Abstract

RNA polymerase II CTD Evolutionary Diversity and Associated Protein Identification in Green and Red Algae

by Chunlin Yang

December, 2014

Director of Dissertation: Dr. John W. Stiller, Department of Biology

Interdisciplinary Doctoral Program in Biological Sciences

In model eukaryotes, the C-terminal domain (CTD) of the largest subunit (RPB1) of DNA-dependent RNA polymerase II is composed of tandemly repeated heptads with the consensus sequence YSPTSPS. Both the core motif and tandem structure generally are highly conserved across many model taxa, including animals, yeasts and higher plants. Broader investigations quickly revealed that the CTDs of many organisms deviate substantially from this canonical structure; however, limited sampling made it difficult to determine whether disordered sequences represent the CTD's ancestral state, or reflect degeneration from an originally repetitive structure. Therefore, I undertook the broadest investigation to date of the evolution of the RNAP II CTD across eukaryotic diversity. The results indicate that a tandem heptad CTD-structure existed in the ancestors of each major taxon, and that degeneration and reinvention of this ordered structure are common features of CTD evolution. Lineage specific modifications of heptads that were amplified initially appear to be associated with greater developmental complexity in multicellular taxa. The pattern has been taken to an extreme in both fungi and red algae. Overall, loss and reinvention of varied repeats have punctuated CTD evolution, occurring independently and sometimes repeatedly in various groups.

Although present in simple, ancestral red algae, CTD tandem repeats have undergone extensive modifications and degeneration during the evolutionary transition to developmentally complex rhodophytes. In contrast, CTD repeats are conserved in both green algae and their more complex land plant relatives. Understanding the mechanistic differences that underlie these variant patterns of CTD evolution requires knowledge of CTD-associated proteins in these two lineages. To provide an initial baseline comparison, potential phospho-CTD associated proteins (PCAPs) were bound to artificially synthesized and phosphorylated CTD repeats from the unicellular green alga *Chlamydomonas reinhardtii* and red alga *Cyanidioschyzon merolae*. My results indicate that red and green algae share a number of PCAPs, including kinases and proteins involved in mRNA export. There also are important taxon-specific differences, including mRNA splicing-related PCAPs recovered from *Chlamydomonas* but not *Cyanidioschyzon*, consistent with the relative intron densities in green and red algae. This work also offers the first experimental indication that different proteins bind the two types of repeats in *Cyanidioschyzon*, suggesting a division of function between the proximal and distal CTD, similar to patterns identified in more developmentally complex model organisms.

RNA POLYMERASE II CTD EVOLUTIONARY DIVERSITY AND ASSOCIATED PROTEIN IDENTIFICATION IN GREEN AND RED ALGAE

A Dissertation

Presented to

The Faculty of the Interdisciplinary Doctoral Program in Biological Sciences

The Brody School of Medicine, East Carolina University

In Association with the Department of Biology, Thomas Harriot College of Arts and Sciences

Submitted in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Interdisciplinary Doctoral Program in Biological Sciences

By

Chunlin Yang

December, 2014

©Copyright 2014
Chunlin Yang

RNA polymerase II CTD Evolutionary Diversity and Associated Protein Identification in Green and Red Algae

by

Chunlin Yang

APPROVED BY:

DIRECTOR OF DISSERTATION:____
John W. Stiller, PhD

COMMITTEE MEMBER:____
Tim Christensen, PhD

COMMITTEE MEMBER:____
Allison Danell, PhD

COMMITTEE MEMBER:____
Paul Hager, PhD

COMMITTEE MEMBER:____
Jingling Huang, PhD

DIRECTOR, INTERDISCIPLINARY DOCTORAL PROGRAM IN BIOLOGICAL SCIENCES: ____
Terry L. West, PhD

DEAN OF THE GRADUATE SCHOOL: _________________________________
Paul J. Gemperline, PhD

Acknowledgements

Time really flies! More than four years have been passed since I first came to ECU. I still remember vividly everything that happened on that day as if it was yesterday. That day I first saw my advisor Dr. John W. Stiller, an energetic and enthusiastic man with gray hair and beard. He warmly received me, showed me his lab and invited me to a lunch. I still remember he told me that he always gives students the highest degree of freedom and independence. During the past four years, just like he said, he granted me the greatest freedom to let me arrange all my time for study and research myself. But whenever I have difficulties, he will stand up timely to help me out. In this way, we had a very happy time to work together and got two projects finished and published. During this period, he provided me greatly on academic guidance, which undoubtedly is invaluable for my further scientific career. I must say his influence on me is not just academic. Actually, his vitality and passion always infected me to be a better person. He is full of energy all the time like always 25 years old, and always passing his positive energy around. I am so lucky to be involved in such an atmosphere, and I appreciate everything he has helped and infected me.

I would like to thank all my committee members for their time and help! Dr. Jingling Huang helped me with Bioinformatics; Dr. Paul Hager helped me with protein purification, which was critical for finishing my second project; Dr. Tim Christensen provided me many valuable suggestions on my projects; Dr. Allison Danell helped me with Mass Spectrometry. Their invaluable help and suggestions made me a more professional and efficient researcher. I thank them all!

I am grateful to Dr. Terry West for his help and suggestions on my research in the past years! Thanks to Dr. Eric Anderson, Dr. Beth Thompson, Dr. Carol Goodwillie, Dr. Charlene Ding, Michael Reubens, Joyce Beatty, Denise Mayer and others in the Biology Department for their generous helps! Many thanks to my lab colleague and office mate Justin Perry.

I would like to thank the IDPBS program, Department of Biology and the Graduate school for providing this study opportunity to me. Also, I am grateful for the International House for providing visa support to me.

At last, I would like to express my appreciation to all my families, especially my wife and parents for their support and consideration.

Table of Contents

Title..i

Copyright...ii

Signature………………………………………………………………………………...iii

Acknowledgements……………………………………………………………………...iv

List of Tables……………………………………………………………………………ix

List of Figures……………………………………………………………………………x

Abbreviations……………………………………………………………………………xi

Chapter 1: Introduction 1

The consensus sequence of the CTD 3

The CTD is essential for life.. 3

The smallest functional units of the CTD... 4

The CTD Post-Transcriptional Modifications and Associated Functions...................... 6

Tyr1 Phosphorylation and Associated Functions... 7

Ser2 and Ser5 Phosphorylations and Associated Functions...................................... 8

Thr4 phosphorylation and associated functions... 11

Ser7 phosphorylation and associated functions 12

Chapter 2: Evolutionary Diversity and Taxon-Specific Modifications of the RNA polymerase II C-Terminal Domain 13

Background 13

Results 15

The CTD Originated with Tandemly Repeated Heptads 15

The CTD Has Expanded and Diversified With Developmental Complexity in Animals and Plants 17

Parallel CTD Evolution in Fungi and Red Algae 19

CTD Diversity across Protist Groups 22

CTD Evolution in the Apicomplexa 23

Discussion 25

The Origin of the CTD 25

The Evolution of the CTD across Eukaryotic Diversity 28

Materials and Methods 33

Data Collection 33

CTD Annotation 33

Character Evolution Analysis 34

Chapter 3: The identification of putative RNA polymerase II C-terminal domain associated proteins in green and red algae 36

Background 36

Results 39

Potential PCAPs with Functions Shared in Both Algae 39

Potential PCAPs Found Only in C. reinhardtii 43

Potential PCAPs Only in C. merolae 44

Discussion 52

Materials and methods 56

Cell Culture and Lysis 56

Ammonium Sulfate Precipitation 57

Ion Exchange Chromatography 57

Affinity Chromatography 58

Protein Annotations 59

Phylogenetic Analyses of CDKs 60

Chapter 4: Conclusion 61

References.. 63

List of Tables

Table 1. The 7 potential *Chlamydomonas* PCAPs identified....................................... 79

Table 2. The 8 potential *Cyanidioschyzon* PCAPs identified. 80

Table 3. The 154 identified proteins from *C. reinhardtii*.. 81

Table 4. The identified 133 proteins from *C. merolae*... 85

Table 5. The identified 55 proteins from *E. coli*. 89

List of figures

Fig. 1. CTD diversity in eukaryotes. 91

Fig. 2. The character state analysis with rooting close to Excavata. 93

Fig. 3. The character state analysis with rooting between Bikonta and Unikonta. 94

Fig. 4. CTD evolution in fungi. 95

Fig. 5. CTD evolution in the Apicomplexa. 96

Fig. 6. Sub-motif SP content comparison. 97

Fig. 7. The CTD in green plants and red algae. 98

Fig. 8. Phylogenetic analyses of CDKs. 99

Fig. 9. PCAP purification process. 101

Fig. 10. The gels of *E. coli* proteins running both heptad and nonatad affinity columns. 102

Fig. 11. The gels of *Cyanidioschyzon* proteins running both heptad and nonatad affinity columns. 103

Fig. 12. The gel of *Chlamydomonas* proteins running heptad affinity column. 104

Fig. 13. Control affinity column gels. 105

Abbreviations

RPB1: Largest subunit of RNA polymerase II

CTD: Carboxyl-terminal domain

PCAPs: Phospho-CTD associated proteins

CDK: Cyclin-dependent kinase

Chapter 1: Introduction

DNA-dependent RNA polymerase, found in all prokaryotic and eukaryotic organisms, is essential for life based on its role in transcribing RNAs from DNA templates (Hurwitz 2005). Prokaryotes, both Eubacteria and Archaea, only contain one type of RNA polymerase, which is responsible for all RNA transcriptions (Ebright 2000; Werner 2007). During the evolution from prokaryotes to eukaryotes, RNA polymerase diverged into several family members with different roles. Mainly, there are three basic types of eukaryotic RNA polymerases, I, II, and III. RNA polymerase I is responsible for transcribing 28S, 18S and 5.8S rRNAs (Grummt 1999); RNA polymerase II is in charge of synthesizing mRNAs, most snRNAs and microRNAs (Lee, Kim et al. 2004); RNA polymerase III takes on the role of synthesizing tRNAs, 5S rRNAs and some other small RNAs (Willis 1993). In land plants, however, there are another two specific RNA polymerases, IV and V, which are involved in non-coding RNA-mediated gene silencing processes (Haag and Pikaard 2011).

Among the three eukaryotic RNA polymerases, RNA polymerase II is the one that has been most studied because of its role mRNA synthesis and processing. RNA polymerase II is a large, ~ 550 kDa, complex containing 12 subunits, named RPB1 through RPB12 based on subunit sizes (Myer and Young 1998). Interestingly, the largest subunit of RNA polymerase II (RPB1) has a special C-terminal extension, which is absent from all other types of RNA polymerases. This C-terminal extension was first discovered in 1985 both in yeast (Allison, Moyle et al. 1985) and mouse cells (Corden, Cadena et al. 1985). The yeast C-terminal extension contains 26 tandemly repeated heptapeptides (heptads), and the mouse contains 52, both with a consensus sequence of $Y_1S_2P_3T_4S_5P_6S_7$ (Tyrosine-Serine-Proline-Threonine-Serine-Proline-Serine). The RPB1 C-terminal extension that contains tandemly repeated heptads was named the C-terminal domain (CTD) by these early researchers. The N-terminal sequence of the CTD is a linker region with about 90 amino acids in both budding yeast and mouse connecting the tandem repeats with the universally conserved H domain of RPB1, and the C-terminal sequence of the CTD is a tip region with about 20 amino acids in budding yeast and about 10 in mouse. Although a C-terminal extension exists in all sequenced eukaryotic RPB1 genes, the cumulative data soon revealed that not all organisms contain noticeable CTD tandem repeats in their C-terminal extensions. This was especially true for a number of parasitic protists, such as *Giardia* and *Entamoeba*, whose C-terminal extensions do not have any identifiable heptads. However, evidence emerged that, despite the absence of tandem repeats, the C-terminal extension is still essential in these eukaryotes (Das and Bellofatto 2009). Therefore, for convenience, more and more of the published literature tended to mention the CTD as the whole C-terminal extension, with the three regions: the linker region, the central region containing heptads when present, and the tip region (Corden 2013). In this dissertation, I also will use the term "CTD" to describe the whole RPB1 C-terminal extension, regardless of its structure. In this first chapter, I would like to review some basic knowledge about the CTD. Chapters 2 and 3 contain details of the two CTD-associated projects that I have finished during my Ph.D. study period. Chapter 4 provides some overall conclusions from the complete study.

The consensus sequence of the CTD

The earliest CTD sequences came mainly from animals and unicellular fungi, and the most common (>50% in yeast and vertebrates) heptad in their CTDs is YSPTSPS. Consequently, this sequence was considered to be the consensus sequence of the CTD, and most CTD-related studies over the past few decades basically have focused on investigating the functions of the consensus CTD heptad repeats. An increasing availability of CTD sequences revealed that ratios of this consensus sequence are very low in many taxa, for example multicellular fungi. In addition, for certain eukaryotic taxa, for example, Stramenopiles (or Heterokonts), the most common heptad is YSPTSPA, and YSPTSPS is rarely seen. Even so, as CTD research has basically focused on vertebrates and unicellular yeasts, and YSPTSPS is still commonly considered as the consensus sequence of the CTD.

The CTD is essential for life

The discovery of the CTD, and especially its unique heptad repeats, inspired the interest of many scientists to investigate its functions. The first CTD functional investigations were carried out immediately following the domain's discovery and involved creating truncation mutants in both budding yeast and mouse to determine their effects on viability. The studies showed that removal of all or most of the CTD resulted in death, and partial deletions showed variable results (Nonet, Sweetser et al. 1987; Allison, Wong et al. 1988). For budding yeast, mutants with fewer than 11 heptads failed to support viability; and for mouse, the minimum heptad number required for viability was shown to be 29 (Nonet, Sweetser et al. 1987). A latter study with improved methods for mutant development showed that viability in budding yeast requires as less as 8 CTD heptads, although the mutants with fewer than 13 repeats are sensitive to temperature and other stresses (West and Corden 1995). CTD truncation investigations were carried out in fission yeast several years ago, and revealed that, of the 29 heptads present in the CTD, only the proximal 16 ones are required for viability (Schneider, Pei et al. 2010). Further mammalian CTD truncation mutants also were conducted and showed that fewer than 23 consensus heptads result in death (Bartolomei, Halden et al. 1988; Meininghaus, Chapman et al. 2000). An interesting CTD truncation study was even conducted in *Trypanosome*, an ancient unicellular parasite without any heptads at all, and showed that complete truncation of the CTD is lethal (Das and Bellofatto 2009). All these studies showed that the CTD is essential for life, and that complete truncations are lethal. Moreover, the fact that partial truncations, with a number of consensus repeats remaining, supported viability suggested that CTD heptads are functionally redundant.

The smallest functional units of the CTD

Because of the repeated nature and redundancy of heptads, the CTD must contain a number of functional units. Stiller and co-workers conducted a study that inserted amino acid(s) between every heptad or every other heptad in budding yeast. This work showed that insertions between every heptad were lethal, whereas insertions between every other heptad supported viability (Stiller and Cook 2004). Therefore, their study revealed that in budding yeast the smallest functional units of the CTD lies within pairs of heptads (Stiller and Cook 2004). Similar results also were obtained later in fission yeast study (Schwer and Shuman 2011; Schwer, Sanchez et al. 2012). Thus, research into CTD function in divergent unicellular fungi all suggested that the smallest functional unit of the CTD requires two tandemly repeated heptads.

Studies also were carried out to investigate the essential amino acids in each consensus heptad by creating various substitution mutants. In budding yeast, these substitution investigations showed that, for each consensus heptad, the substitutions of Tyr1, Ser2 and Ser5 with Ala or Glu were lethal, but that substitutions of Thr4 and Ser7 with Ala supported viability, while the substitution of Ser7 with Glu turned out to be lethal (West and Corden 1995; Stiller, McConaughy et al. 2000; Liu, Greenleaf et al. 2008; Liu, Kenney et al. 2010). Further research combining amino acid substitutions together with insertions revealed that, for each smallest or “core” functional unit, “Y1-Y8’’ and ‘‘S2-S5-S9” are the two essential elements that must be conserved in di-heptads in budding yeast (Liu, Kenney et al. 2010). However, similar studies conducted in fission yeast showed that substitutions of Tyr1 with Phe supported viability, as were substitutions of Ser2 with Ala (Schwer and Shuman 2011; Schwer, Sanchez et al. 2012). Thus, cumulative genetic studies in yeast suggest that core the relative size and spacing of CTD functional motifs is conserved, but that their sequences can vary across organisms.

The studies also have been carried out in mammalian cells. Substitutions of Ser7 with Ala in human cells supported viability, but were lethal for the mutants containing substitutions of Ser7 with Glu or Thr (Egloff, O'Reilly et al. 2007). In chicken DT40 cells, substitutions of all Tyr1 with Phe were lethal (Hsin, Li et al. 2014). Moreover, although chicken CTD mutants with 26 consensus heptads are viable, substitutions of Ser2 or Ser5 with Ala in all 26 repeats were lethal, while universal substitutions of Ser7 with Ala were viable but were lethal with other amino acids, including Glu, Thr and Lys (Hsin, Xiang et al. 2014). Substitutions of Thr4 were also conducted in chicken cells and showed that Thr4 to Val substitutions did not support viability (Hsin, Sheth et al. 2011). All these studies suggest that Tyr1 and two Ser-Pro pairs are the core amino acids of each consensus heptad, and the Thr4 and Ser7 have more important functions in animals, whereas substitution in these two positions are more tolerated in yeasts, especially fission yeast.

The CTD Post-Transcriptional Modifications and Associated Functions

Cumulative studies revealed that CTD heptads adopt different modification patterns to interact with different transcription factors during transcription cycles, and viable phosphorylations are the main modification patterns of the CTD (Corden 2013; Eick and Geyer 2013). For each CTD heptad, there are five amino acid positions that can be phosphorylated, including Tyr1, Ser2,5,7, and Thr4. Besides phosphorylations, the two prolines can adopt *cis* or *trans* isoforms (Zhang, Rodriguez-Molina et al. 2012). Given the large number of heptads and varied modification possibilities of each heptad, it was proposed that the CTD might have codes that use different post-transcriptional modifications to interact with different proteins, and that these codes could be conserved somewhat from yeast to animals (Buratowski 2003). CTD functional studies during the past decade revealed that, although a strict CTD code does not really exist, CTDs do have some very common modification patterns that are conserved from yeast to animals (Corden 2013; Eick and Geyer 2013). Therefore, in this section the different CTD post-transcriptional patterns and associated functions will be reviewed.

Tyr1 Phosphorylation and Associated Functions

Phosphorylation of Tyr1 was first identified in HeLa nuclear extracts using phospho-Tyrosine antibodies, and the kinase c-abl was shown to be associated with Try1 phosphorylation (Tyr1$_{\mathrm{P}}$) (Baskaran, Dahmus et al. 1993). Two years ago, Dirk Eick's laboratory generated a specific monoclonal antibody (3D12) against Tyr1$_{\mathrm{P}}$, and used this antibody to conduct chromatin immunoprecipitation (ChIP) studies. Their work revealed Tyr1$_{\mathrm{P}}$ profiles during transcriptional cycle, which showed that Tyr1$_{\mathrm{P}}$ levels gradually increased from the transcription start site and began to decrease from ~180bp upstream of polyadenylation (pA) site (Heidemann and Eick 2012; Mayer, Heidemann et al. 2012). Further ChIP investigations were conducted to investigate the relationships between Tyr1$_{\mathrm{P}}$ and transcriptional factors, and the results suggested that Tyr1$_{\mathrm{P}}$ impairs recruitment of termination factors such as Nrd1, Pcf11, and Rtt103, but stimulates the interaction with the transcriptional elongation factor Spt6 (Mayer, Heidemann et al. 2012). This study suggested that Tyr1$_{\mathrm{P}}$ might be used to avoid early transcript termination by impairing the CTD ability to interact with termination factors during transcriptional elongation (Mayer, Heidemann et al. 2012). In vitro kinase analysis was also performed in this study and provided further support that c-abl acts as the kinase for Tyr1 phosphorylation. A more recent study showed that Tyr1 functions in protecting the CTD from proteolysis, and that Tyr1 phosphorylation is responsible for regulating uaRNA (upstream antisense RNA) accumulation by ensuring uaRNA turnover (Hsin, Li et al. 2014). Another study published nearly at the same time showed that Tyr1$_P$ is associated with antisense promoter and enhancer transcription (Descostes, Heidemann et al. 2014). As for the phosphatase of Tyr1$_P$, a recent *in vitro* study showed Tyr1$_P$ might be erased by Rtr1, which is a dual specificity phosphatase capable of dephosphorylating both Tyr1$_P$ and Ser5$_P$ (Hsu, Yang et al. 2014).

Ser2 and Ser5 Phosphorylations and Associated Functions

The most thoroughly investigated CTD modification patterns are Ser2 and Ser5 phosphorylations. Cumulative ChIP assays have shown the common Ser2 and Ser5 phosphorylation profiles during transcription cycle. Ser5$_P$ reaches the highest level immediately after RNAP II clears the transcription start site, and decreases gradually during the transcription elongation process. In contrast, Ser2$_P$ level is very low early in transcription, but gradually increases during elongation and reaches its highest level when the polymerase is close to the 3’ UTR starting site, and that relative high levels even last until RNAP II reaches the pA site (Tietjen, Zhang et al. 2010). Thus, Ser5$_P$ usually is dominant early in RNAP transcription, whereas Ser2$_P$ is dominant when transcription is close to ending. During the middle stages of transcription elongation, the most common pattern is bi-phosphorylation of Ser2 and Ser5.

The main kinase responsible for Ser5 phosphorylations is cyclin-dependent kinase 7 (CDK7) in mammalian cells and its counterpart in budding yeast, Kin28 (Bartkowiak and Greenleaf 2011). For kinases that phosphorylate Ser2, cumulative studies showed that, in mammalian cells, the major players are members of the CDK9 subfamily including CDK9, and CDKs12, 13 (Bartkowiak and Greenleaf 2011). In budding yeast there are two CDK9 subfamily members, Bur1 and Ctk1, and studies showed they are in charge of phosphorylations of Ser2 (Bartkowiak and Greenleaf 2011). The main phosphatase that erases Ser5$_{P}$ is Ssu72 (Corden 2013). Recent research showed Rtr1 also serves as a Ser5$_{P}$ phosphatase in yeast (Mosley, Pattenden et al. 2009), however, and suggested that Rtr1 is responsible for removing the phosphate from Ser5 early in transcription elongation, whereas Ssu72 is responsible for erasing the phosphate closer to the transcription termination site (Krishnamurthy, He et al. 2004). For Ser2$_{P}$, the main phosphatase is Fcp1, which performs its function late in transcription elongation and termination (Ghosh, Shuman et al. 2008).

The best established Ser5$_{P}$ function is to promote addition of a m7GpppN cap structure to the 5’ end of new message RNA transcripts by physically interacting with capping enzymes (Ghosh, Shuman et al. 2011). This 5’capping involves three enzymes in yeast, RNA 5’-triphosphatase (RT), Guanylyltransferase (GT) and RNA (guanine-N7) Methyltransferase (MT) (Cho, Takagi et al. 1997; McCracken, Fong et al. 1997). In animals, however, the two enzymes RT and GT have been fused into one enzyme, which is called capping enzyme (CE) (Ho, Sriskanda et al. 1998). Studies showed that in budding yeast, GT (Cet1) and MT (Abd1) both bind directly to the Ser5$_{P}$ CTD (Cho, Rodriguez et al. 1998), and the GT domain of Mammalian CE physically interacts with the Ser5$_P$ CTD (Ho and Shuman 1999; Fabrega, Shen et al. 2003; Schroeder, Zorio et al. 2004). Structure investigations of interactions between the CTD and capping enzymes were carried out both in yeast and mammalian cells, and showed that different CTD conformations interact with capping enzymes (Burley and Sonenberg 2011).

Ser2 and Ser5 bi-phosphorylation is the most common CTD pattern during transcription elongation, and is responsible for interacting with varied elongation factors, chromatin remodeling factors (e.g., set1 and set2), and mRNA splicing factors, such as prp40, U2AF65 (Corden 2013). Based on the large number of transcriptional mRNA processing factors that interact with Ser2 and Ser5 bi-phosphorylated CTD, studies that use a phospho-CTD to pull down CTD associated proteins in vitro are mainly performed by using a Ser2 and Ser5 bi-phospho-CTD (Carty and Greenleaf 2002; Phatnani, Jones et al. 2004). So that my results would be comparable to such previous and ongoing research, I also used this method to identify phospho-CTD associated proteins in green and red algae (see Chapter 3).

Ser2 phosphorylation is barely seen early in transcription. It achieves its highest level during late transcript elongation. Cumulative studies show that Ser2$_P$ is mainly responsible for interacting with mRNA 3' end processing factors, such as Rtt103 and Pcf11 (Corden 2013). Pcf11 is one of the most important termination factors. In vitro assays carried out in early 2000s and revealed that the binding between Pcf11 and the CTD requires Ser2 phosphorylation (Licatalosi, Geiger et al. 2002). Further structure analyses confirmed the presence of a specific CTD interaction domain in Pcf11 and the Ser2 phosphorylated CTD (Lunde, Reichow et al. 2010).

Thr4 phosphorylation and associated functions

Several years ago, James Manley's laboratory constructed three types of CTD mutants using DT40 chicken cells with Rpb1 gene conditional knock-outs (Hsin, Sheth et al. 2011). The mutants were as follows: DT40-Rpb1, which contains a tet-repressive cDNA encoding HA-tagged wild-type human Rpb1; DT40-26r, which contains 26 consensus heptads along with the most C-terminal residues; and DT40-T4V, which contains 30 heptads with all Thr4 residues mutated to Valines. Primary viability analyses of the three types of mutants revealed that DT40-Rpb1 and DT40-26r were both viable, but DT40-T4V was not. Overall transcriptional comparisons conducted among the three mutants showed no significant differences; however, levels of histone mRNAs were significantly reduced for DT40-T4V compared with the other two mutants. Further investigations demonstrated that Thr4 phosphorylation is required specifically for histone mRNA 3' end processing, and also that CTD kinase CDK9 could be responsible for Thr4 phosphorylations (Hsin, Sheth et al. 2011). Another study carried out by Dirk Eick's group further revealed that Thr4 phosphorylation is conducted by Polo-like kinase 3, and also suggested Thr4_{P} is required in transcription elongation (Hintermair, Heidemann et al. 2012). A more recent study from Manley's laboratory showed that Thr4 genetically links with the histone variant Htz1, showed a functional connection between transcription and chromatin remodeling via CTD Thr4 (Rosonina, Yurko et al. 2014).

Ser7 phosphorylation and associated functions

The first study that discovered Ser7 phosphorylation was carried out in Dirk Eick's laboratory several years ago using monoclonal antibodies, and Ser7 phosphorylation was found on polymerase actively transcribing genes (Chapman, Heidemann et al. 2007). Among the mutants constructed in their study, those ones only containing 20 consensus repeats showed no Ser7 phosphorylation, suggesting functional differences between regions of the animal CTD. A study conducted at nearly the same time related $Ser7_P$ to snRNA gene expression based on the fact that mutants with all Ser7 substituted by Alanines were deficient in snRNA gene expression (Egloff, O'Reilly et al. 2007). This study further revealed that phosphorylations of Ser7 facilitate CTD interactions with the snRNA gene-specific Integrator complex. A follow-up study showed that during transcription of snRNA genes, RPAP2 (RNA polymerase II associated protein 2) was recruited by $Ser7_P$, and, in turn, facilitates the recruitment of Integrator (Egloff, O'Reilly et al. 2007). As for the kinases that phosphorylate Ser7 residues, a study in Bentley's laboratory demonstrated that CDK7 functions as one of Ser7 kinases (Glover-Cutter, Larochelle et al. 2009). Another study conducted by Ansari's group found that Ssu72 serves as a $Ser7_P$ phosphatase in budding yeast (Zhang, Mosley et al. 2012)

Chapter 2: Evolutionary Diversity and Taxon-Specific Modifications of the RNA polymerase II C-Terminal Domain

Background

The largest subunit of RNA polymerase II (RPB1) has a unique C-terminal domain (CTD) that, in its canonical form, is composed mainly of tandemly repeated heptads with a consensus sequence YSPTSPS. It has been more than a quarter century since the CTD was first described in yeast (Allison, Moyle et al. 1985; Corden, Cadena et al. 1985), where both global functions and constraints on its evolution are most thoroughly understood (West and Corden 1995; Stiller and Hall 2002; Guo and Stiller 2004; Stiller and Cook 2004; Liu, Greenleaf et al. 2008; Buratowski 2009). In yeast and animals, the CTD mainly functions as a docking platform to recruit transcription and processing factors to RNAPII at appropriate stages of the transcription cycle (Phatnani and Greenleaf 2006; Egloff and Murphy 2008; Buratowski 2009; Bartkowiak, Mackellar et al. 2011). To date, cumulative research has revealed that the factors recruited by the CTD are related to a variety of functions, such as mRNA 5’ capping, mRNA 3’ end processing, pre-mRNA splicing, histone modification and snRNA processing (Hsin and Manley 2012; Corden 2013; Eick and Geyer 2013). Moreover, the CTD uses different codes to recruit different protein factors (Buratowski 2003; Egloff and Murphy 2008; Zhang, Rodriguez-Molina et al. 2012; Jasnovidova and Stefl 2013). Reversible phosphorylation of Ser2 and Ser5 residues are the primary CTD codes, and are crucial for regulating transcription and binding mRNA processing factors (Phatnani and Greenleaf 2006; Heidemann, Hintermair et al. 2013); the major kinases responsible for these phosphorylations are conserved from yeast to metazoans (Bartkowiak and Greenleaf 2011). The CTD adopts additional modifications to enrich its functions, including Tyr1 (Baskaran, Dahmus et al. 1993; Mayer, Heidemann et al. 2012), Ser7 (Chapman, Heidemann et al. 2007), and Thr4 phosphorylations (Hsin, Sheth et al. 2011; Hintermair, Heidemann et al. 2012), as well as *cis/trans* isomerization of Pro3 and Pro6 (Egloff and Murphy 2008; Werner-Allen, Lee et al. 2011).

Despite its essential nature and conservation of multiple core functions across model organisms, when and in what form the CTD originated remains unclear, as do reasons for the remarkable diversity in CTD sequences and structures across eukaryotic species. The last major explicitly phylogenetic treatment of broad scale CTD evolution was published over ten years ago and suggested the presence of a “CTD clade” of associated major taxa, all descended from a common ancestor, in which canonical CTD heptads and functions are invariably conserved (Stiller and Hall 2002; Stiller and Cook 2004). This, in turn, suggested that a “critical mass” of CTD-protein interactions could have coalesced in the common ancestor of this group, after which the canonical CTD became indispensable to cellular function. With the acceleration of DNA sequencing over the last decade, the number of CTD sequences available from diverse organisms has grown substantially. It is now clear that evolutionary processes leading to conservation and degeneration of the CTD are far more complicated than suggested by early evolutionary studies (Chapman, Heidemann et al. 2008; Corden 2013; Stump and Ostrozhynska 2013). Moreover, a recent combined experimental and comparative analysis of mechanistic constraints on the yeast CTD revealed that many fungi have experienced changes across the domain that are incompatible with functional requirements established in the yeast model *Saccharomyces cerevisiae* (Liu, Kenney et al. 2010). Given the CTD's centrality to the entire RNAP II transcription cycle, this degree of degeneration is surprising. Therefore, I undertook a comprehensive investigation of the evolution and diversity of the CTD, both within and among major eukaryotic phyla.

Results

The CTD Originated with Tandemly Repeated Heptads

A global phylogenetic tree reflecting current best estimates of relationships among eukaryotic genera was constructed based on the Tree of Life Web Project and NCBI Taxonomy. The tree included all genera for which CTD sequences were available, and overall CTD structures were mapped onto the tree (Fig. 1). An interesting and consistent pattern emerged: in all major taxa, except the Ciliophora and "supergroup" Excavata, the most deeply branching taxa have the least modified CTD structures; that is, the most basal taxa contain CTDs consisting of simple, tandem repeats with few modifications. In contrast, indels, substitutions or even wholesale degeneration of the CTD's repetitive structure tend to occur in later diverging taxa, particularly in more developmentally complex, multicellular forms. It is interesting to note that maximum-likelihood analyses (see below) inferred the ancestral presence of a repetitive CTD even in groups for which no well-organized CTD has yet been sequenced. For example, although no tandemly repeated CTDs have been found among the handful of ciliates examined to date, the evolutionary pattern still holds when the nearest major sister group to ciliates, the apicomplexans, are considered (Fig. 1). The Excavata is another large super-taxon containing various eukaryotic groups with great diversity. Although the CTD sequences from most excavates sampled have no apparent CTD motifs, the *Naegleria* sequence displays a highly ordered tandem structure, whereas a single canonical heptad is present in the trichomonad *Pentatrichomonas*. Thus, it is reasonable that a tandemly repeated CTD structure was present in the ancestors of all major taxa currently recognized, and that degeneration of this initial tandem structure is a common feature of the CTD evolution.

I addressed this hypothesis more rigorously through maximum-likelihood character evolution analysis, using four assigned states based on the overall structure of each CTD sequence (see methods). Analyses were performed using two commonly suggested roots of the eukaryotic tree, the Excavata and between the Unikonta and Bikonta (Stechmann and Cavalier-Smith 2003). With the former rooting, ML analysis indicated a 49.51% probability that the eukaryotic common ancestor had a CTD with tandemly repeated heptads, versus a 48.52% probability of a random CTD sequence; however, the common ancestors of all other taxa except Excavata had 99.96% or greater probabilities of containing tandemly repeated heptads (Fig. 2). The latter rooting resulted in a 99.79% likelihood that the CTD had a tandemly repeated structure in the eukaryotic common ancestor (Fig. 3). Therefore, contrary to early conclusions based on more limited sampling (Stiller and Hall 1998; Chapman, Heidemann et al. 2008), it appears that the CTD originated as tandemly repeated heptads before the divergence of all (or at least most) extant eukaryotic taxa, and that those taxa with no recognizable CTD repeats have undergone degeneration rather than reflect the ancestral state of the CTD.

The CTD Has Expanded and Diversified With Developmental Complexity in Animals and Plants

Animals and land plants have achieved the greatest developmental diversity and complexity in the eukaryotic world, and interestingly, they have parallel patterns of CTD evolution. The CTD in animals is conserved to different degrees in different taxa. In the phylum Chordata, all 22 genera examined have almost identical CTD sequences with 52 tandem repeats, although serine codon usage (TCx or AGC/T) is slightly different in proximal heptads among more distantly related organisms. Likewise, three nematodes (from *Caenorhabditis* to *Loa*, Fig. 1), two (*Brugia* and *Loa*) from the same family, all have same CTD structures and serine codon usage. Interestingly the two available choanoflagellates (*Monosiga* and *Salpingoeca*), which share the closest common ancestor with metazoans (Lang, O'Kelly et al. 2002), have similar tandemly repeated CTD structures with only subtle differences in codon use. In contrast, in the phylum Arthropoda (*Ixodes* to *Solenopsis*), levels of CTD conservation are variable across orders, families and even within the same genus; for example, *Drosophila* species have several slightly different CTD patterns.

In general, the length of the CTD in animals appears positively correlated with greater evolutionary complexity, but this is not absolute since, for example, the more deeply branching and morphologically simple animal, *Hydra*, has the longest region of heptads among all known CTDs ($\approx$60 repeats). Given the generally dynamic nature of the CTD, however, it is likely that *Hydra* amplified extra repeats recently to acquire its surprisingly long heptad region, and has not yet lost them to a random mutation that could reset the CTD back to a more typical length. In fact, the extremely degenerated far distal region of the inferred *Hydra* CTD appears to reflect this very mutational process. I also found that the pattern of heptad variability first noted within mammalian CTDs, that is, the tendency toward canonical repeats in proximal regions with varied substitutions and/or indels in distal regions, is consistent across metazoan diversity, albeit most prominent in more developmentally complex animals like arthropods and chordates.

Previous broad scale sampling suggested that, in groups like metazoans that require more complex and well-programmed gene expression, a multiplicity of CTD-protein interactions prevent loss of an overall tandem CTD structure (Guo and Stiller 2005); however, recently sequenced CTDs from two flatworms (Platyhelminthes), *Clonorchis* and *Schistosoma*, show this not to be the case. Neither displays almost any vestige of a canonical CTD, so far a unique condition within the Metazoa. Interestingly, the CTD of their nearest available relative, the flatworm *Schmidtea*, is more typical of a metazoan CTD. Both *Clonorchis* and *Schistosoma* are parasitic trematodes, whereas *Schmidtea* is a free-living turbellarian; this highlights another interesting but not absolute association of the CTD, that of parasitic lifestyles with extreme modifications of the ancestral tandem heptads in a given group (see section below).

In general, CTD evolution in green plants has been analogous to that in animals. Five unicellular green algae available (from *Chlamydomonas* to *Bathycoccus*, Fig. 1) show similar tandemly repeated heptads but with largely different serine codon use. Likewise, the CTDs of two early-diverging land plant genera, *Physcomitrella* and *Selaginella*, have few or no substitutions in their distal repeats. More derived and developmentally complex angiosperms (*Sorghum* to *Ricinus*), however, contain longer heptad regions with more frequent substitutions or indels in their distal heptad regions. There is general conservation of CTD structure and serine codon usage in both monocot (*Sorghum* to *Hordeum*) and dicot (*Glycine* to *Ricinus*) taxa, with subtle differences between them. Interestingly this pattern of CTD modification associated with developmental complexity even seems to be followed in more simple green algae; less-derived chlamydomonad unicellular algae (e.g. *Chlamydomonas*) have canonical tandem heptads with nearly no substitutions or indels, whereas the colonial and more developmentally complex genus *Volvox* contains a more modified CTD, similar to derived land plants.

Parallel CTD Evolution in Fungi and Red Algae

Both fungi and red algae show parallel developmental evolution in that they have achieved complex, multicellular forms through the elaboration of filamentous rather than parenchymatous tissue differentiation. Interestingly, the two groups also display similar patterns of CTD evolution with remarkable deviations from the tandem heptad structure found in more developmentally complex forms (Fig. 1).

The CTDs of available chytridiomycetes (e.g., *Batrachochytrium*) and zygomycetes (e.g., *Mucor*), representatives of the ancestors of true fungi, have tandemly repeated heptads nearly without substitutions or indels (Fig. 4). The same is true for all microsporidian parasites (from *Antonospora* to *Nosema*, Fig. 4), although their classification as ancient fungi remains controversial (James, Kauff et al. 2006). In the more derived phylum Ascomycota (*Schizosaccharomyces* to *Claviceps*), unicellular yeasts in the Saccharomycotina display simple tandemly repeated CTDs. In the Pezizomycotina (*Arthrobotrys* to *Claviceps*), however, which form more complicated multicellular fruiting bodies, numerous alterations have occurred that result in regions that would be dysfunctional based on requirements known from mutational experiments in yeast (Liu, Greenleaf et al. 2008; Liu, Kenney et al. 2010). The pattern is especially striking in the Eurotiomycetes (*Exophiala* to *Coccidioides*), where few typical heptads and no CTD functional units (as characterized in yeast) occur. Based on the presence of tandemly repeated CTDs in more ancestral fungi, developmentally complex ascomycetes have taken an evolutionary pathway that resulted in the loss of repeated heptads through modification by individual substitutions and insertions/deletions. This could parallel lineage-specific adaptive modifications in the distal CTD regions of complex animals and plants, only without retention of a more canonical proximal set of tandem repeats in complex fungi. Similar but less extreme patterns of heptad modifications are found in the other pezizomycete classes. Interestingly, with few exceptions the overall structural patterns within these CTDs, even in serine codon use, are highly conserved at the taxonomic level of classes. This conservation is even more striking at the level of orders (Fig. 4). This suggests that co-adapted molecular processes that underlie the conserved developmental patterns reflected in class and lower-level systematic designations, also are reflected in conservation of CTD-protein interactions that regulate RNAPII driven gene expression.

The Basidiomycota (*Malassezia* to *Ceriporiopsis*, Fig. 1) is as comparably diverse as the Ascomycota, but far fewer CTDs have been sequenced. Nevertheless, all available basidiomycete CTD sequences show various degrees of modifications of ancestral heptads and, given the limited sampling, structural patterns and serine codon usage also seem to be conserved at the level of order. For example, members of the Polyporales (including *Trametes*, *Ceriporiopsis* and *Dichomitus*) have highly similar CTD structural patterns and serine codon use (Fig. 4). Thus, despite the paucity of available data, it is reasonable to expect that CTD evolution in basidiomycetes has proceeded comparably to what is observed in the better-sampled Ascomycota.

With respect to broad scale patterns of CTD evolution in fungi, it is intriguing that the basidiomycetes and pezizomycetes are predominantly multicellular fungi with more complex developmental patterns. In contrast, microsporidians, chytrids, zygomycetes and saccharomycetes are relatively simple developmentally, although a few have evolved multicellular forms (Kurtzman and Fell 2006). Thus, my results indicate that there are two distinct evolutionary trajectories for the CTD in fungi. Simple forms tend to retain canonical heptad repeats although varying degrees of differences in serine codon usage, suggesting that specific heptads were lost and regained regularly. In contrast, morphologically complex fungi tend to adopt extreme modifications in their CTDs, which are largely conserved at higher (order) classification levels. This perhaps reflects the evolution of strongly conserved lineage-specific CTD/protein interactions. Unlike in multicellular plants and animals, however, there appears to be no strong selection in developmentally complex fungi to maintain long stretches of tandem heptad repeats.

Based on sequences available from eight genera, it appears that CTD evolution in red algae followed a remarkably similar pattern to what occurred in fungi. The unicellular forms *Glaucosphaera*, *Cyanidioschyzon* and *Galdieria* all have a number of canonical heptad repeats, although *Cyanidioschyzon* has a surprising series of nine amino acid repeats with the sequence YSPSSPNVA, unique in all CTD sequences known. In contrast, the CTDs of five multicellular rhodophytes have almost no canonical heptads. Although taxon sampling is much weaker, this suggests that, as in fungi, large-scale modifications of ancestral heptads, along with reduced purifying selection on maintenance of a tandem structure, are correlated with the evolution of developmental complexity in red algae. It also is interesting that *Pyropia yezoensis* has a highly similar CTD structure to *Porphyra purpurea* and *P. umbilicalis*, although these algae have proven to be genetically distant (Sutherland, Lindstrom et al. 2011). This indicates another interesting parallel with the fungi that, although highly modified, CTD structures are relatively conserved at the level of order (Bangiales). As in fungi, this correlates with conserved life history and developmental similarities that traditionally placed *Pyropia* and *Porphyra* within the same genus (*Porphrya, sensu latu*).

CTD Diversity across Protist Groups

Stramenopiles (from *Aureococcus* to *Phytophthora*, Fig. 1) comprise a large and diverse group of eukaryotes that display a broad range of morphological complexity and ecological habits. The group includes photosynthetic members ranging from unicellular diatoms to giant kelp, as well as heterotrophic oomycetes and various non-photosynthetic protist taxa (Riisberg, Orr et al. 2009). At present, complete and well-annotated RPB1 sequences are available from only six genera; these are the diatom *Phaeodactylum* and pelagophyte *Aureococcus*, the multicellular brown alga *Ectocarpus*, and the filamentous oomycetes *Hyaloperonospora*, *Albugo* and *Phytophthora*. All six of them have long tandemly repeated heptad regions (YSPTSPA) in their CTDs with nearly no substitutions or indels.

Four ciliate CTD sequences are available and none displays a discernible tandem structure, or even recognizable individual heptads. In contrast, of the four CTD sequences available from amoebozoans, only the parasite *Entamoeba* lacks tandemly repeated heptad regions. The Excavata is a diverse eukaryotic supergroup composed of various unicellular species. At present, CTD sequences are available from six genera, five adapted to parasitism and one, *Naegleria*, predominantly free-living. The CTD of *Naegleria* contains 23 canonical heptad repeats, whereas the five CTDs from parasitic excavates have no discernible heptad structures, except for the single YSPASPL motif in trichomonad *Pentatrichomonas* noted earlier.

CTD Evolution in the Apicomplexa

As in most eukaryotic lineages, the most deeply branching apicomplexan, *Cryptosporidium*, has a CTD with a long array of tandemly repeated heptads. Beyond that, CTD evolution has been unusually dynamic in this group. CTDs from *Neospora*, *Theileria* and *Toxoplasma* all are highly degenerate with few canonical heptads, whereas *Babesia* contains numerous tandemly repeated heptads in its middle region with a different consensus sequence from those in *Cryptosporidium*. Most interesting is CTD evolution within the genus *Plasmodium*, for which CTD sequences are available from 10 different species (Fig. 5). Although both the proximal and distal CTD regions are highly conserved across the genus, at least two independent acquisitions of tandem heptads (YSPTSPK) have occurred in primate-infecting species (Kishore, Perkins et al. 2009). One was in the lineage containing *P. fragile*, *P. knowlesi*, and *P. vivax*, the other apparently in the common ancestor of *P. falciparum* and *P. reichinowi*. Even more interesting, the reamplified heptads vary in number (5 to 9) not only between species, but also among different strains of *P. falciparum* and *P. vivax*. Thus, it appears that both tandem heptad degeneration and reinvention have occurred repeatedly in the Apicomplexa, reflecting the global pattern of CTD evolution across the whole of eukaryotic diversity. This suggests that CTD evolution in Apicomplexa can provide, in microcosm, a model for how selective pressures could have shaped CTD evolution more broadly in eukaryotes.

Discussion

Our comprehensive analyses of available CTD sequences show that the phylogenetic distribution of a tandemly repeated structure does not support the earlier hypotheses of a “CTD-clade”, in which some “critical mass” of CTD/proteins coalesced to place strong purifying selection on a canonical and tandemly repeated CTD (Stiller and Hall 2002; Stiller and Cook 2004). In fact, tandemly structured CTDs are scattered across the eukaryotic tree of life, and appear to have been amplified, lost and reamplified from one or more heptads on numerous occasions. It is possible that CTD variation has been impacted by horizontal gene transfer (HGT) of alternative sequences from unrelated taxa; however, such transfers generally are not favored in genes encoding core informational proteins with multiple complex interactions (Jain, Rivera et al. 1999), and I find no empirical evidence of HGT in the sequences I analyzed. Likewise, broader sampling has shown that the CTD can degenerate in members of groups, for example multicellular animals, previously suggested to be incapable of surviving without a well-ordered CTD. My findings demonstrate that the canonical, tandemly repeated CTD has undergone a dynamic process of birth, modification/degeneration and rebirth throughout eukaryotic evolution. Nevertheless, the evolutionary patterns I highlight can provide new clues for understanding what drives CTD diversification.

The Origin of the CTD

Based on a more limited sample of CTD sequences and differences in serine codon use, Chapman and colleagues proposed that the heptads in the CTD were built up initially from smaller motifs (YSPx or SPxY; x represents any amino acid), and then amplified independently in various different eukaryotic lineages (Chapman, Heidemann et al. 2008). My comprehensive investigation of CTD evolution indicates that the extended RPB1 C-terminal domain, present in all RPB1 sequences known to date, originated as tandemly repeated heptads before divergence of extant eukaryotic groups. Therefore, differences in consensus heptads and serine codon use reflect the extremely dynamic evolution of tandem repeats rather than their independent origins.

A very early origin of the RNAP II CTD through relatively rapid amplification of one or a few initial heptad motifs raises a provocative question: what was the initial functional advantage of this new domain? The fact that the extended C-terminal domain was never lost from any lineage that diversified through evolutionary history suggests the CTD was, from its origin, connected to an essential function that also evolved in the common ancestor of extant eukaryotes. Thus, the most likely candidates are those CTD-associated processes that are widely distributed across eukaryotic diversity. It also seems most reasonable that initial selection was on a single function rather than complexes of proteins involved in more complicated pathways, and that it favored longer C-terminal extensions rather than a single binding domain. Given these caveats, I argue that the most likely ancestral function for CTD tandem repeats was as a platform for carrying out co-transcriptional pre-mRNA splicing. It is believed that the last common ancestor of all extant eukaryotes contained an extremely high density of introns in its protein-coding genes (Koonin 2009), apparently the result of a rapid invasion by group II parasitic self-splicing introns at the dawn of the eukaryotic domain. The spliceosome likely evolved as a mechanism to efficiently remove group II introns that lost the ability to self-splice (Rogozin, Carmel et al. 2012). It is reasonable, that the extended CTD evolved to permit spliceosomes to function co-transcriptionally, thereby increasing splicing efficiency and the overall rate of RNAP II transcription. Experimental results linking the CTD to exon recognition and the earliest stages of spliceosome assembly (Hirose, Tacke et al. 1999) suggest the two could have co-evolved in this manner. Effectively, the CTD could have originated as part of a genomic immune response to a massive invasion of genetic parasites.

Another possibility for the ancestral CTD function is as a platform for 5' capping, which appears to be conserved across the breadth of eukaryotes. Lethal CTD substitutions in fission yeast can be complemented by fusing capping enzyme to the CTD, suggesting that 5' capping could be the only essential CTD function in fission yeast (Schwer and Shuman 2011). As a single function, however, capping provides a less compelling explanation than splicing for why an extended array of tandem repeats would have been favored from the outset. In any case, once the domain was in place, it proved to be an attractive binding platform for a wide variety of other protein partners.

I proposed the following scenario for the CTD origin and its early evolution. First, as suggested by Chapman and colleagues (Chapman, Heidemann et al. 2008), submotifs such as YSP and SP evolved at the end of H domain of RPB1 in the eukaryotic ancestor through random mutations, finally in combination resulting in formation of one or more initial heptad (YSPxSPx) motifs. These heptads then were amplified by tandem duplications to create the first major C-terminal extension of RPB1. Such an origin of the original C-terminal extension distal to the H domain is consistent with numerous more recent CTD expansions through tandem duplications, for example those well documented in *Plasmodium* parasites (Kishore, Perkins et al. 2009), as well as nearly identical codon usage in many tandem CTD motifs across the breadth of eukaryotic diversity. The most prominent examples of the latter are proximal tandemly repeated heptad regions of more evolutionarily derived animals and plants. As the CTD grew longer, to extend more prominently from the core of RNAPII, the heptads in the linker region degenerated. The former presence of typical CTD heptads is reflected by the presence of the sub-motif SP, which, on average, is nearly thirty times more abundant in linker regions than in RPB1 from domains A through H (Fig. 6).

The Evolution of the CTD across Eukaryotic Diversity

The remarkable sequence diversity and variable serine codon use in CTD sequences across eukaryotes show that the domain’s evolution has been extraordinarily dynamic. Although CTDs of more deeply branching genera in nearly all major eukaryotic taxa contain clear tandem heptads, it is unlikely that these specific repeats were conserved from the CTD in their ancient common ancestor. Selection appears to have conserved the overall tandem structure of the CTD in ancestral eukaryotes, but not necessarily their underlying sequences at the amino acid or DNA levels. In other words, as long as a structurally unordered and reversibly modifiable docking platform was maintained, slightly different heptapeptides were functionally interchangeable. This has been demonstrated experimentally via evolutionary complementation for CTD function in yeast (Stiller, McConaughy et al. 2000). Once present, tandemly repeated sequences are easy to amplify, lose and reamplify during DNA replication (Corden 2013). The process most likely involves expansion of the CTD by repeated tandem duplications, balanced by degeneration of terminal sequences after random mutations introduced new 3' stop codons.

It appears that in developmentally simple organisms, selection balances replication and loss of heptads, thereby maintaining a given length of tandemly repeated structure. With the evolution of more developmentally complex eukaryotes, selection seemed to favor taxon-specific CTD modification. When accompanied by purifying selection on redundant and overlapping functions, this process also led to retention of tandem repeats and CTD structures like those found in complex land plants and animals. Without purifying selection on greater length and tandem repeats, accumulated modifications of the CTD lead to the appearance of moderate to complete degeneration as in multicellular fungi.

The two recent independent CTD heptad expansions in plasmodium parasites demonstrate how a tandemly structured CTD can be reinvented when required by the addition of new functions. The specific advantage conveyed to plasmodium species that parasitize primates as opposed to birds and rodents is unclear, but could involve the coincident acquisition of chromatin remodeling pathways not present in other apicomplexans (Kishore, Stiller et al. 2013). Regardless, it is clear that the CTD is extremely plastic in response to selection. Given the diversity and variation of CTD protein interactions across the eukaryotes (Corden 2013), it seems unlikely that specific evolutionary modifications from any given lineage will prove generally applicable. Rather, analogous selective pressures likely have yielded parallel patterns of CTD evolution.

The most tantalizing example is the similar patterns of CTD evolution in animals and green plants. The CTD grew longer in both developmentally complex forms in both, with tandemly repeated proximal regions retained along with somewhat modified distal regions. Presumably this was not accomplished by adding distal non-repetitive regions, but by adaptive evolution of the ancestral heptads toward specific functions combined with simultaneous or later additions of new canonical repeats upstream to permit more diverse and overlapping protein binding. In contrast, while CTD heptads underwent various levels of modification in both multicellular fungi and red algae, generally more severe than those in land plants and animals, neither group reamplified proximal tandem repeats. Thus, it appears likely that evolution of developmental complexity is associated with specific alterations of the CTD resulting in deviations from the ancestral tandem structure. In organisms that exhibit the greatest levels of cell and tissue differentiation, such as animals and land plants, transcription and processing functions associated with RNAP II appear to be too varied and complicated to be accommodated without an enlarged CTD, including a repetitive region that permits flexible, redundant function. An association of modified CTD regions with greater transcriptional efficiency required for multicellular development is supported by the observations that only the nonconsensus repeats 1-3 and 52 are essential for proliferation of mammalian cell cultures (Chapman, Conrad et al. 2005), whereas removal of other modified heptads causes retarded growth and increased neonatal lethality in the developing organism (Litingtung, Lawler et al. 1999). In contrast, multicellular fungi and red algae must have evolved lineage-specific functions that modified the ancestral heptads; however, perhaps based on a lesser overall need for complexity in gene expression, they did not re- or co-evolve tandemly repeated regions for more generalized CTD-protein interactions.

Unfortunately there are no comparative empirical data that directly tie specific functions to modified, conserved CTD regions in most organisms. Nevertheless, some studies involving specific CTD alterations provide direct evidence that heptad modifications in animals could be related to conserved, lineage-specific functions. For example, an investigation of the role of R1810 (an Arg7) in the human CTD indicates it is involved specifically in regulating expression of snRNA and snoRNA (Sims, Rojas et al. 2011). This could represent a more broadly applicable lineage-specific function because this Arg7 modification is conserved at a comparable position across chordates. A distal Arg7 also is found in some invertebrate genera, but a conserved specific position within the CTD is not apparent outside the chordate lineage.

It is unknown why developmentally complex fungi and red algae have lost the need for tandemly repeated heptads as their CTDs underwent extensive modifications associated with the evolution of multicellularity. It may not be coincidental, however, that both multicellular fungi and red algae have relatively simpler developmental programs. Although both groups have been considered plant-like historically, unlike land plants they do not exhibit coordinated cellular development required for elaboration of organs such as roots, stems, leaves and vascular tissues. It also is interesting that, thus far, the pattern I highlight is compatible with CTD evolution in stramenopiles, another group that has evolved complex multicellular forms. All unicellular stramenopiles (e.g., *Albugo*, Fig. 1) examined to date have relatively uniform tandemly repeated CTDs, as do mycelial oomycetes and the only multicellular stramenopile alga sequenced, *Ectocarpus,* a structurally simple, filamentous form. The group as a whole, however, has evolved more complex cellular differentiation, including rudimentary vascularization (Charrier, Coelho et al. 2008). I predict that CTD evolution in stramenopiles will prove to be more similar to animals and green plants than to fungi and red algae; that is, more developmentally complex brown algae, such as kelp, will have longer CTDs with proximal tandem repeats and greater numbers of modifications and indels in distal regions.

It is clear that extensive CTD modification and relaxed purifying selection on the CTD can be associated with the transition to a parasitic lifestyle (Stump and Ostrozhynska 2013). Remarkably this extends to parasitic flatworms, even though a closely related free-living flatworm retains a CTD with tandemly repeated structure. Nevertheless, it also is clear that a parasitic lifestyle is not synonymous with CTD degeneration. Microsporidians, which arguably are the most derived of all eukaryotic parasites, with genomes smaller than those of typical bacteria (Keeling and Slamovits 2004), retain CTDs of tandem heptad repeats. Furthermore, the relationship between parasitism and CTD structure is more complicated in apicomplexan parasites, where tandem repeats have been lost and reinvented multiple times.

In conclusion, the CTD most likely originated as a tandemly repeated structure, which has been maintained, modified and/or lost during broad scale evolution of eukaryotes. The result is a remarkable diversity of sequences, which undoubtedly reflect a comparable diversity of underlying CTD-protein interactions. Some CTD-associated proteins surely could have undergone related changes to allow continued interactions with changing CTD structures. For example, although both bind to the CTD, mammalian and yeast capping enzymes read CTD codes differently (Fabrega, Shen et al. 2003; Ghosh, Shuman et al. 2011). Even so, it is likely that only a handful of CTD functions, if any, are conserved across all eukaryotes. Nevertheless, given that parallel patterns of CTD evolution can be found between unrelated taxa, investigations like those in apicomplexan parasites (Kishore, Perkins et al. 2009) can help to elucidate more broadly applicable mechanisms of CTD evolution.

Materials and Methods

Data Collection

RPB1 protein sequences from 205 genera were collected from NCBI and individual genome project databases. I excluded sequences with apparent annotation errors, keeping only reliably interpreted sequences in my analyses. Evolutionary relationships used to interpret patterns of CTD evolution are based on the Tree of Life Web Project and NCBI Taxonomy Database.

CTD Annotation

Previous analyses in both budding and fission yeasts indicated that essential functions of the CTD are conferred by repeated domains, and that minimum essential units of function are contained within heptad pairs (Stiller and Cook 2004; Schwer, Sanchez et al. 2012). To better highlight patterns of CTD conservation and degeneration, I developed graphics for each CTD based on these results with the following color annotations. Green regions contain essential CTD functional units identified in budding yeast (Liu, Greenleaf et al. 2008); that is, paired heptads are present within conserved essential sequence elements (YSPxSPxYSP or SPxYSPxSPxY). Yellow designates individual canonical CTD heptads (YSPxSPx) that are not part of a CTD functional unit (as defined above). Red regions have no conserved heptad structure or contain substitutions that are incompatible with CTD function as defined in yeast. Purple heptads have the sequence FSPxSPx that is lethal (if present universally) in budding yeast but turns out to be very common in many other fungal genera.

Character Evolution Analysis

Each CTD was assigned a character state ranging from 0-3. CTDs containing tandemly repeated canonical heptads (generally not less than 8 heptads) were assigned state 3; examples are the CTDs of yeasts, animals and plants. CTD sequences that have functional heptads but fewer than 8 uninterrupted (the minimum length for viability in yeast) were assigned state 2; examples include CTDs of most sordariomycetes (e.g., *Sordaria*). Sequences with few to no functional regions, but still with recognizable heptads, were assigned state 1 (e.g., eurotiomycetes). CTDs with no discernible heptads were assigned state 0 (e.g., ciliates). The program Mesquite (Maddison and Maddison 2011) was used to carry out maximum-likelihood character state analysis, using the Mk 1

Model, to estimate likelihoods of each state at key nodes and at the root of the eukaryotic tree.

Chapter 3: The identification of putative RNA polymerase II C-terminal domain associated proteins in green and red algae

Background

RNA polymerase II is a large complex containing 12 subunits; the largest (RPB1) has a unique carboxyl-terminal domain (CTD) that has attracted the interest of many scientists since it was discovered in the 1980s (Allison, Moyle et al. 1985; Corden, Cadena et al. 1985). In model systems where most functional studies of the CTD have been carried out, the domain is composed of a varied number of tandemly repeated heptapeptides (yeast 26, human 52, *Arabidopsis* 34) with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$. Initial functional studies of the CTD employed truncation mutants in yeast and human cells (Corden 2013); they showed that the domain is essential for viability and there is functional redundancy amongst CTD repeats (Nonet, Sweetser et al. 1987; Bartolomei, Halden et al. 1988; West and Corden 1995). Genetic substitution screens in yeast revealed that Y_1 residues and the two SP pairs are essential, consistent with their stronger evolutionary conservation than the T_4 and S_7 positions (Liu, Greenleaf et al. 2008; Schwer and Shuman 2011). Further, insertions between individual heptapeptides proved to be lethal in fission and budding yeasts, whereas insertions between paired repeats were not, indicating that the smallest CTD functional unit lies within pairs of heptapeptides (Stiller and Cook 2004; Schwer, Sanchez et al. 2012). Further studies narrowed the smallest functional CTD unit in budding yeast to two Y_1 residues surrounded by three SP pairs; that is, YSPxSPxYSP or SPxYSPxSPxY (x represents any amino acid) (Liu, Kenney et al. 2010). Consistent with genetic analyses, cumulative structural studies indicate that most CTD interactions with binding partners involve motifs between one and two heptapeptides in length, and usually not starting from a Y_1 residue (Jasnovidova and Stefl 2013). Although great insights into the functional significance of CTD residues has been gained from experimental analyses, primarily in yeast and animals, comprehensive evolutionary investigations have shown that CTD sequence diversity precludes broader generalization of these results to many other organisms (Yang and Stiller 2014). This has been borne out by functional studies, for example, the demonstration that the CTD is indispensable in *Trypanosoma brucei* despite the absence of any of the essential motifs or repetitive structures required in yeast and animal models (Das and Bellofatto 2009). Parallel with studies of the CTD sequence itself, further investigations have implicated the domain's role in a wide variety of metabolic pathways in yeasts, animals and *Arabidopsis*, including transcription initiation and elongation, pre-mRNA processing, RNA transport, and chromatin modification among others (Eick and Geyer 2013).

The main way that the CTD performs these functions is by recruiting other proteins involved in the various pathways to create transcription/processing factories. Different modifications of heptapeptide residues provide a code that allows for the widely varied interactions between the CTD and many target proteins (Egloff and Murphy 2008; Zhang, Rodriguez-Molina et al. 2012). Among the possible residue modifications, phosphorylations of S_2 and S_5 are the most common, and mainly relate to co-transcriptional functions like mRNA 5’ capping, mRNA 3’ end processing and pre-mRNA splicing (Eick and Geyer 2013). Interestingly, these core mRNA processing functions are broadly conserved across the eukaryotic domain, as are CTD-directed kinases responsible for these modifications (Bartkowiak and Greenleaf 2011).

Very little empirical evidence exists for CTD functions in most eukaryotic groups. To my knowledge, there has been no previous direct experimental work reported on the CTD in red or green algae. Interestingly, the CTDs of these groups have evolved in very different ways. Comparable to what has been found in animals (Yang and Stiller 2014), simple forms of green algae have CTDs consisting of canonical tandem repeats, whereas developmentally complex land plants display both tandemly repeated proximal regions and more modified distal regions (Fig. 7). Tandem repeats also are present in unicellular red algae; however, multicellular rhodophytes have highly modified CTDs without retention of any tandem repeats (Fig. 7). Why the CTD has adopted such different evolution trajectories in green plants and red algae is unknown, but it undoubtedly relates to underlying differences in the types and numbers of protein partners in the two lineages. Given the limited genetic tools available for investigating CTD function in rhodophytes, I undertook a biochemical comparison of baseline CTD-protein interactions in unicellular green and red algae as a reasonable first step toward elucidating comparative CTD function in the two groups.

Chlamydomonas reinhardtii is a well-studied unicellular green alga with a CTD comprising 20 tandem heptapeptides with the consensus YSPTSPA. The red unicellular alga, *Cyanidioschyzon merolae*, has a CTD with seven proximal tandem heptapeptides

(YSPTSPA) and, surprisingly, 11 distal tandem nonapeptides (YSPSSPNVA). This latter structure is unique among all CTD sequences known (Yang and Stiller 2014). Complete genomes are available for both of these algae, permitting identification of proteins through mass spectrometry. Applying methods used previously to identify PCAPs from both yeast and mammalian cells (Carty and Greenleaf 2002; Phatnani, Jones et al. 2004), I isolated proteins that bind to bi-phosphorylated (S_2 and S_5), tri-heptapeptide CTD repeats from both algae, and tri-nonapeptide CTD repeats from *C. merolae*. I aimed to 1) identify proteins that bind differentially to the two different CTD regions in *C. merolae*, and 2) provide a first view of CTD-protein interactions that were in place before the CTD was modified differently in multicellular green plants and rhodophytes.

Results

Potential PCAPs with Functions Shared in Both Algae

I isolated 154 total proteins from *C. reinhardtii* that bound the phospho-CTD, and 133 from *C. merolae*, yields that are very similar to those reported from yeast using comparable methods (Phatnani, Jones et al. 2004). Through careful screening and annotation, I identified seven proteins from *C. reinhardtii* (Table 1) and eight from *C. merolae* (Table 2) that I consider to be likely PCAPs. Six of the eight red algal proteins were eluted from the nonapeptide affinity column and two from the heptapeptide column. The fact that this group of proteins from *Cyanidioschyzon* bound only to the heptapeptide or nonapeptide repeats, but not to both, suggests they have specific CTD-motif affinities and are not simply binding artifacts on a negatively charged polypeptide. Other reasonable candidate PCAPs were recovered (full lists provided in Tables 3 and 4, and see further discussion below), including a number that were specific to only one set of repeats; however, in this report I provide a thorough comparative discussion of only those proteins for which there is some prior experimental evidence of a CTD-interaction from other organisms. This focuses my results on more central CTD functions that are likely to be conserved broadly across eukaryotic diversity, and are most viable candidates for follow-up experimental investigations in red and green algae.

The proteins from *C. reinhardtii* and *C. merolae* share two functional groups, and co-purification of these proteins from both organisms further implies that they are biologically relevant PCAPs. One shared functional group contains three casein kinases, serine/threonine-targeting enzymes, Q84SA0 and A8IYG9 from *C. reinhardtii* and CMS377C from *C. merolae*. Q84SA0 and CMS377C show significant similarity to casein kinase I (CK1) and A8IYG9 to casein kinase II (CK2). Considering that CMS377C is most similar (1e-152) to Q84SA0 in reciprocal Blast searches, the two appear to be homologous. Inferred homologs of both of these algal proteins in yeast (Hrr25), human and *Arabidopsis* are annotated as CK1 isoforms.

The catalytic domain of CK1 lies in its N-terminus, with variable domains in the C-terminus that confer substrate specificity for protein-protein interactions or subcellular localization (Lee 2009). Budding yeast contains four CK1 isoforms, and Hrr25 is the only one that is localized to the nucleus (Lee 2009). Hrr25 is involved in transcriptional response to DNA damage through physical interactions with the transcription factor Swi6, a component of cell cycle regulatory complex SBF (Ho, Mason et al. 1997). Notably, comparable affinity column assays in yeast also recovered Hrr25 as a PCAP (Phatnani, Jones et al. 2004). The fact that homologs from yeast, and now both *C. reinhardtii* and *C. merolae*, all bind to phospho-CTD repeats, strongly implicates this protein as a conserved functional CTD partner. The third protein in this group, *C. reinhardtii* A8IYG9, is most likely the alpha subunit of CK2, which has been reported to phosphorylate the most C-terminal serine of the mammalian CTD (Payne, Laybourn et al. 1989), although its association with CTD heptapeptides has not been reported previously. Moreover, CK2 has been implicated as the main kinase that phosphorylates FCP1 in *Xenopus*, a CTD phosphatase that binds transcription factor IIF (Palancade, Dubois et al. 2002). Thus, prior evidence indicates at least indirect associations between CK2 and the CTD, and my results suggest that CK2 could serve as a CTD-dependent kinase in *Chlamydomonas*.

The second shared functional group includes A8I1B8 and A8HME6 from *C. reinhardtii* and CMH135C from *C. merolae*. All contain RNA recognition motifs and appear to be related to mRNA export based on similarity scores in reciprocal Blast searches that recovered putative human and *Arabidopsis* homologs. The human homolog is ALY/REF, an mRNA export factor that shuttles between the nucleus and cytoplasm. Previous studies showed that metazoan ALY/REF couples pre-mRNA splicing and mRNA export by associating with spliced mRNPs, and also that ALY/REF co-localizes with splicing factors (Zhou, Luo et al. 2000). The apparent yeast homolog of both CMH135C and ALY/REF is Yral, also an mRNA export factor, and it is perhaps the more likely functional model given the relative paucity of introns in both yeast and red algae. Interestingly, Yra1 is another of the proteins that was recovered from comparable binding experiments with bi-phosphorylated heptapeptides in yeast (Phatnani, Jones et al. 2004). Further, experiments substituting negatively charged glutamates for phospho-serines indicated the interaction between Yra1 and phospho-CTD is specific rather than simply an opposite charge attraction (MacKellar and Greenleaf 2011). In addition, structural analysis revealed that both the RNA binding and CTD interaction domains of Yra1 are located in its N-terminus, and partial N-terminal truncations resulted in a severe decrease of Yra1 recruitment to elongating genes (MacKellar and Greenleaf 2011). Mutations resulting in deficient RNA binding or CTD interactions both negatively impact mRNA export (MacKellar and Greenleaf 2011), indicating that Yra1 is likely recruited to transcriptionally elongating genes by the phospho-CTD.

The closest match from yeast to both *Chlamydomonas* sequences A8HME6 and A8I1B8 is Pab1, a poly(A) binding protein that also functions in mRNA export; however, based on similarity scores, Pab1 is more closely related to A8HME6 (they are reciprocal best hits). Although A8I1B8 does not share significant similarity with Yra1 from yeast (E-value = 0.078), it is the reciprocal match to Yra1 homolog CMH135C (see above) from *Cyanidioschyzon* (3e-04), meaning that A8I1B8 could be a PCAP in *C. reinhardtii* with a similar function in mRNA export as ALY/REF and Yra1. A8HME6 is not only identified as the homolog of Pab1 from yeast, but also from human and *Arabidopsis*. Previous studies have shown that Pab1 binds the poly(A) tail of pre-mRNA and could be involved in final trimming of the tail, mRNA release from transcription sites, and its transport to the cytoplasm (Mangus, Evans et al. 2003). To my knowledge, Pab1 has never been shown to interact with the CTD; however, given the confirmed relationship between other mRNA export factors and the CTD, for example, Npl3 in yeast (Dermody, Dreyfuss et al. 2008), the proteins in this functional class from *Chlamydomonas* and *Cyanidioschyzon* are reasonable candidates for further experimental validation as *bona fide* PCAPs. Taken together, my results suggest that the coupling of mRNA processing and export to the phospho-CTD, previously characterized in animals and yeast, also is conserved in both red and green algae.

Potential PCAPs Found Only in C. reinhardtii

Two of the proteins isolated only from *Chlamydomonas* appear to be related to pre-mRNA splicing. A8J3U2 and A8HRV5 both are most similar to components of the U2 snRNP complex, which combines with pre-mRNAs and other snRNPs to form spliceosomes. The homologs of A8J3U2 in yeast, human and *Arabidopsis* are U2A components, and those of A8HRV5 are U2B components. Previous studies have shown a strong functional link between the CTD and pre-mRNA splicing, including several splicing factors that physically interact with the phospho-CTD; for example, Prp40 in yeast, a component of the U1 snRNP (Morris and Greenleaf 2000). Moreover, a recent study reported that the auxiliary factor 65-kDa subunit (U2AF65) of the U2 snRNP and Prp19 complex is recruited by the CTD to promote splicing activation, and that U2AF65 interacts directly with the CTD (David, Boyne et al. 2011). Although there is no evidence for direct interactions between the CTD and U2 snRNP complex components, considering the importance of the CTD in pre-mRNA splicing, along with established

CTD/spliceosome interactions, broader or even slightly different direct interactions between the CTD and spliceosome components is reasonable.

Another putative PCAP identified in *C. reinhardtii* is A8IDW3, which contains both SANT and MPN domains. The human homolog of A8IDW3 is histone H2A deubiquitinase MYSM1, a chromatin regulator. Domain analysis showed that A8IDW3 shares the SANT and MPN domains with its human counterpart and, therefore, is likely to function as a histone H2A deubiquitinase in *C. reinhardtii*. Human histone H2A deubiquitinase regulates transcriptional activation and elongation of many genes (hormone related genes, for example) by deubiquitination of H2A, which enhances the dissociation of linker histone H1 from the nucleosome (Zhu, Zhou et al. 2007). Previous studies reported that several proteins related to chromatin modifications are associated with the phospho-CTD, including histone methyltransferases set1 and set2 (Corden 2013). Such interactions are consistent with my recovery of a green algal H2A deubiquitinase as a putative PCAP; if demonstrated *in vivo*, this would identify a new function of the CTD in chromatin modification.

Potential PCAPs Only in C. merolae

In addition to the two proteins discussed above (CMS377C bound the nonapeptide and CMH135C the heptapeptide columns, respectively), there are another six likely PCAPs identified only from *C. merolae*; five (CMH210C, CMT578C, CMM263C, CMM087C and CMG052C) bound to nonapeptides and one (CMS144C) to heptapeptides.

CMH210C is a nonapeptide-associated PCAP that is likely to be a peptidyl-prolyl cis/trans isomerase (PPIase), based on its strong similarity to homologs from yeast, human and *Arabidopsis*. The yeast homolog Ess1 and the human homolog Pin1 both have been confirmed experimentally to interact with phosphorylated CTD; their putative function is to help Ssu72 dephosphorylate S_5 on CTD repeats by making the S_{5P}-P_6 bond take on a *cis* conformation (Werner-Allen, Lee et al. 2011). Ess1 and Pin1 interact with the CTD through their WW domains (Corden 2013). Although CMH210C does not have a recognizable WW domain, it does contain a SurA domain with predicted PPIase function as in yeast and human. Instead of a WW domain, however, CMH210C has a FHA domain at its N-terminus, which also is a phospho-peptide (mostly phospho-threonine) interacting domain present in many regulatory proteins (Durocher and Jackson 2002). CMH210C was eluted only from the nonapeptide column, suggesting this protein does not interact strongly with phosphorylated heptapeptides in the CTD of *C. merolae*. This certainly could be explained by the presence of a FHA instead of a WW domain; in both yeast and human homologs of CMH210C, the latter interacts only with phosphorylated heptapeptides.

CMT578C, another potential nonapeptide-associated PCAP, is homologous with Mgt1 from yeast. No reciprocal homolog was found in human, although the nearest match was to MGMT, homologous to yeast Mgt1, with an e-value a little higher than my threshold. Although the similarity between CMT587C and MGMT is not significant based on my *a priori* cutoff, the significant relationships between Mgt1 and MGMT, and between Mgt1 and CMT587C, make it likely that CMT587C also is homologous with

MGMT. Both Mgt1 and MGMT are 6-O-methylguanine-DNA methyltransferases that use cysteine residues to interact with alkyl groups, which are transferred from toxic lesions of alkylated guanine in DNA (Shaiu and Hsieh 1998; Sedgwick, Bates et al. 2007). If CMT578C has the same function in *C. merolae*, it is the first time this methyltransferase has been implicated as having interactions with the RNAP II CTD.

The nonapeptide-associated PCAP CMM263C is likely to be a Topoisomerase I, based on its strong similarity to yeast, human and *Arabidopsis* Top I genes. During transcription, Top I relaxes superhelical stress in unwinding DNA. Early analyses indicated that the N-terminal domain of *Drosophila* Top I could associate with RNA polymerase II (Shaiu and Hsieh 1998), and later work revealed that both human and yeast Top I physically bind the phospho-CTD (Phatnani, Jones et al. 2004). A more recent study demonstrated that both *Drosophila* and human Top I use the proximal half of their N-terminal domain to interact with the CTD (Wu, Phatnani et al. 2010). Therefore, my identification of Top I as a PCAP in *C. merolae* is consistent with established Top I interactions with the phospho-CTD.

Another nonapeptide-associated protein, CMM087C, contains a SWIB/MDM2 domain found in both SWI/SNF complex B and in MDM2, a regulator of the p53 tumor suppressor gene. SWI/SNF components, first characterized in chromatin remodeling complexes in yeast (Winston and Carlson 1992), are widely conserved in eukaryotes. Cumulative studies indicated that they remove nucleosome blocks on interactions between DNA and regulatory proteins like transcription factors (Schwabish and Struhl 2007). In doing so, SWI/SNF complexes regulate many biological processes, including

RNAP II transcription initiation, elongation and associated DNA repair (Euskirchen, Auerbach et al. 2012). To my knowledge, no direct interaction between SWI/SNF complex subunits and phospho-CTD has been established. Nevertheless, given the phospho-CTD's apparent recruitment of histone acetyltransferase (HAT) complexes and deacetylase complexes (HDACs) that remodel nucleosomes around the elongating RNAP II (Spain and Govind 2011), it is reasonable that SWI/SNF components, which also accompany RNAP II transcription factory, could have evolved direct CTD interactions in some organisms. Thus, my recovery of a SWI/SNF-like subunit acting as a PCAP in *C. merolae* could be the first evidence of a more broadly important CTD protein interaction.

The last nonapeptide-associated protein from *Cyanidioschyzon* is CMG052C, which is inferred to be homologous with yeast Bas1, a MYB-related transcription factor required for transcriptional regulation of a number of genes related to the biosynthesis of purine, pyrimidine and several amino acids; for example, the ADE3 gene encoding the purine and glycine biosynthetic enzyme tetrahydrofolate synthase (Joo, Kim et al. 2009). The most similar sequence to CMG052C in *Arabidopsis* is an R2R3 transcription factor, which belongs to a MYB-protein subfamily in plants. Cumulative research on plant R2R3-type MYB factors suggests they are involved in controlling development, determination of cell fate, and transcriptional activation (Stracke, Werber et al. 2001). To date, there is no experimental evidence for a Bas1/CTD interaction in yeast, or any reports of a CTD association with MYB-related transcription factor in plants and animals. Moreover, if the CTD in *C. merolae* is hyperphosphorylated during transcript elongation rather than initiation, as is true in all CTD model organisms (Egloff and Murphy 2008), then a relevant biological interaction between an MYB factor and the phospho-CTD is not immediately apparent.

The same can be said for a potential PCAP, CMS144C, which bound to phospho-heptapeptides. All yeast, human and *Arabidopsis* homologs are identified as TFIID subunit 12 (Taf12), a TATA-binding protein associated factor. Previous studies in yeast have shown an association between the CTD and the TFIID complex (Conaway, Bradsher et al. 1992; Koleske, Buratowski et al. 1992); however, the specific component(s) of TFIID that is/are the target(s) for this interaction remain(s) unclear. Interestingly, a recent study revealed that another TFIID subunit, Taf15, can interact with the unphosphorylated CTD *in vitro* through its polymerized Low Complexity (LC) domain, and that this interaction is deterred by phosphorylation of the CTD (Kwon, Kato et al. 2013). This suggests that recruitment of RNAP II during transcription initiation is facilitated by interactions between the unphosphorylated CTD and Taf15, and that its release from the transcription initiation complex is promoted by CTD phosphorylation (Kwon, Kato et al. 2013). No clear homolog of Taf15 is present in yeast, however, and I likewise found no Taf15 homolog in *C. merolae* through extensive Blast searches using human Taf15 as the query. Taf12 does not contain a LC domain, suggesting it might not interact with the unphosphorylated CTD in *Cyanidioschyzon*. Thus, if the interaction of Taf12 with the phospho-CTD in *C. merolae* is biologically relevant, it suggests a more complicated relationship between the TFIID complex and the CTD, at least in red algae.

Other Proteins That Bound Phospho-CTD Peptides

In addition to the 15 likely PCAPs I singled out for in-depth comparative analyses, many other proteins bound to my phospho-CTD affinity columns. A number have putative functions that are relevant to the CTD's established roles in transcription and mRNA processing, while many others have no recognizable homologs that allow a prediction of function or cellular localization. Like the eight red algal PCAPs discussed above, many of these proteins from *Cyanidioschyzon* bind to either the heptapeptide or nonapeptide column, but not to both (Table 4). Thus, it is reasonable that a number of other proteins I isolated are biological relevant CTD partners.

Interestingly, most of the proteins that can be annotated do not function in the nucleus based on inferred yeast, human and *Arabidopsis* homologs. The largest fractions are ribosomal proteins, consistent with prior results from yeast where numerous ribosomal proteins bound to bi-phosphorylated CTD affinity columns (Phatnani, Jones et al. 2004). This is not surprising, given that these proteins generally interact with uniform, negative phosphate charges on rRNAs within the ribosome. Although individual ribosomal proteins are imported into the nucleus, where major ribosome components are assembled before transport to the cytoplasm, the physical separation between the nucleolus (site of ribosome synthesis) and RNAP II transcription factories suggests that their direct contact with the phospho-CTD (present only where RNAP is actively elongating mRNA transcripts) as individual proteins is unlikely.

I also found a similar result using the *E. coli* proteome, an additional negative control for assessing non-specific binding. Given that the CTD is present only on RNAP II in eukaryotes, there has been no selective pressure on *E. coli* proteins to avoid binding inappropriately to negative charges on a phospho-CTD. Sixty-five percent *E. coli* proteins that bound phospho-CTD peptides were ribosomal proteins (Table 5). Therefore, it appears that their recovery represents the major issue with non-specific protein binding to phospho-CTD peptides.

Despite potential binding artifacts, it has been demonstrated that the methodologies employed here are effective in recovering numerous bona fide PCAPs from both yeast and human cells (Morris, Phatnani et al. 1999; Carty, Goldstrohm et al. 2000; Carty and Greenleaf 2002; Phatnani, Jones et al. 2004). I believe this is because inside the nucleus, where elaborate and intricate regulation of transcription and mRNA processing is carefully orchestrated, there must be strong selection for more highly specific interactions between the phospho-CTD and its binding partners. In other words, transcription-related nuclear proteins are likely to be under strong selection to avoid simple opposite surface charge attractions, whereas cytoplasmic proteins that do not encounter the CTD will have experienced weaker or no selection to avoid non-specific interactions. Because my results were comparable to those reported in previously published investigations, I thought it important to investigate this issue further by examining the proportions of nuclear and cytoplasmic proteins that bound my CTD affinity columns.

Of the 116 *C. reinhardtii* proteins recovered with identifiable functional homologs, 26 (22.4%) are putatively related to processes occurring in the nucleus; for *C. merolae*, 23 of the 113 (20.4%) are nucleus-related. Because of the large numbers of genes without known homologs in their genomes, clear ratios of nuclear to total proteins are difficult to estimate for these two algal species. In more thoroughly characterized budding yeast and *Arabidopsis* genomes, however, the ratios are 35.2% (2070/5887) and 32.4% (9356/28912) respectively, according to GO annotations. I therefore set a conservative estimate of 30% as the fraction of the nuclear localized proteins for both algae, and ran binomial tests to determine whether, as predicted, there is evidence for reduced non-specific binding of artificial phospho-CTD peptides for nuclear proteins. For the purposes of this analysis, I used the highly unlikely and conservative assumption that none of the nuclear proteins isolated were true PCAPs, and that all proteins had an equal probability of binding the phospho-CTD. For *C. reinhardtii*, the 26 (22.4%) nuclear proteins recovered are significantly fewer than 30% (*P* = 0.044, one-tailed), which also was true for the 23 (20.4%) *C. merolae* proteins (*P* = 0.014, one-tailed). Thus, even assuming that no legitimate PCAPs were recovered from either alga, my results are consistent with the argument that natural selection has diminished non-specific CTD-protein interactions within the nucleus. Clearly, if even some of the nuclear proteins isolated are legitimate PCAPs, the differences in non-specific binding compared to cytoplasmic proteins is that much greater.

Based on this result, along with the similarity of my data with those from prior analyses in yeast, I believe that the transcription/mRNA processing related proteins I isolated from *C. reinhardtii* and *C. merolae* can be considered viable candidates for further investigation as biologically relevant PCAPs. The evidence I find for selection against non-specific CTD binding in nuclear proteins also offers further validation of PCAPs inferred in previous studies using comparable methods.

Discussion

Although I am unaware of experimental investigations showing the CTD is phosphorylated at S_2 and S_5 in red or green algae, my phylogenetic analyses revealed that, except for the absence of CDK8 from the two unicellular red algae *Cyanidioschyzon* and *Galdieria*, members of all CDK subfamilies are conserved in both the red and green lineages (Fig. 8). Therefore, the presence of homologs of the CDK7 and CDK9/12/13 subfamilies, those mainly responsible for S_2 and S_5 CTD phosphorylations in human and yeast (Bartkowiak and Greenleaf 2011), suggests that this phosphorylation pattern also is conserved in *Chlamydomonas* and *Cyanidioschyzon*, and is predicted to be present during transcription elongation. Thus, using S_{2P} and S_{5P} CTD peptides as bait for PCAPs appears reasonable for both algae.

Our proteomics analyses provide the first experimental evidence of CTD-protein interactions in red and green algae. Although the potential for nonspecific binding to artificial CTD repeats dictates caution when interpreting results from this sort of assay, a number of factors suggest I have identified viable PCAP candidates in both algal species. First, my data are consistent with natural selection favoring reduced non-specific CTD binding by proteins that function in the nucleus, where they could encounter the CTD by chance. I think this result is important, in itself, given that similar non-specific binding has been reported in prior studies, and is always a concern in any assay of protein binding to a highly charged peptide like the CTD.

Second, a number of the homologs of known PCAPs were recovered from both algal taxa, which is unlikely to be coincidental give the small fractions of the proteomes involved. Third, a number of the proteins I recovered are inferred homologs of yeast and human proteins that have been shown to bind comparable phospho-CTD affinity columns for those organisms (Carty and Greenleaf 2002; Phatnani, Jones et al. 2004), and for which there is additional corroborating evidence of a CTD interaction. Perhaps more compelling, however, is the level of differential binding of nuclear proteins to heptapeptides and nonapeptides from *Cyanidioschyzon*. Of the 23 proteins with nuclear annotations, only five bound to both peptide affinity columns. In contrast, over half (48 of 90) cytoplasmic proteins bound to both versions of the phospho-CTD. This demonstrates a significant (*P* = 0.01, binomial test, one tailed) tendency for nuclear proteins to bind specifically to one or the other type of CTD repeats present in *Cyanidioschyzon*, as would be expected if CTD-protein interactions are spacio-temporally arranged as in model systems (Jasnovidova and Stefl 2013).

Finally, despite focusing on only the 15 proteins for which CTD interactions can be argued from prior research, differences in PCAP functional categories recovered relate to a clear and important biological difference between the two algae; PCAPs associated with re-mRNA splicing were recovered from *Chlamydomonas*, but not from

Cyanidioschyzon. Only 27 introns (in 26 genes) have been identified in the entire *C. merolae* genome and several spliceosome-related proteins appear to be missing (Matsuzaki, Misumi et al. 2004). In *Chlamydomonas*, on the other hand, over 90% of protein-encoding genes contain introns, with 8.3 exons per gene on average (Merchant, Prochnik et al. 2007). Thus, it is unlikely that spliceosomal proteins that could interact with the CTD are expressed as highly in *Cyanidioschyzon* as in *Chlamydomonas*. Comparable methods applied in *S. cerevisiae* recovered splicing-related proteins (Phatnani, Jones et al. 2004), despite the relatively paucity of yeast introns (Spingola, Grate et al. 1999) compared to *Chlamydomonas*, suggesting the possibility that fewer or no splicing factors interact with the CTD in *C. merolae.* Given the likelihood that splicing is an ancient CTD function (Yang and Stiller 2014), it will be interesting to determine whether the CTD remains involved in co-transcriptional splicing in other eukaryotes that, like red algae, are thought to have lost most of their ancestral introns (Csuros, Rogozin et al. 2011).

Such differences highlight the importance of further experimental investigations of CTD function in red algae and other diverse eukaryotes. When considering my results, it is important to note that red algae have been evolving independently from other eukaryotes for well over a billion years (Butterfield 2000), and relatively few gene functions have been determined experimentally. Moreover, patterns of CTD evolution among eukaryotic taxa are far more diverse than was suggested by early comparative studies (Yang and Stiller 2014). Thus, my recovery of two different proteins implicated in transcription initiation among my potential PCAPs, could be a first suggestion that patterns of CTD hypo- and hyper-phosphorylation in at least some red algae differ from those established in model systems (Egloff and Murphy 2008). Interestingly, I found neither CDK8 (Fig.8), nor most components of mediator in the *Cyanidioschyzon* genome, in line with potential differences in CTD phosphorylation during transcription initiation. Although I limited my detailed treatment to proteins with evidence from other organisms to implicate involvement with the CTD, other nuclear proteins were recovered from both species (Tables 3 and 4). Some have inferred functions that are biologically relevant to the CTD's established roles, whereas others have no identifiable homologs to provide predictions of function and cellular localization. Many could prove to be CTD-interacting proteins.

Given the great evolutionary distances between major eukaryotic lineages, the single CTD phosphorylation pattern I examined, the small fractions of the algal proteomes recovered and even smaller fractions that have identifiable homologs, it is interesting that I uncovered as many shared putative homologs and functional categories as I did. Although clearly biased by the fact that I looked for prior evidence of CTD involvement, my results nevertheless suggest there could be functional conservation of a number of core CTD-protein interactions across broad eukaryotic diversity.

In conclusion, my study provides the first experimental evidence of baseline CTD-protein interactions in simple, undifferentiated unicellular green and red algae. They permit an initial comparison of potential PCAPs with those recovered in comparable previous investigations in yeast and mammals. The PCAPs shared among all these groups indicate that a number of CTD-protein interaction are widely conserved, at least among eukaryotic groups that evolved multicellularity. In contrast, differential PCAP binding to heptapeptides and nonapeptides in the red alga further highlights the importance of lineage-specific modifications, which have punctuated CTD evolution during the diversification of major eukaryotic phyla (Yang and Stiller 2014). Indeed, the large number of unclassified proteins that bind specifically to nonapeptide repeats from *Cyanidioschyzon* (Table 4) suggests the presence of a variety of new, taxon-specific CTD-protein interactions. This variation likely reflects differences in how CTD-protein interactions have elaborated and diversified, providing what Zachary Burton (Burton 2014) has called the "New Testament" in the Genesis of organismal complexity through elaborations of CTD-based mechanisms for controlling gene expression. My investigation provides a first glimpse into the chapters of that book on red and green algae.

Materials and methods

Cell Culture and Lysis

C. reinhardtii (CC-503 cw92 mt+) was cultured in TAP medium (Gorman and Levine 1965) at room temperature and 24 hrs light, and *C. merolae* (N-1804) was cultured in Allen Culture medium (Minoda, Sakagami et al. 2004) at 42 ℃ and 24 hrs light. *Escherichia coli* (DH5α) was cultured in LB medium at 37ºC overnight. Harvested algal and *E. coli* cells were suspended in cold BY-AS400 buffer (25 mM HEPES, pH 7.6; 1 mM EDTA; 1 mM PMSF; 400 mM $AmSO_4$; protease inhibitor cocktail for plants 1:100 dilution) using 2-3 ml buffer per gram of cells. A French press (12,000 psi) was used twice to break suspended cells and obtain crude protein extracts. The crude extracts were centrifuged in a SS34 rotor at 20,000 ×g for 45 minutes at 4 ℃, and the supernatant was collected. A flowchart of the protein purification methodology is shown in Fig. 9.

Ammonium Sulfate Precipitation

The detergent NP-40 was added to the SS34-supernatant to a final concentration of 1%, and $(NH_4)_2SO_4$ was gradually added to a final concentration of 50% (~313g/l) while stirring at 4 ℃. The ammonium sulfate suspension was then centrifuged again in a SS34 rotor at 30,000 ×g for 45 minutes at 4 ℃. The $(NH_4)_2SO_4$ pellet was collected and suspended with enough cold BH buffer (25 mM HEPES, pH 7.6; 1 mM EDTA; 1 mM DTT; 1 mM PMSF; 8% glycerol) to bring conductivity in the suspension approximately equal to 0.15 M NaCl.

Ion Exchange Chromatography

In order to increase concentration of proteins with positive surface charges that could bind CTD phosphoserines, I employed two steps of ion-exchange chromatography modified from the protocol of Greenleaf and colleagues (Phatnani, Jones et al. 2004). This both enriched potential PCAPs, and removed remaining cell debris and undesired proteins (e.g. chromoproteins) that were not eliminated by initial centrifugations.

The BH-suspension was passed through a ~21 ml (1.5cm × 12cm) anion exchange column (Q Sepharose Fast Flow, GE Healthcare) at a flow rate of ~1.4 ml/min, and the column was washed with 4 column volumes of BH buffer + 0.15 M NaCl. The flow through from the column was collected and loaded on a same size cation exchange column (SP Sepharose Fast Flow, GE Healthcare) with the same flow rate. The column also was washed with 4 column volumes of BH buffer + 0.15 M NaCl, and eluted with BH buffer + 1 M NaCl. The elution from cation exchange column was collected and desalted by dilution and ultrafiltration.

Affinity Chromatography

One ml CTD affinity columns were constructed using NeutrAvidin Agarose Resin (Thermo Scientific) bound to biotin-labeled, synthetic CTD tri-heptapeptides (Biotin-YSpPTSpPAYSpPTSpPAYSpPTSpPA) or tri-nonapeptides (Biotin-YSpPSSpPNVAYSpPSSpPNVAYSpPSSpPNVA), which were constructed at Eton Bioscience Inc, each containing three repeats phosphorylated at all S_2 and S_5 residues. Because these peptides are very similar in sequence, and in phosphorylation patterns, each represents an excellent negative control for non-specific binding to the other. That is, if a protein cannot bind to one of these very similar phospho-peptides, it is strong evidence of a specific affinity for the other. A 1 ml control column also was made containing only the NeutrAvidin Agarose Resin. I chose this phosphorylation pattern to allow direct comparison to PCAPs isolated previously from the far more thoroughly characterized *Saccharomyces cerevisiae* genome (Phatnani, Jones et al. 2004).

I added a PhosSTOP phosphatase inhibitor cocktail tablet to each cation-elution pool (~4 mg of protein) to avoid de-phosphorylation of the CTD peptides and then passed the pool through the appropriate heptapeptide or nonapeptide affinity column. All columns were washed with 16 column volumes of BH + 0.1 M NaCl. Bound proteins were eluted sequentially with increasing salt concentrations (1ml BH buffer + 0.3, 0.5, 1.0 M NaCl), with each elution step collected in four 250 μl aliquots. To assay the presence and quality of eluted proteins, 25 μl of each aliquot was examined using SDS-PAGE (4-20% Tris HCl gradient gels from Bio-Rad) stained with Coomassie blue (Fig. 10, 11, 12). The control column (resin with no CTD peptides) followed the same procedure as above, and showed no indication of protein binding (Fig. 13). The middle two 250 μl aliquots from each elution concentration were pooled, desalted and concentrated. 10 μg of proteins from each elution pool were subjected to SDS-PAGE, followed by Coomassie blue stain (gels shown in Fig. 9); the rest were submitted to Duke University Proteomics Center for mass spectrometry (LC/ESI/MS/MS) identification.

Protein Annotations

Because functions assigned to genes in both the *C. reinhardtii* and *C. merolae* genome are based primarily upon sequence similarity to genes from more well-developed models, I relied on annotated functions of apparent homologs from yeast, human and *Arabidopsis* to identify potential CTD-binding partners in both algae. Homologs were identified through reciprocal Blast searches between the *C. reinhardtii* or *C. merolae* and each of the three reference genomes (E-value cutoff of 1e-04). Reciprocal best hits were considered to be homologous sequences. Protein domain analyses were based on the National Center for Biotechnology Information (NCBI) structure online service http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi.

Phylogenetic Analyses of CDKs

I performed phylogenetic analyses of putative cyclin-dependent kinases (CDKs) from both algae to verify that appropriate homologs are present to expect the pattern of CTD-phosphorylation analyzed in this study. According to previous investigations, human CDKs can be divided into well-defined subfamilies (Guo and Stiller 2004; Cao, Chen et al. 2014). Therefore, I applied reciprocal Blast searches to identify the homologs of each CDK subfamily from yeast, *Arabidopsis*, *Chlamydomonas*, *Cyanidioschyzon* and two additional complete red algal genomes (*Chondrus crispus* and *Galdieria sulphuraria*). For each organism, the putative CDK homolog with the highest similarity score to each subfamily was chosen for phylogenetic analyses together with the representative human CDKs. A multiple sequence alignment was performed in MUSCLE (Edgar 2004) (online service: http://www.ebi.ac.uk/Tools/msa/muscle/) and Gblocks 0.91b (Castresana 2000) (http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks) was used to select the conserved blocks appropriate for tree-building. Phylogenetic analysis were performed in MrBayes using a WAG + invgamma model (Huelsenbeck and Ronquist 2001) as determined through maximum-likelihood model estimation in MEGA 5.2.2 (Tamura, Peterson et al. 2011). Relative support for the presence of CTK1/CDK9 homologs was inferred from Bayesian posterior probabilities estimated from all trees (10^6 generations) sampled after the average standard deviation of split frequencies had converged on a value < 0.01.

Chapter 4: Conclusion

The C-terminal domain of the largest subunit of RNA polymerase II is responsible for coordinating a wide range of co-transcriptional functions. Although tandem repeats of a seven amino acid motif comprise the CTD in model eukaryotes, the domain is highly unordered in many other organisms. The research presented in chapter 2 represents the most comprehensive investigation of CTD diversity and evolution to date, and finds that the CTD's tandem structure likely existed in the last eukaryotic common ancestor, that unordered CTDs have resulted from extensive, lineage-specific sequence modifications, and that tandem heptads have been lost and reinvented many times. The work also highlights interesting parallels in CTD evolution that appear to be associated with the requirements of developmental complexity. For red algae and fungi, although present in simple, ancestral red algae and fungi, CTD tandem repeats have undergone extensive modifications and degeneration during the evolutionary transition to developmentally complex rhodophytes and fungi. In contrast, CTD repeats are maintained in animals, green algae and their more complex land plant relatives.

The different CTD evolution trajectories in eukaryotes inspired my interest in investigate studying the mechanisms that underlie CTD sequence variation, and investigations of CTD-associated proteins is primarily required to understand these mechanisms. Based on controversial relationships and differences in the pattern of CTD evolution between green plants and red algae, I initiated a baseline comparison of the CTD associated proteins in the unicellular green algae *Chlamydomonas* and red algae

Cyanidioschyzon. The previously established method that uses artificially synthesized and phosphorylated CTD repeats to bind PCAPs was adopted in this study. A number of potential PCAPs were found in this study, and several of them have yeast and human counterparts that have been identified experimentally as PCAPs by previous research. This study represents the first CTD associated functional analyses in both green and red algae. I hope this work will spark broader interest in these organisms and lead to further functional experimentations in both.

References

Allison, L. A., M. Moyle, et al. (1985). "Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases." Cell **42**(2): 599-610.

Allison, L. A., J. K. Wong, et al. (1988). "The C-terminal domain of the largest subunit of RNA polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and mammals: a conserved structure with an essential function." Mol Cell Biol **8**(1): 321-329.

Bartkowiak, B. and A. L. Greenleaf (2011). "Phosphorylation of RNAPII: To P-TEFb or not to P-TEFb?" Transcription **2**(3): 115-119.

Bartkowiak, B., A. L. Mackellar, et al. (2011). "Updating the CTD Story: From Tail to Epic." Genet Res Int **2011**: 623718.

Bartolomei, M. S., N. F. Halden, et al. (1988). "Genetic-Analysis of the Repetitive Carboxyl-Terminal Domain of the Largest Subunit of Mouse Rna Polymerase-Ii." Mol Cell Biol **8**(1): 330-339.

Bartolomei, M. S., N. F. Halden, et al. (1988). "Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II." Mol Cell Biol **8**(1): 330-339.

Baskaran, R., M. E. Dahmus, et al. (1993). "Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain." Proc Natl Acad Sci U S A **90**(23): 11167-11171.

Buratowski, S. (2003). "The CTD code." Nat Struct Biol **10**(9): 679-680.

Buratowski, S. (2009). "Progression through the RNA Polymerase II CTD Cycle." Mol Cell **36**(4): 541-546.

Burley, S. K. and N. Sonenberg (2011). "Gimme phospho-serine five! Capping enzyme guanylyltransferase recognition of the RNA polymerase II CTD." Mol Cell **43**(2): 163-165.

Burton, Z. F. (2014). "The Old and New Testaments of gene regulation: Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD." Transcription **5:e28764. doi.org/10.4161/trns.28674**.

Butterfield, N. J. (2000). "Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes." Paleobiology **26**: 386-404.

Cao, L., F. Chen, et al. (2014). "Phylogenetic analysis of CDK and cyclin proteins in premetazoan lineages." BMC Evol Biol **14**: 10.

Carty, S. M., A. C. Goldstrohm, et al. (2000). "Protein-interaction modules that organize nuclear function: FF domains of CA150 bind the phosphoCTD of RNA polymerase II." Proc Natl Acad Sci U S A **97**(16): 9015-9020.

Carty, S. M. and A. L. Greenleaf (2002). "Hyperphosphorylated C-terminal repeat domain-associating proteins in the nuclear proteome link transcription to DNA/chromatin modification and RNA processing." Mol Cell Proteomics **1**(8): 598-610.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Mol Biol Evol **17**(4): 540-552.

Chapman, R. D., M. Conrad, et al. (2005). "Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation." Mol Cell Biol **25**(17): 7665-7674.

Chapman, R. D., M. Heidemann, et al. (2007). "Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7." Science **318**(5857): 1780-1782.

Chapman, R. D., M. Heidemann, et al. (2008). "Molecular evolution of the RNA polymerase II CTD." Trends Genet **24**(6): 289-296.

Charrier, B., S. M. Coelho, et al. (2008). "Development and physiology of the brown alga Ectocarpus siliculosus: two centuries of research." New Phytol **177**(2): 319-332.

Cho, E. J., C. R. Rodriguez, et al. (1998). "Allosteric interactions between capping enzyme subunits and the RNA polymerase II carboxy-terminal domain." Genes Dev **12**(22): 3482-3487.

Cho, E. J., T. Takagi, et al. (1997). "mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain." Genes Dev **11**(24): 3319-3326.

Conaway, R. C., J. N. Bradsher, et al. (1992). "Mechanism of Assembly of the Rna Polymerase-Ii Preinitiation Complex - Evidence for a Functional Interaction between the Carboxyl-Terminal Domain of the Largest Subunit of Rna Polymerase-Ii and a High Molecular Mass Form of the Tata Factor." J Biol Chem **267**(12): 8464-8467.

Corden, J. L. (2013). "RNA polymerase II C-terminal domain: Tethering transcription to transcript and template." Chem Rev **113**(11): 8423-8455.

Corden, J. L., D. L. Cadena, et al. (1985). "A Unique Structure at the Carboxyl Terminus of the Largest Subunit of Eukaryotic Rna Polymerase-Ii." Proc Natl Acad Sci U S A **82**(23): 7934-7938.

Csuros, M., I. B. Rogozin, et al. (2011). "A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes." PLoS Comput Biol **7**(9): e1002150.

Das, A. and V. Bellofatto (2009). "The Non-Canonical CTD of RNAP-II Is Essential for Productive RNA Synthesis in Trypanosoma brucei." PLoS One **4**(9).

David, C. J., A. R. Boyne, et al. (2011). "The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex." Genes Dev **25**(9): 972-983.

Dermody, J. L., J. M. Dreyfuss, et al. (2008). "Unphosphorylated SR-like protein Npl3 stimulates RNA polymerase II elongation." PLoS One **3**(9): e3273.

Descostes, N., M. Heidemann, et al. (2014). "Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells." Elife **3**: e02105.

Durocher, D. and S. P. Jackson (2002). "The FHA domain." FEBS Lett **513**(1): 58-66.

Ebright, R. H. (2000). "RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II." J Mol Biol **304**(5): 687-698.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Egloff, S. and S. Murphy (2008). "Cracking the RNA polymerase II CTD code." Trends Genet **24**(6): 280-288.

Egloff, S., D. O'Reilly, et al. (2007). "Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression." Science **318**(5857): 1777-1779.

Eick, D. and M. Geyer (2013). "The RNA polymerase II carboxy-terminal domain (CTD) code." Chem Rev **113**(11): 8456-8490.

Euskirchen, G., R. K. Auerbach, et al. (2012). "SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions." J Biol Chem **287**(37): 30897-30905.

Fabrega, C., V. Shen, et al. (2003). "Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II." Mol Cell **11**(6): 1549-1561.

Ghosh, A., S. Shuman, et al. (2008). "The structure of Fcp1, an essential RNA polymerase II CTD phosphatase." Mol Cell **32**(4): 478-490.

Ghosh, A., S. Shuman, et al. (2011). "Structural Insights to How Mammalian Capping Enzyme Reads the CTD Code." Mol Cell **43**(2): 299-310.

Glover-Cutter, K., S. Larochelle, et al. (2009). "TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II." Mol Cell Biol **29**(20): 5455-5464.

Gorman, D. S. and R. P. Levine (1965). "Cytochrome f and plastocyanin: their sequence in the photosynthetic electron transport chain of Chlamydomonas reinhardi." Proc Natl Acad Sci U S A **54**(6): 1665-1669.

Grummt, I. (1999). "Regulation of mammalian ribosomal gene transcription by RNA polymerase I." Prog Nucleic Acid Res Mol Biol **62**: 109-154.

Guo, Z. and J. W. Stiller (2004). "Comparative genomics of cyclin-dependent kinases suggest co-evolution of the RNAP II C-terminal domain and CTD-directed CDKs." BMC Genomics **5**: 69.

Guo, Z. and J. W. Stiller (2005). "Comparative genomics and evolution of proteins associated with RNA polymerase II C-terminal domain." Mol Biol Evol **22**(11): 2166-2178.

Haag, J. R. and C. S. Pikaard (2011). "Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing." Nat Rev Mol Cell Biol **12**(8): 483-492.

Heidemann, M. and D. Eick (2012). "Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code." RNA Biol **9**(9): 1144-1146.

Heidemann, M., C. Hintermair, et al. (2013). "Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription." Biochim Biophys Acta **1829**(1): 55-62.

Hintermair, C., M. Heidemann, et al. (2012). "Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation." Embo Journal **31**(12): 2784-2797.

Hirose, Y., R. Tacke, et al. (1999). "Phosphorylated RNA polymerase II stimulates pre-mRNA splicing." Genes Dev **13**(10): 1234-1239.

Ho, C. K. and S. Shuman (1999). "Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme." Mol Cell **3**(3): 405-411.

Ho, C. K., V. Sriskanda, et al. (1998). "The guanylyltransferase domain of mammalian mRNA capping enzyme binds to the phosphorylated carboxyl-terminal domain of RNA polymerase II." J Biol Chem **273**(16): 9577-9585.

Ho, Y., S. Mason, et al. (1997). "Role of the casein kinase I isoform, Hrr25, and the cell cycle-regulatory transcription factor, SBF, in the transcriptional response to DNA damage in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **94**(2): 581-586.

Hsin, J. P., W. Li, et al. (2014). "RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells." Elife **3**: e02112.

Hsin, J. P. and J. L. Manley (2012). "The RNA polymerase II CTD coordinates transcription and RNA processing." Genes Dev **26**(19): 2119-2137.

Hsin, J. P., A. Sheth, et al. (2011). "RNAP II CTD Phosphorylated on Threonine-4 Is Required for Histone mRNA 3 ' End Processing." Science **334**(6056): 683-686.

Hsin, J. P., K. Xiang, et al. (2014). "Function and control of RNA polymerase II CTD phosphorylation in vertebrate transcription and RNA processing." Mol Cell Biol.

Hsu, P. L., F. Yang, et al. (2014). "Rtr1 Is a Dual Specificity Phosphatase That Dephosphorylates Tyr1 and Ser5 on the RNA Polymerase II CTD." J Mol Biol **426**(16): 2970-2981.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Hurwitz, J. (2005). "The discovery of RNA polymerase." J Biol Chem **280**(52): 42477-42485.

Jain, R., M. C. Rivera, et al. (1999). "Horizontal gene transfer among genomes: The complexity hypothesis." Proc Natl Acad Sci U S A **96**(7): 3801-3806.

James, T. Y., F. Kauff, et al. (2006). "Reconstructing the early evolution of Fungi using a six-gene phylogeny." Nature **443**(7113): 818-822.

Jasnovidova, O. and R. Stefl (2013). "The CTD code of RNA polymerase II: a structural view." Wiley Interdiscip Rev RNA **4**(1): 1-16.

Joo, Y. J., J. A. Kim, et al. (2009). "Cooperative regulation of ADE3 transcription by Gcn4p and Bas1p in Saccharomyces cerevisiae." Eukaryot Cell **8**(8): 1268-1277.

Keeling, P. J. and C. H. Slamovits (2004). "Simplicity and complexity of microsporidian genomes." Eukaryot Cell **3**(6): 1363-1369.

Kishore, S. P., S. L. Perkins, et al. (2009). "An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases." J Mol Evol **68**(6): 706-714.

Kishore, S. P., J. W. Stiller, et al. (2013). "Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite Plasmodium falciparum and other apicomplexans." BMC Evol Biol **13**: 37.

Koleske, A. J., S. Buratowski, et al. (1992). "A Novel Transcription Factor Reveals a Functional Link between the Rna Polymerase-Ii Ctd and Tfiid." Cell **69**(5): 883-894.

Koonin, E. V. (2009). "Intron-dominated genomes of early ancestors of eukaryotes." J Hered **100**(5): 618-623.

Krishnamurthy, S., X. He, et al. (2004). "Ssu72 Is an RNA polymerase II CTD phosphatase." Mol Cell **14**(3): 387-394.

Kurtzman, C. P. and J. W. Fell (2006). "in Biodiversity and Ecophysiology of Yeasts, eds Rosa C, Peter G (Springer, Heidelberg)." pp11-30.

Kwon, I., M. Kato, et al. (2013). "Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains." Cell **155**(5): 1049-1060.

Lang, B. F., C. O'Kelly, et al. (2002). "The closest unicellular relatives of animals." Curr Biol **12**(20): 1773-1778.

Lee, J. Y. (2009). "Versatile casein kinase 1: multiple locations and functions." Plant Signal Behav **4**(7): 652-654.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-4060.

Licatalosi, D. D., G. Geiger, et al. (2002). "Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II." Mol Cell **9**(5): 1101-1111.

Litingtung, Y., A. M. Lawler, et al. (1999). "Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II." Mol Gen Genet **261**(1): 100-105.

Liu, P., A. L. Greenleaf, et al. (2008). "The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation." Mol Biol Evol **25**(4): 719-727.

Liu, P., J. M. Kenney, et al. (2010). "Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain." Mol Biol Evol **27**(11): 2628-2641.

Lunde, B. M., S. L. Reichow, et al. (2010). "Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain." Nat Struct Mol Biol **17**(10): 1195-1201.

MacKellar, A. L. and A. L. Greenleaf (2011). "Cotranscriptional association of mRNA export factor Yra1 with C-terminal domain of RNA polymerase II." J Biol Chem **286**(42): 36385-36395.

Maddison, W. P. and D. R. Maddison (2011). "Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org."

Mangus, D. A., M. C. Evans, et al. (2003). "Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression." Genome Biol **4**(7): 223.

Matsuzaki, M., O. Misumi, et al. (2004). "Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D." Nature **428**(6983): 653-657.

Mayer, A., M. Heidemann, et al. (2012). "CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II." Science **336**(6089): 1723-1725.

McCracken, S., N. Fong, et al. (1997). "5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II." Genes Dev **11**(24): 3306-3318.

Meininghaus, M., R. D. Chapman, et al. (2000). "Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription." J Biol Chem **275**(32): 24375-24382.

Merchant, S. S., S. E. Prochnik, et al. (2007). "The Chlamydomonas genome reveals the evolution of key animal and plant functions." Science **318**(5848): 245-250.

Minoda, A., R. Sakagami, et al. (2004). "Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, Cyanidioschyzon merolae 10D." Plant Cell Physiol **45**(6): 667-671.

Morris, D. P. and A. L. Greenleaf (2000). "The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II." J Biol Chem **275**(51): 39935-39943.

Morris, D. P., H. P. Phatnani, et al. (1999). "Phospho-carboxyl-terminal domain binding and the role of a prolyl isomerase in pre-mRNA 3'-End formation." J Biol Chem **274**(44): 31583-31587.

Mosley, A. L., S. G. Pattenden, et al. (2009). "Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation." Mol Cell **34**(2): 168-178.

Myer, V. E. and R. A. Young (1998). "RNA polymerase II holoenzymes and subcomplexes." J Biol Chem **273**(43): 27757-27760.

Nonet, M., D. Sweetser, et al. (1987). "Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II." Cell **50**(6): 909-915.

Palancade, B., M. F. Dubois, et al. (2002). "FCP1 phosphorylation by casein kinase 2 enhances binding to TFIIF and RNA polymerase II carboxyl-terminal domain phosphatase activity." J Biol Chem **277**(39): 36061-36067.

Payne, J. M., P. J. Laybourn, et al. (1989). "The Transition of Rna Polymerase-Ii from Initiation to Elongation Is Associated with Phosphorylation of the Carboxyl-Terminal Domain of Subunit-Iia." J Biol Chem **264**(33): 19621-19629.

Phatnani, H. P. and A. L. Greenleaf (2006). "Phosphorylation and functions of the RNA polymerase II CTD." Genes Dev **20**(21): 2922-2936.

Phatnani, H. P., J. C. Jones, et al. (2004). "Expanding the functional repertoire of CTD kinase I and RNA polymerase II: novel phosphoCTD-associating proteins in the yeast proteome." Biochemistry **43**(50): 15702-15719.

Riisberg, I., R. J. Orr, et al. (2009). "Seven gene phylogeny of heterokonts." Protist **160**(2): 191-204.

Rogozin, I. B., L. Carmel, et al. (2012). "Origin and evolution of spliceosomal introns." Biol Direct **7**: 11.

Rosonina, E., N. Yurko, et al. (2014). "Threonine-4 of the budding yeast RNAP II CTD couples transcription with Htz1-mediated chromatin remodeling." Proc Natl Acad Sci U S A.

Schneider, S., Y. Pei, et al. (2010). "Separable functions of the fission yeast Spt5 carboxyl-terminal domain (CTD) in capping enzyme binding and transcription elongation overlap with those of the RNA polymerase II CTD." Mol Cell Biol **30**(10): 2353-2364.

Schroeder, S. C., D. A. Zorio, et al. (2004). "A function of yeast mRNA cap methyltransferase, Abd1, in transcription by RNA polymerase II." Mol Cell **13**(3): 377-387.

Schwabish, M. A. and K. Struhl (2007). "The Swi/Snf complex is important for histone eviction during transcriptional activation and RNA polymerase II elongation in vivo." Mol Cell Biol **27**(20): 6987-6995.

Schwer, B., A. M. Sanchez, et al. (2012). "Punctuation and syntax of the RNA polymerase II CTD code in fission yeast." Proc Natl Acad Sci U S A **109**(44): 18024-18029.

Schwer, B. and S. Shuman (2011). "Deciphering the RNA Polymerase II CTD Code in Fission Yeast." Mol Cell **43**(2): 311-318.

Sedgwick, B., P. A. Bates, et al. (2007). "Repair of alkylated DNA: recent advances." DNA Repair (Amst) **6**(4): 429-442.

Shaiu, W. L. and T. S. Hsieh (1998). "Targeting to transcriptionally active loci by the hydrophilic N-terminal domain of Drosophila DNA topoisomerase I." Mol Cell Biol **18**(7): 4358-4367.

Sims, R. J., L. A. Rojas, et al. (2011). "The C-Terminal Domain of RNA Polymerase II Is Modified by Site-Specific Methylation." Science **332**(6025): 99-103.

Spain, M. M. and C. K. Govind (2011). "A role for phosphorylated Pol II CTD in modulating transcription coupled histone dynamics." Transcription **2**(2): 78-81.

Spingola, M., L. Grate, et al. (1999). "Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae." Rna-a Publication of the Rna Society **5**(2): 221-234.

Stechmann, A. and T. Cavalier-Smith (2003). "The root of the eukaryote tree pinpointed." Curr Biol **13**(17): R665-666.

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

Stiller, J. W. and B. D. Hall (1998). "Sequences of the largest subunit of RNA polymerase II from two red algae and their implications for rhodophyte evolution." Journal of Phycology **34**(5): 857-864.

Stiller, J. W. and B. D. Hall (2002). "Evolution of the RNA polymerase II C-terminal domain." Proc Natl Acad Sci U S A **99**(9): 6091-6096.

Stiller, J. W., B. L. McConaughy, et al. (2000). "Evolutionary complementation for polymerase II CTD function." Yeast **16**(1): 57-64.

Stracke, R., M. Werber, et al. (2001). "The R2R3-MYB gene family in Arabidopsis thaliana." Curr Opin Plant Biol **4**(5): 447-456.

Stump, A. D. and K. Ostrozhynska (2013). "Selective constraint and the evolution of the RNA polymerase II C-Terminal Domain." Transcription **4**(2): 77-86.

Sutherland, J. E., S. C. Lindstrom, et al. (2011). "A New Look at an Ancient Order: Generic Revision of the Bangiales (Rhodophyta)." Journal of Phycology **47**(5): 1131-1151.

Tamura, K., D. Peterson, et al. (2011). "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods." Mol Biol Evol **28**(10): 2731-2739.

Tietjen, J. R., D. W. Zhang, et al. (2010). "Chemical-genomic dissection of the CTD code." Nat Struct Mol Biol **17**(9): 1154-1161.

Werner-Allen, J. W., C. J. Lee, et al. (2011). "cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72." J Biol Chem **286**(7): 5717-5726.

Werner, F. (2007). "Structure and function of archaeal RNA polymerases." Mol Microbiol **65**(6): 1395-1404.

West, M. L. and J. L. Corden (1995). "Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations." Genetics **140**(4): 1223-1233.

Willis, I. M. (1993). "RNA polymerase III. Genes, factors and transcriptional specificity." Eur J Biochem **212**(1): 1-11.

Winston, F. and M. Carlson (1992). "Yeast SNF/SWI transcriptional activators and the SPT/SIN chromatin connection." Trends Genet **8**(11): 387-391.

Wu, J., H. P. Phatnani, et al. (2010). "The phosphoCTD-interacting domain of Topoisomerase I." Biochem Biophys Res Commun **397**(1): 117-119.

Yang, C. and J. W. Stiller (2014). "Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain." Proc Natl Acad Sci U S A **111**(16): 5920-5925.

Zhang, D. W., A. L. Mosley, et al. (2012). "Ssu72 phosphatase dependent erasure of phospho-Ser7 marks on the RNA Polymerase II C-terminal domain is essential for viability and transcription termination." J Biol Chem.

Zhang, D. W., J. B. Rodriguez-Molina, et al. (2012). "Emerging Views on the CTD Code." Genet Res Int **2012**: 347214.

Zhou, Z., M. J. Luo, et al. (2000). "The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans." Nature **407**(6802): 401-405.

Zhu, P., W. Zhou, et al. (2007). "A histone H2A deubiquitinase complex coordinating histone acetylation and H1 dissociation in transcriptional regulation." Mol Cell **27**(4): 609-621.

List of Tables

Table 1. The 7 potential *Chlamydomonas* PCAPs identified.

0.3M and 0.5M represent the NaCl elution concentrations. The numbers of the MS/MS identified matching peptides from the elution are shown under each salt step. The annotations are based on their homologs in yeast, human and *Arabidopsis*. The Blast best matches to proteins in yeast, human and *Arabidopsis* are shown with the e-values. The same is true for Table 2 and the supplementary tables.

Protein names	**Wt. (kDa)**	**Heptapeptide column**		**Annotations**	**Best hit in yeast and e-value**		**Best hit in human and e-value**		**Best hit in *Arabidopsis* and e-value**	
		0.3M NaCl	**0.5M NaCl**							
Q84SA0	34.81	19	16	Casein kinase I	HRR25	e-139	P48730	e-165	AT4G26100	e-171
A8IYG9	41.99	8	4	Casein kinase II subunit alpha	CKA2	e-111	E7EU96	e-118	AT2G23070	e-129
A8J3U2	27.67	2		Component of U2 snRNP complex	LEA1	1.00E-08	P09661	3.00E-55	AT1G09760	2.00E-59
A8IDW3	48.49	1	2	MYB-like transcription factor similar	RPN11	2.00E-06	Q5VVJ2	3.00E-14	AT3G09600	4.00E-08
A8HME6	68.7	2		Polyadenylate-binding protein	PAB1	e-127	P11940	e-135	AT1G49760	e-127
A8I1B8	14.68		3	RNA export factor	PAB1	7.00E-07	Q86V81	4.00E-21	AT5G59950	1.00E-24
A8HRV5	26.14	1		U2B component of U2 snRNP	MSL1	7.00E-10	P08579	1.00E-75	AT1G06960	6.00E-88

Table 2. The 8 potential *Cyanidioschyzon* PCAPs identified.

Protein names	Wt. (kDa)	Heptapeptide column		Nonapeptide column		Annotations	Best hit in yeast and e-value		Best hit in human and e-value		Best hit in *Arabidopsis* and e-value	
		0.3M	0.5M	0.3M	0.5M							
CMM263C	81.13			24		TOP1	TOP1	e-155	P11387	e-152	AT5G55300.1	e-180
CMM087C	44.81			1		SWIB/MDM2 domain containing protein	TRI1	1.00E-12	F8VUB0	7.00E-06	AT3G19080.1	7.00E-20
CMT578C	30.16			5		Similar to methylated-DNA--protein-cysteine methyltransferase	MGT1	2.00E-09				
CMH210C	47.26			7		peptidyl-prolyl cis-trans isomerase activity	ESS1	5.00E-24	Q13526	8.00E-18	AT2G18040.1	7.00E-23
CMS377C	44.66			13	6	Casein kinase I isoform	HRR25	e-142	B0QY34	e-160	AT4G26100.1	e-156
CMG052C	44.5				8	Myb-related transcription factor	BAS1	6.00E-13	E9PJ96	6.00E-24	AT3G18100.2	2.00E-33
CMH135C	30.29	4	2			mRNA export	YRA1	2.40E-05	E9PB61	8.00E-09	AT5G59950.2	5.00E-10
CMS144C	17.17	1				TBP-associated factor TAF12	TAF12	1.00E-14	Q16514	5.00E-24	AT3G10070.1	8.00E-14

Table 3. The 154 identified proteins from *C. reinhardtii*.

Localization	**Protein names**	**Annotations**	**Best hit in yeast and e-value**		**Best hit in human and e-value**		**Best hit in *Arabidopsis* and e-value**	
Nucleus	Q84SA0	Casein kinase I	HRR25	e-139	P48730	e-165	AT4G26100	e-171
	A8IYG9	Casein kinase II subunit alpha	CKA2	e-111	E7EU96	e-118	AT2G23070	e-129
	A8J3U2	Component of U2 snRNP complex	LEA1	1.00E-08	P09661	3.00E-55	AT1G09760	2.00E-59
	A8IDW3	Histone H2A deubiquitinase	RPN11	2.00E-06	Q5VVJ2	3.00E-14	AT3G09600	4.00E-08
	A8HME6	Polyadenylate-binding protein	PAB1	e-127	P11940	e-135	AT1G49760	e-127
	A8I1B8	RNA export factor	PAB1	7.00E-07	Q86V81	4.00E-21	AT5G59950	1.00E-24
	A8HRV5	U2B component of U2 snRNP	MSL1	7.00E-10	P08579	1.00E-75	AT1G06960	6.00E-88
	A8JI44	Exportin-7			Q9UIA9	3.00E-48	AT5G06120	2.00E-89
	A8J3F0	High mobility group protein	NHP6B	1.00E-09	E9PES6	2.00E-10	AT4G11080	6.00E-07
	A8J591	Puf protein	PUF6	1.00E-37	Q15397	2.00E-47	AT3G16810	3.00E-49
	A8ITC0	Pumilio domain-containing protein	IPL1	7.00E-34	O14965	1.00E-47	AT2G45490	3.00E-46
	A8IW57	Zinc finger protein	BUD20	4.00E-13	O00488	3.00E-12	AT2G36930	1.00E-21
	A8IV98	DEAD box RNA helicase	DBP3	e-147	P17844	e-121	AT1G31970	e-175
	A8JHA8	ATP-dependent RNA helicase DDX54	DBP10	e-113	Q8TDD1	e-126	AT1G77030	e-157
	A8J0A8	Subunit of U3-containing Small Subunit (SSU) processome complex;	SAS10	4.00E-08	Q9NQZ2	2.00E-11	AT2G43650	5.00E-10
	A8IJG8	WD repeat-containing protein 46	UTP7	e-131	O15213	e-112	AT3G10530	e-147
	A8J763	RNA exonuclease	REX4	8.00E-50	Q9GZR2	2.00E-48	AT3G15080	5.00E-47
	A8JCZ5	Ribosomal RNA small subunit methyltransferase NEP1	EMG1	2.00E-62	Q92979	7.00E-77	AT3G57000	8.00E-88
	A8HPV5	ribosome biogenesis regulatory protein			Q15050	2.00E-06	AT2G37990	1.00E-09
	A8IWU0	Ribosome production factor	RPF2	1.00E-46	Q9H7B2	9.00E-63	AT3G23620	4.00E-84
	A8IED9	Pseudouridine synthase catalytic subunit of box H/ACA snoRNPs	CBF5	0.00E+00	O60832	0	AT3G57150	0
	A8I6R1	Protein required for biogenesis of ribosomal subunit;	SOF1	4.00E-52	Q9NV06	3.00E-53	AT4G28450	2.00E-74
	A8I4A8	Nucleolar component of the spliceosomal ribonucleoprotein complexes	NOP4	1.00E-27	Q9NW13	7.00E-39	AT2G21440	1.00E-59
	A8I0Z4	Nucleolar GTP-binding protein 1	NOG1	e-169	Q9BZE4	0	AT1G50920	0
	A8JB67	Nucleolar protein, small subunit of H/ACA snoRNPs	NHP2	8.00E-43	Q9NX24	9.00E-30	AT5G08180	6.00E-35
	A8IA86	methyltransferase fibrillarin	NOP1	e-121	M0R299	e-131	AT4G25630	e-134
Non-nucleus	A8IWI1	Mitochondrial ribosomal protein L17			Q9H2W6	8.00E-06	AT1G14620	7.00E-15
	A8HXM1	Mitochondrial ribosomal protein L29						
	A8I8Z4	Plastid ribosomal protein L1	MRPL1	3.00E-05			AT3G63490	2.00E-81
	A8HWZ6	Plastid ribosomal protein L13	MRPL23	3.00E-24	Q9BYD1	1.00E-15	AT1G78630	4.00E-59
	A8JAL6	Plastid ribosomal protein L15	MRPL10	2.00E-10			AT3G25920	8.00E-54
	A8I3M4	Plastid ribosomal protein L17	MRPL8	6.00E-11	Q9NRX2	9.00E-12	AT3G54210	7.00E-44
	A8HNJ8	Plastid ribosomal protein L18					AT1G48350	4.00E-41
	A8J9D9	Plastid ribosomal protein L24			Q96A35	5.00E-09	AT5G54600	5.00E-42
	A8INR7	Plastid ribosomal protein L27	MRP7	3.00E-11	Q9P0M9	4.00E-10	AT5G40950	4.00E-27
	A8HWS8	Plastid ribosomal protein L28					AT2G33450	4.00E-24
	A8I1D3	Plastid ribosomal protein L33					ATCG00640	8.00E-15

Q84U22	Plastid ribosomal protein L4			Q9BYD3	8.00E-14	AT1G07320	2.00E-51
A8J503	Plastid ribosomal protein L6	MRPL6	3.00E-25			AT1G05190	3.00E-58
A8HTY0	Plastid ribosomal protein L7/L12	MNP1	1.00E-07			AT3G27850	1.00E-18
A8IYS1	Plastid ribosomal protein L9					AT3G44890	3.00E-24
A8JDP6	Plastid ribosomal protein S13	SWS2	8.00E-17	P62269	2.00E-08	AT5G14320	7.00E-38
A8JDN8	Plastid ribosomal protein S16	MRPS16	3.00E-15	A6ND22	2.00E-09	ATCG00050	2.00E-17
A8JGS2	Plastid ribosomal protein S17	MRPS17	1.00E-10	P62280	1.00E-04	AT1G49400	6.00E-19
A8JDN4	Plastid ribosomal protein S20					AT3G15190	4.00E-15
A8IMN3	Plastid-specific ribosomal protein 6						
A8I645	Ribosomal protein	CIC1	1.00E-05	O76021	8.00E-30	AT3G58660	2.00E-39
A8J597	Ribosomal protein L12	RPL12A	7.00E-74	P30050	3.00E-78	AT2G37190	4.00E-89
A8IQE3	Ribosomal protein L14	RPL14A	5.00E-10	E7EPB3	1.00E-16	AT4G27090	9.00E-28
A8JI94	Ribosomal protein L22	RPL22A	5.00E-12	C9JYQ9	1.00E-20	AT3G05560	1.00E-38
A8HMG7	Ribosomal protein L26	RPL26A	1.00E-31	P61254	8.00E-41	AT3G49910	3.00E-46
A8JF05	Ribosomal protein L28			P46779	4.00E-12	AT2G19730	4.00E-15
A8ICT1	Ribosomal protein L30	RPL30	2.00E-43	P62888	7.00E-58	AT1G36240	8.00E-57
A8HP90	Ribosomal protein L6	RPL6B	3.00E-46	Q02878	3.00E-35	AT1G74050	2.00E-53
A8IVE2	Ribosomal protein L7	RPL7A	2.00E-69	A8MUD9	2.00E-80	AT2G44120	2.00E-90
A8J567	Ribosomal protein L7a	RPL8B	5.00E-71	P62424	1.00E-77	AT3G62870	2.00E-92
A8JDP4	Ribosomal protein L9	RPL9B	2.00E-66	P32969	3.00E-71	AT4G10450	1.00E-81
A8HSU7	Ribosomal protein S16	RPS16A	3.00E-71	P62249	6.00E-73	AT2G09990	8.00E-78
A8J8M9	Ribosomal protein S20	RPS20	1.00E-38	P60866	1.00E-60	AT3G47370	4.00E-59
A8IZ36	Ribosomal protein S25	RPS25A	1.00E-24	P62851	1.00E-25	AT4G39200	7.00E-29
A8IKP1	Ribosomal protein S28	RPS28B	1.00E-19	P62857	1.00E-19	AT5G03850	1.00E-20
A8JGI9	Ribosomal protein S7	RPS7B	2.00E-64	P62081	5.00E-79	AT1G48830	1.00E-89
C5HJB7	Ribosomal protein S9			P82933	6.00E-06	AT1G74970	1.00E-15
RR19	30S ribosomal protein S19, chloroplastic	RSM19	1.00E-15	P62841	1.00E-09	ATCG00820	7.00E-44
Q6Y682	38 kDa ribosome-associated protein	YLL056C	1.00E-07	B3KV61	1.00E-04	AT1G09340	e-174
RS14	40S ribosomal protein S14	RPS14A	6.00E-73	P62263	2.00E-86	AT2G36160	5.00E-85
A8I0I1	40S ribosomal protein S24	RPS24B	4.00E-44	P62847	2.00E-35	AT5G28060	2.00E-57
Q6Y683	41 kDa ribosome-associated protein					AT3G63140	2.00E-90
RK22	50S ribosomal protein L22, chloroplastic					ATCG00810	2.00E-29
RK5	50S ribosomal protein L5, chloroplastic	MRPL7	1.00E-22	Q5VVC9	4.00E-06	AT4G01310	2.00E-74
RL11	60S ribosomal protein L11	RPL11B	5.00E-78	P62913	2.00E-79	AT2G42740	8.00E-93
Q8GUQ9	60S ribosomal protein L38	RPL38	5.00E-15	P63173	2.00E-32	AT3G59540	1.00E-35
A8I232	Eukaryotic initiation factor	GCD11	0.00E+00	P41091	0	AT1G04170	0
A8HX38	Eukaryotic translation elongation factor 1	TEF1	e-105	Q05639	e-113	AT5G60390	e-106
A8HWK8	Subunit of the signal recognition particle			Q9UHB9	4.00E-36	AT5G61970	3.00E-54
A8JH66	Subunit of the signal recognition particle			O76094	3.00E-15	AT1G67680	5.00E-29
A8JG36	Subunit of the signal recognition particle	SEC65	2.00E-11	P09132	1.00E-22	AT1G48160	9.00E-27

	A8IAB5	Flagella associated protein						
	A8IZG0	Flagellar associated protein						
	A8JAC0	Flagellar associated protein						
	A8IAA8	Flagellar associated protein						
	A8IXA1	Protein involved in an early step of 60S ribosomal subunit biogenesis;	MAK11	6.00E-22	O75695	3.00E-44	AT1G65030	3.00E-46
	A8IAF7	RNA pseudouridine synthase	RIB2	4.00E-12	B4DDD1	2.00E-09	AT1G76050	1.00E-99
	A8JA59	ABC transporter G family	YOL075C	2.00E-13	Q9UNQ0	1.00E-07	AT2G29940	3.00E-19
	A8JDV2	Alpha subunit of the nascent polypeptide-associated complex (NAC);	EGD2	8.00E-22	Q13765	4.00E-59	AT3G49470	3.00E-59
	A8JA80	AP-2 complex subunit mu	APM1	2.00E-89	Q96CW1	e-122	AT5G46630	e-160
	A8IL88	Axin interactor, dorsalization-associated protein			F5H715	5.00E-33		
	A8IHL6	Calcium/calmodulin dependent protein kinase II Association;			H0Y9J2	6.00E-39		
	A8IZI4	Carbohydrate sulfotransferase 15			Q7LFX5	3.00E-36		
	A8JIC1	Carbohydrate sulfotransferase 15			Q7LFX5	8.00E-11		
	Q6PLP6	Cell wall protein GP2						
	CB29	Chlorophyll a-b binding protein CP29					AT2G40100	4.00E-83
	A8IIK4	Chloroplast stem-loop-binding protein					AT3G63140	3.00E-91
	Q9XHE2	Class II DNA photolyase					AT1G12370	0
	A8I2M1	Exostosin-like glycosyltransferase			P22105	2.00E-06	AT3G57630	5.00E-28
	A8JHN6	Exostosin-like glycosyltransferase			Q93063	2.00E-05	AT3G57630	2.00E-32
	Q9LD42	Fe-assimilating protein 1						
	ALFC	Fructose-bisphosphate aldolase 1, chloroplastic			P04075	3.00E-90	AT4G38970	e-172
	G3PA	Glyceraldehyde-3-phosphate dehydrogenase A, chloroplastic	TDH3	4.00E-98	P04406	2.00E-86	AT1G42970	0
	A8JFM5	Glycosyltransferase-like protein LARGE2			Q8N3Y3	1.00E-17		
	A8IWB3	Low-CO2-inducible protein						
	A8IGD9	Low-CO2-inducible protein						
	MDHM	Malate dehydrogenase, mitochondrial	MDH1	8.00E-99	P40926	e-121	AT1G53240	e-164
	A8J979	Methylcrotonoyl-CoA carboxylase alpha subunit	DUR1,2	5.00E-94	Q96RQ3	e-161	AT1G03090	e-157
	A8J7A9	Mitogen-activated protein kinase kinase kinase 9	BCK1	2.00E-14	J3KPI6	2.00E-29	AT2G42640	7.00E-28
	A8HYN3	NADP-dependent malic enzyme	MAE1	1.00E-80	P48163	e-153	AT5G25880	e-172
	A8IQU9	Oligopeptidase			P48147	1.00E-20	AT1G50380	1.00E-83
	PSBP	Oxygen-evolving enhancer protein 2					AT1G06680	3.00E-70
	A8HQ69	Protein sel-1 homolog			Q9UBV2	1.00E-21	AT1G18260	2.00E-19
	A8HQL5	Pyridoxal-5'-phosphate-dependent enzyme family protein;					AT3G26115	9.00E-09
	A8HQC9	Rhodanese-like domain;	UBA4	1.00E-05				
	RBL	Ribulose bisphosphate carboxylase large chain					ATCG00490	0
	A8JIC2	Serine/threonine-protein kinase CTR1	CDC15	1.00E-07	Q02779	3.00E-13	AT5G03730	3.00E-15
	A8J3M8	Superoxide dismutase	SOD2	2.00E-19	P04179	1.00E-29	AT3G10920	7.00E-29
	A8HPY3	tyrosylprotein sulfotransferase					AT1G08030	9.00E-41
	A8HN92	Uridine 5'-monophosphate synthase	URA3	2.00E-62	P11172	e-161	AT3G54470	0
Unannotated	A8HNG8							

A8HQC6						
A8HQL4					AT3G26115	4.00E-10
A8HYZ9					AT2G01640	4.00E-06
A8HZK3						
A8I2L9						
A8I363						
A8I4J5						
A8I829	RRP14	1.00E-04			AT5G05210	2.00E-12
A8IAA9						
A8IBT9					AT4G05400	1.00E-04
A8IHD2						
A8IHJ7						
A8IKY2						
A8ITX3						
A8IVS3						
A8IWR4						
A8IY50						
A8IZS7						
A8J0X6						
A8J127					AT2G45830	2.00E-05
A8J148						
A8J290						
A8J2L0						
A8J437						
A8J4A2						
A8J6I0						
A8J7S1						
A8JAA9						
A8JBA6	NHP2	1.00E-06				
A8JBL1			Q6UW63	5.00E-10		
A8JBR8			Q6UW63	2.00E-06		
A8JD45						
A8JE77			Q6UW63	5.00E-08		
A8JGF5						
A8JH42						
A8JH86						
A8JI67						

Table 4. The identified 133 proteins from *C. merolae*.

Localization	Protein names	Heptad column		Nonatad column		Annotations	Best hit in yeast and e-value		Best hit in human and e-value		Best hit in Arabidopsis and e-value	
		0.3M	0.5M	0.3M	0.5M							
Nucleus	CMS377C	0		13	6	Casein kinase I isoform	HRR25	e-142	B0QY34	e-160	AT4G26100	e-156
	CMH135C	4	2			mRNA export	YRA1	2.40E-05	E9PB61	8.00E-09	AT5G59950	5.00E-10
	CMG052C		0		8	Myb-related transcription factor	BAS1	6.00E-13	E9PJ96	6.00E-24	AT3G18100	2.00E-33
	CMH210C			7		peptidyl-prolyl cis-trans isomerase activity	ESS1	5.00E-24	Q13526	8.00E-18	AT2G18040	7.00E-23
	CMT578C			5		Similar to methylated-DNA--protein-cysteine methyltransferase	MGT1	2.00E-09				
	CMM087C			1		SWIB/MDM2 domain containing protein	TRI1	1.00E-12	F8VUB0	7.00E-06	AT3G19080	7.00E-20
	CMS144C	1				TBP-associated factor TAF12	TAF12	1.00E-14	Q16514	5.00E-24	AT3G10070	8.00E-14
	CMM263C	0		24	0	TOP1	TOP1	e-155	P11387	e-152	AT5G55300	e-180
	CMN174C				5	Histone H2A	HTA1	3.00E-56	P0C0S8	4.00E-59	AT1G51060	2.00E-56
	CMN145C	3	6	5		Histone H2B	HTB2	4.00E-55	Q99880	6.00E-59	AT3G45980	2.00E-53
	CMR457C	1	4	3	3	Histone variant H2AZ	HTZ1	1.00E-50	P0C0S5	2.00E-56	AT3G54560	3.00E-57
	CMN183C	4	2	4	3	Histones H1			P07305	1.00E-04		
	CMT575C	0		9	1	3'-5' exonuclease activity	REX4	4.00E-42	Q8WTP8	1.00E-35	AT3G15080	2.00E-28
	CMD071C			1		3'-5' exonuclease activity			Q8N9H8	3.00E-15	AT1G56310	1.00E-21
	CMC063C			5		Methyltransferase for rRNA	EMG1	8.00E-55	Q92979	7.00E-60	AT3G57000	1.00E-56
	CMI184C			4		Protein component of the H/ACA snoRNP pseudouridylase complex	GAR1	4.00E-23	Q9NY12	3.00E-18	AT3G03920	9.00E-23
	CMF022C			9	0	PseudoUridine Synthase	PUS4	2.00E-15	Q8WWH5	4.00E-31	AT5G14460	4.00E-34
	CMP061C	0		9	1	rRNA-processing protein			E5RGP0	3.00E-06	AT2G34570	1.00E-04
	CMN074C	0		8	5	rRNA 2'-O-methyltransferase fibrillarin	NOP1	e-111	P22087	e-124	AT5G52470	e-123
	CMK102C			4		Ribosomal Protein			Q96EU6	1.00E-06	AT1G12650	1.00E-05
	CMT080C			1		Ribosome biogenesis protein UTP30	UTP30	2.00E-10	J3QSV6	2.00E-15	AT2G42650	4.00E-20
	CMP145C	1		5		heat shock protein 70	SSA1	0.00E+00	P11142	0.00E+00	AT3G12580	0.00E+00
	CMQ470C	1		5	0	thioredoxin peroxidase	DOT5	2.00E-19			AT3G26060	8.00E-07
Non-nucleus	CMA082C	8	0	16	10	40S ribosomal protein S2	RPS2	3.00E-91	P15880	e-113	AT1G58684	e-113
	CMG109C	3	1	8	5	40S ribosomal protein S15	RPS15	2.00E-43	P62841	3.00E-57	AT1G04270	5.00E-55
	CMI202C		0	3	2	40S ribosomal protein S15A	RPS22A	4.00E-60	P62244	4.00E-61	AT5G59850	8.00E-60
	CMP007C	0	1		2	40S ribosomal protein S16	RPS16A	6.00E-63	P62249	1.00E-66	AT2G09990	1.00E-69
	CMB004C			2	0	40S ribosomal protein S18	RPS18B	5.00E-68	P62269	1.00E-77	AT4G09800	2.00E-80
	CMR148C	0	0	11	2	40S ribosomal protein S19	RPS19B	5.00E-34	P39019	8.00E-40	AT3G02080	4.00E-48
	CMN125C	1	0			40S ribosomal protein S27A	RPS31	1.00E-16	P62979	6.00E-17	AT2G47110	5.00E-19
	CMO024C	2	0	1		40S ribosomal protein S28	RPS28B	2.00E-21	P62857	6.00E-22	AT5G03850	3.00E-15
	CMN148C			2		40S ribosomal protein S3	RPS3	e-101	P23396	e-116	AT5G35530	e-120
	CMT030C	3	1	4	2	40S ribosomal protein S30	RPS30B	3.00E-15	E9PR30	1.00E-15	AT5G56670	4.00E-18
	CMT627C	3	1	7	8	40S ribosomal protein S5	RPS5	e-101	P46782	e-115	AT2G37270	e-112

CMA122C	5	2	11	5	40S ribosomal protein S7	RPS7B	2.00E-54	P62081	6.00E-64	AT1G48830	2.00E-65
CMT159C		1			40S ribosomal protein S8	RPS8B	5.00E-79	P62241	2.00E-81	AT5G59240	5.00E-89
CMJ109C	6	7	8	7	60S ribosomal protein L10	RPL10	e-102	P27635	e-114	AT1G66580	e-110
CML196C		1			60S ribosomal protein L11	RPL11B	3.00E-61	P62913	2.00E-65	AT5G45775	4.00E-67
CMH065C			3	0	60S ribosomal protein L14	RPL14A	4.00E-28	P50914	2.00E-29	AT2G20450	2.00E-34
CMQ463C		1		0	60S ribosomal protein L17	RPL17A	3.00E-49	J3QQT2	1.00E-61	AT1G27400	7.00E-62
CMO302C	12	6	16	14	60S ribosomal protein L18a	RPL20B	2.00E-39	Q02543	9.00E-44	AT2G34480	9.00E-45
CMP179C			5		60S ribosomal protein L1-A	RPL1A	4.00E-83	P62906	1.00E-90	AT2G27530	4.00E-94
CMR150C	5	1	8	4	60S ribosomal protein L21	RPL21A	1.00E-40	P46778	2.00E-37	AT1G09690	3.00E-48
CMS262C			6		60S ribosomal protein L23	RPL23B	8.00E-74	P62829	3.00E-80	AT3G04400	1.00E-79
CMK273C	4	7		1	60S ribosomal protein L23A	RPL25	5.00E-41	P62750	8.00E-51	AT3G55280	7.00E-43
CMG157C		7		4	60S ribosomal protein L26	RPL26B	1.00E-37	E5RIT6	1.00E-38	AT3G49910	5.00E-43
CML305C		1		2	60S ribosomal protein L27	RPL27A	3.00E-42	P61353	2.00E-39	AT3G22230	4.00E-39
CMM040C	6	7	5	7	60S ribosomal protein L30	RPL30	5.00E-41	P62888	2.00E-48	AT3G18740	3.00E-46
CMP175C	5	1	7	5	60S ribosomal protein L31	RPL31B	2.00E-19	P62899	7.00E-25	AT4G26230	3.00E-18
CMP012C	0	2	2	2	60S ribosomal protein L34	RPL34B	1.00E-37	P49207	1.00E-25	AT1G69620	6.00E-30
CMC053C	4	4	2	5	60S ribosomal protein L35	RPL35A	2.00E-28	P42766	4.00E-39	AT5G02610	1.00E-40
CMN315C	0		2		60S ribosomal protein L37A	RPL43A	8.00E-31	P61513	6.00E-34	AT3G60245	1.00E-32
CMJ170C	1	3	2	3	60S ribosomal protein L38	RPL38	2.00E-17	P63173	2.00E-22	AT3G59540	1.00E-20
CMC044C	1	0	3	1	60S ribosomal protein L44	RPL42A	2.00E-34	P83881	9.00E-33	AT4G14320	4.00E-36
CMH071C	1	8			60S ribosomal protein L5	RPL5	5.00E-96	P46777	e-108	AT5G39740	e-109
CMQ078C		5	3	7	60S ribosomal protein L6	RPL6B	2.00E-35	Q02878	4.00E-38	AT1G74050	2.00E-47
CMO310C		2			60S ribosomal protein L7	RPL7A	9.00E-74	P18124	1.00E-79	AT2G01250	1.00E-79
CML317C			2		60S ribosomal protein L7A	RPL8A	4.00E-82	P62424	1.00E-83	AT2G47610	3.00E-96
CMR287C	2	3		1	60S ribosomal protein L8	RPL2B	5.00E-99	P62917	e-101	AT2G18020	e-102
CMC145C	2		12	9	60S ribosomal protein L9	RPL9B	9.00E-57	P32969	6.00E-63	AT4G10450	6.00E-63
CMV189C			2	2	28S ribosomal protein S12, mitochondrial	MRPS12	7.00E-43	O15235	1.00E-26	ATCG00905	1.00E-63
CMV084C	3	2	2	2	28S ribosomal protein S16, mitochondrial	MRPS16	8.00E-10	A6ND22	9.00E-08	AT4G34620	8.00E-15
CMV173C	2	5	5	3	28S ribosomal protein S17, mitochondrial	MRPS17	2.00E-09			AT1G79850	2.00E-12
CMS081C		1	4	2	28S ribosomal protein S34, mitochondrial			P82930	3.00E-05	AT5G52370	1.00E-06
CMV170C			2		30S ribosomal protein S3, chloroplastic					ATCG00800	1.00E-38
CMV180C		4	8	8	30S ribosomal protein S5, chloroplastic					AT2G33800	2.00E-13
CMV177C	2			5	30S ribosomal protein S8, chloroplastic	RPS22A	5.00E-08	P62244	1.00E-06	ATCG00770	2.00E-21
CMV187C			4	2	30S ribosomal protein S9, chloroplastic	MRPS9	5.00E-05	P82933	2.00E-05	AT1G74970	2.00E-22
CMV168C	2	3	3	3	37S ribosomal protein S19, mitochondrial	RSM19	1.00E-15	K7ELC2	2.00E-08	ATCG00820	3.00E-38
CMV190C	2	2	13	7	37S ribosomal protein S7, mitochondrial	RSM7	1.00E-10	J3KSI8	2.00E-12	ATCG00900	4.00E-48
CMV009C	4		11	10	37S ribosomal protein, mitochondrial	NAM9	5.00E-07			ATCG00380	4.00E-42
CMV183C			4		37S ribosomal protein, mitochondrial	SWS2	2.00E-15			AT5G14320	4.00E-26
CMV171C	1	0	2	5	39S ribosomal protein L16, mitochondrial	MRPL16	8.00E-15	Q9NX20	2.00E-08	ATCG00790	2.00E-57
CMV186C				2	39S ribosomal protein L23, mitochondrial	MRPL23	6.00E-10			AT1G78630	2.00E-19

CMV035C			2	1	39S ribosomal protein L27, mitochondrial	MRP7	4.00E-10	Q9P0M9	3.00E-05	AT2G16930	1.00E-20
CMV164C	3		8	6	39S ribosomal protein L9, mitochondrial	MRPL9	6.00E-20	P09001	1.00E-17	AT2G43030	9.00E-34
CMW038C	1		6		50S ribosomal protein L16, mitochondrial	MRPL16	1.00E-06			ATCG00790	4.00E-13
CMV179C			3	1	50S ribosomal protein L18, chloroplastic					AT1G48350	4.00E-24
CMV166C	1		3		50S ribosomal protein L23, chloroplastic					ATCG01300	3.00E-06
CMV175C	6	2	7	7	50S ribosomal protein L24, chloroplastic					AT5G54600	1.00E-15
CMW036C			2		50S ribosomal protein L5, mitochondrial						
CMV178C			5	0	54S ribosomal protein L6, mitochondrial	MRPL6	2.00E-20			AT1G05190	7.00E-43
CMQ292C	0		9	0	Chloroplast ribosomal protein L15	MRPL10	6.00E-10	E9PLX7	2.00E-05	AT3G25920	5.00E-39
CMB032C	1		1	2	Chloroplast ribosomal protein S21					AT3G27160	8.00E-07
CMP308C	3	0	4	1	Mitochondrial ribosomal protein L27	MRP7	1.00E-16	Q9P0M9	2.00E-14	AT2G16930	1.00E-21
CMH275C			9	0	Mitochondrial ribosomal protein L46	MRPL17	2.00E-05	Q9H2W6	7.00E-16	AT1G14620	7.00E-14
CMT544C	1	0	5	2	Mitochondrial ribosomal protein S16	MRPS16	1.00E-17	A6ND22	1.00E-11	AT5G56940	6.00E-17
CMS212C		0		1	Mitochondrial ribosomal protein S17	MRPS17	4.00E-06			AT1G49400	2.00E-13
CMV108C			3		Ribosomal protein L19 family protein					AT5G11750	2.00E-13
CMV158C	0		3	6	[pt] allophycocyanin (APC) alpha chain						
CMV159C		0		3	[pt] allophycocyanin (APC) beta chain						
CMV162C	4	4	7	5	[pt] DNA-binding protein Hu homolog						
CMV063C			5	3	[pt] phycocyanin (PC) alpha chain						
CMV064C	0	0	4		[pt] phycocyanin (PC) beta chain						
CMQ087C	5	0	16	12	chloroplast ATP synthase	ATP3	2.00E-22	P36542	2.00E-23	AT4G04640	4.00E-85
CMT202C			6		chloroplast phosphatase activity					AT2G25870	1.00E-13
CMQ121C	1		5	4	chloroplast, endonuclease activity					AT1G18680	1.00E-38
CMV163C	0		2		Hsp70 family ATPase chloroplasts	SSC1	0	P38646	0	AT4G24280	0
CMH226C	5	0	20	7	Translation Elongation Factor	TEF1	0.00E+00	Q05639	0.00E+00	AT5G60390	0
CMT223C			8		Translation initiation factor					AT4G30690	2.00E-07
CMV195C	1		2	2	ATPase	ATP16	2.00E-05			ATCG00470	2.00E-29
CMJ015C			5	0	Calcineurin-like metallo-phosphoesterase superfamily protein					AT1G18480	9.00E-42
CMI049C			6	0	Fructose-bisphosphate aldolase A			P04075	1.00E-99	AT2G01140	e-127
CMJ042C			3	0	glyceraldehyde-3-phosphate dehydrogenase	TDH2	5.00E-92	P04406	7.00E-89	AT1G42970	e-145
CMV013C	7	0	7	8	large subunit of RUBISCO					ATCG00490	0
CMN338C			23	14	oxidoreductase			Q9NZC7	9.00E-12	AT1G03630	9.00E-69
CMO306C			1		Oxidoreductase family protein;	YMR315W	4.00E-05	Q9UQ10	9.00E-13	AT4G09670	6.00E-11
CMP166C	4	3	14	16	phycocyanin-associated rod linker protein						
CMN111C	5		14	4	Protein disulfide-isomerase A6	MPD1	2.00E-25	Q15084	1.00E-31	AT2G32920	2.00E-33
CMD190C			14	7	Putative oxidoreductase			Q9NZC7	4.00E-11	AT1G03630	3.00E-66
CMV014C	4		6	3	Rubisco small subunit (RBCS) multigene family					AT1G67090	1.00E-20
CMT279C	8		12	0	similar to prostatic acid phosphatase precursor			P11117	3.00E-21		
CMJ105C	0		16	7	Tic22-like family protein; LOCATED IN: chloroplast,					AT3G23710	1.00E-07

Unannotated	CMB153C	1		5	
	CMC095C	8	1	8	6
	CMD103C			1	
	CMD165C	2	0	3	2
	CME038C			5	
	CMH254C			1	0
	CMK221C			9	
	CML117C	0		4	
	CML294C			13	3
	CMN296C	1		3	0
	CMN330C	3	1	1	1
	CMP346C	1		9	5
	CMQ170C			3	
	CMQ259C			4	
	CMR253C			12	0
	CMT270C			3	
	CMT340C	5	3	4	5
	CMT366C	3		13	4
	CMT392C			9	4
	CMT440C	0	1		1

Table 5. The identified 55 proteins from *E. coli*.

Protein names	**Heptapeptide column 0.3M**	**Nonapeptide column 0.3M**	**Annotations**
AP_004493.1	3		30S ribosomal subunit protein S11
AP_004448.1	5	5	30S ribosomal subunit protein S12
AP_004492.1	7	11	30S ribosomal subunit protein S13
AP_004483.1	3		30S ribosomal subunit protein S14
AP_003710.1	6		30S ribosomal subunit protein S15
AP_003190.1	6	2	30S ribosomal subunit protein S16
AP_004479.1	2		30S ribosomal subunit protein S17
AP_004702.1	4	1	30S ribosomal subunit protein S18
AP_004474.1	7	7	30S ribosomal subunit protein S19
AP_000687.1	4	6	30S ribosomal subunit protein S20
AP_003615.1	5	9	30S ribosomal subunit protein S21
AP_004476.1	13	13	30S ribosomal subunit protein S3
AP_004494.1	16	17	30S ribosomal subunit protein S4
AP_004487.1	11	13	30S ribosomal subunit protein S5
AP_004700.1		3	30S ribosomal subunit protein S6
AP_004449.1	9	14	30S ribosomal subunit protein S7
AP_003772.1	7	7	30S ribosomal subunit protein S9
AP_003773.1	10	8	50S ribosomal subunit protein L13
AP_004480.1	4		50S ribosomal subunit protein L14
AP_004489.1	10	12	50S ribosomal subunit protein L15
AP_004477.1	5	11	50S ribosomal subunit protein L16
AP_004496.1	6	5	50S ribosomal subunit protein L17
AP_004486.1	5	5	50S ribosomal subunit protein L18
AP_003187.1	7		50S ribosomal subunit protein L19
AP_004473.1	15	13	50S ribosomal subunit protein L2
AP_004475.1	8	12	50S ribosomal subunit protein L22
AP_004472.1	3		50S ribosomal subunit protein L23
AP_004481.1	7	8	50S ribosomal subunit protein L24
AP_002783.1	2		50S ribosomal subunit protein L25

AP_003728.1	4	4	50S ribosomal subunit protein L27
AP_004154.1	7	9	50S ribosomal subunit protein L28
AP_004471.1	4		50S ribosomal subunit protein L4
AP_004482.1	6	3	50S ribosomal subunit protein L5
AP_004485.1	6		50S ribosomal subunit protein L6
AP_003833.1	3		50S ribosomal subunit protein L7/L12
AP_004703.1	7		50S ribosomal subunit protein L9
AP_001712.1	17	18	23S rRNA pseudouridylate synthase
AP_001428.1	19	6	RNA helicase
AP_001583.1	3		ribosome modulation factor
AP_002572.1	9		DNA cytosine methylase
AP_003818.1	4		HU, DNA-binding transcriptional regulator, alpha subunit
AP_002332.1	10	3	integration host factor (IHF), DNA-binding protein, alpha subunit
AP_001542.1	3	5	integration host factor (IHF), DNA-binding protein, beta subunit
AP_002192.1	4	2	predicted regulator for DicB
AP_004701.1		4	primosomal protein N
AP_001116.1	6		primosomal replication protein N
AP_000689.1	7		bifunctional riboflavin kinase and FAD synthetase
AP_002941.1	10		fused enoyl-CoA hydratase
AP_004160.1	7		glucosyltransferase I
AP_001586.1	3		hypothetical protein
AP_004162.1	4		lipopolysaccharide core biosynthesis protein
AP_000935.1	16		predicted phage integrase
AP_002948.1	4		predicted prophage CPS-53 integrase
AP_002338.1	5	1	protein chain initiation factor IF-3
AP_004365.1	37	26	sn-glycerol-3-phosphate dehydrogenase, aerobic, FAD/NAD(P)-binding

List of figures

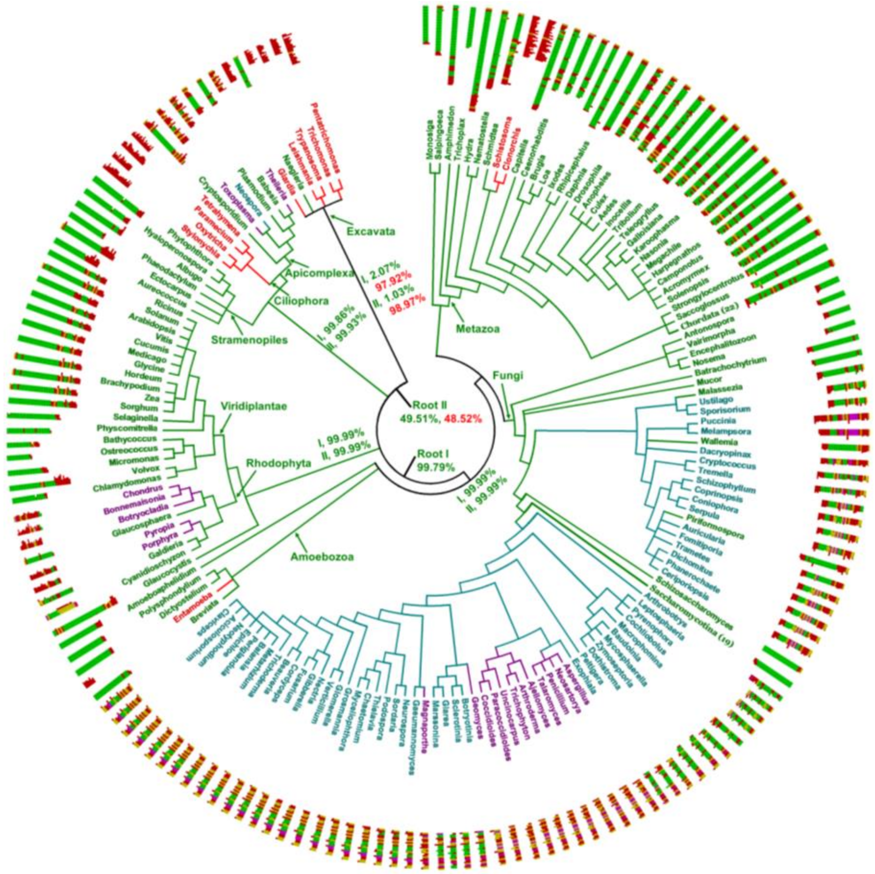
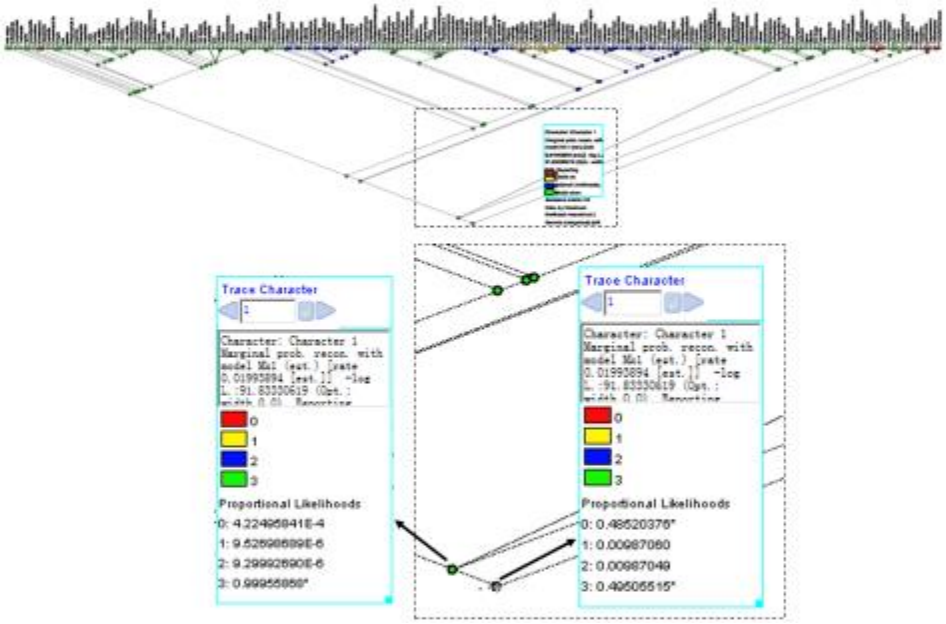


Fig. 1. CTD diversity in eukaryotes.

The tree shows consensus relationships of the 205 eukaryotes with CTD sequences mapped to each taxon. Sequences are oriented with N-termini at the outer edge and C-termini toward the center. Most CTD sequences are shown from the first obvious heptad to the C-terminal end; and for those with few or without heptads are shown from a supposed first heptad position, based on typical linker lengths, to the C-terminal end (the same convention is used in other figures). The 22 chordates are collapsed into one branch as their CTD sequences are nearly identical; the same was done for the 19 saccharomycete species. The annotated CTD structure for each genus is shown around the tree. Genus names and their branches are shown in four different colors based on their CTD states (see methods); 3 = green; 2 = teal; 1 = purple; 0 = red. Roots I and II reflect alternative rootings of the eukaryotic tree for character state analyses. The probability that the ancestor of descending clades in state 0 (completely disorganized CTDs) or state 3 (tandem repeats) are shown separately in red and green.



Trace Character

1

Character: Character 1
Marginal prob. recon. with
model Mk1 (est.) [rate
0.01993894 [est.]] -log
L.:91.83330619 (Opt.:
width 0.0). Reporting

🟥 0
🟨 1
🟦 2
🟩 3

Proportional Likelihoods
0: 4.22495841E-4
1: 9.52696699E-6
2: 9.29992690E-6
3: 0.99955858*

Trace Character

1

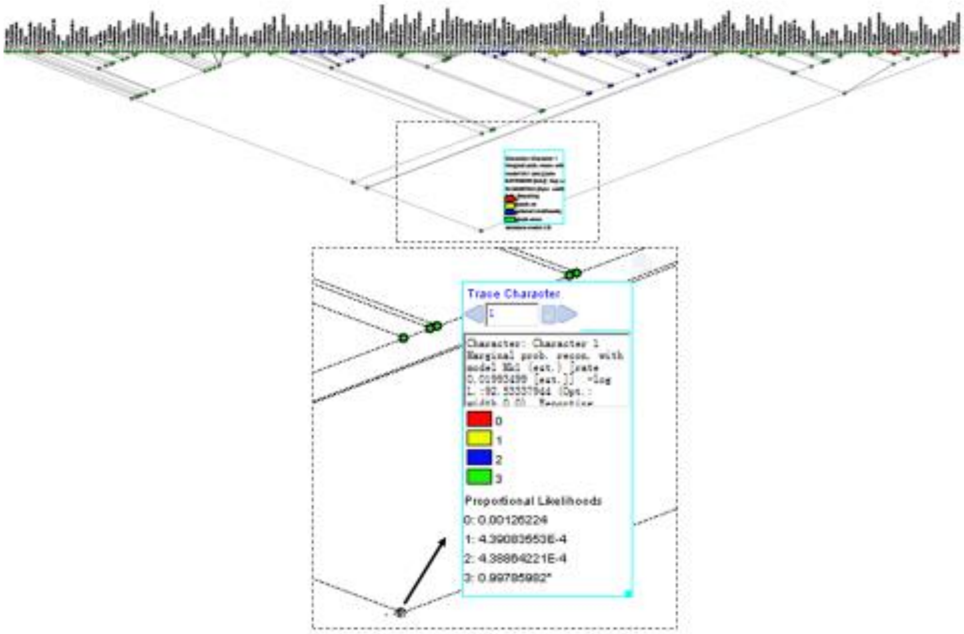
Character: Character 1
Marginal prob. recon. with
model Mk1 (est.) [rate
0.01993894 [est.]] -log
L.:91.83330619 (Opt.:
width 0.0). Reporting

🟥 0
🟨 1
🟦 2
🟩 3

Proportional Likelihoods
0: 0.48520376*
1: 0.00987060
2: 0.00987049
3: 0.49505515*

Fig. 2. The character state analysis with rooting close to Excavata.

The number 0,1,2,3 and corresponding color represent specific CTD states (See Chapter 1 Materials and Methods). The small dash-line framed area are expanded into the big frame. The possibilities of the character states of the eukaryotic common ancestor and the common ancestor of the eukaryotes except excavates are shown separately with arrows directed.



```
Character: Character 1
Marginal prob. recon. with
model Mk1 (est.) [rate
0.01993499 [est.]]  -log
L.:92.53337944 (Opt.:
width 0.01  Repetitive
```

Proportional Likelihoods
0: 0.00126224
1: 4.39083553E-4
2: 4.38864221E-4
3: 0.99785982*

Fig. 3. The character state analysis with rooting between Bikonta and Unikonta.

The annotation is similar with Fig. 2.

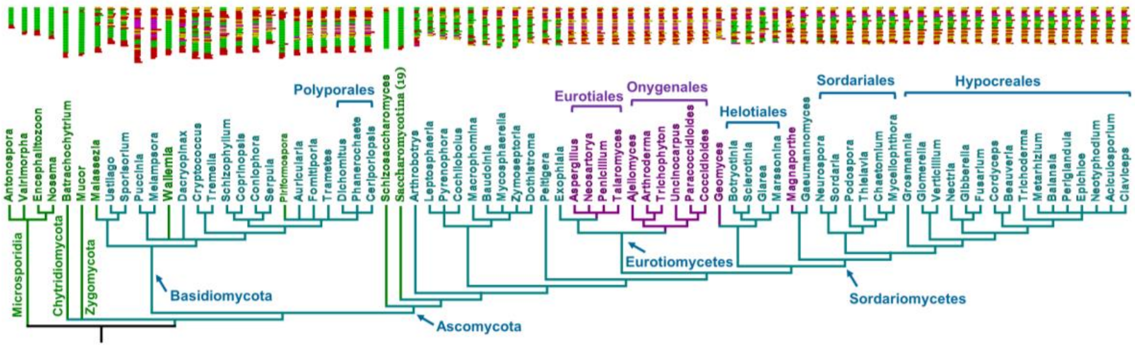


Fig. 4. CTD evolution in fungi.

The tree shows consensus relationships of all fungal genera used in this study. Branch colours are based on the conventions described for Figure 1. The annotated CTD structure for each genus is shown above the tree (CTD N-termini are at the top of each sequence). Each bracket contains all genera belonging to the taxonomic order named above.

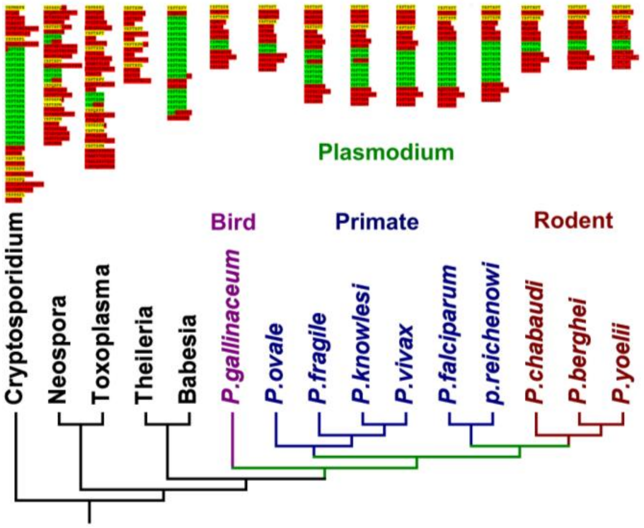


Fig. 5. CTD evolution in the Apicomplexa.

The tree shows the evolutionary relationships of apicomplexans. The 10 *Plasmodium* species are divided into three groups (shown in different colors) according to their hosts: bird, primate and rodent. CTD N-termini are at the top of each sequence.

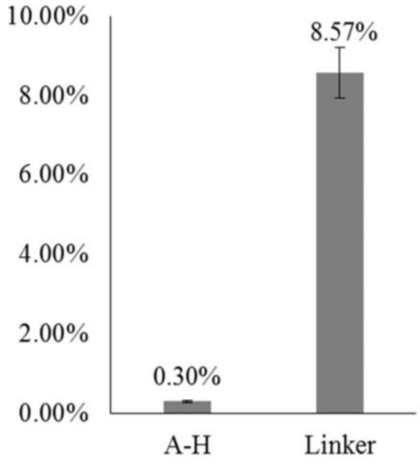


Fig. 6. Sub-motif SP content comparison.

To avoid biases based on imbalances of available RPB1 sequences across eukaryotic taxa, and similarities within closely related genera, I chose 20 RPB1 sequences (6 from Metazoa, 6 from Fungi, 4 from green plants, 3 from Apicomplexa and 1 from Excavata) from distantly related genera across eukaryotic taxa for a sub-motif comparison of Ser-Pro pair content between RPB1 domains A-H and the CTD linker region. Each bar represents the mean percent (standard errors are shown on each bar) of SP sub-motifs for the main body of RPB1 and the CTD linker respectively.

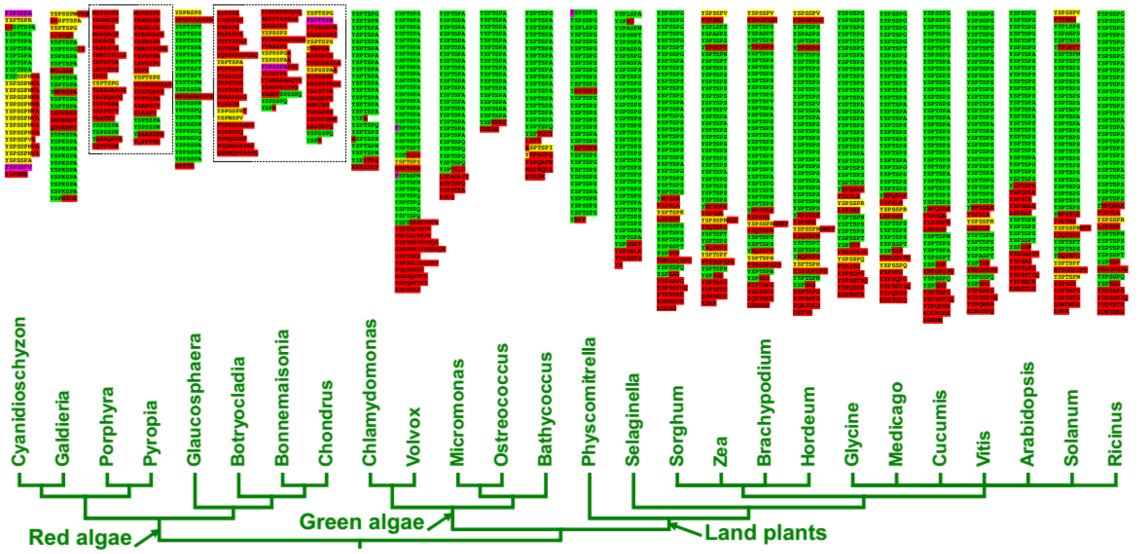


Fig. 7. The CTD in green plants and red algae.

The tree reflecting the relationships of the taxa Green algae/plants and Red algae are constructed based on the Tree of Life Web Project. Annotated CTDs for each genus is shown above the taxa included in the tree (CTD N-termini are at the top of each sequence). Sequences from multicellular red algae are shown in boxes; they have highly modified CTDs with no discernable repetitive structures that are present in unicellular (ancestral) forms. Green indicates regions with at least two continuous canonical (YSPxSPx) heptapeptides; yellow indicates the presence of isolated heptads, not in tandem with another canonical repeat; purple indicates the presence of the non-canonical motif “FSPTSPS”; red regions are without any canonical heptapeptides whatsoever. For more detail on these annotations, see Figure1.

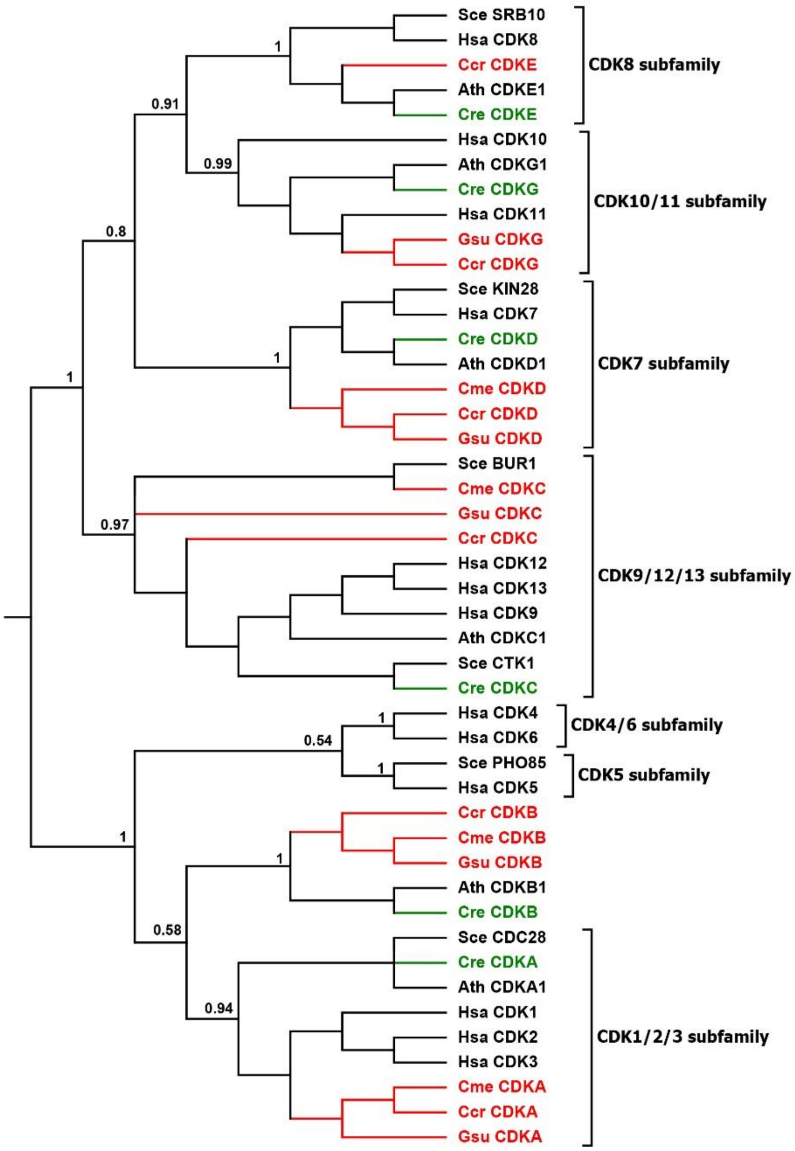


Fig. 8. Phylogenetic analyses of CDKs.

Tree recovered through Bayesian inference showing that CDKs shown previously to phosphorylate the CTD in experimental models, all present in red and green algae. Notably, CDK8 is absent from the two unicellular red algae, but is present in *Chlamydomonas*. Green algal CDKs are shown in green, red algal CDKs in red.

Abbreviations are as follows: Human, Hsa; Yeast, Sce; *Arabidopsis*, Ath; *Chlamydomonas*, Cre; *Cyanidioschyzon*, Cme; *Chondrus*, Ccr; *Galdieria*, Gsu.)

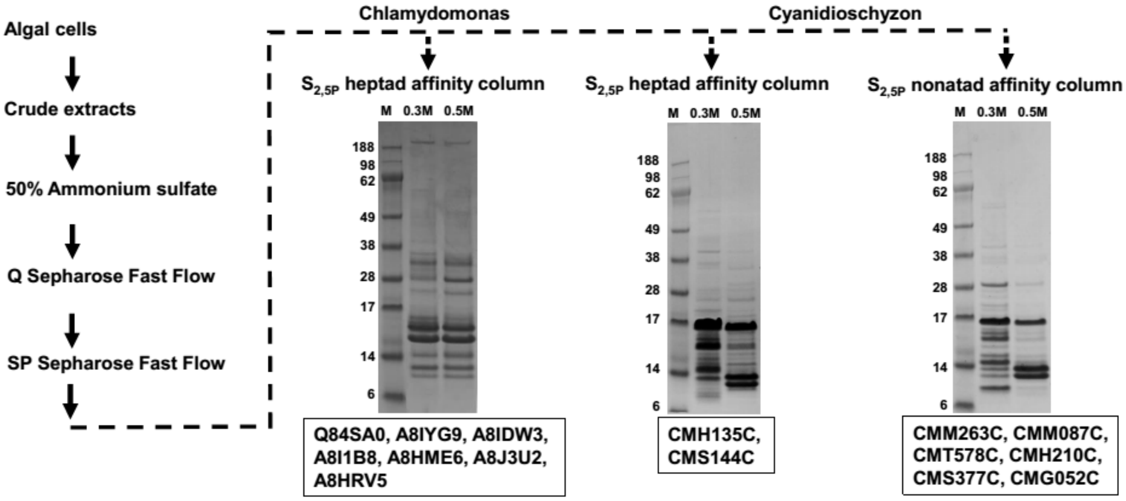

Algal cells
Crude extracts
50% Ammonium sulfate
Q Sepharose Fast Flow
SP Sepharose Fast Flow
Chlamydomonas
Cyanidioschyzon
$S_{2,5P}$ heptad affinity column
$S_{2,5P}$ heptad affinity column
$S_{2,5P}$ nonatad affinity column
M 0.3M 0.5M
M 0.3M 0.5M
M 0.3M 0.5M
188
98
62
49
38
28
17
14
6
Q84SA0, A8IYG9, A8IDW3, A8I1B8, A8HME6, A8J3U2, A8HRV5
CMH135C, CMS144C
CMM263C, CMM087C, CMT578C, CMH210C, CMS377C, CMG052C

Fig. 9. PCAP purification process.

The PCAP purification process is shown step by step as indicated by the direction of the arrows. The elution from each affinity column was subjected to SDS-PAGE followed by staining with Coomassie blue. The gels run on elutions are shown for each affinity column. M represents molecular weight (KDa) marker, 0.3 M and 0.5 M indicate elution with those concentrations of NaCl in BH buffer. The putative PCAPs from each elution highlighted in my results section are shown under the respective gels.

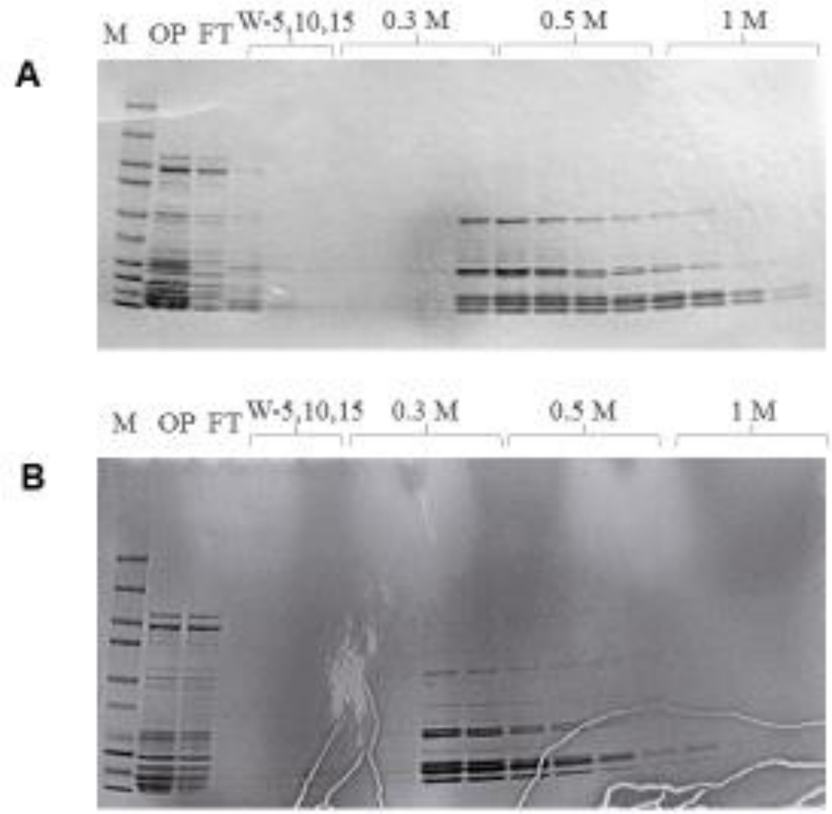


Fig. 10. The gels of *E. coli* proteins running both heptad and nonatad affinity columns.

A, *E. coli* proteins run heptad affinity column; B, *E. coli* proteins run nonatad affinity column. M represents molecular weight (KDa) ladder; OP, onput; FT, flow through; W-5, 10, 15 indicate the 5th, 10th and 15th ml wash buffer collections; 0.3 M, 0.5 M and 1 M indicate elution with those concentrations of NaCl in BH buffer; and for each concentration, 4 × 250ul elution was collected. For gel running, 5ul marker, OP and FT was separately used, and all other wells were added with 25ul samples.

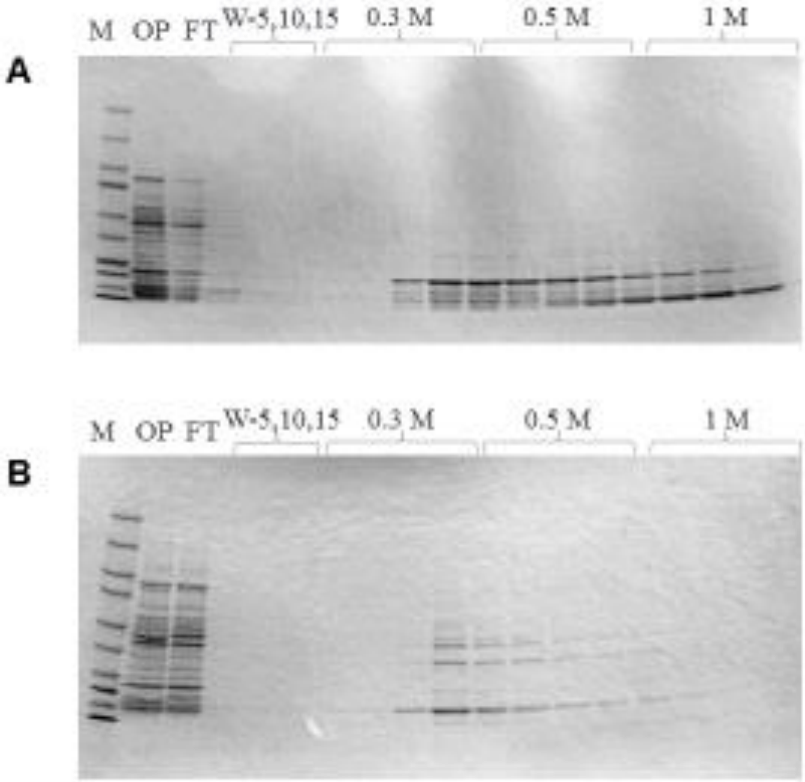


Fig. 11. The gels of *Cyanidioschyzon* proteins running both heptad and nonatad affinity columns.

The annotations are the same as Fig. 10. A, Heptad affinity column gel; B, Nonatad affinity column gel.

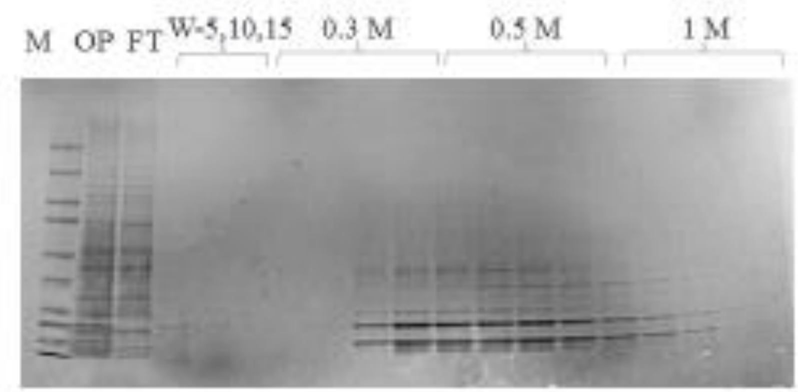


Fig. 12. The gel of *Chlamydomonas* proteins running heptad affinity column.

The annotations are the same as Fig. 10.

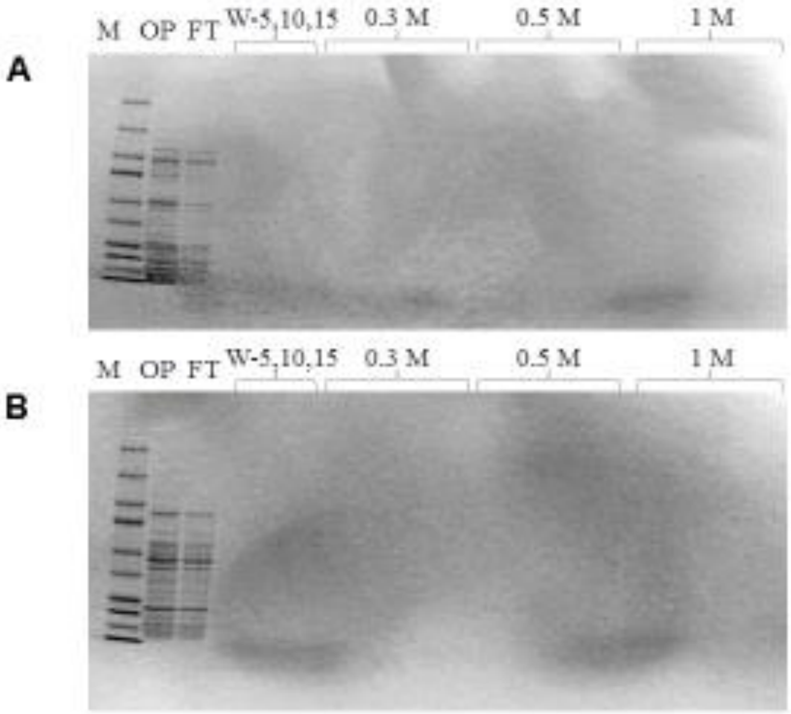


Fig. 13. Control affinity column gels.

The control affinity column was constructed by using NeutrAvidin resin without artificially synthesized peptides attached. A, *E. coli* proteins run control column. B, *Cyanidioschyzon* proteins run control column. The annotations are the same as Fig. 10.