleosomes in establishing paused Pol II. Adelman and colleagues also found that NELF RNAi mediated loss of paused Pol II lead to increased nucleosome occupancy on many promoters, and that these promoters have sequences that would be predicted to
favor nucleosome occupancy. Together these observations suggesting a competition exists between paused Pol II and nucleosomes for promoter proximal DNA.

Work by Jian Li from the Gilmour lab provides a means of unifying the observations that nucleosomes may both contribute to pausing while also being largely absent from paused genes. Jian discovered that pausing is kinetically regulated by the recruitment of the NELF complex to Pol II(180). Further, many paused genes in Drosophila fall into two, largely mutually exclusive, categories based on whether their promoters are bound by GAGA factor or M1BP(180, 185). GAGA factor aids in recruiting NELF to Pol II, leading to focused pausing close to the transcription start site. Notably, GAGA factor bound genes have low nucleosome occupancy and weak positioning, so focused pausing is likely achieved independent of collision with the +1 nucleosome(180, 185). GAGA genes are also enriched for stress and stimulus response genes, and are not highly transcribed under normal conditions(180, 185). In contrast M1BP bound genes are enriched for constitutively transcribed housekeeping genes, have high nucleosome occupancy and well positioned nucleosome arrays. M1BP genes have less focused pausing that occurs further from the transcription start site compared to GAGA factor genes. Thus, the +1 nucleosome may contribute to pausing on M1BP genes, perhaps by imposing a barrier that slows the polymerase, providing time for the association of NELF(185).

Pol II ChiP-chip of MNase digested chromatin suggests a physical interaction between Pol II and +1 nucleosomes on many genes(183). NELF
knockdown revealed a relationship between the occupancy of paused Pol II and nucleosomes on many genes(184). The caveat to the RNAi experiment is that NELF plays a role in regulating transcription elongation on thousands of genes and the RNAi treatment takes place over days(175, 184). Thus, the likelihood of secondary effects stemming from the aberrant regulation of thousands of genes is high.

Staining *Drosophila* polytene chromosomes with antibodies directed against pol II subunits before and after heat shock reveals a genome-wide loss of Pol II (**Fig. B-1**). Jian undertook a more refined analysis using genome wide permanganate footprinting followed by ChIP-seq (KMnO4 ChIP-seq)(180). This analysis reveals a reduction of paused Pol II on nearly all genes after heat shock.

**Figure B-1** *Immunofluorescence of polytene chromosomes before and after heat shock.* (a) Staining of polytene chromosomes from 2X FLAG Pol II lines before (a) and after (b) heat shock. FLAG-Rpb1 and Rpb3 co-localize. Staining is largely reduced to the heat shock loci after 15 minutes of heat shock.
I wanted to use the physiological response to heat stress as a means to interrogate the interplay between nucleosomes and paused Pol II. Based on the results from Adleman and colleagues, I posited that a genome wide loss of Pol II would be accompanied by an upstream shift in the position of the +1 nucleosome. Based on observations made in the Pugh study, which noted that the register of the nucleosome arrays on paused genes is shifted 10bp downstream relative to non-paused genes; I hypothesized that a shift in the +1 nucleosome would be accompanied by an upstream shift in the nucleosome array. If these changes were observed, this might represent a chromatin based mechanism for the widespread transcriptional repression observed during heat shock for nearly all non-heat shock genes. To this end, I performed MNase ChIP-seq in *Drosophila* tissue culture cells before and after heat shock using an antibody directed against histone H3. I found that, in contrast to my hypothesis, genome-wide nucleosome organization is maintained during heat stress. This stable nucleosome organization may aid in recovery by providing “memory” of the chromatin status of the cell prior to transcriptional reprogramming by heat stress, and thus may aid in re-establishing transcriptional homeostasis.

While I was undertaking analysis of the MNase ChIP-seq data, Shiela Teves and Steven Henikoff published a study with a similar basis (186). They mapped nucleosome position and dynamics following heat shock in *Drosophila*. Their published work corroborates our observation that paused Pol II is reduced by heat shock and nucleosome organization is maintained (186). The Teves work
went a step beyond my study in that it also mapped nucleosome turnover, which was found to be reduced during heat stress (186).

**B-2 Results**

**Figure B-2 MNase ChIP-seq before and after heat shock.** MNase ChIP-seq was carried out on S2R+ cells and the data were mapped to +/-1kb of active transcription start sites as defined and annotated by Gilchrist et al (184). (a) My data and analysis recapitulate the pattern of nucleosome organization identified
previously by Mavrich et al.(183). (b) My data and analysis affirm the nucleosome organization characteristic of M1BP and GAGA factor bound genes noted by Li and Gilmour(185). GAGA bound genes are depleted of nucleosomes. M1BP bound genes have well positioned nucleosomes. (c) Nucleosome organization is maintained after 15 minute heat shock on ~7800 genes that were active prior to heat shock. (d) A heatmap of MNase ChIP-seq reads +/-1kB from the TSS before and after heat shock. The pattern of nucleosome organization is similar before and after heat shock.

**B-3 Materials and methods.**

**MNase ChIP-seq sample preparation.**

Four S2R+ cultures were grown in 15ml Shields and Sang M3 media (Sigma Aldrich) supplemented with 10% FBS (HyClone) to a total cell number of 30-40 x 10^6. For heat shocked cells, the media was replaced with pre-warmed 37°C media containing 10% FBS and the flask was placed on an aluminum block in a 37°C water bath for 15 minutes. For mock heat shock cells, the media was replaced with pre-warmed 24°C media and the cultures were returned to the 24°C incubator for 15 minutes.

For crosslinking, the media was removed from each flask and replaced with 15ml filter sterilized PBS containing 1% formaldehyde at room temperature. After 1 minute of crosslinking, the reaction was quenched by the addition of 770µl 2.5M glycine and the cells were incubated for 5 minutes. The cells were removed from the flask with a scraper, harvested by centrifugation, washed with 1x BPS, pelleted, and resuspended in Adelman’s cell suspension buffer + PI cocktail.(184) The cells were transferred to a glass dounce and homogenized
with 10 strokes in a glass dounce. The lysate was pelleted and the nuclei resuspended in Adelman’s buffer B + PI cocktail at $5 \times 10^6$ cells/100µl. To each aliquot, 40 units of MNase (Worthington) was added and the nuclei were digested in a thermomixer at 25°C for 30 minutes. The reaction was stopped with EDTA to a final concentration of 10mM and SDS to 0.05%. To solubilize the digested chromatin, the samples were sonicated in the Biorupter for 5 cycles at 30s on/30s off on low intensity. The resulting sample was subject to MNase ChIP-seq using an antibody against histone H3 (Abcam #1791) according the Pugh lab protocol(183).
References


219


Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*


139. Flory, P. J. (1953) Principles of polymer chemistry


142. Albert, A., Lavoie, S., and Vincent, M. (1999) 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. *Journal of Cell Science*. **112**, 2493–2500


cells. *eLife Sciences*


VITA

EDUCATION
The Pennsylvania State University, Department of Biochemistry and Molecular Biology
B.S. in Microbiology 2008

PUBLICATIONS


SELECTED PRESENTATIONS
Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Invited talk at the lab of Aaron Giltler, Department of Genetics, Stanford University School of Medicine. Stanford, California, April 2017

Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Invited talk at the first Invited Postdoctoral Candidate Symposium. Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine. Charlottesville, Virginia, November 2016

Bede Portz, Feiyue Lu, Eric B. Gibbs, Scott A. Showalter and David S. Gilmour. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Poster at the Gordon Research Conference on Intrinsically Disordered Proteins. Les Diablerets, Switzerland, June 2016* *Poster Award


Bede Portz, Eric B. Gibbs, Scott A. Showalter and David S. Gilmour. A structural interrogation of the RNA polymerase II C-terminal domain. Poster at the 34th Symposium in Molecular Biology, Chromatin and Epigenetic Regulation of Transcription. Pennsylvania State University, University Park, Pennsylvania, July 2015

Bede Portz, Jian Li and David S. Gilmour. Mapping the promoter proximal nucleosome landscape in response to the heat shock induced loss of paused RNA polymerase II in Drosophila S2R+ cells. Poster at the Keystone Symposium on Eukaryotic Transcription. Snowbird, Utah, April 2012* *Wrote conference proceedings for Keystone Symposia and sponsors

SELECTED AWARDS
The Pennsylvania State University, Eberly College of Science.
Robert T. Simpson Graduate Student Award for Innovative Science 2016

Gordon Research Conference on Intrinsically Disordered Proteins
Gordon Research Conference Poster Award 2016

The Pennsylvania State University, Department of Biochemistry and Molecular Biology
Paul M. Althouse Outstanding Teaching Award 2012