RESEARCH PAPER

Transcription 4:2, 77–86; March/April 2013; © 2013 Landes Bioscience

Selective constraint and the evolution of the RNA Polymerase II C-Terminal Domain

Aram D. Stump* and Khrystyna Ostrozhynska

Adelphi University; Garden City, NY USA

Keywords: RNA Polymerase II, CTD, purifying selection, eukaryote evolution, selective constraint, dN/dS

Abbreviations: CTD, C-Terminal Domain; RNAP II, RNA Polymerase II; CTE, C-Terminal Extension; RV, Repeat Variability

The C-Terminal Domain (CTD) of the large subunit (Rpb1) of RNA Polymerase II has a Tyrosine-Serine-Proline-Threonine-Serine-Proline-Serine repeat structure in many eukaryotes. Chemical modifications of these residues play a central role in the regulation and coordination of the events of transcription. However, substantial variability in the presence and regularity of repeat arrays exists between eukaryote taxa. Following a survey of CTD structure from diverse eukaryote species, two hypotheses were tested relating to repeat structure and the action of selection on the CTD. First, it was found that degenerated repeat structure is associated with lower serine and proline frequencies in some eukaryote taxa but not in others. Second, maximum likelihood models of the evolution of Rpb1 in a number of species groups found that purifying selection on the non-repetitive CTD of several *Leishmania* species was substantially lower than for the rest of Rpb1, whereas purifying selection in a number of species groups containing repeat arrays was usually as high or nearly as high as for the rest of Rpb1. Characterization of CTD structure for a larger number of species than has been completed previously also revealed a greater diversity of CTD structures in eukaryotes than previously known, along with loss of repeat structure in the animals and fungi, two taxa where it has not previously been known. These results suggest that loss of CTD repeat structure has been an important aspect of RNA Polymerase II evolution in diverse eukaryotes.

Introduction

RNA Polymerase II (RNAP II), which carries out the transcription of all protein-coding genes in eukaryotes, has a multisubunit structure that is largely conserved among all eukaryotic RNA polymerases.[1,2] However, the largest subunit of RNAP II (called Rpb1) includes an extension at the C-terminal end that is not found in other RNA polymerases. In many eukaryotes, this region of Rpb1 includes a large array of heptapeptide repeats with a consensus sequence of Tyrosine-Serine-Proline-Threonine-Serine-Proline-Serine (YSPTSPS), which has been called the C-Terminal Domain (CTD).[3-5] The CTD undergoes phosphorylation at serine, threonine, and tyrosine residues, and isomerization at proline residues, at different stages of transcription.[6,7] Different chemical modification patterns lead to the recruitment and binding of different proteins involved in various stages and processes of mRNA production.[8] Through this binding, the CTD plays a central role in the regulation and coordination of the events of gene expression.

There is substantial variability in CTD structure between different eukaryote groups. The CTDs of human and *Saccharomyces cerevisiae* have very similar repeat structures, but the human CTD contains 52 repeat units and that of yeast contains only 26.[3,4] The CTD of *Drosophila melanogaster* contains 44 repeat units, but with much greater divergence at individual units from the YSPTSPS consensus than is found in either the human or yeast CTD.[9] A survey of CTD structure from a larger number of species showed that these differences reflect broad patterns of variability between and sometimes within major eukaryote taxa.[10] Within the animals, sequenced representatives of the deuterostomes, protostomes, and cnidarians have CTD repeat structures of various lengths and that match a YSPTSPS consensus to varying degrees. In the Fungi, representatives of some groups (Saccharomycetes, Schizosaccharomycetes, and Microsporidia) have very regular CTD repeat structure, whereas those of others (Leotiomycetes, Dothideomycetes, Eurotiomycetes, Basidiomycetes, and Sordariomycetes) are much less regular.

Other eukaryote groups are known to contain even greater variability in CTD structure.[10] In the Plantae, representatives of land plants, green algae, glaucophytes, and some red algae have fairly regular repeat structure, while other red algae species have Rpb1 C-terminal extensions that cannot be parsed into a repetitive structure. Within the Alveolata, apicomplexans have been found to contain repeat structure, whereas ciliates do not. The Amoebozoa also contains species with repeat structure and at least one without. CTD structures are also known for representatives of other protistan taxa: two stramenopile species have very regular YSPTSPA repeat arrays, whereas all sequenced representatives of the Excavata lack repeat structure. Finally, the plastid nucleomorphs of the cryptophyte species *Guillardia theta* and

*Correspondence to: Aram D. Stump; Email: astump@adelphi.edu
Submitted: 09/28/12; Revised: 12/14/12; Accepted: 12/17/12
http://dx.doi.org/10.4161/trns.23305

Hemiselmis andersenii, which are derived from the nucleus of an ancestral red alga, both contain an Rpb1 gene completely lacking a CTD or any C-terminal extension.[11,12]

The variability in CTD structure is likely to reflect important differences in how the CTD functions in different eukaryotes, as suggested by two lines of evidence. First, in *S. cerevisiae*, replacement of the native CTD with the less regularly structured CTD of *D. melanogaster* is lethal when homozygous, while replacement with a mammal CTD results in a normal phenotype.[9] Second, the critical functional unit of the CTD of *S. cerevisiae* has been found to be composed of a set of three serine-proline pairs within a nine-residue region, along with a pair of tyrosines seven residues apart.[10] The functional unit thus spans heptad repeats, and a long YSPTSPS repeat array will contain many overlapping units. This pattern can be found in the CTDs of diverse eukaryote taxa, however in some taxa these units are combined in an array with many repeat sequences that do not match this pattern and which would be lethal in yeast.[10] In other eukaryotes, this pattern is completely absent from the CTD. It can therefore be concluded that there are many eukaryotes with different critical functional units than that of yeast, although it is not currently known what they are. Differences in CTD structure between eukaryotes may thus reflect differences in how the CTD serves as a binding site for proteins involved in transcription-related processes, which could include interacting with different suites of proteins, or undergoing different patterns of chemical modification.

One aspect of Rpb1 CTD structure that is important when considering its evolution is the regularity of repeat structures. Tandem amplification of repeat units, which produces very precise repeat arrays, has been found to play a major role in the evolution of the CTD.[13] Once produced, such a repeat sequence will degenerate into a less precise array, unless there is strong selection to maintain it. Thus, one could hypothesize that the presence of a precise repeat array is an indication of a history of strong purifying selection (that is, selection against new non-silent mutations) maintained by requirements for such a precisely repetitive structure. These requirements may relate to the ability to undergo complex patterns of chemical modification, or to accommodate the binding of a diverse suite of CTD-binding proteins. Conversely, a less precise repeat structure may have been the result of a history of weaker purifying selection to maintain the original precise array, leading to loss of precise repeat structure. Indeed, CTDs with less precise repeat structure have previously been considered to have degenerated from a more regular structure.[10] Such weaker purifying selection could be the result of a CTD that is involved in fewer aspects of gene expression, and thus not needing to undergo such complex patterns of chemical modification, or needing to accommodate the binding of a smaller number of proteins. Thus, a hypothesis for one aspect of differences in CTD structure between eukaryotes is that a very precise repeat structure is an indication of a recent history of strong purifying selection on the CTD, while less precise repeat structure is an indication of relatively weaker purifying selection.

One approach to testing this hypothesis is to investigate the recent molecular evolution of the Rpb1 gene in different eukaryotes in order to estimate the parameter ω, also called dN/dS, for a variety of CTD structures. This parameter is the number of non-synonymous changes per non-synonymous site (dN) divided by the number of synonymous changes per synonymous site (dS) in a coding sequence, and it can be informative about the type of selection acting on the sequence. Under purifying selection, non-synonymous differences are largely selected against, while synonymous changes are largely neutral. This makes ω less than one, and the stronger the selection against new non-synonymous variants, the smaller ω will be. With a goal of comparing purifying selection across a variety of CTD repeat structures, one approach would be to directly compare estimates of ω for different CTDs to each other. However, such an approach could suffer from the fact that some underlying patterns of molecular evolution may vary between eukaryote groups. For example, in some groups synonymous changes may be less neutral than in others, lowering dS and thus raising ω even with equal non-synonymous rates of change. An alternative approach would be, for each of several species groups, to compare ω for the CTD to ω for some other coding sequence that would be expected to be under fairly consistent purifying selection across different eukaryotes, as would be expected of the region of Rpb1 that is conserved across all RNA polymerases (which we will call the core of Rpb1). The core of Rpb1 would be evolving under similar mutation pressure, with similar patterns of neutral evolution, and in a similar recombination environment as the CTD. Thus, if ω_{CTD} is higher relative to ω_{core} for species with less precise repeat structure than is ω_{CTD} relative to ω_{core} for species with very precise repeat structure, then one could conclude that purifying selection acting on less repetitive CTDs is weaker than that acting on more precisely repetitive CTDs. Thus, one test of the hypothesis that repeat structure reflects purifying selection is to compare ω_{core} to ω_{CTD} for a number of taxa with different CTD structures.

Another aspect of the CTD that may reflect purifying selection is the frequency of the residues that undergo chemical modification (particularly serines and prolines). If CTD structure commonly evolves by tandem amplification of repeat units rich in these residues, then purifying selection will serve to maintain a high frequency of these residues. Weaker selection might allow a reduction in serine and proline frequencies as the repeat structure decays. However, it is also possible that selection could maintain high serine and proline frequencies but not a precise repeat structure. Thus, two different patterns of selection would result in two different patterns in the CTD: first, a correlation between repeat structure and serine and proline frequencies, and a lack of correlation between the two.

To answer these questions relating to repeat structure and purifying selection, we surveyed CTD structure from 116 Rpb1 sequences, of which 43 have not previously been studied, from diverse eukaryote species. To facilitate comparison between species where repeat structure is absent with those where it is present, the entire C-terminal end of Rpb1 following the last conserved domain (which we will call the C-Terminal Extension, or CTE) was characterized, rather than just repetitive regions. We tested the prediction that serine and proline frequencies will be correlated with the regularity of repeat structure within different

CTEs. When Rpb1 sequences from closely related species were available, a maximum-likelihood model comparison was implemented to determine if ω for the CTE was significantly different than ω for the conserved core of the subunit. These results were then compared across taxa in order to test the hypothesis that repeat structure reflects levels of purifying selection on the CTE. Finally, the determination of CTE structure for a number of new species provided an opportunity to test the consistency of the patterns of phylogenetic distribution of different CTE structures that have previously been described.

Results

Complete Rpb1 sequences were collected for 114 species, including 43 sequences whose CTEs to our knowledge have not previously been characterized in publication, coming from representatives of the Excavata, Stramenopiles, Alveolata, Plantae, Fungi, Animalia, and the nucleomorphs of Rhizaria and Cryptophyta (**Tables S1** and **S2**). It was found that two species, *Physcomitrella patens* and *Acyrthosiphon pisum*, have two distinct Rpb1 genes. A phylogenetic analysis supported all 116 sequences being Rpb1 orthologs, with their presence in a monophyletic clade with a bootstrap support value of 84 (**Fig. S1**).

CTE lengths, along with the location of repeat structure (if present according to our criteria), are shown for bikont species in **Figure 1**, and unikont species in **Figure 2**. The frequencies of serine and proline residues in the CTE, along with characteristics of the repeat array if present, are shown for bikont species in **Table S3**, and unikont species in **Table S4**. As indicated in these tables, when CTE structure was very similar for members of a taxonomic group, one representative of that group was chosen for display, in order to avoid bias introduced by the inclusion of multiple similar sequences from closely related species.

Many of the 43 new sequences included here are consistent with the taxonomic patterns of CTE structure that have been described previously. For example, in the stramenopiles the CTE of *Phytophthora infestans* has a similar length, repeat structure, and serine and proline frequencies as the previously described CTE of *Phaeodactylum tricornutum*. However, there are a number of new sequences that demonstrate exceptions to these patterns. For example, in the same group, the CTE of *Thalassiosira pseudonana* is substantially shorter, with many fewer repeat units, and lower serine and proline frequencies than the other stramenopiles (**Fig. 1**, **Table S3**).

There are many other examples of new exceptions to previously described patterns as well. The CTE of *Naegleria gruberi* has higher serine and proline frequencies than those of previously considered excavate species, although like the others it lacks repeat structure (**Table S3**). The repeat array of *Toxoplasma gondii* is shorter and more variable than those of other apicomplexans. *Perkinsus marinus*, a representative of a new group in the Alveolata, the Perkinsozoa, lacks repeat structure, similar to previously described members of the Ciliophora, but has much lower frequencies of serine and proline residues. In the plants, three monocot species (*Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor*) were found to have somewhat higher repeat variability (RV, defined as the number of different repeat sequences present divided by the total number of repeat units) than previously described dicot species (*Arabidopsis thaliana*, *Ricinus communis*, and *Vitis vinifera*).

The Rpb1 encoded in the nucleomorph of the cryptophyte species *Cryptomonas paramecium* has virtually no CTE, consistent with observations for two other cryptophyte species (**Fig. 1**). The nucleomorph of *Bigelowiella natans*, a member of the Rhizaria, has a CTE that is longer than that of the cryptophyte species (70 residues), and has a number of serine and proline residues, but is still much shorter than any other eukaryote CTE (**Fig. 1**). Two new fungal species, one a member of the Dikarya (*Phaeospheria nodorum*) and one a microsporidian (*Enterocytozoon bieneusi*) had CTEs with no repeat structure and lower serine and proline frequencies than other members of their groups (**Table S4**), although they were not significantly shorter (**Fig. 2**). One new animal species, *Schistosoma mansoni*, was also found to have a CTE with no repeat structure and somewhat lower serine and proline frequencies in the CTE, and the placozoan *Trichoplax adhaerens* had a shorter CTE than other animals although with a much lower RV.

The relationships between RV and serine and proline frequencies within the CTE for stramenopiles, Alveolata, Amoebozoa, Plantae, Fungi, and Animalia are shown in **Figure 3**. In Fungi there is a significant negative relationship between RV and serine frequency, and in Alveolata, Fungi, and Animalia there is a significant negative relationship between RV and proline frequency, as indicated by the Spearman rank order correlation test (**Table 1**). This non-parametric test is not sensitive to outliers, so while it did not find a significant negative relationship between RV and serine frequency for the Alveolata and Animalia, outliers in these two groups follow this pattern as well. Specifically, in the Alveolata, serine frequencies in the non-repetitive CTEs of ciliophorans are slightly lower than in the repetitive CTEs of the apicomplexans, but it is far lower in the non-repetitive CTE of the perkinsozoan *P. marinus* (**Table S4**). In fact, the *P. marinus* CTE has by far the lowest serine frequency of any eukaryote Rpb1 so far found (with the exception of the cryptophyte nucleomorph Rpb1s). In the Animalia, serine frequency in the non-repetitive CTE of *Schistosoma mansoni* is substantially lower than in the repetitive CTEs of all other animal species (**Table S4**). Additionally, while it was not possible to run the statistical tests for the representatives of the Amoebozoa or stramenopiles (because of small sample numbers), in both cases the species available fit this pattern as well, with species with high RV or no repeat structure having substantially lower serine and proline frequencies than those with more regular repeat arrays (**Fig. 3**).

Maximum likelihood models of CTE evolution. Of the ten groups of closely related species used for Maximum Likelihood model testing of CTE evolution (**Table 2**), one group, Leishmania species (in the Excavata), has a CTE with no repeat structure. For this group, a likelihood ratio test found that Model E (in which κ and ω are allowed to differ between the CTE and the conserved core) was very significantly better than Model C (in which the two parameters both had to be the same in the two regions),

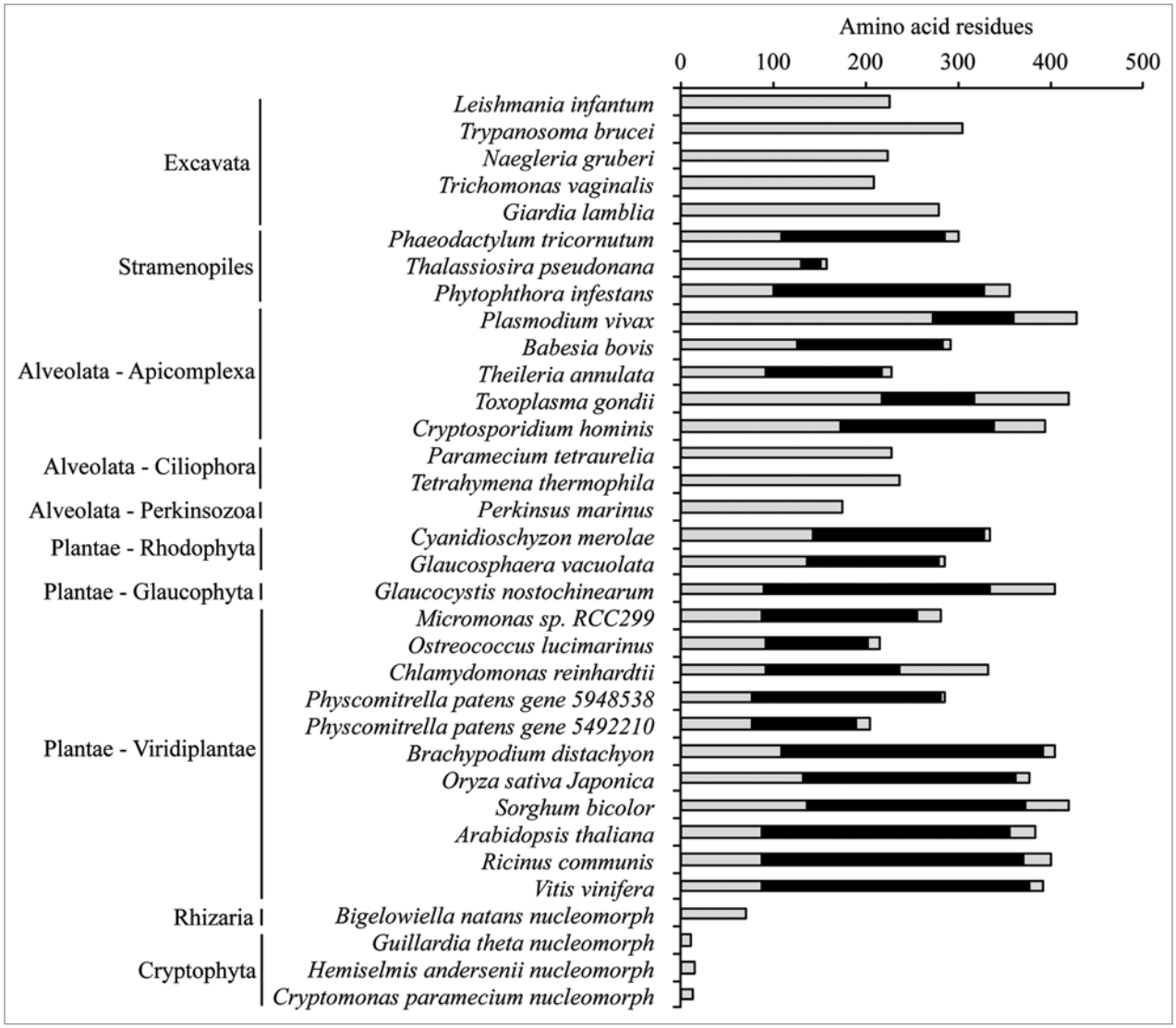

Figure 1. Length of *Rpb1* C-terminal extensions (defined in text) from bikont species. Black regions show the location of repetitive sequences, as described in **Table S3**.

with a p value less than 0.00005 (**Table S4**). From Model E, the maximum likelihood estimate of ω_{CTE} was substantially larger by nearly an order of magnitude than ω_{core}.

The other nine species groups, including apicomplexan, plant, fungi, and animal species, are all characterized by the presence of CTE repeat structure. For four of these species groups, Model E was not significantly more likely than Model C (**Table 2**). For four other species, the p value for the likelihood ratio test was less than 0.05 but greater than 0.005, suggesting that Model E was moderately better than Model C. For one of these groups, Theileria species, the improvement is likely to be more due to differences in κ between the two regions than in ω (**Table 2**). However, for the other three (*Aspergillus* species, *Drosophila* species, and mammals), Model E estimates of ω_{CTE} were modestly larger (less than an order of magnitude) than ω_{core}. For the remaining group, *Candida* species, the likelihood ratio test was significant with a p value less than 0.00005, and the Model E estimate of ω_{CTE} was larger by nearly an order of magnitude than ω_{core}.

Discussion

We used two approaches to test hypotheses relating to repeat structure and purifying selection acting on the CTE: testing for correlations between a measure of repeat structure and serine and proline frequencies within major eukaryote groups, and a comparison of levels of purifying selection acting on the CTE between different groups. We found that the CTE of *Leishmania* species, which lacks repeat structure, is evolving under significantly weaker purifying selection than the rest of Rpb1, as indicated by a substantial difference in ω between the two regions of the subunit. In contrast, the common pattern found for species groups with CTEs containing repeat structure was greater similarity in ω between these two regions of Rpb1, indicating a

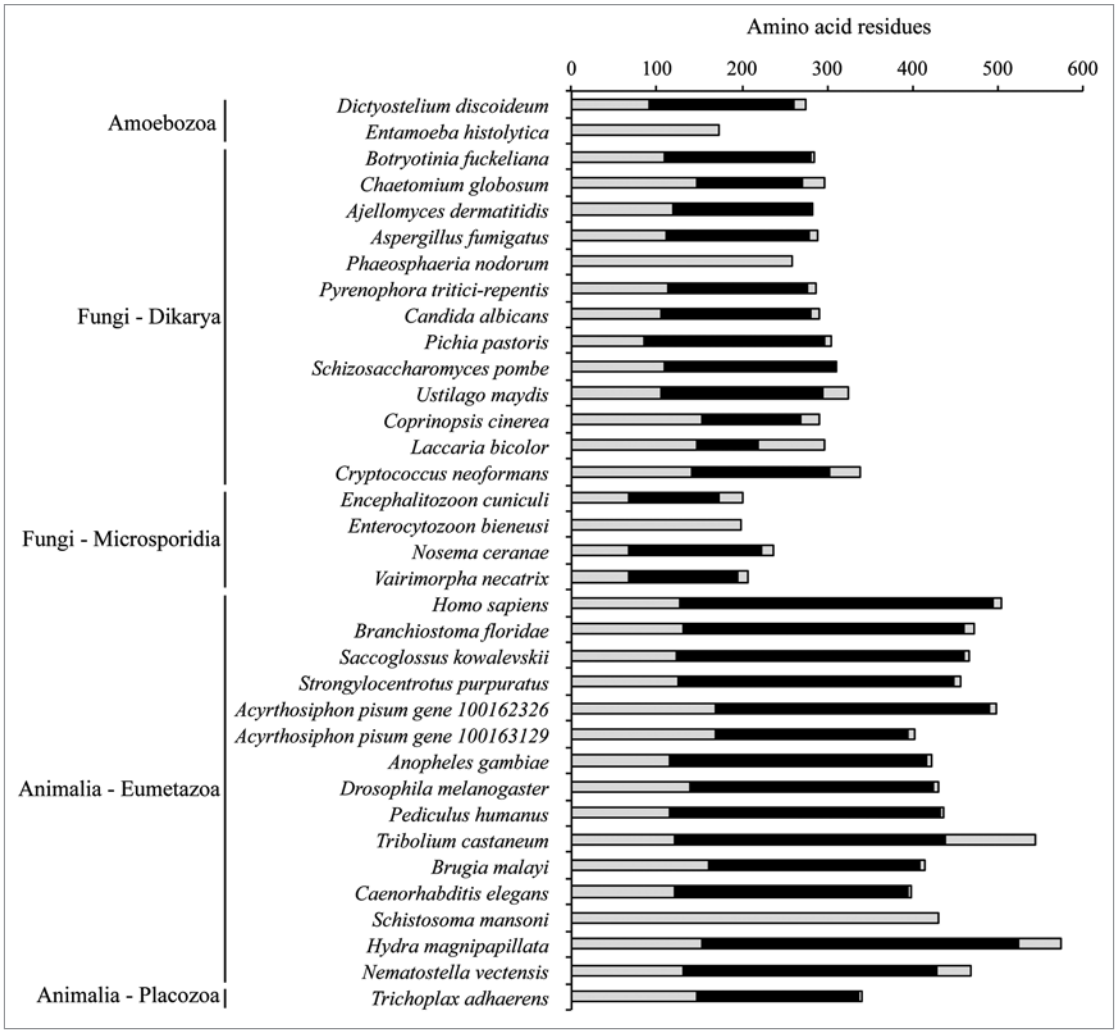

Figure 2. Length of *Rpb1* C-terminal extensions (defined in text) from unikont species. Black regions show the location of repetitive sequences, as described in **Table S4**.

level of negative selection on the CTE fairly similar to that acting on the conserved core of the subunit. The rationale for our approach of comparing ω_{CTE} to ω_{core} for each species group is that some general patterns of molecular evolution may vary between eukaryote groups, making direct comparisons of ω from one taxa to another of questionable value. These background patterns need not be the same in all groups for our comparative approach to be informative. Nevertheless, the estimates of ω_{core} for the different species groups are fairly consistent, with the exception of the estimate for mammals, which is much lower than the others (**Table 2**). Estimates of ω_{CTE} are somewhat less consistent, with the estimates for *Leishmania* species and *Candida* species substantially larger than the others; this matches the results from the comparison of ω_{CTE} to ω_{core}.

While these results generally support our hypothesis, the conclusions would be significantly strengthened with the addition of more species groups. Similar analyses for more species groups lacking CTE repeat structure would be especially valuable, as the conclusion of less negative selection in CTEs lacking repeat structure is currently based only on one group of three *Leishmania* species. A maximum likelihood analysis was performed for a group of two Trypanosoma species, which are also excavates and thus lack repeat structure, which seems to produce the same result as for *Leishmania*: much higher ω_{CTE} than ω_{core}. However, the dS between these two sequences was 3.03, which is much higher than our cutoff of one and which raises the possibility that saturation of dS may make estimates unreliable. For this reason, the results are not included in **Table 2**, nevertheless they may offer some very limited support that the *Leishmania* results are not anomalous. Further analyses would also be stronger if they included species groups lacking repeat structure outside of the Excavata as well. Additional analyses of species groups that do contain repeat structure would also be valuable, to determine how common it is for repetitive CTEs to evolve under relatively weaker purifying selection, as we found with Candida species (**Table S4**). It is also important to note that while estimates of ω might be relatively higher in some cases, they are still very low in absolute terms, indicating the CTE is still evolving under strong purifying selection. This is an indication that the CTE is still functionally important, and indeed it has been found that

Table 1. Spearman rank order correlation coefficiants between CTE Repeat Variability (defined in text) and CTE serine and proline frequencies, for each of four eukaryote groups

Taxon	Sample size	Degrees of freedom	Spearman's Correlation Coefficient	
			Serine frequency	Proline frequency
Alveolata	n = 8	df = 6	ρ = -0.670	ρ = -0.791 *
Plantae	n = 14	df = 12	ρ = 0.127	ρ = -0.222
Fungi	n = 17	df = 15	ρ = -0.805 **	ρ = -0.567 *
Animalia	n = 16	df = 14	ρ = -0.049	ρ = -0.559 *

Asterisks indicate statistically significant correlations. * $p < 0.05$; ** $p < 0.005$.

the non-repetitive CTE of *Trypanosoma brucei* is critical for cell viability.[14]

We found that in the *Drosophila* species used in our ω comparisons, ω_{CTE} was only modestly larger than ω_{core}. This was somewhat unexpected, because *Drosophila* species have relatively disordered repeat structures in the CTE (RV = 0.91, contrasted with R.V. = 0.37 for humans and other vertebrates), so according to our hypothesis, their CTE was predicted to be subjected to substantially weaker selection. However, there is reason to suspect that the ω results for *Drosophila* are somewhat misleading. Llopart and Aguadé[15] investigated synonymous substitution rates for the Rpb1 gene in several *Drosophila* lineages. Their results suggest that in one of these lineages, dS for the CTD was higher than in the rest of the gene. If this was the case for our group of Drosophila species, then ω would be lowered, so purifying selection acting on the CTE might actually be stronger than the ω comparison suggests.

A comparison of serine and proline frequencies across CTEs with a variety of repeat structures reveals that less precise or absent repeat structure is associated with lower serine and proline frequencies in some taxa but not in others (**Fig. 2**). Three distinct patterns can be seen. First, a consistent decline in residue frequency across a range of increasing RV values is seen very strongly for serine in CTEs of the Fungi, and to a lesser degree for proline frequency in the Alveolata, Fungi, and Animalia. A second pattern is consistently high serine and proline frequencies for most CTEs within a group, but substantially lower for at least some CTEs that lack repeat structure, which is seen for serine frequency in Animalia and Alveolata. The small number of stramenopile and amoebozoan species examined here mean that it is not possible to say which of these two patterns these taxa fit, although in both cases a species with reduced or absent repeat structure had a substantially lower serine and proline frequency than one or two with strong repeat structure.

A third pattern, no correlation between RV and serine and proline frequencies, was found in the plant species included here. However, it must be noted that partial Rpb1 sequences for two rhodophyte species are known that lack CTE repeat structure (*Porphyra yezoensis*, AAC17924; and *Bonnemaisonia hamifera*, AAC18416;[10]) (these sequences were not included in this analysis because one of our criteria for inclusion was a complete Rpb1 sequence). In the incomplete CTE sequences that are available for these species, serine and proline frequencies are lower (data not shown) than in the complete CTE sequences of other plants, however it is possible that the missing sequences could raise the overall frequencies. Thus, we cannot say how the non-repetitive CTEs of these two rhodophyte Rpb1 sequences compare with the repetitive CTEs of other plants.

The description of CTEs from a number of new species has also revealed new patterns of variability of CTE structure between species within major eukaryote groups, further emphasizing the evolutionary lability of the CTE in many taxa. Perhaps most dramatic are the instances of species having greatly reduced or absent repeat structure in groups where this has not previously been known: stramenopiles, animals, and both dikaryan and microsporidian fungi. Added to the previously known examples of this from the Amoebozoa and Plantae, it can be seen that despite its importance, CTE repeat structure has been lost numerous times, and in diverse taxa. These species would be very interesting subjects for future research into what roles the CTE has lost and what are retained in these cases. While a reduction of serine and proline frequencies is associated with the loss of repeat structure in these species (with the possible exception of the two rhodophyte sequences as discussed above), these CTEs remain fairly serine and proline rich, suggesting patterns of chemical modification of these residues remains an important aspect of their function. The same may not be true for the CTE of the perkinsozoan alveolate *P. marinus*, which is not serine rich, containing only six serines out of a total of 175 residues, the lowest serine frequency of any eukaryote CTE of substantial length. This example raises the possibility of a CTE that undergoes much less chemical modification during RNAP II activity, raising the question of how its role has changed.

The most dramatic losses of CTE structure have occurred in the Rpb1 genes encoded in the plastid nucleomorphs of the Rhizaria and Cryptophyta. All three cryptophyte nucleomorph Rpb1s are essentially completely lacking any CTE (according to our criteria, they were all 12 to 15 residues long, which is presumably not long enough to form a functional structure). Whereas the cryptophyte plastid has been derived from an ancestral red alga, the plastid of the Rhizaria has been independently derived from a green alga.[16] Thus, there has been convergent evolutionary loss of CTE structure in the nucleomorph of the Rhizaria, although not as complete: a 70 residue CTE remains. It is intriguing that this CTE is still fairly serine rich, raising the possibility that it retains some function involving phosphorylation patterns. It should be noted that each of these cases of "atypical" CTEs could potentially be the result of incorrect gene annotations. However, several facts suggest that at least some of these are correctly annotated. First, many are still fairly serine and proline rich and fairly similar in size to more typical CTEs. Second, in at least some cases (such as the Rpb1 CTEs of the cryptophyte nucleomorph), several independent annotations of closely related species have produced the same result. Third, the more examples of atypical CTEs that are found, the clearer it is that these are biologically plausible.

The presence of two distinct Rpb1 genes in two species (the aphid *Acyrthosiphon pisum* and the moss *Physcomitrella patens*) is

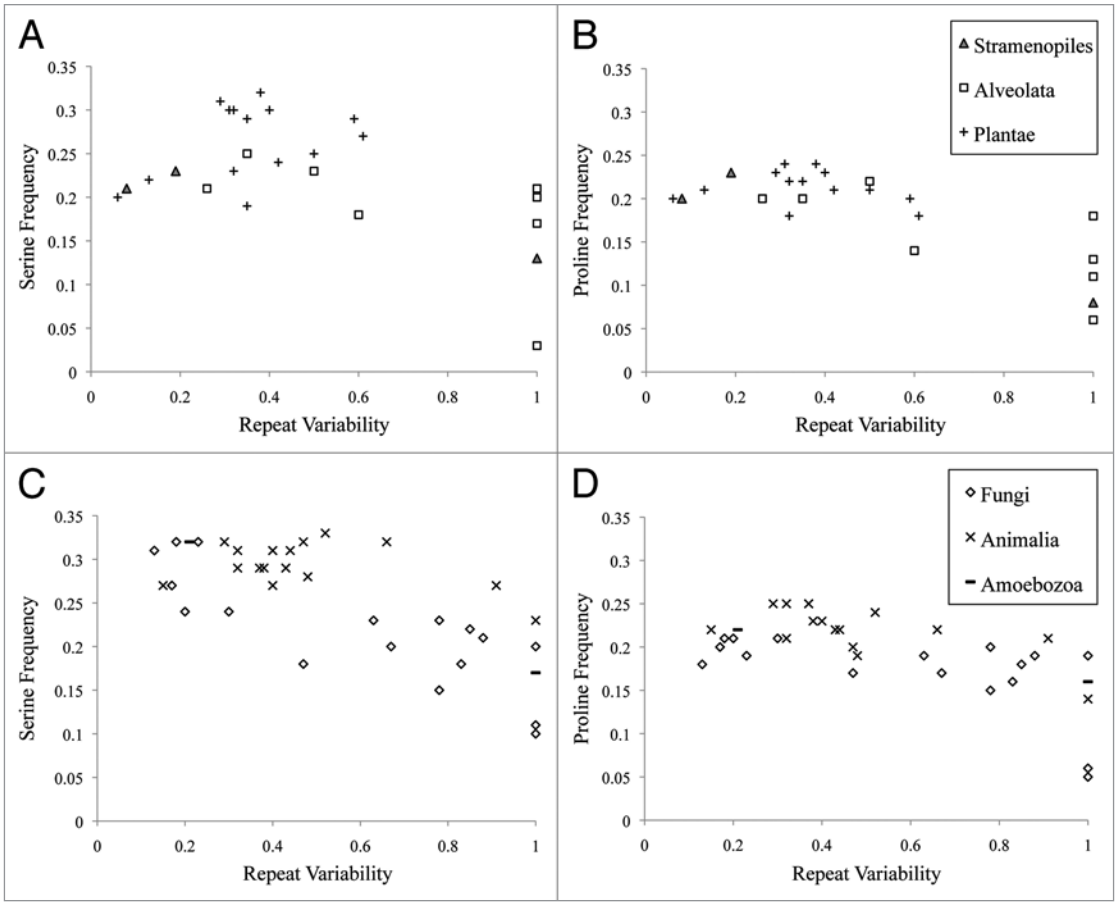

Figure 3. Relationship between CTE Repeat Variability (defined in text) and frequency of serine and proline residues, as shown in **Tables S3** and **S4**. (**A**) Serine frequency in bikont species; (**B**) proline frequency in bikonts; (**C**) serine frequency in unikont species; (**D**) proline frequency in unikonts. For the purposes of this figure, CTEs with no repeat structure are considered to be equivalent to RV = 1. CTEs of the Excavata, of the nucleomorphs of Cryptophyta and Rhizaria are not plotted here because all lack repeat structure.

interesting as well. In both species, the CTE and repeat region of one gene is substantially longer than the other, suggesting a shortening of the CTE in one of the genes following duplication. The two *P. patens* genes encode proteins that are 95% identical after the removal of alignment gaps (including the longer part of the CTE), and the two *A. pisum* genes encode proteins that are 97% identical after removal of gaps. In both species, the two genes cluster together in the phylogenetic analysis with high bootstrap support (**Fig. S1**). Thus, in both cases, while the duplications have occurred relatively recently, there has been enough time for the two paralogs to diverge to a degree at the amino acid level. This suggests that selection may have played a role in maintaining two functional Rpb1 genes in both of these species over this time. If so, it raises the intriguing possibility that the two genes may not be completely functionally redundant, but rather may have started to diverge to produce further functional specialization between the two paralogs, possibly related to the CTE length differences between paralogs.

In summary, our results provide support for two hypotheses relating to the presence of repeat structure, and the precision and regularity of that repeat structure. First, the absence of repeat structure is associated with weaker purifying selection acting on the CTE relative to typical levels of selection acting on CTEs with repeat structure. Second, the absence or reduction of repeat structure is often associated with reduced serine and proline frequencies within the CTE. These results suggest a further hypothesis: that the presence and precision of CTE repeat structure is positively correlated with greater requirements for CTE functional complexity. These requirements could take two forms. Some CTEs could be constrained by requirements to interact with a larger number of different CTD-binding proteins, and a more precise repeat structure could maximize this ability. Alternatively, some CTEs may be constrained by requirements to undergo more complex patterns of chemical modification, which again may benefit from a more precise repeat structure. The results presented here also point the way to future comparative research that has the potential to answer questions about the importance of CTE repeat structure. For example, comparative genome surveys could determine if *Schistosoma mansoni* lacks genes homologous to any of the various CTD-binding proteins found in other metazoans, and if so, what functions these genes relate to. Similarly, functional studies on the chemical modifications that the *S. mansoni* CTE undergoes during transcription could help determine what modification patterns are completely

Table 2. Likelihood values for fixed-site maximum likelihood models of the evolution of *Rpb1* in various eukaryote species groups (see text for details)

Taxon	Species	Likelihood values		Likelihood ratio Test (df = 2)	Model E estimated parameters	
		Model C	Model E		Rpb1 core	CTE
Excavata	*Leishmania infantum*, *Leishmania major*, *Leishmania braziliensis*	-8371.92 (24)	-8340.58 (26)	$\chi2 = 62.69$ ****	$\omega = 0.015 \pm 0.002$ $\kappa = 7.814 \pm 0.854$	$\omega = 0.135 \pm 0.029$ $\kappa = 6.493 \pm 1.339$
Alveolata - Apicomplexa	*Plasmodium vivax*, *Plasmodium knowlesi*	-12245.16 (23)	-12243.97 (25)	$\chi2 = 2.38$	$\omega = 0.042 \pm 0.004$ $\kappa = 2.734 \pm 0.257$	$\omega = 0.059 \pm 0.012$ $\kappa = 3.031 \pm 0.586$
	Theileria annulata, *Theileria parva*	-8553.35 (23)	-8549.77 (25)	$\chi2 = 7.16$ *	$\omega = 0.008 \pm 0.002$ $\kappa = 1.746 \pm 0.199$	$\omega = 0.010 \pm 0.004$ $\kappa = 4.463 \pm 1.378$
	Cryptosporidium hominis, *Cryptosporidium parvum*	-7964.68 (23)	-7964.37 (25)	$\chi2 = 0.63$	$\omega = 0.034 \pm 0.012$ $\kappa = 6.599 \pm 2.006$	$\omega = 0.021 \pm 0.012$ $\kappa = 5.609 \pm 1.782$
Plantae - Viridiplantae	*Arabidopsis thaliana*, *Arabidopsis lyrata*	-7986.95 (23)	-7985.74 (25)	$\chi2 = 2.44$	$\omega = 0.034 \pm 0.010$ $\kappa = 2.459 \pm 0.495$	$\omega = 0.070 \pm 0.034$ $\kappa = 1.719 \pm 0.591$
Fungi - Dikarya	*Ajellomyces dermatitidis*, *Ajellomyces capsulatus*	-8653.05 (23)	-8652.04 (25)	$\chi2 = 2.02$	$\omega = 0.019 \pm 0.004$ $\kappa = 2.631 \pm 0.316$	$\omega = 0.023 \pm 0.009$ $\kappa = 3.930 \pm 1.026$
	Candida albicans, *Candida dubliniensis*	-8529.69 (23)	-8502.82 (25)	$\chi2 = 53.76$ ****	$\omega = 0.014 \pm 0.003$ $\kappa = 2.321 \pm 0.303$	$\omega = 0.122 \pm 0.025$ $\kappa = 1.130 \pm 0.252$
	Aspergillus fumigatus, *Aspergillus clavatus*	-8995.70 (23)	-8991.54 (25)	$\chi2 = 8.34$ *	$\omega = 0.013 \pm 0.002$ $\kappa = 2.633 \pm 0.314$	$\omega = 0.036 \pm 0.010$ $\kappa = 2.314 \pm 0.577$
Animalia - Eumetazoa	*Homo sapiens*, *Rattus norvegicus*, *Bos taurus*, *Ailuropoda melanoleuca*, *Loxodonta africana*	-13228.73 (28)	-13224.60 (30)	$\chi2 = 8.26$ *	$\omega = 0.0007 \pm 0.0003$ $\kappa = 4.113 \pm 0.294$	$\omega = 0.0044 \pm 0.0014$ $\kappa = 4.267 \pm 0.421$
	Drosophila melanogaster, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*	-9906.71 (26)	-9901.68 (28)	$\chi2 = 10.07$ *	$\omega = 0.014 \pm 0.003$ $\kappa = 2.805 \pm 0.328$	$\omega = 0.050 \pm 0.013$ $\kappa = 2.680 \pm 0.503$

The number of parameters for each model is shown in parentheses. Chi square values are shown for likelihood ratio tests comparing models, with asterisks showing statistically significant differences between models. Model E estimated parameters are estimates of ω and κ for the core of the subunit, and for the CTE. * $p < 0.05$; ****$p < 0.00005$.

indispensible for cell functioning, and what can be evolutionarily lost. These questions will be important for gaining a comprehensive view of the evolution of this important aspect of transcription.

Materials and Methods

Sequence collection. Rpb1 protein sequences were identified by BLASTP searches against the NCBI reference protein data set, using the human Rpb1 sequence (NCBI accession number NP_000928.1) as a query, and an initially defined e-value cutoff of 1e-30. (In actuality, all Rpb1 sequences were identified with an e-value reported as 0.0, with one exception: XP_001704218.1, the Rpb1 from *Giardia lamblia*, which was identified with an e-value of 2e-67). For each, the associated mRNA sequence and Gene ID numbers were also collected, and sequences were excluded from analysis if the cds was reported as incomplete in GenBank. Complete cds sequences are available from the corresponding author on request.

Verification of orthologs. To verify that all collected sequences were Rpb1 orthologs, a phylogenetic analysis was performed with paralogous sequences. From each of four species (*Homo sapiens*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Plasmodium falciparum*), we collected the protein sequence for Rpa1 (the large subunit of RNA Polymerase I; NP_056240.2, NP_014986.1, NP_191325.1, XP_001351652.1, respectively), and Rpc1 (the large subunit of RNA Polymerase III; NP_008986.2, NP_014759.1, NP_001190573.1, XP_001350009.1, respectively). These were aligned with all collected Rpb1 protein sequences using MUSCLE v3.7[17] in full mode without finding diagonals, with a maximum of 16 iterations, as implemented on the Phylogeny.fr server.[18] Poorly aligning sites were removed from the alignment using Gblocks v0.91b[19] with default settings, leaving a curated alignment of 520 positions (alignment available from the corresponding author on request).

The best model of protein evolution for the curated alignment was determined using ProtTest v2.4,[20] set to compare all models, with default settings for the slow optimization strategy. The AIC

criterion was used to select the most likely model. A maximum likelihood phylogenetic analysis was performed using PhyML v3.0,[20] with the LG substitution model (based on the ProtTest results), four substitution rate categories, a gamma distribution parameter estimated from the data, an estimated proportion of invariable sites, and estimated equilibrium frequencies. A BioNJ tree was used as the starting tree, and both NNI and SPR were used to search among trees, optimizing both topology and branch length. Branch support was determined using 100 bootstrap replicates. The resulting tree and bootstrap values were visualized using iTOL v2.1.1.[21]

Characterization of CTEs. For each Rpb1 protein sequence, the C-Terminal Extension (CTE) was defined as the entire sequence following the residue aligning with residue 1436 of human Rpb1 (defined by Cramer et al.[2] as the last residue of the domains conserved between all RNA polymerases, and the last residue before the linker domain). Repeat structure within the CTE was determined by looking for heptapeptide repeat units. To be considered a repeat unit, a set of seven residues had to meet two criteria: having a majority of residues matching a consensus sequence found within the CTE; and being part of at least two sequential repeat units. For sequences with three- or nine-peptide repeat units, the same criteria applied. CTE sequences with repeat structure parsed according to these criteria are available as a supplemental data file.

For CTEs containing repeat structure, we determined the length and location of the repeat array (defined as starting with the first repeat unit and ending with the last), the consensus sequence from all repeat units within the repeat array, the number of repeats with that consensus sequence, and the total number of repeat units. Repeat Variability (RV), defined as the number of different repeat unit sequences present divided by the total number of repeat units, was also determined for each CTE with repeat structure. The frequencies of serine and proline residues within the CTE (the number of serine and proline residues divided by the total number of residues) were determined and plotted against RV. To facilitate comparison between CTEs with and without repeat structure, CTEs without repeat structure were plotted at RV = 1 for this figure. For members of each of several eukaryote groups (Alveolata, Plantae, Fungi, and Animalia) Spearman's rank order correlation was used to test for a relationship between RV and both the frequencies of serine and proline residues, using SPSS version 19.

Maximum likelihood models of CTE evolution. Groups of closely related species (congeneric species, plus placental mammals; species listed in **Table 2**) for which complete Rpb1 sequences were available were identified. For each group, the TranslatorX server[22] was used to produce cleaned CDS alignments while ensuring the retention of codon information. (Briefly, full CDS sequences were translated to amino acid sequences, which were then aligned by MUSCLE. This alignment was then cleaned by Gblocks with standard settings, before back-translating the amino acids to their original codons). For all species groups, over 96% of the positions in the original alignment were retained in the cleaned alignment.

Because accurate estimation of ω (dN/dS as described previously) requires that dS has not become saturated, a preliminary analysis was performed to determine dS between each pair of sequences within each group. For the cleaned alignment for each group, codeml with runmode = -2 was run in PAML v4.4.[23] Using these results, sequences were removed as necessary so that within each group, dS was less than one for each pair of sequences. This left the ten groups of species shown in **Table 2** for maximum likelihood analysis.

PAML was used to implement fixed-site partition models in codeml, with the F3x4 codon frequency model, ω the same in all branches and sites within partitions, κ (the transition/transversion rate ratio) and ω to be estimated (with starting values 1.5 and 0.1 respectively), Small_Diff = 3e-7, and method = 0. An unrooted tree based on the known phylogenetic relationships between species within each group was provided for the runmode = 0 option. Sequence and tree input files are available from the corresponding author on request. For each species group, two partition models were implemented: Model C (different codon frequencies but equal κ and ω in the core and CTE), and Model E (different κ, ω, and codon frequencies in the core and CTE).[22] A likelihood ratio test with df = 2 was used to determine if Model E was significantly more likely than Model C for each species group.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We are grateful to Dr. Jason A. Wilder and two anonymous reviewers for helpful comments on an earlier draft of this paper.

Supplemental Material

Supplemental material may be downloaded here: http://www.landesbioscience.com/journals/transcription/article/23305/

References

1. Archambault J, Friesen JD. Genetics of eukaryotic RNA polymerases I, II, and III. Microbiol Rev 1993; 57:703-24; PMID:8246845.
2. Cramer P, Bushnell DA, Kornberg RD. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science 2001; 292:1863-76; PMID:11313498; http://dx.doi.org/10.1126/science.1059493.
3. Allison LA, Moyle M, Shales M, Ingles CJ. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell 1985; 42:599-610; PMID:3896517; http://dx.doi.org/10.1016/0092-8674(85)90117-5.
4. Corden JL, Cadena DL, Ahearn JM Jr., Dahmus ME. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proc Natl Acad Sci U S A 1985; 82:7934-8; PMID:2999785; http://dx.doi.org/10.1073/pnas.82.23.7934.
5. Stiller JW, Hall BD. Evolution of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A 2002; 99:6091-6; PMID:11972039; http://dx.doi.org/10.1073/pnas.082646199.
6. Bartkowiak B, Mackellar AL, Greenleaf AL. Updating the CTD story: From tail to epic. Genet Res Int 2011; 2011:623718; PMID: 22567360; http://dx.doi.org/10.4061/2011/623718.
7. Zhang DW, Rodriguez-Molina JB, Tietjen JR, Nemec CM, Ansari AZ. Emerging views on the CTD code. Genet Res Int 2012; 2012:347214; http://dx.doi.org/10.1155/2012/347214.
8. Buratowski S. Progression through the RNA polymerase II CTD cycle. Mol Cell 2009; 36:541-6; PMID:19941815; http://dx.doi.org/10.1016/j.molcel.2009.10.019.
9. Allison LA, Wong JK, Fitzpatrick VD, Moyle M, Ingles CJ. 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 1988; 8:321-9; PMID:3122024.
10. Liu P, Kenney JM, Stiller JW, Greenleaf AL. Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. Mol Biol Evol 2010; 27:2628-41; PMID:20558594; http://dx.doi.org/10.1093/molbev/msq151.
11. Dacks JB, Marinets A, Ford Doolittle W, Cavalier-Smith T, Logsdon JM Jr. Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. Mol Biol Evol 2002; 19:830-40; PMID:12032239; http://dx.doi.org/10.1093/oxfordjournals.molbev.a004140.
12. Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, et al. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci U S A 2007; 104:19908-13; PMID:18077423; http://dx.doi.org/10.1073/pnas.0707419104.
13. Chapman RD, Heidemann M, Hintermair C, Eick D. Molecular evolution of the RNA polymerase II CTD. Trends Genet 2008; 24:289-96; PMID:18472177; http://dx.doi.org/10.1016/j.tig.2008.03.010.
14. Das A, Bellofatto V. The non-canonical CTD of RNAP-II is essential for productive RNA synthesis in *Trypanosoma brucei*. PLoS One 2009; 4:e6959; PMID:19742309; http://dx.doi.org/10.1371/journal.pone.0006959.
15. Llopart A, Aguadé M. Synonymous rates at the RpII215 gene of Drosophila: variation among species and across the coding region. Genetics 1999; 152:269-80; PMID:10224259.
16. Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci U S A 2006; 103:9566-71; PMID:16760254; http://dx.doi.org/10.1073/pnas.0600707103.
17. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792-7; PMID:15034147; http://dx.doi.org/10.1093/nar/gkh340.
18. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 2008; 36(Web Server issue):W465-9; PMID:18424797; http://dx.doi.org/10.1093/nar/gkn180.
19. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 2000; 17:540-52; PMID:10742046; http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334.
20. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 2005; 21:2104-5; PMID:15647292; http://dx.doi.org/10.1093/bioinformatics/bti263.
21. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 2011; 39(Web Server issue):W475-8; PMID:21470960; http://dx.doi.org/10.1093/nar/gkr201.
22. Yang Z, Swanson WJ. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 2002; 19:49-57; PMID:11752189; http://dx.doi.org/10.1093/oxfordjournals.molbev.a003981.