ABSTRACT

A unique feature of RNA polymerase II (RNA pol II) is its long C-terminal extension, called the carboxy-terminal domain (CTD). The well-studied eukaryotes possess a tandemly repeated 7-amino-acid sequence, called the canonical CTD, which orchestrates various steps in mRNA synthesis. Many eukaryotes possess a CTD devoid of repeats, appropriately called a non-canonical CTD, which performs completely unknown functions. Trypanosoma brucei, the etiologic agent of African Sleeping Sickness, deploys an RNA pol II that contains a non-canonical CTD to accomplish an unusual transcriptional program; all protein-coding genes are transcribed as polygenic precursor mRNAs (pre-mRNAs) that are initiated within a several-kilobase-long region, called the transcription start site (TSS), upstream of the first protein-coding gene in the polygenic array. In this report, we show that the non-canonical CTD of T. brucei is important for normal protein-coding gene expression, likely directing RNA pol II to the TSSs within the genome. Our work reveals the presence of a primordial CTD code within eukarya and indicates that proper recognition of the chromatin landscape is a central function of this RNA pol II-distinguishing domain.

INTRODUCTION

Eukaryotic mRNA synthesis is orchestrated by the RNA polymerase II (RNA pol II) carboxy-terminal domain (CTD) (1–5). There are two types of CTDs: the canonical CTD, consisting of tandem repeats of the heptapeptide, Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 and the non-canonical CTD, devoid of any repeats. All well-studied model organisms contain a CTD of the canonical type within their RNA pol II (6,7). This CTD is known to undergo a multitude of post-translational modifications, which help signal the co-transcriptional activities that occur during the synthesis of monogenic mRNAs (8). In addition, the CTD modification codes communicate with the chromatin landscape during the transcription cycle (9).

While the myriad roles played by the canonical CTD have been illuminated over the past three decades, functions of non-canonical CTDs are largely unexplored (7,10). Trypanosoma brucei, the etiologic agent of African Sleeping Sickness, deploys an RNA pol II containing a non-canonical CTD to accomplish an unusual mode of transcription; all protein-coding genes are transcribed as polygenic precursor mRNAs (pre-mRNAs) that are initiated within several-kilobase-long regions, called transcription start sites (TSSs), upstream of the first protein-coding gene in arrays of genes organized head-to-tail within 11 large (1–5 Mb) chromosomes (11). This assertion is based on recent transcriptome analysis and 5’ tri-phosphorylated RNA mapping data (12–14). TSSs appear to be devoid of conserved sequences and do not appear to be bound by normal transcription factors. These findings are consistent with our inability to detect conventional pre-mRNA gene promoter sequences. Our minimal understanding of trypanosome RNA pol II promoters derives from studies defining the spliced-leader (SL) RNA gene promoter, which appears to be an atypical promoter element (15).

In the case of the ∼167 head-to-head polygenic arrays present in the T. brucei genome, TSS regions are marked by histone variants H2Az and H2BV, as well as by a modified...
H4, H4K10ac (reviewed in (16)). It is possible that RNA pol II uses these histone marks to recognize TSS regions within the genome and ensure proper polygenic transcription. The resulting long, polycistronic pre-mRNAs are rapidly processed to produce stable mRNAs by 5′ trans-splicing of a hypermethylated cap derived from a separately transcribed universal SL RNA and 3′ polyadenylation (17).

The T. brucei RNA pol II non-canonical CTD is 284 amino-acids-long and contains an indispensable 90-residue central region (Figure 1A) (18). Like the canonical CTD, the non-canonical CTD is serine-rich (17%, compared to 7% for the entire polypeptide) and is post-translationally phosphorylated (18,19). However, the significance of these modifications in mRNA production is unknown. As protein-coding mRNAs are highly expressed by bacteriophage RNA polymerases driving their cognate promoters in transgenic trypanosomes, the non-canonical CTD is not crucial for pre-mRNA maturation. Thus, the promoters in transgenic trypanosomes, the non-canonical CTD, are serine-rich (17%, compared to 7% for the entire polypeptide) and is post-translationally phosphorylated (18,19). However, the significance of these modifications in mRNA production is unknown. As protein-coding mRNAs are highly expressed by bacteriophage RNA polymerases driving their cognate promoters in transgenic trypanosomes, the non-canonical CTD is not crucial for pre-mRNA maturation. Thus, the essentiality of the non-canonical CTD likely lies in its role in TSS recognition.

In this report, we used mRNA-seq and RNA pol II chro-matin immunoprecipitation to observe the global effects of mutations to the non-canonical CTD on TSS recognition in vivo. Clusters of serine-to-alanine mutations in the CTD adversely affect cell growth and decreased mRNA production from TSS-proximal genes. Other RNA pol II-dependent transcripts were mostly unaffected, consistent with previously observed mRNA production initiation from internal sites within transcription units (12). Interestingly, RNA pol II occupancy in the CTD-mutant M2\textsuperscript{Ty1} and M9\textsuperscript{Ty1} parasites was shifted downstream from TSS regions, which was not the case for the WT\textsuperscript{Ty1}RNA pol II occupancy. These data indicate that the non-canonical CTD of T. brucei RNA pol II is essential for the proper positioning of polymerase to ensure normal mRNA production.

**MATERIALS AND METHODS**

**Plasmid constructs, mutagenesis and T. brucei strain constructs**

A plasmid containing a tetracycline-inducible RNAi gene fragment that targets the endogenous RPB1 3′UTR was described previously (18). pAD76, a pAD74 derivative, was constructed by insertion of a DNA fragment (primer pair AD271/AD368) encoding the A-H domains of RPB1 (amino acid 1–1481) between the unique XbaI and BstZ1I sites, followed by insertion of a DNA fragment (primer pair AD441/AD442), encoding the entire CTD (amino acids 1482–1766), between the unique AvrII and NsiI sites, of pAD74 (18). pAD76 derivatives that expressed mutant RPB1 proteins were constructed by swapping the wild-type CTD with an altered CTD between the AvrII and NsiI sites. CTD alterations were generated by site-directed mutagenesis of the wild-type CTD within pCR2.1-TOPO\textsuperscript{TM} using a Change-IT\textsuperscript{TM} Multiple Mutagenesis Site Directed Mutagenesis kit (Stratagene) before they were introduced into pAD76. pAD76 and its derivatives were linearized using NotI to facilitate homologous recombination of the construct into an rRNA loci. All exogenous tagged-RPB1 proteins are expressed from a tetracycline-inducible ribosomal RNA gene promoter, which is transcribed by RNA pol I. A blasticidin-resistant gene within the constructs allowed for the selection of stable transformants. Molecular cloning otherwise was done using standard procedures and *Escherichia coli* strains.

**Trypanosome culture and growth analysis**

The T. brucei Lister 427 procyclic cell line 29-13 and all stable transgenic cell lines were grown in SDM-79 media supplemented with 10% fetal bovine serum (tetracycline-free) at 27°C in a humidified chamber containing 5% CO\textsubscript{2} (30). AD-101 cells were produced by introducing a tetracycline-inducible construct containing the RNAi sequence that targets the 3′UTR of RPB1. All other cell lines were derived from AD-101 by introducing a tetracycline-inducible exogenous RPB1 gene (that produces the tagged proteins) with an RNAi-resistant 3′UTR. Clonal selections of all cell lines were done by serial dilution, using 2.5 μg/ml Phleomycin and 10 μg/ml Blasticidin. All cell lines were verified by genomic PCR and DNA sequence analyses.

**Antibodies and western analysis**

Anti-CTD antibody, produced in rabbit against a recombinant CTD peptide, detects endogenous and exogenous RPB1 (tagged proteins). Anti-Ty1 antibody (31), a mouse monoclonal antibody that detects the tagged proteins, was obtained from the Antibody and Bioresource Core Facility of The Rockefeller University and Memorial Sloan-Kettering Cancer Center. Anti-RPB4 antibody was described previously (18). Expression of RPB1 was determined by western blot analysis of whole cell extracts prepared from parasites, before and after one, two and three days of tetracycline induction. Protein extracts from 2 × 10\textsuperscript{6} cells, prepared by boiling in SDS-sample buffer, were separated by 8% SDS-PAGE and analyzed by immunoblotting using the ECL\textsuperscript{TM} kit from Pierce. Inclusion of the tagged proteins into the 12-subunit-RNA pol II enzyme was verified by western blot analysis of antibody-captured RNA pol II with anti-RPB4 antibody. Anti-EF2 antibody was from Santa Cruz (sc-13004). Anti-2,2,7-trimethylguanosine mouse mAb (K121) agarose conjugate was from Calbiochem (NA02A).

**RNA isolation and semi-quantitative RT-PCR analysis**

Total cellular RNA (from 10\textsuperscript{7} cells) was isolated using TRIzol\textsuperscript{TM} reagent (Invitrogen) and treated with RQ1 RNase-free DNase (Promega) to remove DNA contamination. Transcript levels of genes were compared by reverse transcription (RT) followed by semi-quantitative PCR analysis. 7SL RNA, a small nuclear RNA transcribed by RNA pol III, was used as an internal control. Briefly, cDNAs were synthesized from 1–2 μg of total RNA in a 20 μl RT reaction using 5 μM random hexamers and Superscript III\textsuperscript{TM} (Invitrogen) enzyme. Following heat inactivation (75°C, 15 min) and RNA removal using RNase H, 1 μl RT reaction was used for PCR amplification. Each 25 μl PCR reaction contained 40 μM of dATP, dTTP, and dGTP, and 4 μM dCTP along with 0.02 μM [\textsuperscript{32}P]-dCTP, 1U of LongAmp Taq DNA
polymerase and 0.4 μM primer pairs for the target mRNA and 7SL RNA (Supplementary Table S4). An equimolar ratio of Competimer™ pair (3'-end blocked by ddNTP) for 7SL RNA was used to prevent saturation amplification. PCR products, separated by 6% PAGE, were visualized by exposing to a PhosphoImager™ screen and quantified using Imagequant software.

mRNA-Seq

mRNA-Seq libraries were prepared following the Illumina® small RNA library preparation method. Poly(A)⁺-containing RNA from 10 μg of total RNA was captured by two rounds of oligo-d(T)ₙ-bead selection and fragmented by base treatment. Following adaptor ligation to fragmented RNA, cDNA libraries were prepared by reverse transcription and amplified by PCR. Purified cDNA libraries were sequenced on the Illumina HiSeq™ platform. Sequence reads were mapped to T. brucei 927 genome assembly version 5 using TopHat2 (v2.0.8b) set at the following parameters: no-coverage-search, no-novel-juncs, no-novel-indels. The mapped reads were then converted to gene expression values using Cuffdiff2 (v2.1.1) with the default parameters (32). The gene annotation consisted of the coding regions. Log₂-fold changes were displayed on IGV browser. Volcano plots were generated using FPKM value (log₂-fold change) versus P-value in R.
RNA pol II-ChIP and ChIP-Seq

Chromatin immunoprecipitation (ChIP) was performed on extracts from all four cell lines (WT<sup>Ty1</sup>, M2<sup>Ty1</sup>, M3<sup>Ty1</sup> and M9<sup>Ty1</sup>), prepared after two days of tetracycline addition, as described (33). Briefly, 5 × 10<sup>8</sup> cells were cross-linked with 1% formaldehyde, directly added into cell cultures, via mixing by constant shaking at 25°C for 20 min. Formaldehyde was quenched with glycine (0.125 M), and cell pellets were washed twice with ice-cold phosphate buffer saline and collected by centrifugation. To prepare for chromatin fragmentation, cell pellets were washed once each with Lysis Buffer 1 and 2 and chromatin was fragmented in Lysis Buffer 3 by sonication using a Bioruptor<sup>™</sup> (Diagenode) at high setting for 15 min (30 s on, 30 s off). Sheared chromatin samples were clarified by centrifugation to remove particulate debris and cleared supernatant was subjected to immunoprecipitation by mixing with anti-Ty1 antibodies for 12–15 h at 4°C. Antibody-bound chromatin was captured using anti-mouse antibody-coated magnetic beads (Dynabeads<sup>™</sup> M-280), reverse cross-linked to separate DNA from protein, and DNA was purified by extracting twice with phenol/chloroform/isoamyl alcohol solution and concentrated by ethanol precipitation.

ChIP-PCR of RNA pol II within the chromosome 3 and chromosome 7 regions was determined by 30 cycles of amplification of ChIP DNA using [α-<sup>32</sup>P]-dCTP as tracer and primers listed in Supplementary Table S4. Input reactions contain 1/100th fraction of chromatin DNA used in the immunoprecipitation. PCR products were separated by gel electrophoresis (6% PAGE), and visualized by PhosphorImager scanning of dried gels.

ChIP-Seq libraries were constructed using NEBNext<sup>™</sup> Ultra DNA Library Prep Kit for Illumina<sup>™</sup> (NEB). Libraries were size selected for 200 bp DNA insert using AMPure<sup>™</sup> XP beads (Beckman Coulter), PCR amplified and sequenced on HiSeq<sup>™</sup> platform (Illumina). Raw reads were filtered and adapters were removed by FASTQC v0.10.1 using default parameters. The reads were mapped to the reference genome using bowtie v2.1.0. The non-unique reads were randomly distributed. Binding enrichment was called from the aligned reads using MACS2 2.0.10 and SPP (http://compbio.med.harvard.edu/Supplements/ChIP-Seq) using default parameters. All statistical analyses were performed using R.

SL RNA cap analysis

The 5′ end hypermethylated spliced leader (SL) RNA Cap structure in T. brucei [m<sup>7</sup>G(5′)ppp(5′)jm<sub>2</sub> ApmApmCmpm<sub>4</sub>Ump] was detected using primer extension of a [γ-<sup>32</sup>P] 5′-labeled oligonucleotide 5′-CTGGGAGGCTTCTCATACCAA-3′ after hybridization to 2 μg total RNA from transgenic parasites before and after 2 days of tetracycline induction. Extension products, which are a mix of hypermethylated Cap1-4 species, were resolved by electrophoresis on a 10% polyacrylamide–7 M urea gel. Unmodified SL RNA (Cap0) was obtained from transgenic parasites treated with sinfungin for 3 h before isolation.

Nascent RNA synthesis analysis

T. brucei parasites (4 × 10<sup>7</sup> cells/assay), collected before and after 2 days of tetracycline induction, were permeabilized using lysolecithin. [α-<sup>32</sup>P]UTP was added to actively transcribing cells and transcription reactions were performed for 10 min. Nascent radiolabeled RNA was isolated using a phenol–chloroform–isoamyl alcohol solution, resolved by electrophoresis on a 6% polyacrylamide–7 M urea gel and visualized by PhosphorImaging<sup>™</sup>.

RNA polymerase II activity assay

RNA pol II transcription assays were performed following the method described in (34). Briefly, nuclear extracts from all four cell lines (WT<sup>Ty1</sup>, M2<sup>Ty1</sup>, M3<sup>Ty1</sup> and M9<sup>Ty1</sup>), prepared after one day of tetracycline induction, were used to immune-capture the tagged-RNA pol II complexes using anti-Ty1 antibodies and anti-mouse antibody-coated magnetic beads (Dynabeads<sup>™</sup> M-280). Bead-bound, immune-captured RNA pol II complexes were used in transcription assays on Calf Thymus DNA (Sigma) as template and [α-<sup>32</sup>P]-UTP as tracer. Mouse pre-immune IgG was used as control for immune-capture. The resulting RNAs were separated on a denaturing 6% polyacrylamide gel and detected by PhosphorImaging<sup>™</sup>.

RESULTS

Clusters of serine-to-alanine mutations in the essential 90-amino acid central region of the non-canonical CTD negatively impact both cell growth and long RNA abundance

To investigate how the T. brucei RNA pol II machinery uses its non-canonical CTD, we have set up a genetic system in which cell growth depends upon exogenously-expressed Ty1-tagged versions of RPB1 that contain an RNAi-resistant 3′ UTR (18) (Figure 1B and C). The endogenous RPB1 (endoRPB1) subunit of RNA pol II was depleted using tetracycline-induced RNA interference (RNAi) that targeted its RNAi-sensitive 3′ UTR. After 1–2 days of RNAi induction, endoRPB1 was largely depleted in a control cell line, ultimately causing cell death (Supplementary Figure S1A). Derivative cell lines were produced that contained either an exogenous, tetracycline-inducible, Ty1-tagged wild-type CTD-containing-RPB1 protein (designated as tagged-wild-type CTD) or a mutant CTD-containing-Rpb1 protein (designated as tagged-mutant CTD) (Supplementary Table S1). We measured the growth of these cell lines to assess if mutant CTDs enable cell viability (Figure 1D).

Two cell lines (WT<sup>Ty1</sup>, including the tagged-wild-type CTD, and M3<sup>Ty1</sup>, including the tagged-mutant CTD with S1652A and S1653A substitutions) grew normally, whereas two other cell lines (M2<sup>Ty1</sup>, including the tagged-mutant CTD with S1591A, S1594A, S1595A and S1597A substitutions and M9<sup>Ty1</sup>, containing the tagged-mutant CTD with S1651A, S1653A, S1662A and S1663A substitutions) were unable to grow (Figure 1D). Each of these four cell lines expressed functional RNA pol II complexes (Supplementary Figures S1B and S2) with comparable amounts of tagged RPB1 (Figure 1E and Supplementary Table S2) that could transcribe the RNA pol II-dependent SL RNA genes (Sup-
were observed when measured by \(^{32}\text{P}\)-UTP incorporation. To determine the transcriptional effect of mutations within the M2\(^{Ty1}\) and M9\(^{Ty1}\) cell lines, we measured gene expression levels before and after their tagged-mutant CTDs were produced (Figure 1C). We also measured expression levels in the WT\(^{Ty1}\) before and after its tagged-wild-type CTD was produced and M3\(^{Ty1}\) cell lines before and after its tagged-mutant CTD was produced. cDNA libraries, representing poly(A)\(^+\) RNA from 16 individual cell cultures, were deep sequenced to obtain quantitative mRNA measurements (mRNA-Seq). Heat maps were used to visualize gene expression changes on day 1, 2 and 3, in WT\(^{Ty1}\) or M2\(^{Ty1}\) cells, compared to day 0, when no tagged protein was produced (Figure 2A and B, Supplementary Figure S5A and B). Reliance on the tagged proteins did not markedly alter overall mRNA production levels in any of the four cell lines. Specifically, less than 2% of protein-coding genes in WT\(^{Ty1}\) and M3\(^{Ty1}\) cells, and \(\sim 4\%\) of protein-coding genes in M2\(^{Ty1}\) and M9\(^{Ty1}\) cells, showed \(\geq 4\)-fold change in expression on day 2 compared to day 0. However, in both M2\(^{Ty1}\) and M9\(^{Ty1}\) cells, genes specifically situated immediately downstream of TSSs showed markedly reduced expression. This reduced expression of TSS-proximal genes became progressively more pronounced during the three-day experiment (Figure 2A and B, Supplementary Figure S5A and B).

To examine statistically whether the reduced expression of TSS-proximal genes is a general feature of M2\(^{Ty1}\) and M9\(^{Ty1}\) cells, we selected a representative TSS-proximal set of genes for analysis (Supplementary Figure S4). This set includes several-kilobase-long TSS regions, constitutively transcribed polygenic pre-mRNAs, and trans-spliced \(7\text{mG}\) capped mRNAs (17).

Mutations within the non-canonical CTD result in reduced expression of TSS-proximal genes

In M2\(^{Ty1}\) and M9\(^{Ty1}\) cells, TSS-proximal genes were poorly expressed because they were either poorly transcribed into mRNAs or efficiently transcribed into highly unstable mRNAs. These two possibilities were distinguished by assessing RNA pol II engagement across the genome. Specifically, we performed chromatin immunoprecipitation (ChIP) assays in WT\(^{Ty1}\) and M2\(^{Ty1}\) cells after 2 days of tagged-protein expression, and identified RNA pol II-associated DNA by deep sequencing (ChIP-Seq). RNA pol II ChIP-Seq of two biological replicate samples from both WT\(^{Ty1}\) and M2\(^{Ty1}\) cells shows high correlations (Supplementary Figure S7 and S10). Analyses were done using combined replicate datasets.

Wild-type RNA pol II transcribes the SL RNA genes in T. brucei (15). As expected, RNA pol II in WT\(^{Ty1}\) cells was highly associated (enrichment of \(\sim 6\)-fold) with the transcribed region of the SL RNA gene clusters (Figure 4A). Similarly, RNA pol II in M2\(^{Ty1}\) cells was highly associated (enrichment of \(\sim 8\)-fold) with the transcribed regions of the SL RNA gene clusters. Enrichment analyses in both cases were defined as the fold-change of RNA pol II ChIP DNA relative to total DNA (see Materials and Methods).

Wild-type RNA pol II does not transcribe the RNA pol I-dependent pre-rRNA gene cluster, nor does it transcribe the limited set of RNA pol I-dependent protein coding genes (20). As expected, RNA pol II in WT\(^{Ty1}\) or M2\(^{Ty1}\) cells was not associated with the pre-rRNA gene cluster (data not shown) or the RNA pol I-dependent protein coding genes (Figure 4B), but does associate with RNA pol II-dependent protein coding genes (Figure 4C).

Wild-type RNA pol II's main function is to transcribe the polygenic protein-coding regions within T. brucei's 11 large chromosomes. RNA pol II occupancy, represented as positive log2-fold changes, is visible as orange peaks in the chromosome-wide line plots of WT\(^{Ty1}\) and M2\(^{Ty1}\) ChIP-Seq analysis (Figure 5, Supplementary Figures S8 and S9). As expected, WT\(^{Ty1}\) RNA pol II occupies the coding and intergenic regions of polygenic arrays, producing wide areas of relatively modest (\(\sim 1.2\)-fold, log2-fold changes), though clearly positive, ChIP-Seq enrichment distributed within the 167 polygenic arrays. In contrast to the RNA pol II patterns observed within the more typical monogenetic transcription units of other eukaryotes, we did not see enzyme enrichment at TSSs. This likely reflects the trypanosome-specific gene expression program that includes several-kilobase-long TSS regions, constitutively synthesized polygenic pre-mRNAs, and trans-spliced \(7\text{mG}\) capped mRNAs (17).
Figure 2. mRNA-Seq shows reduced expression of TSS-proximal genes in M2Ty1 and M9Ty1 cells. (A and B) Heat map views of gene expression changes (Fragments Per Kilobase of transcript per Million mapped reads, log_{2}-fold change) in WT Ty1 and M2Ty1 cells on chromosomes 3 (A) and 7 (B) after 1 day (Day 1), 2 days (Day 2) and three days (Day 3) of tetracycline induction (compared to untreated cells). Data are shown using the IGV™ browser. Yellow and black show increases and decreases in expression, respectively. Black arrows indicate the length and direction of polycistronic transcription units. Black bars show coding sequences. Pre-ribosomal RNA gene clusters, tRNAs and snoRNAs are shown by asterisks (*). (C) Volcano plots for WTTy1, M2Ty1, M9Ty1 and M3Ty1 cells, showing gene expression (log_{2}-fold changes) versus adjusted significance P-values (log_{10}) for all protein coding genes, before and after two days of tetracycline addition. The set of 482 TSS-proximal genes is highlighted in blue; all other genes (∼8000) are shown in red.

There was a striking difference between the WT^{Ty1} RNA pol II and M2Ty1 RNA pol II occupancy within the chromosomal regions that define the probable promoters of polygenic pre-mRNA transcription. T. brucei’s probable RNA pol II promoters are recognized by H2Az, H2BV and H4K10ac occupancy (12,13). These overlapping histone marks are broad; each one is ∼10 kb, covers a TSS region, and is upstream from, as well as within, the first several protein-coding regions of polygenic arrays. In WT^{Ty1} cells, RNA pol II occupancy begins within these histone marks, consistent with its proper chromatin recognition of the TSS regions upstream for each polygenic gene array. In M2Ty1 cells, RNA pol II occupancy shows a downstream shift relative to the WT^{Ty1} RNA pol II pattern. To probe the differences between the WT^{Ty1} and M2Ty1 RNA pol II patterns within the TSS regions, we performed semi-quantitative PCR analyses on two representative TSS regions (Supplementary Figure S11). Consistent with the RNA pol II occupancy data shown in Figure 5 and Supplementary Figures S9 and S10, these data show that the TSS regions harbor much less mutant enzyme than wild-type enzyme.

Although it is established that co- and post-transcriptional processing of pre-mRNA plays a major role in linking RNA pol II activity to mRNA steady state populations, we assessed the relationship between the absence of mutant pol II within TSS regions and the levels of TSS-proximal genes determined in our RNA-seq data sets (Supplementary Figure S8). Although poor TSS recognition might be expected to lead to decreased TSS-proximal steady state mRNA levels, this will not always be the case as TSS-proximal genes are various distances from their cognate TSS regions, and post-transcriptional processes are known to play significant roles in mRNA production. Moreover, to provide an unbiased presentation of our RNA-seq data (Figure 5 and Supplementary Figures S9 and S10), we designated ‘TSS-proximal genes’ as the...
**Figure 3.** Small clusters of mutations in the RNA Pol II CTD result in reduced RNA abundance within TSS-proximal genes. (A) Representative examples of transcript production at TSS-proximal and -distal regions on chromosomes 3, 7, 6 and 10 in cell lines WT Ty1, M2 Ty1, M9 Ty1 and M3 Ty1. Black bar plots show gene expression changes after 2 days of tetracycline-induced tagged RNA Pol II production relative to uninduced cells (log2-fold changes in Fragments Per Kilobase of transcript per Million mapped reads). The blue blocks denote TSS-proximal protein-coding genes. The red blocks denote TSS-distal protein-coding genes. The black arrows indicate the direction of transcription of polygenic units. The two sets of blue blocks in opposite orientations flank two divergently arranged TSSs. (B) Representative data showing that the ~10 kb genomic loci, which include TSS regions, are marked by specific histone variants and histone modifications in wild-type *T. brucei*. Asterisk (*) indicates data is from Siegel et al. (2009). The variant histone H2Az occupancy (gray histograms) on regions shown corresponds to those in panel (A).
Figure 4. ChIP-Seq analyses of RNA pol II engagement in WT^Ty1 and M2^Ty1 cells. In all panels data from WT^Ty1 are red and data from M2^Ty1 are blue. (A) Histograms of RNA pol II ChIP-enrichment relative to total DNA, on SL RNA genes in WT^Ty1 and M2^Ty1 cells. A 4-kb region on chromosome 9, covering two SL RNA gene repeats (green boxes) plus flanking sequences (black line), is shown. Black arrows indicate direction of SL RNA transcription. The y-axis represents estimated ChIP enrichment (log2) relative to total DNA at each position based on smoothed tag density in each dataset. (B) Line plots show RNA pol II occupancy did not occur within RNAP I-dependent protein coding genes in either WT^Ty1 or M2^Ty1 cells. An ∼10 kb region on chromosome 10, containing procyclins (EP1 and EP2; 10.10260 and 10.10250), procyclin-associated gene (PAG1; 10.10240) and flanking sequences (black line) is shown. The y-axis is as in (A). (C) Line plots show RNA pol II occupancy relative to total DNA, as in (A), on nine protein-coding genes that are expressed at high levels (above 90 percentile in RNA-Seq analysis; source is TriTrypDB) and are not within the TSS regions of any polygenic array. Gene IDs are shown as chromosome number followed by gene number. Each plot covers the entire coding region of the respective gene plus 200 base pair flanking sequences. Each of these genes are between 60 and 300 kb downstream from the first protein coding gene of the polygenic array in which they are embedded. These data show examples that WT^Ty1 and M2^Ty1 RNA pol II occupancy are often relatively the same at regions 60–300 kb downstream from the first protein-coding gene.
first three protein-coding regions of a polygenic unit, regardless of their position within a several-kilobase-long TSS region and regardless of their position relative to the H2Az, H2BV and H4K10ac marks. Thus, the first three mRNA-coding regions, color-coded as the ‘TSS-proximal gene set’ to describe our RNA-seq findings, map at various positions relative to the promoter-associated histone marks. The data analysis presented in Supplementary Figure S8 indicates that the downstream shift of mutant RNA pol II occupancy, relative to the WT\(^{Ty1}\) RNA pol II pattern, only partially accounts for the alterations in steady state mRNA levels observed between the mutant and wild type cells.

An unexpected enrichment of RNA pol II occupancy was at the start of the downstream shift of the M2\(^{Ty1}\) RNA pol II pattern compared to the WT\(^{Ty1}\) RNA pol II pattern (for example, compare ChIP-seq data corresponding to genes Tb927.3.2209 and Tb927.3.2210 in Figure 5 and all ‘zoomed in’ panels on Supplementary Figure S8). These 2–3 kb ChIP-enrichment regions in the M2\(^{Ty1}\) cells may contain non-processive enzyme, as they do not correspond to increased transcripts (mRNA-seq) from these regions.

**DISCUSSION**

Our study, the first detailed analysis of the global effects of mutating an RNA pol II non-canonical CTD, shows that the non-canonical CTD of \(T. \) brucei RNA pol II plays a critical role in navigating the genome. We have previously shown that the non-canonical CTD is essential for cell viability (18). We now demonstrate that minor alterations in the essential central region of the non-canonical CTD in \(T. \) brucei RNA pol II specifically affect the ability of RNA pol II to properly recognize the polygenic coding regions of the genome (Figure 2C). These same mutations do not affect the transcription of SL RNA genes. Thus, \(T. \) brucei RNA pol II likely uses its non-canonical CTD to discriminate polygenic mRNA gene arrays from SL RNA gene arrays in chromatin.

Having previously divided the CTD into thirds and stud-
Our model suggesting how RNA pol II reads the chromatin landscape is shown in Figure 6. The inability of the M2\textsuperscript{Ty1} and M9\textsuperscript{Ty1} RNA pol II to be recruited to TSSs likely prevents the subsequent CTD modifications (see Figure 1E) that are hallmarks of transcribing RNA pol II.

It is unlikely that the histone marks or RNA pol II occupancy data reveal the precise dynamics of RNA pol II transcription within the T. brucei genome. For example, WT\textsuperscript{Ty1} RNA pol II must transcribe the sequences upstream from the first protein-coding gene of a polygenic array as this upstream region contains essential SL RNA transcribing sites and a 5’-untranslated region (5’ UTR). It appears that our ChIP assays do not capture the polymerase as it rapidly moves across non-coding gene regions after engaging chromatin within the TSS regions. Nevertheless, poor TSS-recognition by the mutant enzyme most likely accounts for the large downstream shift of enzyme engagement adjacent to the polygenic arrays within all chromosomes.

It is also unlikely that the lack of proper TSS recognition alone can fully explain the abnormal steady state mRNA pattern seen in the mutant parasites. For example, distance from TSS regions is known to influence gene expression levels in T. brucei (29). Our data show that the steady state levels of the genes far away from TSS regions appear to be increased (see Figure 2A and B) in mutant cells compared to wild type cells. If RNA pol II is blind to TSS regions, polymerase may be redirected to the internal regions of polygenic arrays. Internally initiated transcripts, recognized in T. brucei and the related parasite Leishmania major (12,27,28), may produce steady state mRNAs if they contain sufficient post-transcriptional signals. Additionally, a mRNA is more likely to be produced if it is far away from a TSS region, and thus more likely to benefit from an internal initiation event. Finally, a lack of proper TSS recognition by RNA pol II is expected to generate abnormal pre-mRNA transcripts which are unrecognizable to the cell’s co and post-transcriptional machinery. These compromised machineries would produce the abnormal mRNA pattern seen in the M2\textsuperscript{Ty1} and M9\textsuperscript{Ty1} mutants and the loss of parasite viability.

In conclusion, the T. brucei RNA pol II non-canonical CTD permits polymerase to be properly recruited to TSS regions within chromatin. We speculate that in primitive organisms the CTD encodes information that enables RNA pol II to read the chromatin landscape present in the nucleus of growing parasites. Molecular genetic studies on non-canonical CTD-containing RNA pol II machinery will illuminate our speculation.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank James F. Theis, George A.M. Cross and F. Nina Papavasiliou for critical reading of the manuscript.

FUNDING
National Institutes of Health [AI108290, AI111453]. Funding for open access charge: NIH-NIAID [111453].
Conflict of interest statement. None declared.

REFERENCES