THE CHROMATIN MODIFICATIONS BASE J AND HISTONE VARIANT H3.V PROMOTE TRANSCRIPTION TERMINATION AND REPRESS GENE EXPRESSION IN *TRYPANOSOMA BRUCEI* AND *LEISHMANIA MAJOR*

by

DAVID LEE REYNOLDS

(Under the Direction of Robert Sabatini)

ABSTRACT

Kinetoplastid parasites are early-diverged protozoans responsible for multiple diseases in humans. The entire genome of kinetoplastids is unusually organized into gene clusters containing functionally unrelated genes that are transcribed polycistronically by RNA Polymerase (RNAP) II. This genome arrangement has led to the assumption that these early-diverged eukaryotes lack RNAP II transcriptional regulation and instead regulate gene expression solely through post-transcriptional mechanisms. However, several chromatin modifications are enriched in the regions flanking gene clusters, including post-translationally modified histones, histone variants, and the DNA modification base J, suggesting that chromatin modifications impact the process of transcription in kinetoplastids. Here, two chromatin modifications enriched at transcription termination sites, base J and histone variant H3.V, and their effect on transcription and gene expression were investigated in the kinetoplastid species *Trypanosoma brucei* and *Leishmania major*. In *T. brucei*, we find that both J and H3.V independently promote transcription termination within gene clusters, thereby repressing the expression of genes located near the end of clusters. In *L. major*, base J has a more extensive role in promoting termination, where the

loss of J between convergently transcribed gene clusters results in read through transcription and the production of antisense RNAs. Interestingly, this production of antisense RNAs does not negatively affect the abundance of sense mRNAs. J also promotes termination prior to the end of gene clusters in *L. major*, indicating that this function of J is conserved between *L. major* and *T. brucei*. Conversely, H3.V does not promote termination in *L. major*, despite its enrichment at termination sites. These findings provide the first evidence of RNAP II transcriptional regulation by chromatin modifications in kinetoplastids, specifically the promotion of termination within gene clusters to repress gene expression.

INDEX WORDS: Kinetoplastids, Trypanosomes, Base J, Histone variant H3.V, RNA Polymerase II, Transcription termination, Gene expression, Epigenetic, Chromatin, Polycistronic transcription

THE CHROMATIN MODIFICATIONS BASE J AND HISTONE VARIANT H3.V PROMOTE TRANSCRIPTION TERMINATION AND REPRESS GENE EXPRESSION IN *TRYPANOSOMA BRUCEI* AND *LEISHMANIA MAJOR*

by

DAVID LEE REYNOLDS

BS, University of Virginia, 2010

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

THE CHROMATIN MODIFICATIONS BASE J AND HISTONE VARIANT H3.V PROMOTE TRANSCRIPTION TERMINATION AND REPRESS GENE EXPRESSION IN *TRYPANOSOMA BRUCEI* AND *LEISHMANIA MAJOR*

by

DAVID LEE REYNOLDS

Major Professor:	Robert Sabatini
Committee:	David Garfinkel
	Lance Wells
	Zachary Lewis

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

ACKNOWLEDGEMENTS

Firstly, I want to thank my advisor, Bob Sabatini, for his mentorship and support over the years. I also want to thank the members of my committee for their support and for providing their unique perspectives and insight. Members of the Sabatini Lab, past and present, thank you for your support and for making my time in the lab such an enjoyable and stimulating experience. Collaborators over the years, much of this work could not have been completed without your assistance and expertise, so thanks so much for your help. Of course I also want to thank family and friends for all of their constant love and support. The past five years have been truly enriching and I am incredibly grateful for everything that I've learned and accomplished. I look forward to my continued career in science and I hope I will be lucky enough to work with such talented researchers as the ones I have met here.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION: EPIGENETIC REGULATION OF GENE EXPRESSION IN KINETOPLASTID PARASITES

ABSTRACT

Kinetoplastid parasites branched early during the evolution of eukaryotes, which is reflected in the many unusual biological features of these organisms. Among other peculiarities, the entire genome of kinetoplastids is organized into gene clusters that are transcribed polycistronically by RNA Polymerase (RNAP) II. This genome arrangement superficially resembles bacterial operons, but kinetoplastid gene clusters do not contain functionally related genes. Kinetoplastid genome structure, combined with the apparent lack of RNAP II promoter sequences has led to the assumption that these early-diverged eukaryotes lack RNAP II transcriptional regulation and instead regulate gene expression through post-transcriptional mechanisms. Several chromatin modifications are enriched in the regions flanking gene clusters however, including post-translationally modified histones, histone variants, and the DNA modification base J, suggesting that chromatin modifications impact the process of transcription in kinetoplastids. Growing evidence has implicated epigenetic mechanisms in the regulation of surface protein expression from subtelomeric regions in the kinetoplastid *Trypanosoma brucei*, however it is becoming clearer that similar mechanisms are also utilized to regulate RNAP II transcribed gene clusters genome-wide in kinetoplastids. Here, the role of epigenetic mechanisms on transcriptional regulation in kinetoplastids is discussed, including evidence that

chromatin modifications promote termination within gene clusters as a mechanism to regulate gene expression in the context of the unusual genome organization of kinetoplastids.

INTRODUCTION

Kinetoplastids are a group of early-diverged eukaryotes that include the protozoan parasites *Trypanosoma brucei*, *T. cruzi*, and *Leishmania* spp. These pathogens are responsible for multiple neglected tropical diseases in humans that affect millions worldwide, including African sleeping sickness, Chagas disease, and a collection of diseases called leishmaniasis that range in severity from skin lesions to potentially fatal infections of internal organs (1-3). Kinetoplastids are also responsible for diseases of agricultural importance such as nagana in livestock. All kinetoplastid parasites cycle between insect vectors and mammalian hosts and progress through multiple life stages in response to various environmental stimuli including changes in pH, temperature, and nutrients. Both *T. cruzi* and *Leishmania* spp. are obligate intracellular parasites in the mammalian host. *T. cruzi* mammalian stages invade a variety of host cells, including macrophages, muscle cells, and fibroblasts (4) whereas *Leishmania* spp. invade host macrophages (5). Unlike *T. cruzi* and *Leishmania* spp., *T. brucei* parasites remain extracellular throughout its life stages, residing in the blood and interstitial spaces of the mammalian host, and therefore are constantly exposed to the host innate and adaptive immune systems (6). *T. brucei* parasites escape host immune detection through several mechanisms, one of which includes the tightly regulated transcription of genes encoding surface proteins in a process called antigenic variation (7,8) (discussed below). All kinetoplastid parasites thus share the challenge of surviving and propagating in hostile and changing host conditions, necessitating the ability to sense and respond to environmental changes through regulated gene expression.

In addition to the focus on kinetoplastids as pathogens, these organisms have also served as model organisms in the study of eukaryotic evolution, particularly *T. brucei* due to its experimental tractability (9), and to a lesser extent *L. major*. Unusual features of kinetoplastid biology have also contributed to their use as model organisms in the study of monoallelic expression, RNA editing, flagellar biology, and trans-splicing, among other biological phenomena. Another unusual aspect of kinetoplastids is the arrangement of functionally unrelated genes in RNAP II polycistronically transcribed clusters (10-12) (Figure 1.1). Along with the lack of identified RNAP II promoter sequences and the paucity of sequence specific transcription factors, these observations have led to the study of post-transcriptional mechanisms as the focus of gene expression regulation in kinetoplastids, including RNA stability, translation efficiency, and post-translational modifications (13,14). The assumption of solely post-transcriptional regulation of gene expression has been challenged within the past decade by the discovery of multiple chromatin modifications that are enriched at RNAP II transcription initiation and termination sites including histone variants, histone post-translational modifications, and the DNA modification base J, which consists of a glucosylated thymidine (15-18). The presence of these chromatin modifications at sites of RNAP II transcription initiation and termination raises the possibility that they are used to regulate RNAP II transcription. Several chromatin modifications are also enriched within centromeric, telomeric, and subtelomeric regions, the latter of which often contain surface protein genes important for host immune evasion and that are known to undergo high rates of DNA recombination and gene evolution (19-21). Investigation of transcriptional regulation in kinetoplastids could therefore shed new light on the regulation of gene expression in these parasites and potentially reveal novel drug targets, particularly transcriptional regulatory mechanisms unique to kinetoplastids.

Numerous mechanisms of post-transcriptional regulation in kinetoplastids have been described and reviewed previously (22-25). Here, current understanding of transcriptional regulation in kinetoplastids and recent findings that chromatin modifications promote transcription termination to regulate gene expression will be discussed.

GENOME ORGANIZATION

Kinetoplastids are diploid organisms. The haploid nuclear genome of *T. brucei* is approximately 30-40 megabases (Mb), depending on the strain, and is distributed among 11 chromosomes (10). In comparison, the haploid nuclear genomes of *L. major*, a *Leishmania* spp. that causes cutaneous lesions in humans, and *T. cruzi* are distributed among a greater number of chromosomes that are smaller in size, 32.8 Mb on 36 chromosomes in *L. major* and 60.3 Mb on 41 chromosomes in *T. cruzi* (11,12). Kinetoplastid genomes contain approximately 8000-9000 protein-coding genes. Very few genes, including a gene encoding a poly-A polymerase and another encoding an RNA helicase, have been shown to undergo cis-splicing (26,27), and all others lack introns.

Because of the large evolutionary distance from other eukaryotes, only around 35% of genes can be assigned a biological function based on sequence similarity and the remaining genes are annotated as coding for hypothetical proteins. Examination of the distribution of genes with an identifiable biological function has revealed that functionally unrelated genes are arranged into gene clusters that are transcribed polycistronically by RNAP II. This genome organization therefore differs from bacterial genomes, in which functionally related genes are arranged into operons. Some polycistronic gene clusters are found in other eukaryotes, including nematodes and *Drosophila* (28), however kinetoplastid genomes are unique in that the

vast majority of protein coding genes are organized into clusters. RNAP I transcribed rRNA genes and RNAP III transcribed tRNA and 5S rRNA genes are located between gene clusters.

Recent analysis of gene ontology (GO) annotations has revealed some positional bias within gene clusters in *T. brucei* (29). Four GO term categories enriched near the beginning of gene clusters were related to translation and two GO term categories enriched towards the end of gene clusters were involved in transcription. It was further shown that positional bias within gene clusters could impact mRNA abundance throughout the cell cycle and during stress responses (29). It is not known if these results in *T. brucei* are general to all kinetoplastid species, but kinetoplastid genomes do contain a relatively high degree of synteny (30), supporting this possibility. Thus, gene clusters in *T. brucei*, and perhaps other kinetoplastids, contain some functional organization, but not to the same extent that is observed in bacterial or nematode operons (31).

TRANSCRIPTION IN KINETOPLASTIDS

RNAP II initiates transcription at regions called divergent strand switch regions (dSSRs), elongates polycistronically through tens to hundreds of genes, and terminates at convergent strand switch regions (cSSRs) (32-36). RNAP II transcription also initiates at non-strand switch regions called head-tail (HT) sites, where transcription of an upstream gene cluster terminates and initiation of a downstream independent gene cluster on the same strand occurs (15,36). Nascent transcripts are processed by trans-splicing, in which a 39 nucleotide spliced leader (SL) sequence is added to the 5' end of all mRNAs, followed by 3' polyadenylation (37-42). Trans-splicing is mechanistically coupled to mRNA 3' cleavage and polyadenylation of the upstream

transcript (43). RNA processing separates nascent transcripts into individual mRNAs that are then exported from the nucleus for translation.

Sequence analysis of dSSRs has not identified any conserved sequences or motifs indicative of an RNAP II promoter, however guanosine rich tracts were found upstream of transcription start sites within many dSSRs, which has been proposed to provide promoter directionality (15). Higher levels of DNA curvature are also predicted within dSSRs (44). The single RNAP II promoter that has been identified in kinetoplastids is found at the SL RNA gene, which is where SL RNA that is trans-spliced to the 5' end of all mRNAs is transcribed (45). Approximately 100 copies of SL RNA genes are arranged in tandem in kinetoplastid genomes and transcription is monocistronic, such that each SL RNA gene copy has its own RNAP II promoter. Characterization of the SL RNAP II promoter identified two sequence elements upstream of the transcription start site that are involved in the recruitment of RNAP II (46,47).

Relative to other eukaryotes, kinetoplastids appear to possess fewer general transcription factors, some of which are highly diverged (reviewed in (48)). Kinetoplastid genomes also encode few sequence specific transcription factors. In addition, the C-terminal domain of RPB1, the largest subunit of the RNAP II complex, is truncated relative to other eukaryotes and lacks the heptameric repeats that are differentially phosphorylated at different stages of transcription in most eukaryotes (49,50). Thus, transcriptional machinery and the process of transcription are highly unusual in the early-diverged kinetoplastids.

Because of the organization of largely functionally unrelated genes into co-transcribed clusters, it is unclear how regulated gene expression can be accomplished at transcription initiation in kinetoplastids. It is therefore a generally held assumption that RNAP II transcription is a constitutive process across the genome in kinetoplastids and that solely post-transcriptional

processes influence mRNA and protein abundance (13,14). In support of this, it has been shown that transcription initiates randomly on plasmids transfected into kinetoplastids (51-53). Although a low level of non-specific transcription initiation has been shown on *L. major* chromosomes, RNAP II initiates at eight to tenfold higher rates specifically at dSSRs compared to other chromosomal regions and terminates specifically at cSSRs (34,35). RNAP II transcription is clearly not a random process in kinetoplastids and no direct evidence exists to indicate that transcription rates are the same across all chromosomes. A major unresolved question therefore is what mechanisms are utilized to direct RNAP II transcription initiation and termination in the context of polycistronic gene clusters and whether they might contribute to the regulation of transcription and gene expression. A major breakthrough was the finding that multiple chromatin modifications are enriched at both RNAP II transcription initiation and termination sites (15-18). Evidence is now emerging that several of these chromatin modifications do alter the process of transcription, and consequentially impact gene expression in kinetoplastids (54-57) (discussed below).

Although very little is known about the regulation of RNAP II transcription at any stage (initiation, elongation, or termination) in kinetoplastids, the organization of genes into polycistronically transcribed gene clusters suggests that the process of termination in particular differs from other eukaryotes. In most eukaryotes the process of RNAP II termination is tightly coupled to the process of mRNA 3' end formation following transcription of a gene (reviewed in (58,59)). 3' end formation and termination cannot be linked in kinetoplastids as this would result in premature termination within a gene cluster. As will be discussed below, premature termination near the end of gene clusters provides some level of transcriptional regulation of gene expression, but clearly RNAP II cannot terminate shortly after initiation, as this would

potentially leave tens to hundreds of genes untranscribed. Two major models of RNAP II termination at mRNA coding genes have been proposed in eukaryotes, the allosteric and torpedo models (58,59). In both models recognition of the poly-A signal sequence in the nascent RNA by the cleavage and polyadenylation factor (CPF) and cleavage factor (CF) complexes, which associate with the elongating RNAP II via interaction with the C-terminal domain of RBP1, results in the cleavage of the nascent RNA. The 3' end of the RNA upstream of the cleavage site is polyadenylated and the resulting pre-mRNA is further processed into mature mRNA. The 5' end of the RNA downstream of the cleavage site remains associated with the elongating RNAP II. In the allosteric model of termination, restructuring of the elongating RNAP II complex following recognition of the poly-A signal sequence by the CPF and CF complexes triggers RNAP II pausing and loss of associated elongation factors, thereby promoting termination. In the torpedo model, the generation of an unmodified free 5' phosphate following cleavage of the nascent RNA creates a substrate for a 5'-3' exonuclease, Xrn2 in humans, that degrades the nascent RNA, eventually overtaking the elongating RNAP II and promoting its release from the DNA. Pausing of the RNAP II may facilitate termination by allowing the exonuclease to more quickly reach the RNAP II. These two models are not mutually exclusive and in fact, a unified model has been proposed in which both factors involved in the allosteric and torpedo models function in concert to promote termination. In any of the models proposed it remains unclear how exactly RNAP II is dissociated from the DNA template.

Like other eukaryotes, 3' mRNA end formation occurs co-transcriptionally in kinetoplastids; however, in kinetoplastids 3' end formation is mechanistically coupled to trans-splicing (43) and not to RNAP II termination. Because of the decoupling between 3' end formation and termination, RNAP II continues to elongate and transcribe downstream genes

within each cluster, thus avoiding premature termination after transcription of a single gene that would occur if termination was linked to 3' end formation as in other eukaryotes (Figure 1.2). Signals that promote termination in kinetoplastids and the mechanisms involved remain unclear. It is likely that kinetoplastids utilize unique mechanisms to terminate RNAP II transcription at the end of gene clusters, further highlighting the significance of kinetoplastid-specific chromatin modifications enriched at RNAP II termination sites.

It is unclear if co-transcriptional RNA processing contributes to the regulation of gene expression. No consensus sequences have been identified at splice or polyadenylation sites, though splicing usually occurs downstream of a polypyrimidine tract, followed by polyadenylation of the upstream transcript usually within 100 base pairs (bp) upstream of the splice site (60-62). Developmental regulation of splice and polyadenylation site choice has been observed in both *T. brucei* and *L. major* (60,61). Several sequences within mRNA 3' untranslated regions that impact mRNA abundance have been identified in kinetoplastids (reviewed in (22)). Thus, further analysis is needed to determine the role of alternative mRNA processing in regulating transcript abundance and translation, and to further elucidate the regulation of alternative mRNA processing events.

RNAP I TRANSCRIPTION AND ANTIGENIC VARIATION

Another unusual feature of kinetoplastid transcription is the fact that RNAP I transcribes not only rRNAs, but also some protein coding genes (63). In *T. brucei*, RNAP I transcribes all of the major cell surface proteins. Surface protein expression is developmentally regulated such that *procyclin* genes are transcribed only during the insect stage of the parasite and *variant surface glycoprotein* (*VSG*) genes solely during the mammalian stage (63). The unique extracellular

nature of *T. brucei* sets it apart from other kinetoplastids and has led to a mammalian immune evasion mechanism of regulated surface protein expression called antigenic variation (7,8). The *T. brucei* genome contains a repertoire of about 2500 *VSG* genes (64); however, RNAP I transcription is tightly regulated such that a single *VSG* is transcriptionally active from one of 15 specialized subtelomeric expression sites (ES) in a given cell. This monoallelic expression ensures that the entire cell surface is covered by a single VSG that is not recognized by the immune system. Rapid VSG turnover from the parasite cell surface also prevents detection by components of the immune system. This rapid VSG turnover may also explain the use of RNAP I transcription, which presumably allows higher rates of transcription than RNAP II (65). Eventually the adaptive immune system generates antibodies that recognize the single expressed VSG, which necessitates periodic *VSG* switching through several mechanisms that involve switching the actively RNAP I transcribed ES or through homologous recombination to swap out the expressed *VSG* for a previously silent *VSG* that is unrecognized by the immune system (66,67). The process of antigenic variation ensures that some parasites in the population outpace the production of specific antibodies by the adaptive immune system, allowing the infection to persist.

The 15 ESs contain multiple genes called expression site associated genes (ESAGs) that are transcribed polycistronically by RNAP I (68) (Figure 1.3). ESAGs encode for mostly membrane-associated proteins including receptors and transporters (69-77). Because ESAGs are present in multiple copies, both within ESs and in some cases internally within chromosomes, the function of most ESAGs remains unknown (10,68,78,79). A 50bp repeat sequence ranging in size from 10-50kb is present upstream of each ES, followed by an RNAP I promoter. A 70bp repeat sequence is located downstream of the ESAGs within each ES, followed by a single *VSG*

gene prior to the telomeric repeats. In addition to the 15 ESs, *VSG* genes are also located in transcriptionally repressed subtelomeric arrays and on approximately 100 minichromosomes that range in size from 50-150kb (80,81). Repeat sequences associated with ESs and minichromosomes are presumed to facilitate homologous recombination and activation of silent *VSG*s (66).

CHROMATIN MODIFICATIONS IN KINETOPLASTIDS

Epigenetics in kinetoplastids

Epigenetic mechanisms are broadly defined as those that impart mitotically and/or meiotically heritable changes in gene or genome function that do not involve changes in DNA sequence (82), and their impact on transcription and gene expression regulation have been increasingly appreciated in both higher eukaryotes and kinetoplastids (83-85). Epigenetic modifications include histone post-translational modifications (PTM), histone variants, and DNA modifications (86). Through altering chromatin structure, epigenetic modifications impact genome accessibility and therefore influence a variety of biological processes in addition to gene expression, such as DNA replication, DNA recombination, and silencing of transposable elements.

Because of its important role in mammalian immune evasion, much of the knowledge of epigenetic regulation and chromatin modifications in kinetoplastids has come from studies investigating the regulation of antigenic variation in *T. brucei*. Numerous studies have now been published that link epigenetic mechanisms to transcriptional repression of silent ESs (reviewed in (87-89)). Telomeric repression also contributes to the maintenance of silent ESs (90). Notably, evidence has emerged that epigenetic mechanisms utilized to regulate the expression of

subtelomeric ESs are also employed to regulate RNAP II transcription of chromosome internal gene clusters (55-57).

Histone modifications

Early characterization of kinetoplastid chromatin structure revealed that, like higher eukaryotes, nucleosomes are the basic units of chromatin, consisting of histone H3-H4 and H2A-H2B dimers that form a histone octamer (91,92). 146 bp of DNA is wrapped around each histone octamer and histone H1 binds the DNA between nucleosomes, providing further compaction. The primary amino acid sequence of histones is typically very well conserved in eukaryotes, but this is not the case in kinetoplastids, which contain histones with relatively large sequence variation, both compared to other eukaryotes and, to a lesser extent, between kinetoplastid species. Much of the sequence divergence is within the N-terminal region of histones, which extend from the core nucleosome and are the target of a variety of PTMs. Histone methylation, acetylation, and phosphorylation have been characterized in kinetoplastids, but overall, histone PTMs appear to be less extensive than those in higher eukaryotes (84,93-95). Consistently, kinetoplastid genomes also encode fewer histone binding and modifying proteins, the readers and writers of the histone code hypothesis proposed by the Allis lab (96). Sequence divergence makes it difficult to compare histone PTMs between kinetoplastids and other eukaryotes, and some modifications appear unique to kinetoplastids, such as mono-methylation of N-terminal alanines and extensive acetylation of C-terminal lysines of H2A (93). Overall, little is known about the function of histone PTMs in the regulation of gene expression in kinetoplastids, though much insight has been gained by mapping the localization of modified histones by chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) and high throughput

sequencing (ChIP-seq). Histone H3 trimethylated at K4 (H3K4me3) and histone H4 acetylated at K10 (H4K10ac) are both enriched at gene cluster transcription start sites (TSSs) (15,17). Bromodomain-containing factors bind to acetylated histones (97), and the bromodomain-containing factor 3, BDF3, is enriched at TSSs in *T. brucei* (15). BDF3 is essential in *T. brucei*, but whether BDF3 binds to H4K10ac and its molecular role remain unclear. H3K76 methylation in *T. brucei*, which corresponds to H3K79 in other eukaryotes, has been localized to some TSSs and TTSs where it appears to function in DNA replication control. Localization of additional kinetoplastid histone PTMs has been hindered by a lack of specific antibodies. Many of the modifications found in other eukaryotes, for which specific antibodies exist, are absent or present in a different amino acid sequence context in kinetoplastids. For example H3K23, which has been identified as trimethylated in kinetoplastids, may correspond to H3K27 in other eukaryotes.

Histone variants

Histones variants are histones that differ in their amino acid sequence from canonical histones. The sequence variation can range from as little as four amino acids in the case of the canonical H3 and the variant H3.3, to much greater, such as H2A and the variant H2A.Z, which share about 60% sequence identity. Many histone variants have been identified in eukaryotes, mainly histone H2A and H3 variants, though variants of H2B and H4 also exist (reviewed in (98)). Unlike canonical histones, histone variants are incorporated into DNA outside of the S phase of the cell cycle by specialized histone chaperones and chromatin remodelers at specific genomic locations. Like other chromatin modifications, incorporation of histone variants has been shown to influence many processes including transcription, formation of repressive chromatin, and DNA repair in eukaryotes (99,100).

Four histone variants have been identified in kinetoplastids: H2A.Z, H2B.V, H3.V, and H4.V (15,19,101). Only H2A.Z is found in other eukaryotes, which share about 51-58% sequence identity with the *T. brucei* H2A.Z. In other eukaryotes, H2A.Z has been linked to both transcriptional activation and formation of heterochromatin, which consists of chromatin that is highly compacted and refractory to transcription. Differential H2A.Z PTMs may account for these seemingly contradictory functions. Kinetoplastid histone variants have largely been studied only in *T. brucei*. The *T. brucei* H2A.Z has been shown to specifically associate with H2B.V (101), where together they are enriched at TSSs (15). Co-immunoprecipitation studies have revealed that nucleosomes containing H2A.Z and H2B.V are less stable than those containing canonical histones (15), suggesting that the H2A and H2B variants form less stable nucleosomes at TSSs, perhaps facilitating a more open chromatin structure amenable to RNAP II initiation. Both H2A.Z and H2B.V are essential in kinetoplastids (101,102), but whether they are utilized to regulate transcription initiation and gene expression remains unknown. PTMs have not been extensively characterized on any histone variants in kinetoplastids.

In contrast to H2A.Z and H2B.V, H3.V and H4.V are enriched at transcription termination sites (TTSs) in *T. brucei* (15). It is unknown if H3.V and H4.V occupy the same nucleosomes, as is observed for H2A.Z and H2B.V. In addition to TTSs, H3.V is also enriched in telomeric and subtelomeric regions (15,19), whereas H4.V is much less enriched in these regions, suggesting that H3.V can form nucleosomes with the canonical histone H4. In other eukaryotes an essential H3 variant, CenH3, is involved in kinetochore formation and chromosome segregation. While H3.V is enriched at centromeres (15,21), H3.V lacks some conserved features of most centromeric H3 variants and is not essential for proper cell division (or parasite viability). Given that no other H3 variants have been identified in kinetoplastids, it

appears that a centromere specific H3 is absent from these organisms, consistent with other unusual features of the kinetoplastid kinetochore and chromosome segregation (21,103). H4.V is also not essential. The localization of H3.V and H4.V to TTSs in *T. brucei* implicates these histone variants in the regulation of RNAP II termination. As discussed below and in more detail in Chapter 4, H3.V does indeed promote transcription termination, both at RNAP II transcribed gene cluster termination sites and within the RNAP I transcribed subtelomeric ESs (56,57).

Base J

Kinetoplastids contain a unique nuclear DNA modification called base J, beta-D-glucosyl-hydroxymethyuracil, which consists of a glucosylated thymidine. In addition to kinetoplastids, this modified DNA base has only been found in *Euglena* and *Diplonema*, which are also early-diverged flagellated protozoans. J was initially discovered in *T. brucei* where it is enriched in repetitive regions of the genome including telomeres, and the 50bp and 70bp repeats located within ESs (Figure 1.3) (104).

Base J is synthesized in a two-step pathway in which the methyl group on thymidine in the context of DNA is hydroxylated by a thymidine hydroxylase to yield hydroxymethyluridine (105). A glucosyltransferase subsequently transfers a glucose molecule to yield base J (106,107). Two thymidine hydroxylases have been identified, J-binding protein (JBP) 1 and JBP2 (108,109). The JBPs belong to a subclass of 2-oxoglutarate dependent hydroxylases called the TET/JBP subclass (110). These enzymes utilize 2-oxoglutarate, oxygen, and Fe^{2+} in the hydroxylation reaction. A single glucosyltransferase, J-associated glucosyltransferase (JGT), is responsible for the final step in the J synthesis pathway (106,107). Base J synthesis is described in more detail in Chapter 2. *JBP1/2 T. brucei* double knockouts (KO) are viable, as are *JGT* KO

cells, indicating that J is non-essential in *T. brucei*. In *T. cruzi*, individual *JBP1* or *JBP2* KO cells have significantly reduced J and are viable, but attempts to generate *JBP1/2* double KOs have been unsuccessful. Similarly, *JBP2* KO cells are viable in *L. tarentolae*, a *Leishmania* species that infects lizards, but a *JBP1* knockout cell line cannot be generated unless an ectopic copy of *JBP1* is expressed. The inability to completely remove base J from *T. cruzi* and *Leishmania* spp. strongly suggests that the modified DNA base is essential in these kinetoplastids. Efforts (currently underway) to knock out *JGT* in *T. cruzi* and *L. major* will further clarify the essential nature of J. Significantly greater reduction in J levels can be achieved in *JBP* KO cells through pharmacological means, including the use of dimethyloxalylglycine (DMOG), which is a structural analog of 2-oxoglutarate that acts as a competitive inhibitor of the JBPs, as well as bromodeoxyuridine (BrdU), a thymidine analog that inhibits J by an unknown mechanism. These independent approaches to reduce J levels have greatly facilitated investigations of the role of J in transcriptional regulation in kinetoplastids.

Several observations have linked J with the process of antigenic variation in mammalian stage *T. brucei* parasites. One of the first indications that chromatin modifications may contribute to the regulation of antigenic variation in fact came with the discovery of differential restriction enzyme digestion of the actively transcribed ES compared to the 14 silent ESs (111,112). Differences in digestion were eventually linked to the presence of base J specifically in the silent ESs, where the glucose group of J inhibits complete digestion. In silent ESs, base J is most enriched in the telomeres and extends into silent ES gene clusters in a gradient that decreases closer to the beginning of the ES gene cluster (113). Following the activation of a silent ES, and concomitant repression of the previously active ES, base J is lost from the silent (now active) ES and is synthesized in the previously active ES. Furthermore, J is developmentally regulated in *T.*

brucei and is only present in the mammalian stage of the parasite and not in the insect stage (114), which does not undergo antigenic variation. These observations strongly implicate base J in the regulation of antigenic variation during mammalian infections; however, removal of the enzymes that synthesize base J have yet to uncover any alterations in *VSG* silencing or switching (115). Chapter 4 presents evidence that loss of J in combination with H3.V loss does lead to increased expression of silent *VSG* genes. These findings indicate that H3.V, and likely other factors, can compensate for the loss of base J alone. Despite its large enrichment within telomeres, no telomeric defects have been identified following J loss (116).

Additional insight into the function of base J was gained through J IP-seq in *T. brucei*, which confirmed the enrichment of J at repetitive regions of the genome, but also identified J at SSRs flanking gene clusters (18). Similar observations were made in *T. cruzi* and *Leishmania* spp (54,117). These findings implicated J in the regulation of RNAP II transcription of gene clusters. Subsequent work has identified a role of J in the regulation RNAP II transcription in each kinetoplastid species examined, though the specific effect of J on transcription differs between kinetoplastids. In *T. cruzi*, the reduction of J leads to the formation of more active chromatin, including decreased nucleosome abundance and increased histone acetylation, increased RNAP II recruitment and gene cluster transcription rates, and global gene expression changes (54,118). No evidence of a termination defect at cSSRs was found following J loss. In contrast, in *L. tarentolae* the major effect of J loss is a transcriptional defect at cSSRs (117). Reduction of base J in *JBP2* KO *L. tarentolae* cells leads to the production of antisense RNAs, which was attributed to the process of read through transcription past the TTS within cSSRs, and thus continued RNAP II elongation into the opposing gene cluster upon J loss. Treating the *JBP2* KO cells with BrdU causes further J loss, exacerbating the termination defect, and results in cell

death. As seen in *T. cruzi*, J loss leads to global gene expression changes in *L. tarentolae* (117). In Chapters 3 and 5 I present findings that confirm the role of J in promoting RNAP II termination in *L. major* and *T. brucei*. In *T. brucei* however, loss of base J does not result in read through transcription into opposing gene clusters past cSSRs and instead has a more specific role in the promotion of RNAP II termination prior to the end of gene clusters (55,56). J is enriched prior to the end of some gene cluster and the gene(s) downstream of the J peak are typically lowly expressed. The loss of J results in the RNAP II read through transcription and upregulation of the gene(s) downstream of base J. Thus, the loss of J leads to increased transcription of normally lowly transcribed genes located near the end of gene clusters (55,56). This process is therefore referred to as gene cluster internal termination (Figure 1.4). A similar role of J in promoting gene cluster internal termination has been identified in *L. major*. These investigations of the function of base J have collectively provided the strongest evidence thus far of RNAP II transcriptional regulation by epigenetic mechanisms in kinetoplastids.

TRANSCRIPTION TERMINATION

Generally, transcription initiation is the primary focus of transcriptional regulation of gene expression in eukaryotes. Multiple layers of regulation have been characterized for transcription initiation, but much less is known about termination and how it contributes to the regulation of gene expression. The recent realization that much of the genome is pervasively transcribed in eukaryotes has led to a greater appreciation of the role of termination in what is referred to as transcriptome surveillance. For example, many eukaryotic promoters give rise to bidirectional transcription. Termination of transcription upstream of promoters prevents the formation of non-coding RNAs and the disruption of mRNA production that can be caused by

transcription interference and antisense repression (119). Termination of protein-coding genes has also been linked to the regulation of promoter directionality through gene looping (120). In addition to transcriptome surveillance, termination can also contribute to regulated gene expression whereby termination prior to the end of a gene, referred to a premature termination, limits gene expression (121). Defects in transcription termination and subsequent read through transcription have been linked to cancer severity, in part by altering the expression of oncogenes (122).

Several termination pathways have been characterized in eukaryotes, and their usage largely depends on the RNA species being transcribed (58). As discussed above, in both yeast and metazoans termination of RNAP II transcribed protein-coding genes is closely linked to 3' mRNA processing. Although the precise molecular details are still being defined, it is clear that 3' mRNA processing initiates RNAP II termination. In the context of polycistronic transcription however, a link between termination and 3' mRNA processing would lead to termination after transcription of a single gene in a cluster and thus the mechanism of termination must be independent of 3' end processing in kinetoplastids. Indeed, 3' mRNA processing is instead linked to 5' trans-splicing of the downstream gene in the cluster. Although kinetoplastids contain homologs of termination factors in other eukaryotes, these have largely been uncharacterized and the mechanism of RNAP II termination remains unclear. RNAP II termination has been characterized at SL RNA genes, which also contain the only known RNAP II promoter. Termination at SL genes is linked to a T tract, where at least six T's located downstream of the 3' end of the SL RNA promotes termination (123). Because polypyrimidine tracts direct 5' trans-splicing, tracts of six or more T's are not unusual within gene clusters, and therefore the

mechanism of RNAP II termination at the end of gene clusters must be distinct from that at SL RNA genes.

As described above, base J promotes RNAP II termination at most TTSs in *Leishmania* spp. and at specific sites within gene clusters in *T. brucei*. The few cSSRs where read through transcription was not observed following J loss in *Leishmania* spp. contain RNAP III transcribed genes, including tRNAs and 5S rRNA. These findings suggest a J independent termination mechanism that involves RNAP III genes. tRNA genes have been linked to transcriptional repression in other eukaryotes and the mechanism involves the subnuclear localization of tRNA genes near the nucleolus and requires chromatin structural changes and the activity of nucleosome remodelers (124,125). The mechanism by which RNAP III genes promote termination in *Leishmania* spp. (and perhaps other kinetoplastids) has not been investigated. In contrast to other eukaryotes however, the orientation of the RNAP III transcribed gene does impact RNAP II termination. RNAP III transcribed genes encoded on the opposite strand of a gene cluster more strongly promote RNAP II termination, implicating RNAP II and III transcriptional collision as the potential mechanism of termination.

Because the majority of RNAP II termination sites in kinetoplastid genomes do not contain RNAP III transcribed genes, additional termination mechanisms exist, one of which involves base J. As described above, the importance of J in promoting RNAP II termination is most clearly seen in *Leishmania* spp., where J loss leads to read through transcription at most TTSs. In *T. brucei* J does not appear to be involved in promoting termination at cSSRs and instead promotes termination at some sites within gene clusters. Given the developmental regulation of J synthesis in *T. brucei*, it is perhaps not surprising that this kinetoplastid species is less dependent on J for RNAP II termination. Proper termination may also be more critical in *T. brucei* because

unlike most *Leishmania* spp., *T. brucei* has retained functional RNAi machinery. Therefore, read through transcription and subsequent formation of antisense RNAs could be more detrimental in *T. brucei*, suggesting that additional J independent termination mechanisms are important. Nevertheless, J has a conserved role in promoting termination within gene clusters to repress gene expression in both *T. brucei* and *L. major*. Although a role of J in promoting termination has not been identified in *T. cruzi*, only three cSSRs have been examined thus far, including one that contains a tRNA gene (118). Whether J promotes termination within gene clusters in *T. cruzi* has not been examined.

In addition to base J, histone variant H3.V has also been shown to promote RNAP II termination in *T. brucei*. H3.V and base J co-localize at TTSs, subtelomeres, and telomeres. Similar to J loss in *T. brucei*, loss of H3.V leads to read through transcription at gene cluster internal termination sites and increased expression of genes at the end of clusters (Figure 1.4). H3.V loss also leads to derepression of silent *VSG* genes. Removal of both J and H3.V results in a synergistic effect on gene derepression, both within gene clusters and on normally silent *VSG* genes. *H3.V* KO cells also generate more siRNAs derived from cSSRs, where increased dual strand transcription at the end of gene clusters upon H3.V loss gives rise increased double stranded RNAs, which are subsequently processed into siRNAs in *T. brucei*.

Interestingly, the function of H3.V in promoting termination is not conserved in *L. major*, but H3.V does localize to TTSs where it promotes J synthesis. *L. major H3.V* KO cells have reduced J levels at TTSs, but also at TSSs where H3.V is not enriched, suggesting an indirect link between H3.V and J in *L. major*. These findings indicate that the role of base J in promoting transcription termination is conserved among *T. brucei* and *Leishmania* spp., but the role of H3.V is not. Overall, it is becoming clear that chromatin modifications are utilized in

kinetoplastids to promote transcription termination and repress the expression of genes at the end of gene clusters. A remaining challenge will be to identify other transcription termination factors that likely exist in kinetoplastids and how these work in conjunction with base J and H3.V. Additionally, investigating whether other mechanisms are utilized to promote termination within gene clusters will provide further insight into the extent of transcriptional regulation of gene expression in kinetoplastids.

CONCLUSION

The identification of chromatin modifications enriched at TSSs and TTSs in kinetoplastids has cast doubt on the notion that these early-diverged eukaryotes lack RNAP II transcriptional regulation. Investigations of the function of chromatin modifications in kinetoplastids have revealed that at least two modifications, base J and H3.V, promote RNAP II termination and repress genes at the end of gene clusters. It will be important for future studies to further elucidate the biological role of gene cluster internal termination, addressing for example the extent to which these epigenetic mechanisms are utilized by kinetoplastids to effect gene expression changes during developmental progression. Similarly, additional studies are needed to determine whether other chromatin modifications impact transcription in these parasites and how they might effect meaningful changes in gene expression. This will require continued efforts to determine the localization of other chromatin modifications, which in some cases will require the production of specific antisera. Comparisons between kinetoplastid species will also be important, particularly given the observation that the extent to which chromatin modifications impact transcriptional processes in *T. brucei*, *T. cruzi*, and *Leishmania* spp. appear to differ. Future investigations should also aim to better understand how multiple chromatin regulatory

mechanisms are integrated to influence gene expression. For example, cross talk between chromatin modifications has largely been unexplored in kinetoplastids and it is unknown if modifications function to recruit other chromatin modifications to establish epigenetic states. How environmental signals may impinge on epigenetic modifications and whether this effects changes in gene expression is also unknown. Given the prior lack of obvious transcriptional regulation, analysis of epigenetic mechanisms in kinetoplastids is still in its nascent stages. Yet, in a relatively brief period, in addition to their impact on transcription and gene expression, chromatin modifications have been linked to DNA replication, recombination, and telomeric repression in kinetoplastids. Therefore, continued analysis of epigenetic mechanisms will provide insight into the biology of these unusual organisms, potentially informing new therapeutic interventions to combat the diseases caused by kinetoplastid parasites.

REFERENCES

1. Desjeux, P. (2004) Leishmaniasis: current situation and new perspectives. *Comparative immunology, microbiology and infectious diseases*, **27**, 305-318.
2. Hide, G. and Tait, A. (2009) Molecular epidemiology of African sleeping sickness. *Parasitology*, **136**, 1491-1500.
3. Reithinger, R., Tarleton, R.L., Urbina, J.A., Kitron, U. and Gürtler, R.E. (2009) Eliminating Chagas disease: challenges and a roadmap. *BMJ*, **338**.
4. Souza, W. (2002) Basic Cell Biology of Trypanosoma cruzi. *Current Pharmaceutical Design*, **8**, 269-285.
5. Handman, E. (1999) *Cell Biology of Leishmania.* In J.R. Baker, R. M. and Rollinson, D. (eds.), *Advances in Parasitology*. Academic Press, Vol. Volume 44, pp. 1-39.

6. Matthews, K.R. (2005) The developmental cell biology of Trypanosoma brucei. *Journal of cell science*, **118**, 283-290.

7. Borst, P. and Ulbert, S. (2001) Control of VSG gene expression sites. *Molecular and biochemical parasitology*, **114**, 17-27.

8. Borst, P. (2002) Antigenic Variation and Allelic Exclusion. *Cell*, **109**, 5-8.

9. Matthews, K.R. (2015) 25 years of African trypanosome research: From description to molecular dissection and new drug discovery. *Molecular and biochemical parasitology*, **200**, 30-40.

10. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B. *et al.* (2005) The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416-422.

11. Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R. *et al.* (2005) The genome of the kinetoplastid parasite, Leishmania major. *Science*, **309**, 436-442.

12. El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G. *et al.* (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science*, **309**, 409-415.

13. Clayton, C.E. (2002) Life without transcriptional control? From fly to man and back again. *EMBO J*, **21**, 1881-1888.

14. Campbell, D.A., Thomas, S. and Sturm, N.R. (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.*, **5**, 1231-1240.

15. Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S. and Cross, G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev.*, **23**, 1063-1076.

16. Thomas, S., Green, A., Sturm, N.R., Campbell, D.A. and Myler, P.J. (2009) Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*, **10**, 152.

17. Respuela, P., Ferella, M., Rada-Iglesias, A. and Aslund, L. (2008) Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. *J. Biol. Chem.*, **283**, 15884-15892.

18. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A. and Sabatini, R. (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res*, **38**, 3923-3935.

19. Lowell, J.E. and Cross, G.A. (2004) A variant histone H3 is enriched at telomeres in Trypanosoma brucei. *Journal of cell science*, **117**, 5937-5947.

20. Ekanayake, D.K., Cipriano, M.J. and Sabatini, R. (2007) Telomeric co-localization of the modified base J and contingency genes in the protozoan parasite Trypanosoma cruzi. *Nucleic Acids Res*, **35**, 6367-6377.

21. Akiyoshi, B. and Gull, K. (2014) Discovery of unconventional kinetochores in kinetoplastids. *Cell*, **156**, 1247-1258.

22. Haile, S. and Papadopoulou, B. (2007) Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Current opinion in microbiology*, **10**, 569-577.

23. Clayton, C. and Shapira, M. (2007) Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Molecular and biochemical parasitology*, **156**, 93-101.

24. De Gaudenzi, J.G., Noe, G., Campo, V.A., Frasch, A.C. and Cassola, A. (2011) Gene expression regulation in trypanosomatids. *Essays in biochemistry*, **51**, 31-46.

25. Kramer, S. (2012) Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. *Molecular and biochemical parasitology*, **181**, 61-72.

26. Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H.O., Bishop, J.R., Falcone, F.H., Gavrilescu, C., Montgomery, J.L., Santori, M.I. *et al.* (2000) A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *Rna*, **6**, 163-169.

27. Jaé, N., Wang, P., Gu, T., Hühn, M., Palfi, Z., Urlaub, H. and Bindereif, A. (2010) Essential Role of a Trypanosome U4-Specific Sm Core Protein in Small Nuclear Ribonucleoprotein Assembly and Splicing. *Eukaryotic cell*, **9**, 379-386.

28. Blumenthal, T. (2004) Operons in eukaryotes. *Briefings in functional genomics & proteomics*, **3**, 199-211.

29. Kelly, S., Kramer, S., Schwede, A., Maini, P.K., Gull, K. and Carrington, M. (2012) Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. *Open biology*, **2**, 120033.

30. El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404-409.

31. Blumenthal, T. and Gleason, K.S. (2003) Caenorhabditis elegans operons: form and function. *Nature reviews. Genetics*, **4**, 112-120.

32. Johnson, P.J., Kooter, J.M. and Borst, P. (1987) Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*, **51**, 273-281.

33. Mottram, J.C., Murphy, W.J. and Agabian, N. (1989) A transcriptional analysis of the Trypanosoma brucei hsp83 gene cluster. *Mol. Biochem. Parasitol.*, **37**, 115-127.

34. Martinez-Calvillo, S., Nguyen, D., Stuart, K. and Myler, P.J. (2004) Transcription initiation and termination on Leishmania major chromosome 3. *Eukaryotic cell*, **3**, 506-517.

35. Martinez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K. and Myler, P.J. (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Mol. Cell*, **11**, 1291-1299.

36. Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S. and Tschudi, C. (2010) The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog.*, **6**, e1001090.

37. Boothroyd, J.C. and Cross, G.A. (1982) Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. *Gene*, **20**, 281-289.

38. Van der Ploeg, L.H., Liu, A.Y., Michels, P.A., De Lange, T.D., Borst, P., Majumder, H.K., Weber, H., Veeneman, G.H. and Van Boom, J. (1982) RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes. *Nucleic Acids Res*, **10**, 3591-3604.

39. De Lange, T., Liu, A.Y., Van der Ploeg, L.H., Borst, P., Tromp, M.C. and Van Boom, J.H. (1983) Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? *Cell*, **34**, 891-900.

40. Nelson, R.G., Parsons, M., Barr, P.J., Stuart, K., Selkirk, M. and Agabian, N. (1983) Sequences homologous to the variant antigen mRNA spliced leader are located in tandem repeats and variable orphons in Trypanosoma brucei. *Cell*, **34**, 901-909.

41. Sutton, R.E. and Boothroyd, J.C. (1986) Evidence for Trans splicing in trypanosomes. *Cell*, **47**, 527-535.

42. Agabian, N. (1990) Trans splicing of nuclear pre-mRNAs. *Cell*, **61**, 1157-1160.

43. LeBowitz, J.H., Smith, H.Q., Rusche, L. and Beverley, S.M. (1993) Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev.*, **7**, 996-1007.

44. Smircich, P., Forteza, D., El-Sayed, N.M. and Garat, B. (2013) Genomic analysis of sequence-dependent DNA curvature in Leishmania. *PloS one*, **8**, e63068.

45. Saito, R.M., Elgort, M.G. and Campbell, D.A. (1994) A conserved upstream element is essential for transcription of the Leishmania tarentolae mini-exon gene. *The EMBO Journal*, **13**, 5460-5469.

46. Günzl, A., Ullu, E., Dörner, M., Fragoso, S.P., Hoffmann, K.F., Milner, J.D., Morita, Y., Nguu, E.K., Vanacova, S., Wünsch, S. *et al.* (1997) Transcription of the Trypanosoma brucei spliced leader RNA gene is dependent only on the presence of upstream regulatory elements. *Molecular and biochemical parasitology*, **85**, 67-76.

47. Luo, H. and Bellofatto, V. (1997) Characterization of Two Protein Activities That Interact at the Promoter of the Trypanosomatid Spliced Leader RNA. *Journal of Biological Chemistry*, **272**, 33344-33352.

48. Martinez-Calvillo, S., Vizuet-de-Rueda, J.C., Florencio-Martinez, L.E., Manning-Cela, R.G. and Figueroa-Angulo, E.E. (2010) Gene expression in trypanosomatid parasites. *J Biomed Biotechnol*, 525241.

49. Lee, T.I. and Young, R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, **34**, 77-137.

50. Evers, R., Hammer, A., Köck, J., Jess, W., Borst, P., Mémet, S. and Cornelissen, A.W.C.A. (1989) Trypanosoma brucei contains two RNA polymerase II largest subunit genes with an altered C-terminal domain. *Cell*, **56**, 585-597.

51. Bellofatto, V., Torres-Munoz, J.E. and Cross, G.A. (1991) Stable transformation of Leptomonas seymouri by circular extrachromosomal elements. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6711-6715.

52. Papadopoulou, B., Roy, G. and Ouellette, M. (1994) Autonomous replication of bacterial DNA plasmid oligomers in Leishmania. *Molecular & Biochemical Parasitology*, **65**, 39-49.

53. Marchetti, M.A., Tschudi, C., Silva, E. and Ullu, E. (1998) Physical and transcriptional analysis of the Trypanosoma brucei genome reveals a typical eukaryotic arrangement with close interspersionof RNA polymerase II- and III-transcribed genes. *Nucleic Acids Res*, **26**, 3591-3598.

54. Ekanayake, D. and Sabatini, R. (2011) Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryotic cell*, **10**, 1465-1472.

55. Reynolds, D., Cliffe, L., Forstner, K.U., Hon, C.C., Siegel, T.N. and Sabatini, R. (2014) Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res*, **42**, 9717-9729.

56. Reynolds, D., Hofmeister, B.T., Cliffe, L., Alabady, M., Siegel, T.N., Schmitz, R.J. and Sabatini, R. (2016) Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005758.

57. Schulz, D., Zaringhalam, M., Papavasiliou, F.N. and Kim, H.-S. (2016) Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005762.

58. Porrua, O. and Libri, D. (2015) Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature reviews. Molecular cell biology*, **16**, 190-202.

59. Kuehner, J.N., Pearson, E.L. and Moore, C. (2011) Unravelling the means to an end: RNA polymerase II transcription termination. *Nature reviews. Molecular cell biology*, **12**, 283-294.

60. Dillon, Laura A.L., Okrah, K., Hughitt, V.K., Suresh, R., Li, Y., Fernandes, M.C., Belew, A.T., Corrada Bravo, H., Mosser, D.M. and El-Sayed, N.M. (2015) Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. *Nucleic Acids Res*, **43**, 6799-6813.

61. Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I. and Ochsenreiter, T. (2010) Spliced leader trapping reveals widespread alternative

splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS pathogens*, **6**, e1001037.

62. Rastrojo, A., Carrasco-Ramiro, F., Martín, D., Crespillo, A., Reguera, R.M., Aguado, B. and Requena, J.M. (2013) The transcriptome of Leishmania majorin the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. *BMC Genomics*, **14**, 1-13.

63. Gunzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L.C., Chung, H.M., Lee, P.T. and Lee, M.G. (2003) RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Eukaryotic cell*, **2**, 542-551.

64. Cross, G.A.M., Kim, H.-S. and Wickstead, B. (2014) Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. *Molecular and biochemical parasitology*, **195**, 59-73.

65. Biebinger, S., Rettenmaier, S., Flaspohler, J., Hartmann, C., Pena-Diaz, J., Wirtz, L.E., Hotz, H.R., Barry, J.D. and Clayton, C. (1996) The PARP promoter of Trypanosoma brucei is developmentally regulated in a chromosomal context. *Nucleic Acids Research*, **24**, 1202-1211.

66. McCulloch, R., Morrison, L.J. and Hall, J.P. (2015) DNA Recombination Strategies During Antigenic Variation in the African Trypanosome. *Microbiology spectrum*, **3**, Mdna3-0016-2014.

67. Horn, D. and McCulloch, R. (2010) Molecular mechanisms underlying the control of antigenic variation in African trypanosomes. *Current opinion in microbiology*, **13**, 700-705.

68. Hertz-Fowler, C., Figueiredo, L.M., Quail, M.A., Becker, M., Jackson, A., Bason, N., Brooks, K., Churcher, C., Fahkro, S., Goodhead, I. *et al.* (2008) Telomeric expression sites are highly conserved in Trypanosoma brucei. *PloS one*, **3**, e3527.

69. Redpath, M.B., Windle, H., Nolan, D., Pays, E., Voorheis, H.P. and Carrington, M. (2000) ESAG11, a new VSG expression site-associated gene from Trypanosoma brucei. *Molecular and biochemical parasitology*, **111**, 223-228.

70. Pays, E., Lips, S., Nolan, D., Vanhamme, L. and Perez-Morga, D. (2001) The VSG expression sites of Trypanosoma brucei: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Molecular and biochemical parasitology*, **114**, 1-16.

71. Paindavoine, P., Rolin, S., Van Assel, S., Geuskens, M., Jauniaux, J.C., Dinsart, C., Huet, G. and Pays, E. (1992) A gene from the variant surface glycoprotein expression site encodes one of several transmembrane adenylate cyclases located on the flagellum of Trypanosoma brucei. *Molecular and cellular biology*, **12**, 1218-1225.

72. Ligtenberg, M.J., Bitter, W., Kieft, R., Steverding, D., Janssen, H., Calafat, J. and Borst, P. (1994) Reconstitution of a surface transferrin binding complex in insect form Trypanosoma brucei. *EMBO Journal*, **13**, 2565-2573.

73. Salmon, D., Hanocq-Quertier, J., Paturiaux-Hanocq, F., Pays, A., Tebabi, P., Nolan, D.P., Michel, A. and Pays, E. (1997) Characterization of the ligand-binding site of the transferrin receptor in Trypanosoma brucei demonstrates a structural relationship with the N-terminal domain of the variant surface glycoprotein. *EMBO J*, **16**, 7272-7278.

74. Hoek, M., Engstler, M. and Cross, G.A. (2000) Expression-site-associated gene 8 (ESAG8) of Trypanosoma brucei is apparently essential and accumulates in the nucleolus. *Journal of cell science*, **113 (Pt 22)**, 3959-3968.

75. Hoek, M., Zanders, T. and Cross, G.A. (2002) Trypanosoma brucei expression-site-associated-gene-8 protein interacts with a Pumilio family protein. *Molecular and biochemical parasitology*, **120**, 269-283.

76. Gottesdiener, K.M. (1994) A new VSG expression site-associated gene (ESAG) in the promoter region of Trypanosoma brucei encodes a protein with 10 potential transmembrane domains. *Molecular and biochemical parasitology*, **63**, 143-151.

77. Florent, I.C., Raibaud, A. and Eisen, H. (1991) A family of genes related to a new expression site-associated gene in Trypanosoma equiperdum. *Molecular and cellular biology*, **11**, 2180-2188.

78. Carruthers, V.B., Navarro, M. and Cross, G.A. (1996) Targeted disruption of expression site-associated gene-1 in bloodstream-form Trypanosoma brucei. *Molecular & Biochemical Parasitology*, **81**, 65-79.

79. Morgan, R.W., El-Sayed, N.M., Kepa, J.K., Pedram, M. and Donelson, J.E. (1996) Differential expression of the expression site-associated gene I family in African trypanosomes. *The Journal of biological chemistry*, **271**, 9771-9777.

80. Callejas, S., Leech, V., Reitter, C. and Melville, S. (2006) Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length. *Genome research*, **16**, 1109-1118.

81. Melville, S.E., Leech, V., Navarro, M. and Cross, G.A. (2000) The molecular karyotype of the megabase chromosomes of Trypanosoma brucei stock 427. *Molecular and biochemical parasitology*, **111**, 261-273.

82. Felsenfeld, G. (2014) A Brief History of Epigenetics. *Cold Spring Harbor Perspectives in Biology*, **6**.

83. Figueiredo, L.M., Cross, G.A. and Janzen, C.J. (2009) Epigenetic regulation in African trypanosomes: a new kid on the block. *Nat Rev Microbiol*, **7**, 504-513.

84. Horn, D. (2007) Introducing histone modification in trypanosomes. *Trends in parasitology*, **23**, 239-242.

85. Schenkman, S. and da Cunha, J.P. (2007) Response to Horn: Introducing histone modifications in trypanosomes. *Trends in parasitology*, **23**, 242-243.

86. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693-705.

87. Alsford, S., duBois, K., Horn, D. and Field, M.C. (2012) Epigenetic mechanisms, nuclear architecture and the control of gene expression in trypanosomes. *Expert reviews in molecular medicine*, **14**, e13.

88. Croken, M.M., Nardelli, S.C. and Kim, K. (2012) Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives. *Trends in parasitology*, **28**, 202-213.

89. Rudenko, G. (2010) Epigenetics and transcriptional control in African trypanosomes. *Essays in biochemistry*, **48**, 201-219.

90. Yang, X., Figueiredo, L.M., Espinal, A., Okubo, E. and Li, B. (2009) RAP1 is essential for silencing telomeric variant surface glycoprotein genes in Trypanosoma brucei. *Cell*, **137**, 99-109.

91. Astolfi Filho, S., Martins de Sa, C. and Gander, E.S. (1980) On the chromatin structure of Trypanosoma cruzi. *Molecular and biochemical parasitology*, **1**, 45-53.

92. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**, 251-260.

93. Janzen, C.J., Fernandez, J.P., Deng, H., Diaz, R., Hake, S.B. and Cross, G.A. (2006) Unusual histone modifications in Trypanosoma brucei. *FEBS Lett*, **580**, 2306-2310.

94. da Cunha, J.P., Nakayasu, E.S., de Almeida, I.C. and Schenkman, S. (2006) Post-translational modifications of Trypanosoma cruzi histone H4. *Molecular and biochemical parasitology*, **150**, 268-277.

95. Mandava, V., Fernandez, J.P., Deng, H., Janzen, C.J., Hake, S.B. and Cross, G.A. (2007) Histone modifications in Trypanosoma brucei. *Molecular and biochemical parasitology*, **156**, 41-50.

96. Strahl, B.D. and Allis, C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41-45.

97. Winston, F. and Allis, C.D. (1999) The bromodomain: a chromatin-targeting module? *Nature structural biology*, **6**, 601-604.

98. Law, C. and Cheung, P. (2013) *Histone Variants and Transcription Regulation.* In Kundu, T. K. (ed.), *Epigenetics: Development and Disease*. Springer Netherlands, Vol. 61, pp. 319-341.

99. Talbert, P.B. and Henikoff, S. (2010) Histone variants--ancient wrap artists of the epigenome. *Nature reviews. Molecular cell biology*, **11**, 264-275.

100. Weber, C.M. and Henikoff, S. (2014) Histone variants: dynamic punctuation in transcription. *Genes & development*, **28**, 672-682.

101. Lowell, J.E., Kaiser, F., Janzen, C.J. and Cross, G.A. (2005) Histone H2AZ dimerizes with a novel variant H2B and is enriched at repetitive DNA in Trypanosoma brucei. *Journal of cell science*, **118**, 5721-5730.

102. Anderson, B.A., Wong, I.L., Baugh, L., Ramasamy, G., Myler, P.J. and Beverley, S.M. (2013) Kinetoplastid-specific histone variant functions are conserved in Leishmania major. *Mol. Biochem. Parasitol.*, **191**, 53-57.

103. Akiyoshi, B. and Gull, K. (2013) Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open biology*, **3**, 130023.

104. Borst, P. and Sabatini, R. (2008) Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol*, **62**, 235-251.

105. Cliffe, L.J., Hirsch, G., Wang, J., Ekanayake, D., Bullard, W., Hu, M., Wang, Y. and Sabatini, R. (2012) JBP1 and JBP2 Proteins Are Fe2+/2-Oxoglutarate-dependent Dioxygenases Regulating Hydroxylation of Thymidine Residues in Trypanosome DNA. *J. Biol. Chem.*, **287**, 19886-19895.

106. Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y. and Sabatini, R. (2014) Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. *J. Biol. Chem.*, **289**, 20273-20282.

107. Sekar, A., Merritt, C., Baugh, L., Stuart, K. and Myler, P.J. (2014) Tb927.10.6900 encodes the glucosyltransferase involved in synthesis of base J in Trypanosoma brucei. *Mol. Biochem. Parasitol.*, **196**, 9-11.

108. Cross, M., Kieft, R., Sabatini, R., Wilm, M., de Kort, M., van der Marel, G., van Boom, J., van Leeuwen, F. and Borst, P. (1999) The modified base J is the target for a novel DNA-binding protein in kinetoplastid protozoans. *EMBO Journal*, **18**, 6573-6581.

109. DiPaolo, C., Kieft, R., Cross, M. and Sabatini, R. (2005) Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Molecular cell*, **17**, 441-451.

110. Iyer, L.M., Tahiliani, M., Rao, A. and Aravind, L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698-1710.

111. Pays, E., Delauw, M.F., Laurent, M. and Steinert, M. (1984) Possible DNA modification in GC dinucleotides of Trypanosoma brucei telomeric sequences; relationship with antigen gene transcription. *Nucleic Acids Res*, **12**, 5235-5247.

112. Bernards, A., van Harten-Loosbroek, N. and Borst, P. (1984) Modification of telomeric DNA in Trypanosoma brucei; a role in antigenic variation? *Nucleic Acids Research*, **12**, 4153-4170.

113. Bernards, A., De Lange, T., Michels, P.A., Liu, A.Y., Huisman, M.J. and Borst, P. (1984) Two modes of activation of a single surface antigen gene of Trypanosoma brucei. *Cell*, **36**, 163-170.

114. van Leeuwen, F., Wijsman, E.R., Kieft, R., van der Marel, G.A., van Boom, J.H. and Borst, P. (1997) Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes & development*, **11**, 3232-3241.

115. van Leeuwen, F., Kieft, R., Cross, M. and Borst, P. (1998) Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Molecular & Cellular Biology*, **18**, 5643-5651.

116. Genest, P.A., Ter Riet, B., Cijsouw, T., van Luenen, H.G. and Borst, P. (2007) Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania. *Nucleic Acids Res*, **35**, 2116-2124.

117. van Luenen, H.G., Farris, C., Jan, S., Genest, P.A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G. *et al.* (2012) Glucosylated

hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, **150**, 909-921.

118. Ekanayake, D.K., Minning, T., Weatherly, B., Gunasekera, K., Nilsson, D., Tarleton, R., Ochsenreiter, T. and Sabatini, R. (2011) Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. *Mol. Cell. Biol.*, **31**, 1690-1700.

119. Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J. and Cramer, P. (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, **155**, 1075-1087.

120. Grzechnik, P., Tan-Wong, S.M. and Proudfoot, N.J. (2014) Terminate and make a loop: regulation of transcriptional directionality. *Trends in biochemical sciences*, **39**, 319-327.

121. Kim, K.-Y. and Levin, David E. (2011) Mpk1 MAPK Association with the Paf1 Complex Blocks Sen1-Mediated Premature Transcription Termination. *Cell*, **144**, 745-756.

122. Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vítor, A.C., Desterro, J.M.P., Carmo-Fonseca, M. and de Almeida, S.F. (2015) Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *eLife*, **4**, e09214.

123. Sturm, N.R., Yu, M.C. and Campbell, D.A. (1999) Transcription termination and 3'-End processing of the spliced leader RNA in kinetoplastids. *Molecular and cellular biology*, **19**, 1595-1604.

124. Hull, M.W., Erickson, J., Johnston, M. and Engelke, D.R. (1994) tRNA genes as transcriptional repressor elements. *Molecular and cellular biology*, **14**, 1266-1277.

125. Good, P.D., Kendall, A., Ignatz-Hoover, J., Miller, E.L., Pai, D.A., Rivera, S.R., Carrick, B. and Engelke, D.R. (2013) Silencing near tRNA genes is nucleosome-mediated and distinct from boundary element function. *Gene*, **526**, 7-15.

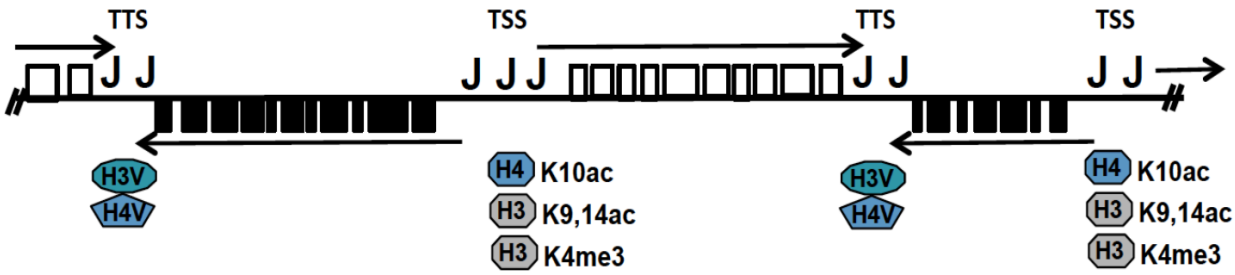


Figure 1.1 The organization of genes into polycistronically transcribed gene clusters in kinetoplastid genomes. Genes are indicated by boxes, those encoded on the top DNA strand in white and those on the bottom strand in black. TSS, RNAP II transcription start site; TTS, RNAP II transcription termination site. The direction of RNAP II transcription is indicated by solid arrows. The DNA modification base J is enriched in TSSs and TTSs. Also enriched at TSSs are histone H4K10 acetylation, histone H3K9 and H3K14 acetylation, and H3K4 trimethylation. The histone variants H3.V and H4.V are enriched at TTSs.

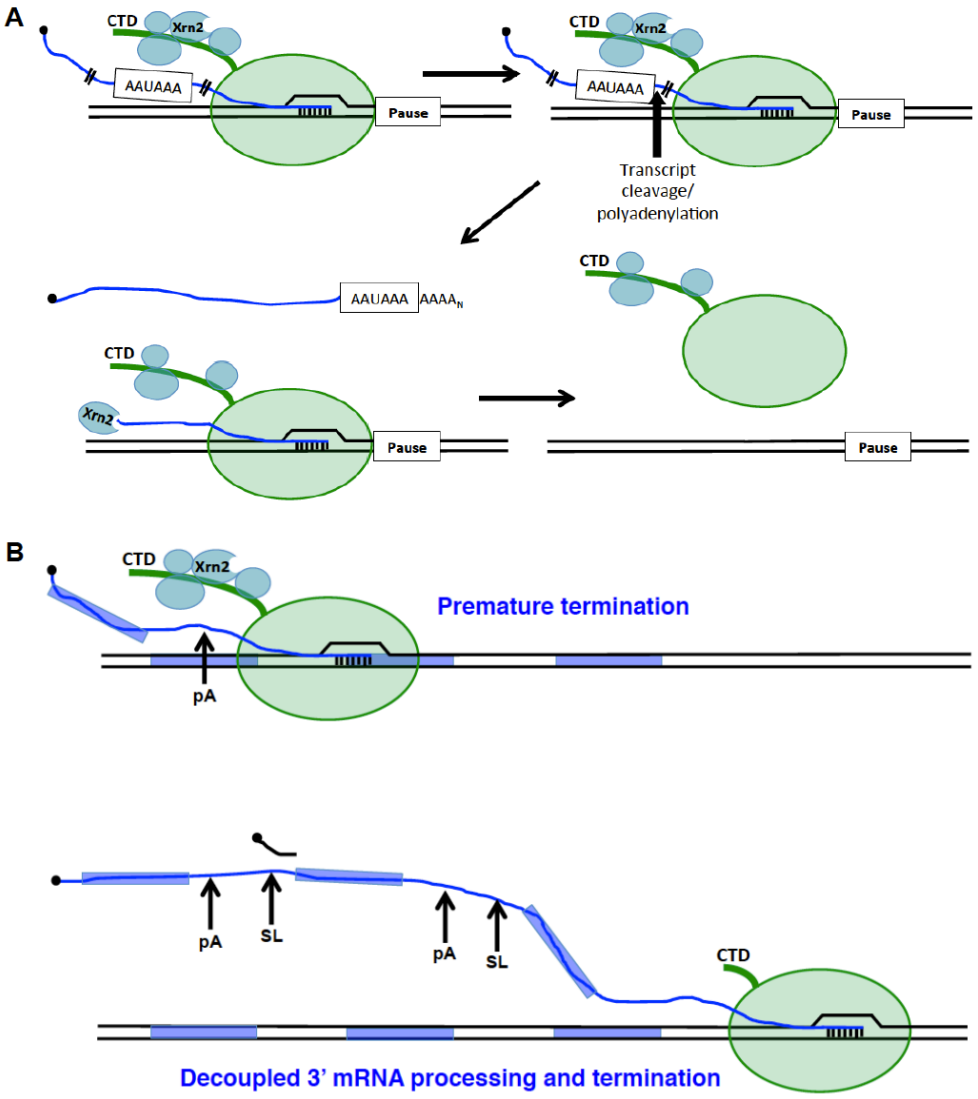


Figure 1.2 RNAP II transcription termination. (A) In humans, RNAP II (green oval) transcribes through the polyadenylation signal sequence (AAUAAA in the RNA in blue), which is recognized by 3' end formation factors (not shown) associated with the RNAP II CTD. The RNA is cleaved and polyadenylated. In the allosteric model of termination, events associated with 3' end formation promote RNAP II pausing and termination. In the torpedo model, the unmodified 5' phosphate on the nascent RNA still associated with the RNAP II complex is targeted by the 5'-3' exonuclease Xrn2, which degrades the nascent RNA and eventually promotes the release of RNAP II from the DNA template. (B) In kinetoplastids, if RNAP II termination was linked to 3' end formation, then premature termination would occur after transcription of a single gene within a cluster. Instead, 3' end formation is decoupled from termination and is linked to the process of trans-splicing of the 5' end of the downstream gene.

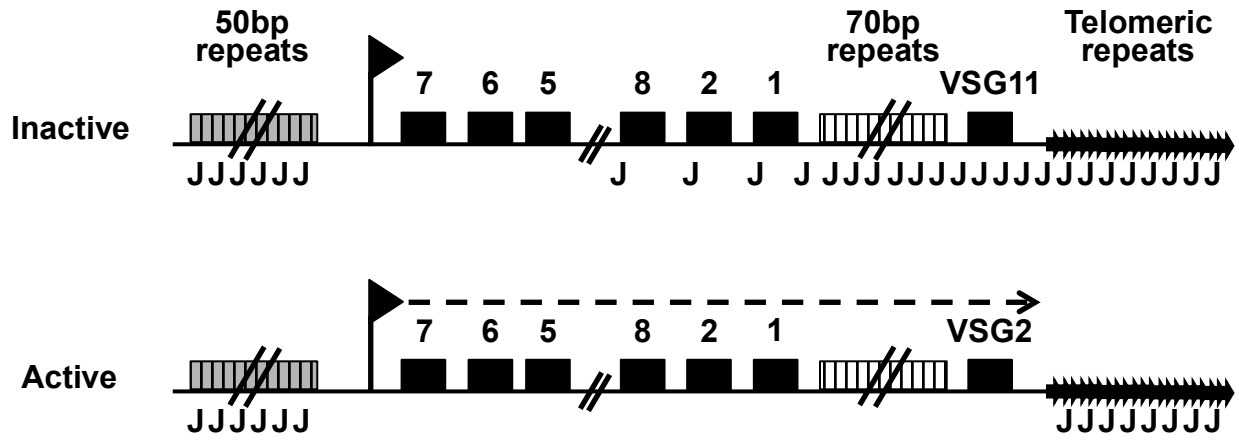


Figure 1.3 Subtelomeric expression sites in *T. brucei*. Top, a representative inactive expression site is shown. Expression site associated genes and the silent *VSG* are indicated by black boxes. The location of 50bp, 70bp, and telomeric repeats are also shown. The RNAP I promoter is indicated by a flag downstream of the 50bp repeats. Bottom, in the active expression site RNAP I initiates at the promoter (dashed arrow indicates transcription) and productively transcribes through the expression site associated genes and the active *VSG* located adjacent to the telomeric repeats.

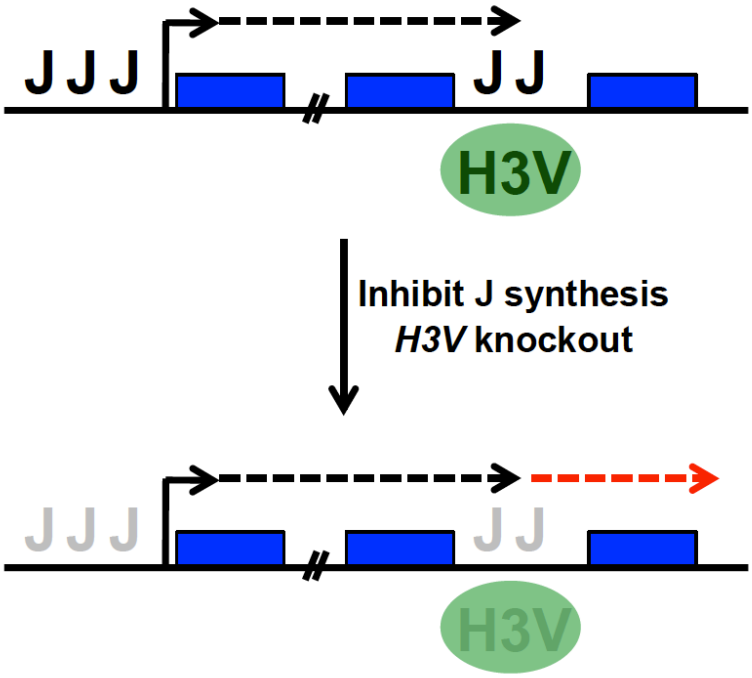


Figure 1.4 Gene cluster internal termination. Top, in *T. brucei* both J and H3.V independently promote RNAP II termination prior to the end of some gene clusters, repressing the expression of the downstream gene(s). Transcription is indicated by a dashed arrow. Bottom, upon the loss of J or H3.V in *T. brucei*, RNAP II read through transcription occurs, indicated by a red dashed arrow, increasing the expression of the downstream gene(s). The combined loss of J and H3.V results in the greatest read through and increase in gene expression. J similarly promotes gene cluster internal termination in *L. major,* but not H3.V.

CHAPTER 2

2-OXOGLUTARATE-DEPENDENT HYDROXYLASES INVOLVED IN DNA BASE J (β-D-GLUCOPYRANOSYLOXYMETHYLURACIL) SYNTHESIS[1]

---

[1]**Reynolds D.**, Cliffe L., and Sabatini R. 2015. *In* C. J. Schofield and R. P. Hausinger (Eds.) 2-Oxoglutarate-Dependent Dioxygenases, Royal Society of Chemistry, Cambridge, U.K.

Reprinted here with permission of publisher

INTRODUCTION TO BASE J LOCALIZATION AND FUNCTION

Base J (β-D-glucopyranosyloxymethyluracil) is an evolutionarily conserved nuclear DNA modification consisting of an *O*-linked glucosylation of thymidine (Figure 2.1). This DNA modification is found within members of the kinetoplastid family, a group of early-diverged unicellular eukaryotes including *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania* spp. Base J was initially discovered in the African trypanosome *T. brucei*, the causative agent of Human African Trypanosomiasis (sleeping sickness).(1) The key regulatory step of base J synthesis is the initial hydroxylation of thymidine at specific sites within the genome by Fe(II)/2-oxoglutarate (2OG)-dependent dioxygenases (JBP1 and JBP2). In this chapter we highlight our current knowledge of the structure and function of these two dioxygenases, focusing on how they regulate base J synthesis and how this regulation has been pivotal in unraveling the function of this unusual epigenetic mark.

The early history of base J discovery, biosynthesis, and function have been reviewed in detail.(2,3) Therefore, we begin by briefly discussing recent advances in J localization and function focusing primarily on *T. brucei*, the organism in which the majority of the work on J biosynthesis has been carried out. We also briefly describe work performed in two related kinetoplastids, *T. cruzi* and *L. tarentolae*, which has helped to define the function of the modified base in the regulation of gene expression.

Base J Localization

Base J localizes to repetitive DNA sequences, namely telomeric repeats, as well as the 5S rRNA repeats, mini-exon repeats, and centromeric DNA repeats.(2,3) Specific to *T. brucei*, base J localizes to specialized subtelomeric sites, termed expression sites (ESs) (Figure 2.2), which

are important for immune evasion in the mammalian host. J is also enriched at other repetitive sequences in the *T. brucei* genome including the 50-bp and 70-bp repeats within ESs and the 177-bp repeats in minichromosomes.(4)

Technological advances in DNA sequencing have enabled the genome-wide characterization of J localization by anti-base J immunoprecipitation (IP) and high-throughput sequencing. This analysis led to the discovery of a minor fraction of base J at sites of RNA polymerase II (Pol II) transcription initiation and termination (Figure 2.2) in *T. brucei*, *T. cruzi*, *L. major*, and *L. tarentolae*.*(5-7)* The resolution of J localization has been further improved using single-molecule, real-time (SMRT) DNA sequencing, which allows for strand-specific, base-resolution detection of modified bases in DNA.(8) The localization of J determined by SMRT sequencing of the *T. brucei* genome closely matches the J-IP sequencing analyses, thus localizing the modified base at regions involved in transcription initiation, elongation, and termination as well as repetitive DNA sequences, including telomeric and centromeric DNA repeats (Figure 2.2 and unpublished data). Although no conserved DNA motif associated with J synthesis was identified, SMRT sequencing results closely match previously identified DNA strand and sequence bias of J within the telomeric repeat.(9)

Base J Function

Base J was initially described in *T. brucei* based on its localization within the variant surface glycoprotein (*VSG*) gene ESs, which are crucial for parasite growth in the mammalian host.(10) African trypanosomes are extracellular parasites that replicate in the bloodstream of their mammalian host, and are therefore under constant exposure from the host adaptive immune response. *T. brucei* evades elimination by the host through antigenic variation, a process which

involves switching the *VSG* produced by the parasite.(11) The *T. brucei* genome contains over 1000 *VSG* genes, but only one of these genes is expressed at a time. Periodic switching of the actively expressed *VSG* gene maintains a population of trypanosomes capable of avoiding antibody recognition, thus avoiding the adaptive immune system of the host. Monoallelic *VSG* expression is achieved through regulated transcription from ESs. Although there are 15 ESs, only one is active at a time, whilst the remainder are silenced. Interestingly, J is found in the silent, but not the active, ES.(12) Furthermore, transcriptional activation of a silent ES results in the loss of J from the site, while the previously active ES is modified to contain J. The presence of J only in mammalian, but not insect stage, parasites,(13) as well as its association with silent *VSG* gene expression sites supports the hypothesis that J plays a role in the regulation of antigenic variation in *T. brucei*. However, to date little experimental evidence exists to support this hypothesis.

Unlike *T. brucei*, base J is found in all developmental stages of *T. cruzi* and *Leishmania* spp.(14) This finding, along with the apparent essential nature of J in kinetoplastids that do not undergo antigenic variation,(6,15) indicates a broad functional role for the modified base. Although the reduction of base J has not revealed any telomeric defects, the identification of J at Pol II initiation and termination sites suggests the modification is involved in regulating gene expression. Indeed, the loss of base J from these chromosome-internal regions in *T. cruzi* and *L. tarentolae* alters Pol II transcription and leads to global gene expression changes.(6,7)

Reduction of J in *T. cruzi* upon deleting one of the enzymes involved in J synthesis results in chromatin changes at Pol II initiation sites including decreased nucleosome abundance and increased histone acetylation, a mark associated with active chromatin.(16) Pol II recruitment to initiation sites and transcription rates both increase upon J loss, resulting in global changes in gene expression and parasite virulence.(6,16) Similar decreases of base J in *L.*

tarentolae also lead to genome-wide changes in gene expression, possibly due, in part, to altered transcription initiation.(7) However, the loss of base J at transcription termination sites clearly leads to a termination defect, with subsequent read-through transcription and the generation of antisense RNAs. Thus, base J represents a novel epigenetic modification of kinetoplastid DNA involved in regulating transcription and gene expression.

Unique for eukaryotes, transcription in kinetoplastids is polycistronic; i.e., multiple genes are arranged in long units that are co-transcribed. Individual mRNAs are processed from the pre-mRNA through the addition of a mini-exon sequence, called the spliced leader, and polyadenylation.(17) Unlike bacterial operons, however, kinetoplastid polycistronic gene clusters do not contain functionally related genes. While the loss of base J clearly affects gene expression in *T. cruzi* and *L. tarentolae*, how and whether kinetoplastids employ J to fine-tune Pol II transcription remains unknown. Since the two dioxygenases JBP1 and JBP2 appear to be the key regulators of J synthesis, understanding how they govern the synthesis and localization of J in the genome is important in fully understanding the role of J in controlling kinetoplastid gene expression.

THE TWO-STEP BIOSYNTHESIS PATHWAY

J is synthesized in a two-step pathway (Figure 2.1). The first step involves the oxidation of specific thymidine residues in DNA by a thymidine hydroxylase (TH, JBP1 or JBP2), forming the intermediate base hydroxymethyluracil (hmU). This intermediate is converted into base J by addition of a glucose molecule by a glucosyltransferase (GT). Once the two-step mechanism of J synthesis was defined it became clear that the enzymes involved are unique to the J synthesis pathway. The J-specific GT is unusual in that it operates in the nucleus, unlike the predominantly

cytoplasmic localization of other known GTs. Thymine hydroxylases that oxidize the free base in the pyrimidine salvage pathway are well characterized, but THs that oxidize the base in DNA were unknown prior to the discovery of base J. Following characterization of the JBP dioxygenases, homology based searches resulted in the identification of the ten eleven translocation (TET) oxygenases in mammalian cells.(18) The TETs hydroxylate 5-methylcytosine (5meC) in DNA by a mechanism analogous to the JBPs (see Chapter 11).(19) Based on their sequence similarity, TH and TET proteins have been grouped together in the TET/JBP subfamily of dioxygenases (see Chapter 11).(20,21)

There is an abundance of evidence in support of the two-step synthesis pathway. The specific localization of J within the genome indicates that thymidine residues are modified in DNA, rather than at the nucleotide level followed by incorporation during DNA replication. This result is in contrast to the related glucosyl-hydroxymethylcytosine (hmC) base in T-even bacteriophage where synthesis of the modified hmC at the nucleotide level results in glucosylation of up to 90% of the thymidine in the genome.(22) In addition, the detection of hmU in the DNA of bloodstream form trypanosomes by post-labeling and thin layer chromatographic analysis,(23) and by mass spectrometry (unpublished data), indicates it is a freely available intermediate. Finally, growth of trypanosomes in media containing hmU allows cells to bypass the first step of the synthesis pathway.(5,24) This circumvention results in the synthesis of J at random sites within the genome, implying that the J-specific GT is non-specific for DNA sequence and is able to glycosylate hmU present anywhere in the genome. Furthermore, it follows that the regulation of J biosynthesis and localization occurs primarily at the level of the TH enzymes. While these data provide indirect support for the two-step pathway,

unambiguous support is provided by recent detailed *in vitro* and *in vivo* characterization of the enzymes that catalyze each step in J synthesis.

Characterization of the Two Dioxygenases in J Biosynthesis

JBP1 was initially discovered based on its ability to bind J-containing DNA substrates.(25) The ability of JBP1 to bind base J is dependent on the *C*-terminal domain (Figure 2.3). *In silico* screening of the *T. brucei* genome led to the identification of JBP2, based on its homology to the *N*-terminus of JBP1.(26) Unlike JBP1, JBP2 is unable to bind base J; instead, JBP2 contains a domain at its *C*-terminus homologous to the SWI2/SNF2 family of chromatin remodeling ATPases. The importance of both JBP1 and JBP2 for J synthesis has been confirmed *in vivo*, through gene deletion studies in bloodstream form *T. brucei*. The loss of either JBP1 or JBP2 results in a 20- and 8-fold reduction in J levels in the genome, respectively.(27,28) Re-expression of the *JBP* gene in the corresponding knock out cell line rescued J levels. The simultaneous deletion of genes encoding both JBP1 and JBP2 generated a trypanosome cell line unable to synthesize base J.(29) JBP null trypanosomes fed hmU can convert this base to J, illustrating that the JBP enzymes are critical for the first step of J biosynthesis.

These results suggest JBP1 and JBP2 are directly involved in thymidine oxidation and their related *N*-terminus regions might contain a TH domain. Upon close examination it is found that the conserved *N*-terminus shares weak homology with enzymes of the Fe(II)/2-oxoglutarate (2OG)-dependent dioxygenase superfamily,(30) where hydroxylation is driven by the oxidative decarboxylation of 2OG to form succinate and CO_2. These enzymes catalyze the oxidation of a wide variety of substrates using ferrous ion as cofactor and 2OG plus molecular oxygen as co-substrates. Enzymes in the dioxygenase superfamily are typically identified on a structural level

by the presence of a jelly roll or double-stranded beta helix fold (see Chapter 2), which contains 4 key conserved residues involved in the binding of Fe(II) and 2OG that are essential for catalytic activity (Figure 2.3).(31,32) Substitution of any of the four conserved residues for alanine abolishes the ability of both JBP1 and JBP2 to stimulate *de novo* J synthesis *in vivo*.(29,30,33) This loss of activity is not due to the inability of the mutant JBP to enter the nucleus or, in the case of JBP1, to bind J-DNA.

To unambiguously characterize the JBPs as Fe(II)/2OG-dependent dioxygenases we developed an *in vitro* TH assay, using recombinant JBP1 produced in *E. coli*. This assay demonstrated that JBP1 stimulates the hydroxylation of thymidine specifically in the context of dsDNA and, as expected with this family of enzymes, was dependent on Fe(II), 2OG, and O_{2}.(34) Under anaerobic conditions, the addition of Fe(II) to JBP1 and 2OG results in the formation of a broad absorption spectrum centered at 530 nm attributed to metal chelation by 2OG that is bound to JBP, a spectroscopic signature of Fe(II)/2OG-dependent dioxygenases. The *N*-terminal TH domain of JBP1 is sufficient for full activity and mutation of residues involved in coordinating Fe(II) inhibit iron binding and the formation of hmU.(34) Thymidine oxidation is inhibited both *in vitro* and *in vivo* by using previously identified 2OG-dependent dioxygenase inhibitors. For example dimethyloxalylglycine (DMOG) is taken up by cells and undergoes ester hydrolysi to form *N*-oxalylglycine, a well-characterized inhibitor of these enzymes.(35). *In vitro* TH reactions performed in the presence of *N*-oxalylglycine demonstrate a significant reduction in thymidine-to-hmU conversion. Moreover, the growth of *T. brucei* in the presence of DMOG results in a complete loss of detectable J.(34) The inhibition of *in vivo* J synthesis by DMOG was recapitulated in *T. cruzi* and *L. major*.

Identification of the Glucosyl Transferase

The culturing of insect-stage *T. brucei* or the bloodstream-form J null (JBP1 and JBP2 KO) in media containing hmU results in the synthesis of base J randomly throughout the genome.(24) Therefore, the GT functions regardless of the DNA sequence context and is expressed in both trypanosome life-stages. The significantly higher levels of J stimulated by hmU addition to bloodstream J-null cells compared to insect-stage cells suggests down-regulation of the GT upon differentiation along with the JBPs. While attempts to purify the GT from nuclear extracts have failed,(2,36) a recent computational screen identified a GT-like sequence in the kinetoplastid genome as a strong candidate for the base J-associated GT. Iyer et al. (2013) identified gene sequences related to GT-A/fringe-like glycosyltransferases with operonic association with TET/JBP-like encoding genes that are predicted to modify DNA or RNA in certain phages.(21) The operonic association with TET/JBP enzymes suggests that these TET/JBP-associated glycosyltransferases could glycosylate hydroxylated bases generated by the former enzyme. We have recently confirmed that the GT homologue in kinetoplastids is the base J-associated GT, which we now refer to as JGT (unpublished data). We have found that recombinant JGT whose gene was cloned from *T. brucei* and *L. major* utilizes UDP-Glc to transfer glucose to hmU in the context of dsDNA. The deletion of both alleles of *JGT* from *T. brucei* generates a cell line that completely lacks base J and re-expression of *JGT* restores J synthesis. These studies confirm the identity of the J-specific GT and the two-step J synthesis model. As discussed below, the identification of the GT provides a useful tool in understanding the regulation of DNA modification by the two dioxygenases.

REGULATION OF J SYNTHESIS BY TWO THYMIDNE HYDROXYLASES

JBP Structure

The genomes of all J containing organisms encode two distinct thymidine 2OG-dependent dioxygenases; JBP1 and JBP2. Although both proteins are capable of hydroxylating thymidine, the *C*-terminal domain differs significantly (Figure 2.3) which is strongly suggestive of a differential role for each in the J biosynthesis pathway. JBP1 has a J-DNA binding domain in the *C*-terminal half of the protein that is essential for function *in vivo*. Both gel shift and fluorescence anisotropy measurements reveal that JBP1 binds to J only in the context of double stranded DNA.(37,38) Optimal DNA binding requires at least five base pairs flanking J, where recognition is dependent upon J itself and the base immediately 5' of J.(39) JBP1 does not make any sequence specific contacts with the bases surrounding the modified base, but rather it recognizes J when presented in any sequence context. However, it appears that interactions with the base immediately 5' of J is essential for the proper orientation of the glucose moiety of the modified base. Analysis of JBP1 binding to various modified DNA substrates indicates that the phosphoryl oxygen of the base upstream of J locks the glucose moiety, via hydrogen bonds of the essential 2- and 3-hydroxyl groups, into an 'edge on' confirmation necessary for optimal JBP1 binding.(38) A crystal structure of the 160-residue DNA binding domain reveals a novel "helical bouquet" fold where a single Asp residue is essential for recognition of base J in DNA and JBP1 function *in vivo*.(40) Recent fluorescent polarization measurements indicate that JBP1 binds J-DNA in a two-step reaction where the protein undergoes a conformational change upon DNA binding.(41) Presumably, this conformational change allows the TH domain to come in proximity to the DNA.

JBP2 does not bind the modified base directly, but is able to bind chromatin in a base J independent manner.(26) Re-expression of *JBP2* in bloodstream-form cells that lack JBP1 and JBP2 leads to *de novo* J synthesis at specific sites of the genome.(5) We believe that interaction of JBP2 with its DNA substrate is driven by the *C*-terminal SWI2/SNF2 domain. Mutation of key residues within the ATPase region of the SWI2/SNF2 domain eliminates JBP2 function,(26) but whether ATP hydrolysis is required for JBP2 to bind to/remodel chromatin structure to potentiate oxidase activity is unknown. Moreover, it is unknown whether recognition of chromatin by JBP2 is driven at the level of DNA sequence or structure, or potentially via interactions with histone variants that co-localize with J.(5,42)

Base J synthesis is developmentally regulated in *T. brucei*; it is detectable in bloodstream-form, but not insect-stage, parasite DNA. Characterization of different life cycle stages reveals that developmental regulation of J synthesis is controlled at both steps of J synthesis. Insect-stage cells down-regulate the production of both dioxygenases and the GT enzyme. However, hmU feeding or ectoptic expression of JBP2 in insect-stage trypanosome results in J synthesis.(24,29) This result implies that the developmental regulation of J synthesis is governed primarily by the dioxygenases.

The results from ectopic expression of *JBP*s in insect-stage trypanosomes allow the proposition of a model for J synthesis where JBP2 is the key regulator of site-specific *de novo* J synthesis.(3,27) According to this model the role of JBP1 is to amplify and maintain the levels of the modified base. However, this model indicating a separation of function was modified upon realization that both JBP1 and JBP2 have *de novo* J synthesis capability in the bloodstream form trypanosome. Re-expression of each *JBP* in the JBP1/JBP2 KO cells stimulated high levels of base J.(5) Interestingly, the localization of J stimulated by each JBP is different. JBP2 stimulates

J at high levels within the telomeres whereas JBP1 stimulates *de novo* synthesis primarily at genome internal sites. Optimal J synthesis (level and localization) occurs only upon repression of both JBPs. Whilst both dioxygenases are able to stimulate *de novo* synthesis of J, the inability of JBP2 to bind J may be essential for maintaining *de novo* J synthesis. Analysis of JBP1 concentrations in the wild-type trypanosome nucleus indicates there are 30-60 fold more J residues in the genome than molecules of JBP1 per cell.(43) The large number of high affinity binding sites would restrict JBP1 function to specific regions within the genome, thus explaining the inability of JBP1 to stimulate *de novo* J synthesis in a telomere fragmentation assay.(27) In wild-type cells, telomeric cleavage results in the growth of a new telomere that contains J; however, in a JBP2 KO cell line, the new telomere lacks base J despite the presence of endogenous JBP1. Presumably, the remaining large number of high affinity JBP1 binding sites precludes any interactions with the newly generated telomeric array. In a wild-type cell, therefore, we believe that JBP2 provides specific basal J-DNA for high affinity JBP1 binding, in turn directing the localization of JBP1-stimulated J synthesis.

The proposed functional separation of each JBP, related to chromatin substrate preference and ability to bind base J, may help explain the evolutionary conservation of two dioxygenase enzymes in the biosynthesis pathway.

Replication-Independent Oxidation

Based upon its ability to bind base J, it was proposed that one key function of JBP1 is to maintain J following DNA replication.(25,26) SMRT-sequencing of J within the trypanosome genome indicates sequences are primarily hemi-modified (Figure 2.2 and unpublished data). This result indicates there is no strict co-replicative maintenance of J and that its genomic localization

occurs at the level of *de novo* J synthesis. This conclusion is supported by the finding that JBP stimulation of *de novo* J formation is replication independent.(5)

To date, no enzyme capable of removing the modified base has been identified.(44) This finding, in combination with hmU feeding experiments, has led to the proposal that J is lost via passive dilution during replication, rather than by active removal.(26,28) However, dynamic regulation of J is still possible by regulating hmU formation and its conversion to J. In combination with thymine to uracil (U) conversion in the pyrimidine salvage pathway, the recently identified activity of TET proteins has provided insight into how the JBPs could regulate hmU levels. In the thymine salvage pathway, the Fe(II)/2OG-dependent dioxygenases thymine hydroxylase converts the base to hmU, formyluracil (fU), and carboxyluracil (caU) through three successive oxidation reactions.(45) Thymine-to-uracil conversion is completed by isoorotate decarboxylase-mediated decarboxylation of caU.(45) Similarly TET-mediated oxidation of 5mC to hmC, fC and caC (Chapter 11), has led to the possibility that these cytosine analogs may play a role in dynamic regulation of 5mC.(46-48) The similarity in chemistry between thymine-to-uracil conversion, TET oxidation of 5mC, and oxidative modification of T in trypanosomes suggests the possibility that JBP enzymes can mediate iterative oxidation of thymidine (Figure 2.4). Preliminary (unpublished) data indicate they do. We have detected low levels of fU in the genome of wild-type trypanosomes. In addition, during the *in vitro* JBP1 TH assay, hmU is rapidly generated and then lost over time. Inducible ablation of GT mRNA *in vivo* by using RNAi leads to a similar increase and decrease of hmU. While further work is needed, we believe JBP iterative oxidation could provide a key regulatory step. For example, regulated interactions between GT and the dioxygenases could allow for the conversion of hmU to fU/caU, leading to thymine. In the thymine salvage pathway the conversion is completed by decarboxylation, and

an isoorate decarboxylase is present in the trypanosome genome. The possibility that JBP enzymes mediate iterative oxidation of thymidine has important consequences on understanding the role of DNA modification in trypanosomes. All of the studies of J function to date have utilized parasites with reduced levels of both J and its intermediate hmU through deletion of the dioxygenases. Future studies of the GT KO cell lines will allow us to distinguish between the role of J and hmU (and potentially fU and caU).

REGULATION OF THYMIDINE OXIDATION BY METABOLISM AND HOST-PARASITE INTERACTIONS

As discussed in additional chapters in this book, the Fe(II)/2OG-dependent dioxygenase enzyme family encompasses a large group of enzymes that catalyze the hydroxylation of a diverse variety of substrates, including but not limited to DNA, protein, RNA, and lipid. Most dioxygenases utilize Fe(II) as a cofactor and 2OG and O_2 as co-substrates. Succinate and carbon dioxide are released as byproducts. As expected, succinate is known to inhibit this class of enzymes by product inhibition. In addition, the requirement of O_2 for enzyme activity has led to the suggestion that dioxygenases function as direct O_2 sensors. It is noteworthy that kinetoplastid parasites experience large environmental differences during their life cycle progression between the insect vector and mammalian hosts. In *T. brucei* for example, these differences include, among others, a change in temperature from 37 °C in the mammalian host to 27 °C in the insect vector as well as a change in the availability of glucose, the preferred energy source of the parasite. Within the various hosts, these pathogens are also exposed to changing O_2 conditions. Oxygen concentrations range from 0 to 21% within human host tissues, depending upon the proximity to blood vessels and the O_2 consumptive activity of the cell. For example, *T. cruzi*

experiences varying O_2 concentrations from high levels on the skin to low levels within various tissues (i.e. gut and muscle) of the human host and insect vector. Accordingly, these organisms are capable of tightly regulating their gene expression, enabling them to adapt to a diverse array of environmental conditions including O_2 concentration. The potential role of dioxygenases as O_2 and metabolic sensors has exciting implications for JBP1/2 as key epigenetic regulators of gene expression in response to the parasite environment and different host niches.

JBP as Oxygen Sensors

We have shown that *T. cruzi* decreases its J levels after the loss of JBP1, resulting in increased Pol II recruitment and transcription initiation. The resultant changes in gene expression have direct effects on the host-parasite relationship, as indicated by enhanced mammalian cell invasion and delayed egress.(6) Growth of *T. cruzi* in physiologically relevant low O_2 tensions (hypoxia) leads to reduced levels of J, presumably through inhibition of both JBP1/2 activities. This reduce-J state resulted in similar phenotypic changes in parasite virulence as previously characterized in the JBP1 KO cells. These results clearly demonstrate a O_2 dependence for epigenetic regulation of virulence gene expression. Conclusive demonstration of the oxygen sensing capabilities of the JBPs, however, will require measuring the K_m for O_2 coupled to appropriate in vivo studies. Future studies are also needed to better describe the mechanistic link between hypoxic signaling and JBP function, exploring the parasite response to different levels of O_2, and examining the corresponding changes in J synthesis, transcription, and gene expression profiles.

JBP Regulation by Parasite Metabolism

In addition to O_2 concentrations, the metabolic state of the cell may be important in regulating JBP1/2 activity. As previously discussed, base J is developmentally regulated in *T. brucei* through down regulation of the biosynthesis machinery. This down-regulation may not fully explain the loss of J synthesis however, since JBP1 and JBP2 are not fully active when over-expressed in insect-stage cells compared to similar expression levels in bloodstream-form cells.(5,29) While it is not clear to what extent this expression is affected by the reduced levels of the GT, hmU feeding leads to significant levels of J in insect-stage parasites,(24) suggesting an additional mechanism exists to reduce JBP function in this life-stage.

In the bloodstream of the mammalian host, trypanosomes utilize free glucose as their major carbon source. During differentiation to the insect form, however, parasites undergo a metabolic shift and utilize amino acids available in the insect vector as their primary carbon source. As a consequence, rather than pyruvate as the major metabolic end product, insect-stage parasites produce high levels of succinate.(49-51) Succinate is known to inhibit some 2-oxoglutarate dioxygenases and, consistent with this, we have shown succinate significantly inhibits JBP1 activity *in vitro*.*(34)* Accordingly, upon re-expression of *JBP1* to achieve identical levels of the enzyme, the ~800-fold decrease in succinate/2OG ratio in bloodstream versus insect-stage parasites corresponds to an ~15-fold increase in JBP1 activity.(5) Thus, the high levels of succinate may inhibit the activity of any low level of JBP1/2 expressed in the insect-stage parasite as well as help explain the origins of developmental regulation of J synthesis in *T. brucei.* Adaptation to life without J in the insect stage may have necessitated the development of new mechanisms to regulate chromatin and gene expression, which might explain the apparent lack of a phenotype of the bloodstream form *T. brucei* upon loss of J.

CONCLUSIONS AND FUTURE GOALS

Since the initial discovery of base J in 1993, significant progress has been made in elucidating its synthesis pathway. Two distinct thymidine 2OG-dependent dioxygenases, JBP1 and JBP2, have been identified that perform the initial oxidation of thymidine in DNA. The characterization of the THs has been crucial in developing our understanding the function of this important regulatory epigenetic mark.

It is clear however, that despite our recent progress, further work remains to fully understand the mechanism of JBP function and the biological role of base J. For example, it is unclear how the two dioxygenases work together to regulate the genomic distribution of J. What is the basis of the apparent chromatin substrate specificity? What aspects of chromatin do they recognize and bind, and does this occur at the level of DNA sequence or structure? Are there JBP-associated proteins that are involved in these interactions? How do the *C*-terminal domains regulate thymidine oxidase activity in the *N*-terminal domain? Finally, are the dioxygenases able to stimulate iterative oxidation of thymidine and formation of additional analogs in the kinetoplastid genome?

Thus, continued analysis of the THs will continue to shed light on the biological function of base J and its role in the parasite during an infection of a mammalian host. The recent identification of the GT in the synthesis pathway provides an additional tool to study J function.

REFERENCES

1. Gommers-Ampt, J.H., Van Leeuwen, F., de Beer, A.L., Vliegenthart, J.F., Dizdaroglu, M., Kowalak, J.A., Crain, P.F. and Borst, P. (1993) beta-D-glucosyl-

hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. *Cell*, **75**, 1129-1136.

2. Borst, P. and Sabatini, R. (2008) Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol*, **62**, 235-251.

3. Sabatini, R., Cliffe, L., Vainio, L. and Borst, P. (2009) *Enzymatic Formation of the Hypermodified DNA Base J.* In Grosjean, H. (ed.), *DNA and RNA Modification Enzymes: Comparative Structure, Mechanism, Function, Cellular Interactions and Evolution*. Landes Biosciences, Texas, pp. 120-131.

4. van Leeuwen, F., Kieft, R., Cross, M. and Borst, P. (2000) Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosal-hydroxymethyluracil in *Trypanosoma brucei*. *Molecular and biochemical parasitology*, **109**, 133-145.

5. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A. and Sabatini, R. (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res*, **38**, 3923-3935.

6. Ekanayake, D.K., Minning, T., Weatherly, B., Gunasekera, K., Nilsson, D., Tarleton, R., Ochsenreiter, T. and Sabatini, R. (2011) Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. *Mol. Cell. Biol.*, **31**, 1690-1700.

7. van Luenen, H.G., Farris, C., Jan, S., Genest, P.A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G. *et al.* (2012) Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, **150**, 909-921.

8. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, **7**, 461-465.

9. van Leeuwen, F., de Kort, M., van der Marel, G.A., van Boom, J.H. and Borst, P. (1998) The modified DNA base beta-D-glucosylhydroxymethyluracil confers resistance to micrococcal nuclease and is incompletely recovered by 32P-postlabeling. *Analytical Biochemistry*, **258**, 223-229.

10. Bernards, A., van Harten-Loosbroek, N. and Borst, P. (1984) Modification of telomeric DNA in Trypanosoma brucei; a role in antigenic variation? *Nucleic Acids Research*, **12**, 4153-4170.

11. Pays, E. (2005) Regulation of antigen gene expression in Trypanosoma brucei. *Trends in parasitology*, **21**, 517-520.

12. van Leeuwen, F., Wijsman, E.R., Kieft, R., van der Marel, G.A., van Boom, J.H. and Borst, P. (1997) Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes & development*, **11**, 3232-3241.

13. Gommers-Ampt, J., Lutgerink, J. and Borst, P. (1991) A novel DNA nucleotide in Trypanosoma brucei only present in the mammalian phase of the life-cycle. *Nucleic Acids Research*, **19**, 1745-1751.

14. van Leeuwen, F., Taylor, M.C., Mondragon, A., Moreau, H., Gibson, W., Kieft, R. and Borst, P. (1998) beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. *Proc. Natl. Acad. Sci. U S A*, **95**, 2366-2371.

15. Genest, P.A., ter Riet, B., Dumas, C., Papadopoulou, B., van Luenen, H.G. and Borst, P. (2005) Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. *Nucleic Acids Res*, **33**, 1699-1709.

16. Ekanayake, D. and Sabatini, R. (2011) Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryotic cell*, **10**, 1465-1472.

17. Martinez-Calvillo, S., Vizuet-de-Rueda, J.C., Florencio-Martinez, L.E., Manning-Cela, R.G. and Figueroa-Angulo, E.E. (2010) Gene expression in trypanosomatid parasites. *J Biomed Biotechnol*, 525241.

18. Iyer, L.M., Tahiliani, M., Rao, A. and Aravind, L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698-1710.

19. Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930-935.

20. Pastor, W.A., Aravind, L. and Rao, A. (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews. Molecular cell biology*, **14**, 341-356.

21. Iyer, L.M., Zhang, D., Maxwell Burroughs, A. and Aravind, L. (2013) Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Research*, **41**, 7635-7655.

22. Gommers-Ampt, J.H. and Borst, P. (1995) Hypermodified bases in DNA. *FASEB Journal*, **9**, 1034-1042.

23. Gommers-Ampt, J.H., Teixeira, A.J., van de Werken, G., van Dijk, W.J. and Borst, P. (1993) The identification of hydroxymethyluracil in DNA of Trypanosoma brucei. *Nucleic Acids Research*, **21**, 2039-2043.

24. van Leeuwen, F., Kieft, R., Cross, M. and Borst, P. (1998) Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Molecular & Cellular Biology*, **18**, 5643-5651.

25. Cross, M., Kieft, R., Sabatini, R., Wilm, M., de Kort, M., van der Marel, G., van Boom, J., van Leeuwen, F. and Borst, P. (1999) The modified base J is the target for a novel DNA-binding protein in kinetoplastid protozoans. *EMBO Journal*, **18**, 6573-6581.

26. DiPaolo, C., Kieft, R., Cross, M. and Sabatini, R. (2005) Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Molecular cell*, **17**, 441-451.

27. Kieft, R., Brand, V., Ekanayake, D.K., Sweeney, K., DiPaolo, C., Reznikoff, W.S. and Sabatini, R. (2007) JBP2, a SWI2/SNF2-like protein, regulates de novo telomeric DNA glycosylation in bloodstream form Trypanosoma brucei. *Molecular and biochemical parasitology*, **156**, 24-31.

28. Cross, M., Kieft, R., Sabatini, R., Dirks-Mulder, A., Chaves, I. and Borst, P. (2002) J binding protein increases the level and retention of the unusual base J in trypanosome DNA. *Molecular microbiology*, **46**, 37-47.

29. Cliffe, L.J., Kieft, R., Southern, T., Birkeland, S.R., Marshall, M., Sweeney, K. and Sabatini, R. (2009) JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J

biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res*, **37**, 1452-1462.

30. Yu, Z., Genest, P.A., ter Riet, B., Sweeney, K., DiPaolo, C., Kieft, R., Christodoulou, E., Perrakis, A., Simmons, J.M., Hausinger, R.P. *et al.* (2007) The protein that binds to DNA base J in trypanosomatids has features of a thymidine hydroxylase. *Nucleic Acids Res*, **35**, 2107-2115.

31. Schofield, C.J. and Zhang, Z. (1999) Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. *Curr Opin Struct Biol*, **9**, 722-731.

32. Hausinger, R.P. (2004) FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. *Critical reviews in biochemistry and molecular biology*, **39**, 21-68.

33. Vainio, S., Genest, P.A., ter Riet, B., van Luenen, H. and Borst, P. (2009) Evidence that J-binding protein 2 is a thymidine hydroxylase catalyzing the first step in the biosynthesis of DNA base J. *Mol. Biochem. Parasitol.*, **164**, 157-161.

34. Cliffe, L.J., Hirsch, G., Wang, J., Ekanayake, D., Bullard, W., Hu, M., Wang, Y. and Sabatini, R. (2012) JBP1 and JBP2 Proteins Are Fe2+/2-Oxoglutarate-dependent Dioxygenases Regulating Hydroxylation of Thymidine Residues in Trypanosome DNA. *J. Biol. Chem.*, **287**, 19886-19895.

35. Nathan R. Rose, M.A.M., Oliver N. F. King, Akane Kawamura, Christopher J. Schofield. (2011) Inhibition of 2-oxoglutarate dependent oxygenases. *Chem. Soc. Rev*, **40**, 4364–4397.

36. Ulbert, S. (2003), University of Amsterdam, Amsterdam.

37. Sabatini, R., Meeuwenoord, N., van Boom, J.H. and Borst, P. (2002) Recognition of base J in duplex DNA by J-binding protein. *J. Biol. Chem.*, **277**, 958-966.

38. Grover, R.K., Pond, S.J., Cui, Q., Subramaniam, P., Case, D.A., Millar, D.P. and Wentworth, P., Jr. (2007) O-glycoside orientation is an essential aspect of base J recognition by the kinetoplastid DNA-binding protein JBP1. *Angew Chem Int Ed Engl*, **46**, 2839-2843.

39. Sabatini, R., Meeuwenoord, N., van Boom, J.H. and Borst, P. (2002) Site-specific interactions of JBP with base and sugar moieties in duplex J-DNA. *Journal of Biological Chemistry*, **277**, 28150-28156.

40. Heidebrecht, T., Christodoulou, E., Chalmers, M.J., Jan, S., Ter Riet, B., Grover, R.K., Joosten, R.P., Littler, D., van Luenen, H., Griffin, P.R. *et al.* (2011) The structural basis for recognition of base J containing DNA by a novel DNA binding domain in JBP1. *Nucleic Acids Res*, **39**, 5715-5728.

41. Heidebrecht, T., Fish, A., von Castelmur, E., Johnson, K.A., Zaccai, G., Borst, P. and Perrakis, A. (2012) Binding of the J-binding protein to DNA containing glucosylated hmU (base J) or 5-hmC: evidence for a rapid conformational change upon DNA binding. *Journal of the American Chemical Society*, **134**, 13357-13365.

42. Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S. and Cross, G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev.*, **23**, 1063-1076.

43. Toaldo, C.B., Kieft, R., Dirks-Mulder, A., Sabatini, R., van Luenen, H.G. and Borst, P. (2005) A minor fraction of base J in kinetoplastid nuclear DNA is bound by the J-binding protein 1. *Molecular and biochemical parasitology*, **143**, 111-115.

44. Ulbert, S., Eide, L., Seeberg, E. and Borst, P. (2004) Base J, found in nuclear DNA of Trypanosoma brucei, is not a target for DNA glycosylases. *DNA Repair (Amst)*, **3**, 145-154.

45. Smiley, J.A., Kundracik, M., Landfried, D.A., Barnes, V.R., Sr. and Axhemi, A.A. (2005) Genes of the thymidine salvage pathway: thymine-7-hydroxylase from a Rhodotorula glutinis cDNA library and iso-orotate decarboxylase from Neurospora crassa. *Biochimica et biophysica acta*, **1723**, 256-264.

46. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. and Song, H. (2011) Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, **145**, 423-434.

47. Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J. and Helin, K. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, **473**, 343-348.

48. Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. and Reik, W. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, **473**, 398-402.

49. Tielens, A.G. and van Hellemond, J.J. (2009) Surprising variety in energy metabolism within Trypanosomatidae. *Trends in parasitology*, **25**, 482-490.

50. van Grinsven, K.W., Van Den Abbeele, J., Van den Bossche, P., van Hellemond, J.J. and Tielens, A.G. (2009) Adaptations in the glucose metabolism of procyclic Trypanosoma brucei isolates from tsetse flies and during differentiation of bloodstream forms. *Eukaryotic cell*, **8**, 1307-1311.

51. Cazzulo, J. (1992) *Energy Metabolism in Trypanosoma cruzi.* In Avila, J. L. and Harris, J. R. (eds.), *Intracellular Parasites*. Springer US, Vol. 18, pp. 235-257.

52. Shen L, Zhang Y. *Curr Opinion Cell Biol* 2013; **25**:289-296

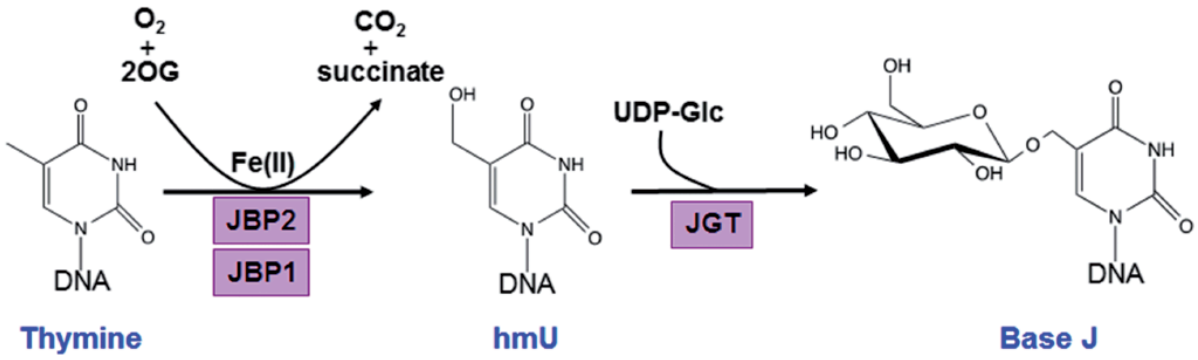


Figure 2.1 The biosynthesis of base J by a two-step modification of a specific thymidine base in the DNA. Base J is synthesized in DNA by a two-step mechanism involving thymidine hydroxylation and glycosylation. JBP1 and JBP2 are members of the Fe(II)/2OG dioxygenases family that utilize 2OG and O_2 as cosubstrates to hydroxylate thymidine bases in dsDNA, releasing succinate and CO_2 as by-products. The intermediate hmU is then glycosylated by a base J-associated glucosyltransferase (JGT) forming base J. UDP-glucose (UDP-Glc) is the activated sugar donor.

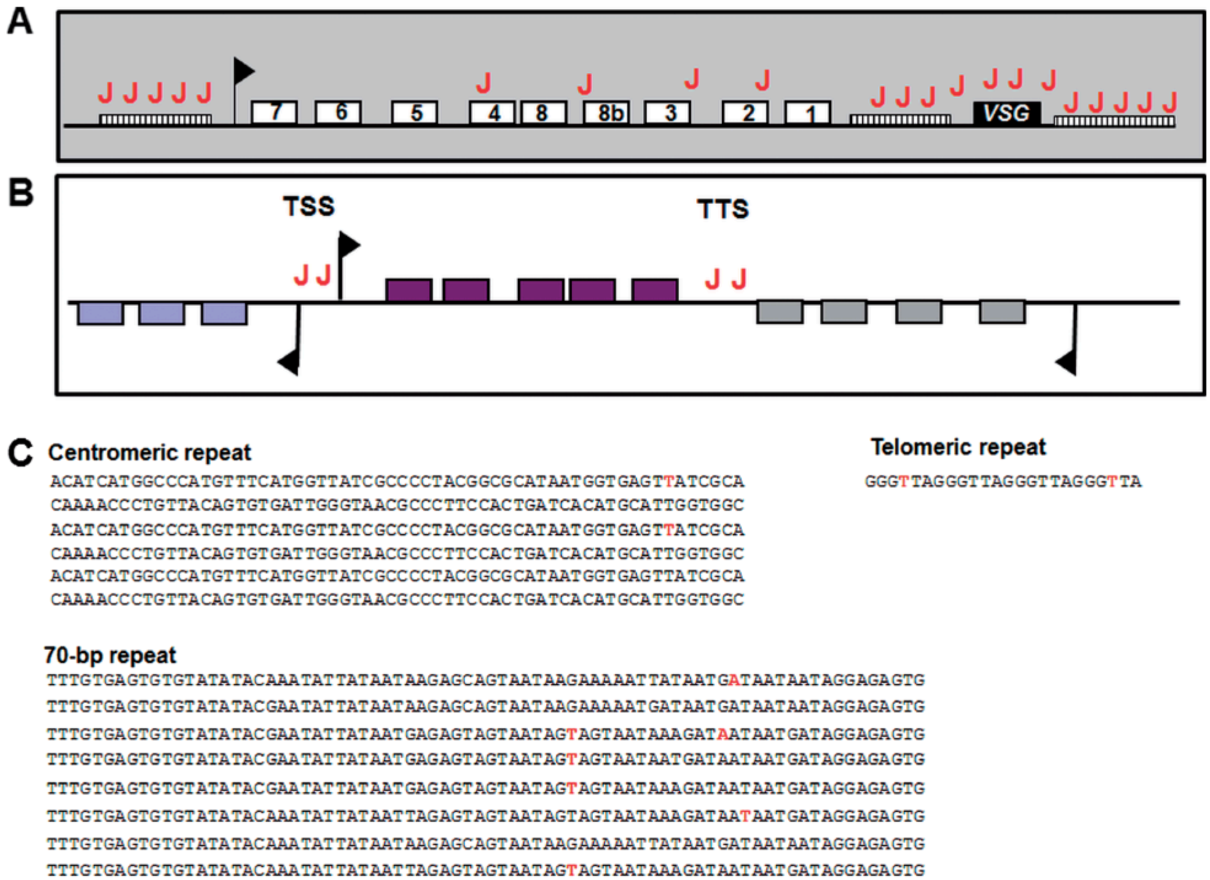


Figure 2.2 Genomic localization of base J. (A) Base J is found within the silent telomeric VSG expression sites in *T. brucei*. The presence of J was determined by immunoprecipitation of J containing DNA fragments by using an antibody against the modified base followed by a combination of DNA hybridization with various probes or high-throughput DNA sequencing approaches. The hatched boxes represent the 50 bp upstream of the promoter, 70 bp upstream of the *VSG* gene, and telomeric DNA repeats. (B) Base J is also found at internal sites in the genome, including RNA polymerase II transcription start sites (TSS) and transcription termination sites (TTS). (C) Strand-specific, base-resolution detection of base J within repetitive DNA sequences as identified by SMRT DNA sequencing. A portion of one DNA strand for each repeat is indicated. Thymidines that are modified to base J are highlighted in red. Highlighted A's indicate the modified T is located on the opposing DNA strand.

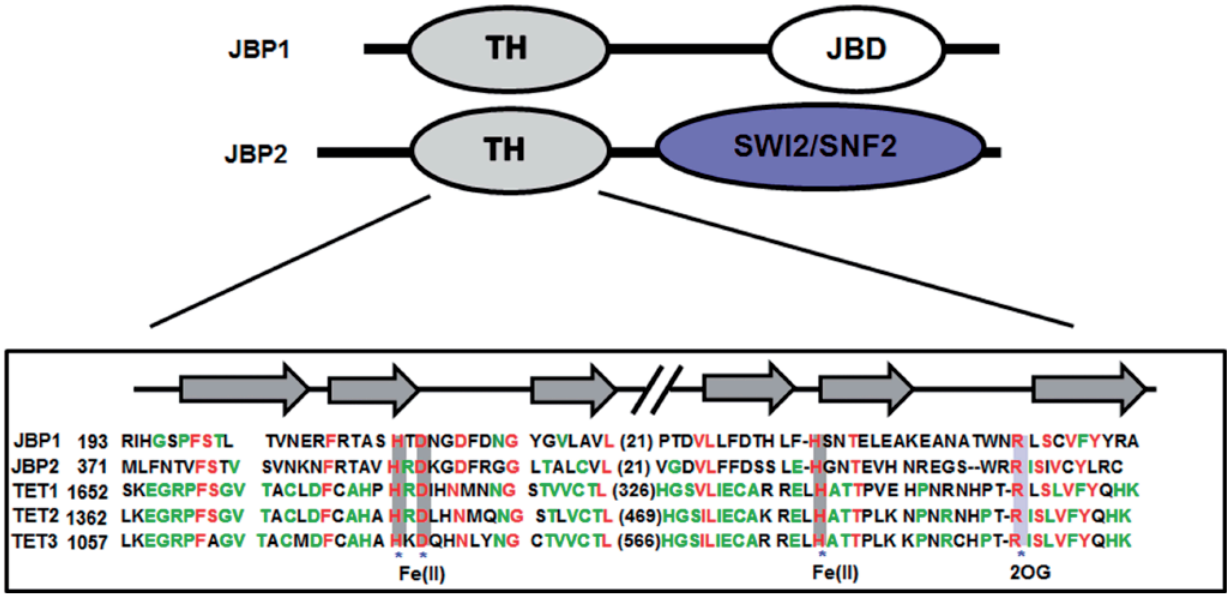


Figure 2.3 Functional domains of JBP1 and JBP2. The white oval within the C-terminus of JBP1 represents the 20 kDa minimal J-binding domain (JBD). The blue oval at the C-terminus of JBP2 represents the domain homologous to the SWI2/SNF2 family of chromatin remodelling ATPases. The region shared between JBP1 and JBP2 at the N-terminus is indicated by the grey oval (TH). Within this region is the ~70-residue motif which is related to the double-stranded β-helix domain of the members of the Fe(II)/2OG-dependent dioxygenase family. For the dioxygenases domain, a multiple sequence alignment of selected JBP/TET family proteins (*T. brucei* JBP1/2 and human TET1–3) is shown. The key residues conserved in members of the JBP/TET family predicted to be involved in iron (HXD-H) and 2OG binding (R) are indicated. Sequences that constitute the conserved β-helix fold are shown above the alignment (adapted from Shen and Zhang[52]). Numbers represent the amino acid numbers. TH, thymidine hydroxylase.

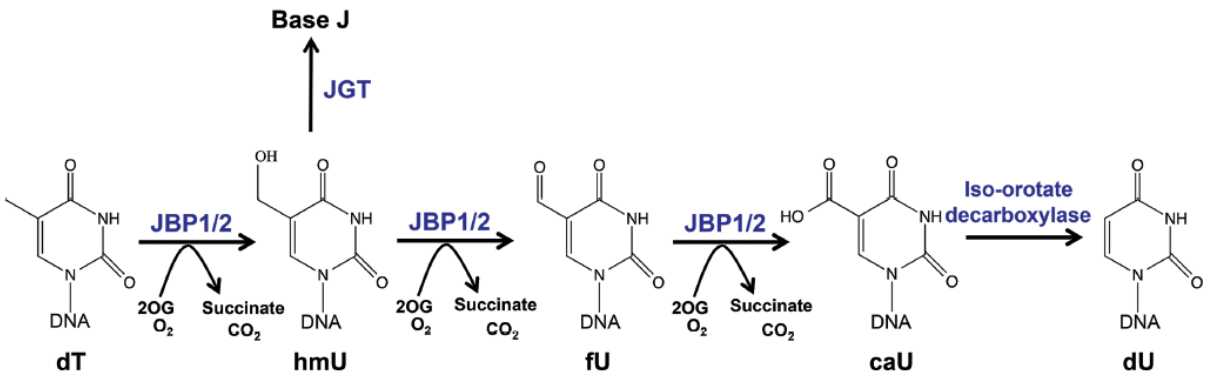


Figure 2.4 Proposed mechanism of iterative oxidation of thymidine initiated by the JBP enzymes. Similar to thymine hydroxylase activity in fungi and the human TET enzymes, the JBP proteins can potentially catalyse sequential conversion of thymine in DNA to 5-hydroxymethyluracil (5-hmU), 5-formyluracil (5-fU) and 5-carboxyuracil (5-caU). Each consecutive oxidation reaction would require O_2 and 2OG and release CO_2 and succinate. 5-caU would then be converted to uracil (dU) by an iso-orotate decarboxylase that is encoded in the kinetoplastid genome. Uracil would presumably be removed by base excision repair.

CHAPTER 3

REGULATION OF TRANSCRIPTION TERMINATION BY GLUCOSYLATED HYDROXYMETHYLURACIL, BASE J, IN *LEISHMANIA MAJOR* AND *TRYPANOSOMA BRUCEI*[1]

---

[1] **Reynolds D.**, Cliffe L., Förstner K., Hon C., Siegel N., and Sabatini R. 2014. *Nucleic Acids Res.*, 42, 9717-9729

Reprinted here with permission of publisher

ABSTRACT

Base J, β-D-glucosyl-hydroxymethyluracil, is an epigenetic modification of thymine in the nuclear DNA of flagellated protozoa of the order Kinetoplastida. J is enriched at sites involved in RNA Polymerase (RNAP) II initiation and termination. Reduction of J in *Leishmania tarentolae* via growth in BrdU resulted in cell death and indicated a role of J in the regulation of RNAP II termination. To further explore J function in RNAP II termination among kinetoplastids and avoid indirect effects associated with BrdU toxicity and genetic deletions, we inhibited J synthesis in *L. major* and *T. brucei* using DMOG. Reduction of J in *L. major* resulted in genome-wide defects in transcription termination at the end of polycistronic gene clusters and the generation of antisense RNAs, without cell death. In contrast, loss of J in *T. brucei* did not lead to genome-wide termination defects; however, the loss of J at specific sites within polycistronic gene clusters led to altered transcription termination and increased expression of downstream genes. Thus, J regulation of RNAP II transcription termination genome-wide is restricted to *Leishmania* spp., while in *T. brucei* it regulates termination and gene expression at specific sites within polycistronic gene clusters.

INTRODUCTION

Members of the Kinetoplastida order include the human parasites *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, which cause African sleeping sickness, Chagas' disease, and leishmaniases, respectively. Kinetoplastids are early-diverged protozoa with unique genome arrangements where genes are organized into long clusters, or polycistronic transcription units (PTUs), that are transcribed by RNA polymerase (RNAP) II (1-3). RNAP II transcription initiation and termination occurs at regions flanking PTUs called divergent strand

switch regions (dSSRs) and convergent strand switch regions (cSSRs), respectively (4). Transcription also terminates and initiates at head-tail (HT) sites, where transcription of an upstream PTU terminates and transcription of a downstream PTU on the same strand initiates (5-7). Pre-messenger RNAs (mRNA) are processed to mature mRNA with the addition of a 5' spliced leader sequence through *trans*-splicing, followed by 3' polyadenylation (8-13). In other eukaryotes RNAP II termination is generally coupled to 3' mRNA processing (14). In kinetoplastids however, 3' mRNA processing is instead coupled to *trans*-splicing of the 5' end of the adjacent gene (15), preventing premature termination within a PTU. The mechanism of RNAP II termination in kinetoplastids remains unknown.

While little is known regarding the regulation of gene expression in kinetoplastids, the unique genome arrangement and polycistronic transcription of functionally unrelated genes has led to the assumption that transcription is an unregulated process in these organisms and that most gene regulation occurs post-transcriptionally (16,17). Recently however, numerous epigenetic markers have been found enriched specifically at sites of transcription initiation and termination, including histone methylation and acetylation, histone variants, and base J, which could function to regulate gene expression at the level of transcription (5,6,18,19).

Base J, β-D-glucosyl-hydroxymethyluracil, is a thymidine modification found in kinetoplastids and closely related unicellular flagellates (20,21). J is largely a telomeric modification, but is also found internally within chromosomes at RNAP II transcription initiation and termination sites (18,22-25). Using high-throughput sequencing and ChIP analysis we identified base J localized within dSSRs and cSSRs in the genomes of *T. brucei*, *T. cruzi* and *L. major* (18,26). Recent high-throughput sequencing studies in *L. major* and *L. tarentolae* confirmed this internal J localization at RNAP II transcription regulatory sites (27). Because

base J is a conserved DNA modification specific to kinetoplastids (not present in the mammalian host) with a possible role in key regulatory processes, it represents a potential drug target to treat the diseases caused by these pathogens (28).

As reviewed in Borst and Sabatini (2008), base J is synthesized in a two-step pathway in which a thymidine hydroxylase, JBP1 or JBP2, hydroxylates T residues at specific positions in DNA to form hydroxymethyluracil (HOMedU), followed by the transfer of glucose to HOMedU by a glucosyltransferase (28). Both JBP1 and JBP2 belong to the new TET/JBP subfamily of dioxygenases, which require Fe^{2+} and 2-oxoglutarate (2-OG) for activity (29-31). The synthesis of base J can be inhibited by knocking out JBP1 and JBP2 or by competitive inhibition of the thymidine hydroxylase domain of JBP1 and JBP2 by dimethyloxalylglycine (DMOG), a structural analog of 2-OG (29,32,33). Removal of both JBP1 and JBP2 in *T. brucei* or growth in the presence of DMOG results in cells devoid of base J (J null) (29,34); however, studies thus far have not identified defects associated with the loss of J in *T. brucei*. In contrast, the inability to delete both JBP enzymes from *T. cruzi* or *Leishmania* spp. suggests the modification is essential in these organisms (26,35).

Base J reduction in *T. cruzi*, following the deletion of either JBP1 or JBP2, resulted in the formation of more active chromatin at transcription initiation sites, increased RNAP II recruitment and PTU transcription rate, and global changes in gene expression (26,36). The loss of J at cSSRs had no detectable effect on RNAP II transcription termination at the end of PTUs (26). The specific J dependent loss of nucleosomes at dSSRs and not cSSRs is consistent with a unique role of J in regulating RNAP II initiation in *T. cruzi* (36). However, van Luenen et al. (2012) recently found that reduction of base J in a JBP2 KO cell line of *L. tarentolae*, a pathogen of lizards, resulted in the generation of antisense RNAs (27). Treating the JBP2 KO

cells with bromodeoxyuridine (BrdU) further reduced J by an unknown mechanism and resulted in increased levels of antisense small RNAs and cell death. It was argued that J loss resulted in readthrough transcription at cSSR termination sites, which led to the production of RNAs antisense to the genes on the opposite strand. Direct evidence of readthrough transcription was not shown however, and it remains possible that antisense RNAs were produced through RNAP re-initiation within the cSSR, as opposed to readthrough transcription. Additionally, despite the known toxicity of BrdU to eukaryotic cells (37-40), it was suggested that massive transcriptional readthrough is lethal in *Leishmania* spp. (27).

To explore the functional conservation of J function among kinetoplastids and avoid indirect effects associated with the use of BrdU and genetic deletions, we utilized DMOG to examine the role of J in regulating RNAP II termination in *L. major* and *T. brucei*. We show that reduction of base J using DMOG in *L. major* resulted in genome-wide transcriptional readthrough at cSSRs and HT sites, without cell death. Strand-specific RT-PCR detection of the nascent transcript confirmed that we are measuring J dependent defects in transcriptional termination, rather than RNAP II re-initiation events. Complete loss of J in *T. brucei* failed to indicate any defect in termination within cSSRs or HT sites. However, we localized base J at sites prior to the end of a PTU where the loss of J led to up-regulated expression of the downstream genes within the same PTU. For one of these sites we show that the gene expression changes occurred at the level of transcription. Therefore, while base J regulates RNAP II termination in both *L. major* and *T. brucei*, it does so to different degrees and at different genomic locations. In *L. major* J regulates termination at the end of each PTU to prevent the generation of antisense RNAs genome-wide. In contrast, while termination occurs at

the end of each PTU in *T. brucei* in a J-independent manner, J-dependent termination within a PTU allows developmentally regulated expression of downstream genes.

MATERIAL AND METHODS

Enzymes and chemicals

All restriction enzymes were purchased from New England Biolabs. Prime-It II random primer labeling kit was purchased from Stratagene. ECL (enhanced chemiluminescence) and Hybond-N+ were from Amersham. Goat anti-rabbit HRP (horseradish peroxidase) was purchased from Southern Biotec Inc. All other chemicals were purchased from Sigma Aldrich.

Parasite cell culture

Bloodstream form *T. brucei* cell line 221a of strain 427 were cultured in HMI-9 medium as described previously (41). *L. major* parasites were grown at 26°C in M199 media supplemented with 10%FBS as described (42). DMOG treatment of cells was performed by supplementing media with 1mM DMOG for 5 days in *T. brucei* or at 5mM for 10 days in *L. major*. BrdU was supplemented into media at 10uM or 100uM for 6 days in *L. major*.

Determination of the genomic level of J

To quantify the genomic J levels, we used the anti-J DNA immunoblot assay as described (24,34) on total genomic DNA, which was isolated as described (43). Briefly, serially diluted genomic DNA was blotted to nitrocellulose followed by incubation with anti-J antisera. Bound antibodies were detected by a secondary goat anti-rabbit antibody conjugated to HRP

and visualized by ECL. The membrane was stripped and hybridized with a probe for the beta-tubulin gene to correct for DNA loading.

The localization of base J

To quantify the levels of base J at specific regions of the *T. brucei* or *L. major* genome, genomic DNA was sonicated and anti J immunoprecipitation was performed as described (23,24,33,34). Immunoprecipitated J containing DNA was used for quantitative PCR analysis. Input DNA was used as a positive control for qPCR (10% of the IP). Quantification of selected genes were performed on an iCycler with an iQ5 multicolor real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA). Primer sequences used in the analysis are available upon request. The reaction mixture contained 5 pmol forward and reverse primer, 2x iQ SYBR green super mix (Bio-Rad Laboratories, Hercules, CA), and 2 µl of template DNA. Standard curves were prepared for each gene using 5-fold dilutions of known quantity (100 ng/µl) of WT DNA. The quantities were calculated using iQ5 optical detection system software.

Strand-specific RNA-Seq library construction

Small RNAs were isolated from *T. brucei* or *L. major* using a Qiagen miRNeasy kit according to the manufacturer's instructions. $5x10^7$ cells were used per sample, and isolated at the log phase of parasite growth. Total RNA was isolated from log phase *T. brucei* cultures using a Qiagen RNeasy kit according to the manufacturer's instructions. $5x10^7$ cells were used for each prep.

All four small RNA-seq libraries were prepared using approximately 250ng small RNA using the TruSeq small RNA kit (Illumina) according to the manufacturers instructions with the

following exception: the PCR amplification was performed for 12-16 cycles using the KAPA HiFi DNA polymerase (Kapabiosystems). The PCR product was purified and concentrated using AMPure XP beads (Beckman Coulter). Quality and concentration of all libraries was determined using a Bioanalyzer 2100 (Agilent) and high throughput sequencing was performed on a HiSeq2000 (Illumina).

The two full-length RNA-seq libraries were constructed from approximately 1 μg of total RNA by Vertis Biotechnology AG. Briefly, total RNA was polyA-enriched using oligo(dT) chromatography and fragmented by ultrasound. Next, first strand cDNA was synthesized using N6 random primers followed by a strand-specific ligation of sequencing adapter to the 3´ and 5´ ends of the first stranded cDNA and PCR amplification of 10-20 cycles depending on the amount of starting material. High throughput sequencing was performed on a HiSeq2000 (Illumina).

Mapping of sequence reads, calculations of expression levels and meta-coverage plot

Sequencing reads were mapped to the respective reference genomes (see Supplemental Table 2) using bowtie-2 with default 'local-sensitive' mode (44) and further processed using samtools (45). For the small RNA libraries, reads shorter than 18 bp were discarded before mapping. To express the transcripts levels for individual genes as shown in Table 1, we determined the number of reads per kilobase per million reads (RPKM) (46). Briefly, we counted the number of reads mapped to all annotated transcriptomic features (e.g. mRNA) on the same strand (i.e. sense) and opposite strand (i.e. antisense). Both the sense and antisense read numbers were normalized by length of the feature (in kilobase) and the total number of reads (in millions) mapped to non-structural RNAs in the corresponding library (i.e. number of

mappable reads excluding rRNA and tRNA reads). For the metaplot of small RNA coverage at TTS, cSSRs with a TTS located at closer than 50kb to the edge of the scaffold were discarded. To eliminate positions with abruptly high coverage, positions with coverage greater than 5-fold of the mean sense coverage upstream of the TTS were ignored. The pooled meta-coverage was then smoothed using a sliding window of 500bp at step size of 100bp. The mean sense coverage upstream of a TTS was normalized to 1 within each sample.

Nuclear run on

Wild type *T. brucei* cultures were grown in the presence of DMSO or 1mM DMOG for 5 days. Nuclear run on was then performed using $1x10^9$ cells as described (47). The hot RNAs isolated were incubated with membranes containing double stranded DNA probes generated by TA cloning of the genomic region of interest. Each probe was an average of 1,000bp in length. An empty TA vector (pCR2.1) was used as a negative control. RNAs were incubated with membranes containing probes for 48 hrs and then subjected to washing with 0.1xSSC, 0.1% SDS at 65°C and then analyzed by phosphoimager. The signal for each probe was background subtracted and normalized to the spliced leader signal.

Strand-specific RT-PCR analysis of read through transcription

Total RNA was isolated using the hot phenol method, as described previously (48). The RNA pellet was re-suspended in water and further purified using a Qiagen RNeasy kit according to the manufacturer's instructions. To ensure complete removal of contaminating genomic DNA, the on column DNase I digest step was performed twice. Purified RNA was eluted from the column by water and the concentration was determined using a spectrophotometer. Strand

specific RT-PCR was performed as previously described (49). ThermoScript™ Reverse Transcriptase from Life Technologies was used for cDNA synthesis at 60-65°C. 2 μg of RNA were used to make cDNA using the reverse primers shown in Figure 3.3A. PCR was performed using GoTaq DNA Polymerase from Promega. A minus-RT control was used to ensure no contaminating genomic DNA was amplified. Primer sequences used in the analysis are available upon request.

Reverse transcription quantitative PCR (RT qPCR)

Total RNA was obtained using Qiagen RNeasy kits according to manufacturers instructions. First-strand cDNA was synthesized from 1 μg of total RNA using an iScript cDNA synthesis kit (Bio-Rad Laboratories, Hercules, CA) per the manufacturer's instructions. Quantification of selected genes were performed on an iCycler with an iQ5 multicolor real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA). Primer sequences used in the analysis are available upon request.

RESULTS

J regulation of RNAP II termination in *L. major*

J regulation of RNAP II termination at cSSRs and HT sites in L. major

We have previously shown that growth of kinetoplastids in DMOG effectively inhibits J synthesis and leads to a large reduction in the total level of base J (29). Wild type *L. major* grown in medium containing 5mM DMOG for 10 days resulted in approximately a 32-fold reduction in the total level of base J (Figure 3.1A) and a modest growth phenotype (Figure 3.S1). In comparison, inhibition of J synthesis by BrdU led to at most an 8-fold reduction in the

total level of J and cell death within 6 days (Figure 3.S1). Given the more significant reduction in total J observed following DMOG treatment compared to BrdU treatment, and the known toxicity of BrdU to other eukaryotic cells (that lack base J) (37,40,50), it is possible that the BrdU treatment resulted in cell death independent of J loss.

To analyze J levels at specific cSSRs, we performed anti-base J immunoprecipitation (IP) followed by quantitative PCR (qPCR). We found that DMOG reduced chromosome internal J, though J loss was more pronounced at some cSSRs than others (Figure 3.1B). While both DMOG and BrdU significantly reduced internal J, DMOG inhibition led to greater reductions of base J at telomeres, explaining the enhanced total J loss observed following DMOG treatment (Figure 3.S1C). Thus, to reduce J at cSSRs while avoiding genetic deletions and potential toxicity of BrdU, we utilized DMOG to study the specific role of base J in transcription termination.

Similar to the method previously utilized to characterize transcription termination in *L. tarentolae*, we performed high-throughput sequencing on small RNAs from wild type *L. major* treated with DMOG to identify the effect of J loss at termination sites and evidence of readthrough transcription. This provides a rough picture of the primary transcript map of *L. major* (i.e. mapping TTS), as reflected in small RNA degradation products. As expected, RNAP II transcription generally terminated at a defined location within the cSSR in wild type cells. As described in *L. tarentolae* (27), the reduction of base J at cSSRs in *L. major* resulted in the production of antisense RNAs corresponding to genes in the opposing PTU presumably due to readthrough transcription (Figure 3.1C).

A metaplot summarizing the readthrough defect for all cSSRs is shown in Figure 3.1D and the amount of readthrough at each cSSR is summarized in Figure 3.2. Readthrough

occurred at 30 of the 37 convergent transcription termination sites (TTS) when J was reduced, though the extent of readthrough varied and was often asymmetric at a given site (Figure 3.2). Interestingly, the degree of readthrough transcription often correlated with the reduction in J level following DMOG treatment at the cSSRs examined (Figures 1B and C, Figure 3.2 and S1D). As was observed in *L. tarentolae*, readthrough transcription was observed in wild type cells at the single cSSR (28.2) that normally lacks J (Figures 1B and 2), further supporting the link between J levels and transcription termination. Another variable is the presence of RNA genes transcribed by RNAP III (i.e. tRNAs and snRNAs) in the cSSR (Figures 1B and C and Figure 3.2). Readthrough transcription was modest at cSSRs containing genes transcribed by RNAP III. For example, DMOG treatment significantly reduced J levels at cSSR 36.3 that contains tRNA genes on both DNA strands, with little to no effect on transcription termination (Figure 3.1C, Figure 3.2 and 3.S1C). Additionally, the orientation of the RNAP III gene relative to the direction of RNAP II transcription influenced the extent of readthrough observed. At several cSSRs containing a RNAP III gene on the opposing strand, readthrough was modest upon the loss of J. However, at sites where the RNAP III genes are transcribed in the same direction as the adjacent PTU, readthrough was more extensive. For example, see cSSRs 30.3, 21.3 and 5.2 33.2). While there were a few exceptions, these results suggest that RNAP II can bypass an RNAP III transcription unit more readily if the polymerases are traveling in the same direction and that opposing RNAP III transcription allows J-independent termination. Asymmetric readthrough was also seen at cSSRs without RNAP III genes; therefore, additional factors likely contribute to the extent of the readthrough defect observed.

In addition to cSSRs, termination at head-tail sites was also affected by the loss of base J (Figure 3.1C). Head-tail (HT) regions are enriched for chromatin marks such as base J and

acetylated histone H3, as well as the histone variants H3V and H4V and H4ac at similar sites in *T. brucei*, which indicate termination and re-initiation on the same DNA strand (5-7,27). Small RNA-seq analysis further supports transcription termination and re-initiation at head-tail sites, given the lack of RNAs detected in the region between the upstream and downstream PTUs in wild type *L. major* (see HT 20.3 Figure 3.1C). Upon the loss of J we detected RNAs throughout the head-tail region, suggesting that J is also involved in regulating RNAP II termination within these sites (Figure 3.1C and data not shown). Additionally, consistent with a role of RNAP III genes in preventing readthrough transcription, little readthrough was observed at a head-tail site containing tRNA genes on both DNA strands (Figure 3.S2).

Detection of nascent RNA generated by RNAP II readthrough

As previously discussed (27), the generation of small RNAs downstream of putative TTSs following the loss of base J is presumably due to readthrough transcription and continued elongation of the RNAP II machinery rather than transcription re-initiation events. To better define the observed defect as transcriptional readthrough, we utilized strand-specific RT-PCR to directly detect the extension of nascent RNA transcripts. Figure 3.3A illustrates the orientation and position of primers relative to the TTS. Primers 1 and 2 spanned two adjacent ORFs and ensured our ability to analyze unprocessed, nascent RNA transcripts. Primers 6 and 7 were positioned upstream and downstream, respectively, of the TTS, as defined by small RNA-seq of wild type cells. Consistent with the TTS mapping, we detected a nascent RNA transcript upstream of the TTS (using primers 5 and 6), but not one that spanned the TTS (using primers 5 and 7) in wild type *L. major* (Figure 3.3B). Consistent with J-dependent transcriptional readthrough, for each of the cSSRs tested we detected a nascent RNA transcript upstream of the

TTS in both wild type and DMOG treated *L. major*; however, a readthrough product that spanned the TTS was present only when J levels were reduced. Any possible transcriptional re-initiation events occurring downstream of the TTS would not generate an RNA molecule that spans the TTS. As an additional control, a readthrough product was not observed in wild type or DMOG treated cells at a (tRNA containing, 36.3) site that lacked a detectable readthrough defect by small RNA-seq. We conclude that the small RNA-seq analysis accurately measures termination defects where base J regulates RNAP II termination at the end of PTUs to prevent transcriptional readthrough in *L. major*. Overall these results suggest that J functions to regulate termination throughout the genome of *L. major*, but that this significant readthrough termination defect and generation of antisense RNAs is not lethal to the cell.

J regulation of RNAP II termination in T. brucei

J regulation of RNAP II termination at cSSRs and HT sites in T. brucei

The termination defects observed in *Leishmania* spp. prompted us to examine *T. brucei* for similar transcription termination defects upon the loss of base J. Base J is not essential in *T. brucei*, as both JBP enzymes can be knocked out without obvious phenotypic effects (34). Consistent with this, wild type *T. brucei* treated with 1mM DMOG reduced global J levels beyond the limits of detection by anti-J dot blot (Figure 3.4A), with no significant growth effect (Figure 3.S3). Consistent with total J levels, all cSSRs examined had significantly reduced levels of J following DMOG treatment (Figure 3.4B).

To measure readthrough defects we performed small RNA-seq analysis as described above. In contrast to *Leishmania* spp., we found no evidence of termination defects in *T. brucei* at cSSRs or at head-tail regions following J loss. Antisense small RNAs, indicative of

readthrough transcription at cSSRs into the downstream PTU, were not increased following the loss of base J (Figure 3.4C). Similarly, no significant changes in small RNAs corresponding to readthrough transcription at head-tail sites were detected (Figure 3.S4). A metaplot summarizing the lack of readthough defects at the approximately 100 cSSRs is shown in Figure 3.4D. During this analysis, unique to *T. brucei*, we identified peaks of sense and antisense small RNAs that mapped near and within some cSSRs of wild type parasites in absence of DMOG (Figure 3.4C; 10.3 and 11.9). These presumably represent the previously identified Argonaute associated siRNAs derived from cSSRs (51). Regardless of their source, the level of (siRNA-like) sense/anti-sense small RNAs that map to particular cSSRs was not changed by the loss of J (Figure 3.4C and D). Thus, in contrast to the readthrough transcription observed in *Leishmania* spp., the loss of J did not affect RNAP II termination at cSSRs or head-tail regions in *T. brucei*.

J regulation of RNAP II termination within PTUs in T. brucei

To further explore the role of J in regulating RNAP II transcription in *T. brucei*, we investigated the effect of J loss on transcript abundance using total RNA-seq. In contrast to the global changes detected in *T. cruzi* and *L. tarentolae* following decreased levels of base J (26,27), we identified very limited gene expression changes in *T. brucei* cells incubated with DMOG (Table 1). RNA-seq indicated only 36 transcripts were changed more than two-fold following the loss of base J, the majority of which (30 transcripts) had increased expression. It is interesting to note that the majority of genes with increased mRNA levels are annotated as VSGs, RHS proteins, ESAGs, and pseudogenes and are lowly expressed (or silent) in wild type *T. brucei*. We confirmed many of these changes by RT qPCR (Figure 3.5) and removal of DMOG restored base J synthesis and wild type expression levels (Figure 3.S5A). Similar

changes in transcript abundance were also observed in the J null cell line (JBP1/JBP2 KO) indicating the expression changes were not due to indirect effects of the DMOG treatment (Figure 3.S5B). While total RNA-seq revealed six genes with at least a two-fold reduction in mRNA upon the loss of base J (Table 1), we were only able to confirm the transcript changes for one out of three transcripts analyzed by RT qPCR (Figure 3.5 and data not shown). We have not followed up further on the few genes that may have decreased expression following loss of J.

Interestingly, 13 out of the 27 upregulated genes (48%) are located at or near the end of an annotated PTU (i.e. a cSSR or HT site) and downstream of a J peak, suggesting that base J may facilitate termination prior to the end of a gene cluster and attenuate the transcription of downstream genes. Consistent with RNAP II termination prior to the end of a PTU, H3V is also enriched upstream of many of the up-regulated genes (ref). An example is shown in Figure 3.6A, where a peak of base J (and H3V) is found upstream of the last two genes in a PTU (list gene names and #). Both of the downstream genes are lowly expressed in the presence of base J, but increased upon the loss of J (Figure 3.6A and B) (for additional examples see Figure 3.S6). As described above, removal of DMOG restored base J synthesis and wild type expression levels (Figure 3.6B). To examine whether the up-regulation was due to increased transcription, as opposed to post-transcriptional mechanisms, we performed nuclear run-ons. In the absence of J, we found increased transcription of the region downstream of the peak of J (Figure 3.6C), suggesting that increases in mRNA abundance occurred at the level of transcription. Overall these results indicate that while base J may not regulate RNAP II termination in *T. brucei* at previously defined sites at cSSRs and head-tail regions, it may attenuate transcription elongation within specific PTUs and enable regulated expression of downstream genes.

DISCUSSION

We have clearly demonstrated an important function for base J in the regulation of RNAP II termination in *L. major*, as was observed in *L. tarentolae*. Whereas previous analyses utilized a JBP2 KO treated with BrdU, the use of DMOG as an alternative method to reduce J levels in wild type cells strongly suggests the epigenetic modification itself regulates RNAP II termination and that termination defects were not due to the loss of JBP2 protein or an indirect effect of BrdU or DMOG. We also demonstrate that J functions to regulate termination at head-tail sites in *L. major* in addition to cSSRs. Finally, we confirmed the termination defects identified by small RNA-seq as readthrough transcription using strand-specific RT-PCR. These findings suggest antisense RNAs were produced as a result of readthrough transcription, though we cannot exclude the possibility that some RNAP II re-initiation contributed to the production of additional antisense RNAs.

The lethality of BrdU treatment in *L. tarentolae* JBP2 KO cells led to the conclusion that readthrough transcription and subsequent generation of antisense RNAs were responsible for cell death (27). However, BrdU is known to be toxic to other eukaryotic cells (that lack base J) and it has been shown that BrdU incorporation into DNA affects nucleosome positioning on plasmids in yeast cells (38) and gene expression in mammalian cells (52,53). We found that drastic reduction of base J in wild type *L. major* via DMOG caused global readthrough transcription without a significant effect on cell growth. While we were not able to directly compare the extent of readthrough between the two studies, our findings suggest *Leishmania* spp. can tolerate the production of antisense RNAs and implicate BrdU toxicity as a potential cause of cell death. Despite the large decrease in total J, and comparable reduction in J by BrdU and DMOG at several of the cSSRs analyzed, DMOG treatment did not reduce J at all cSSRs.

Importantly, at the three cSSRs analyzed that had minimal loss in base J, less readthrough transcription was observed compared to the cSSRs analyzed with greater J loss, strengthening the correlation between base J and termination defects. However, the lack of readthrough at some cSSRs could have also influenced cell viability.

In addition to base J, RNAP III transcribed genes also appear to facilitate RNAP II termination in an orientation dependent fashion. One possibility is that RNAP II and III complexes transcribing in the opposite direction of each other collide, thus preventing further transcription, whereas RNAP complexes transcribing in the same direction are able to bypass each other. Such RNAP complex collision, and subsequent effects on transcription, has been observed in yeast (54). Further work in wild type and DMOG treated *L. major* will allow us to explore the mechanism by which RNAP III transcribed genes facilitate J-independent RNAP II termination.

In contrast to *Leishmania* spp., the loss of base J did not affect RNAP II termination at the majority of transcription termination sites in *T. brucei*. However, we found that J attenuated RNAP II transcription at some specific genomic loci to regulate the expression of downstream genes. A major difference observed between *Leishmania* spp. and *T. brucei* was the presence of small dsRNA peaks within cSSRs, which we assume represent siRNAs, as previously documented (51,55). *T. brucei* has a functional RNAi pathway (56), though the function of cSSR-derived siRNAs and whether they play a role in regulating RNAP II termination remains unknown. Little is known about how the siRNAs are generated from cSSRs, such as which RNAP is responsible for their transcription. The lack of chromatin modifications associated with transcription initiation at siRNA generating loci is suggestive of continued RNAP II transcription from the upstream PTU. Nonetheless, base J did not affect the production of cSSR-

derived siRNAs, nor did the presence or absence of siRNAs within a cSSR influence whether readthrough transcription occurred upon J loss. Furthermore, significant increases in cSSR-derived siRNAs in *T. brucei,* induced by the deletion of a chromatin bound factor, were not associated with changes in RNAP II termination (unpublished data).

Although small RNA seq analysis in *T. brucei* failed to detect any termination defect, total RNA-seq analysis identified several sites where J appeared to promote transcription termination prior to the end of a PTU. The loss of base J within a PTU resulted in increased transcript abundance of genes downstream of the J peak. The majority of these sites are devoid of chromatin modifications associated with transcription initiation, such as histone variants H2AZ, H2BV, or histone acetylation, but are enriched for histone variants H3V and H4V, which are associated with transcription termination in *T. brucei* (5). Nuclear run-on analysis supports the conclusion that the loss of base J within a gene cluster led to readthrough transcription and increased mRNA abundance of downstream genes. Several of the up-regulated genes were not downstream of base J however, and therefore we cannot exclude the possibility that some of the observed changes in transcript abundance were due to increased RNAP II initiation or altered post-transcriptional regulatory processes. Many of the up-regulated genes were internal VSGs, ESAGs, RHS proteins, and pseudogenes that are normally lowly expressed in wild type *T. brucei*. While only modest up-regulation was observed for most of the genes following J loss, these findings are consistent with a bloodstream stage-specific function of base J in the attenuation of RNAP II transcription and the promotion of transcription termination. The functional significance of this J-dependent regulation during trypanosome bloodstream infections is currently unclear.

The arrangement of functionally unrelated genes within the same PTU seemingly precludes regulated transcription as a mechanism to control gene expression. This work however suggests that placement of genes downstream of base J within a PTU can effectively reduce their expression in *T. brucei*. It is not currently clear whether *Leishmania* spp. regulate gene expression similarly. These findings are consistent with work by Kelly et al. (2012) that has indicated the location of a gene within a PTU can impact its expression (57). Thus, regulated expression of genes within PTUs can be achieved through their spatial organization and position relative to base J.

Based on a model in which base J attenuates RNAP II elongation in *T. brucei*, we would expect not only increases in the transcript abundance of downstream genes, but also an increase in the small RNAs, reflecting transcription of these regions upon the loss of base J. Upon J loss however, we found no significant increase in the small RNAs in the region downstream of the peak of J (data not shown). Our nuclear run-on analysis supports the conclusion that increases in transcript abundance were due to increased transcription; therefore, it is possible that small RNA-seq analysis does not provide a complete picture of the primary transcriptome in *T. brucei*. It is plausible that the presence of the RNAi machinery affects the pool of small RNAs detectable by the deep sequencing approach used here, which requires a 5'-P and a 3'-OH. Consistent with this possibility, in *L. braziliensis*, which contains a functional RNAi pathway, small RNAs indicative of readthrough are not observed at the single cSSR that lacks base J, in contrast to other RNAi negative *Leishmania* spp. ((27) and demonstrated here). Whether this is because the RNAi machinery differently processes the small RNAs or because readthrough does not occur at this region is not known.

The non-essential role of base J in *T. brucei* is consistent with the developmentally regulated synthesis of base J and absence of the modification in the procyclic stage of the parasite (24). It is possible that alternative J-independent mechanisms evolved to regulate RNAP II termination during the procyclic life-stage of the parasite within the Teste fly. Additionally, unlike *Leishmania* spp., where base J is found in all but one cSSR (27), J is not found at all cSSRs in *T. brucei* (18). Thus, base J plays an important function in regulating RNAP II termination genome-wide in *Leishmania* spp., but in *T. brucei* J has a more specialized function attenuating transcription and regulating gene expression at specific genomic loci. Additional experiments are underway to explore the mechanism through which J promotes RNAP II termination, specifically investigating whether the modification directly inhibits RNAP II elongation.

Accession number

All sequencing data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE57621.

Supplementary data

Supplementary Data are available at NAR online: Supplementary Tables S1-S2 and Supplementary Figures S1-S6.

Funding

This work was supported by the National Institutes of Health [grant number AI063523-03]; a grant jointly funded by the University of Georgia Franklin College of Arts & Sciences,

Office of the Vice President for Research, and Department of Biochemistry and Molecular Biology to [RS]; and funding by the Human Frontier Science Program to [TNS]. Funding for open access charge: National Institute of Health 063523.

Acknowledgement

We are grateful to Jessica Lopes da Rosa-Spiegler and Whitney Bullard for critical reading of the manuscript. Anti-J antiserum was kindly provided by the Borst laboratory.

REFERENCES

1. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B. *et al.* (2005) The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416-422.
2. Jackson, A.P., Sanders, M., Berry, A., McQuillan, J., Aslett, M.A., Quail, M.A., Chukualim, B., Capewell, P., MacLeod, A., Melville, S.E. *et al.* (2010) The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. *PLoS neglected tropical diseases*, **4**, e658.
3. El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404-409.
4. Martinez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K. and Myler, P.J. (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Mol. Cell*, **11**, 1291-1299.

5. Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S. and Cross, G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev.*, **23**, 1063-1076.

6. Thomas, S., Green, A., Sturm, N.R., Campbell, D.A. and Myler, P.J. (2009) Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*, **10**, 152.

7. Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S. and Tschudi, C. (2010) The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog.*, **6**, e1001090.

8. Boothroyd, J.C. and Cross, G.A. (1982) Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. *Gene*, **20**, 281-289.

9. Van der Ploeg, L.H., Liu, A.Y., Michels, P.A., De Lange, T.D., Borst, P., Majumder, H.K., Weber, H., Veeneman, G.H. and Van Boom, J. (1982) RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes. *Nucleic Acids Res*, **10**, 3591-3604.

10. De Lange, T., Liu, A.Y., Van der Ploeg, L.H., Borst, P., Tromp, M.C. and Van Boom, J.H. (1983) Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? *Cell*, **34**, 891-900.

11. Nelson, R.G., Parsons, M., Barr, P.J., Stuart, K., Selkirk, M. and Agabian, N. (1983) Sequences homologous to the variant antigen mRNA spliced leader are located in tandem repeats and variable orphons in Trypanosoma brucei. *Cell*, **34**, 901-909.

12. Sutton, R.E. and Boothroyd, J.C. (1986) Evidence for Trans splicing in trypanosomes. *Cell*, **47**, 527-535.

13. Agabian, N. (1990) Trans splicing of nuclear pre-mRNAs. *Cell*, **61**, 1157-1160.

14. Gromak, N., West, S. and Proudfoot, N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Molecular and cellular biology*, **26**, 3986-3996.

15. LeBowitz, J.H., Smith, H.Q., Rusche, L. and Beverley, S.M. (1993) Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev.*, **7**, 996-1007.

16. Clayton, C.E. (2002) Life without transcriptional control? From fly to man and back again. *EMBO J*, **21**, 1881-1888.

17. Campbell, D.A., Thomas, S. and Sturm, N.R. (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.*, **5**, 1231-1240.

18. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A. and Sabatini, R. (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res*, **38**, 3923-3935.

19. Respuela, P., Ferella, M., Rada-Iglesias, A. and Aslund, L. (2008) Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. *The Journal of biological chemistry*, **283**, 15884-15892.

20. van Leeuwen, F., Taylor, M.C., Mondragon, A., Moreau, H., Gibson, W., Kieft, R. and Borst, P. (1998) beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. *Proc. Natl. Acad. Sci. U S A*, **95**, 2366-2371.

21. Dooijes, D., Chaves, I., Kieft, R., Dirks-Mulder, A., Martin, W. and Borst, P. (2000) Base J originally found in kinetoplastid is also a minor constituent of nuclear DNA of Euglena gracilis. *Nucleic Acids Res*, **28**, 3017-3021.

22. Gommers-Ampt, J.H., Van Leeuwen, F., de Beer, A.L., Vliegenthart, J.F., Dizdaroglu, M., Kowalak, J.A., Crain, P.F. and Borst, P. (1993) beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. *Cell*, **75**, 1129-1136.

23. van Leeuwen, F., Kieft, R., Cross, M. and Borst, P. (2000) Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosal-hydroxymethyluracil in *Trypanosoma brucei*. *Molecular and biochemical parasitology*, **109**, 133-145.

24. van Leeuwen, F., Wijsman, E.R., Kieft, R., van der Marel, G.A., van Boom, J.H. and Borst, P. (1997) Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes & development*, **11**, 3232-3241.

25. van Leeuwen, F., Wijsman, E.R., Kuyl-Yeheskiely, E., van der Marel, G.A., van Boom, J.H. and Borst, P. (1996) The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands. *Nucleic Acids Res*, **24**, 2476-2482.

26. Ekanayake, D.K., Minning, T., Weatherly, B., Gunasekera, K., Nilsson, D., Tarleton, R., Ochsenreiter, T. and Sabatini, R. (2011) Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. *Mol. Cell. Biol.*, **31**, 1690-1700.

27. van Luenen, H.G., Farris, C., Jan, S., Genest, P.A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G. *et al.* (2012) Glucosylated

hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, **150**, 909-921.

28. Borst, P. and Sabatini, R. (2008) Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol*, **62**, 235-251.

29. Cliffe, L.J., Hirsch, G., Wang, J., Ekanayake, D., Bullard, W., Hu, M., Wang, Y. and Sabatini, R. (2012) JBP1 and JBP2 Proteins Are Fe2+/2-Oxoglutarate-dependent Dioxygenases Regulating Hydroxylation of Thymidine Residues in Trypanosome DNA. *J. Biol. Chem.*, **287**, 19886-19895.

30. Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930-935.

31. Iyer, L.M., Tahiliani, M., Rao, A. and Aravind, L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698-1710.

32. Kieft, R., Brand, V., Ekanayake, D.K., Sweeney, K., DiPaolo, C., Reznikoff, W.S. and Sabatini, R. (2007) JBP2, a SWI2/SNF2-like protein, regulates de novo telomeric DNA glycosylation in bloodstream form Trypanosoma brucei. *Molecular and biochemical parasitology*, **156**, 24-31.

33. Cross, M., Kieft, R., Sabatini, R., Dirks-Mulder, A., Chaves, I. and Borst, P. (2002) J binding protein increases the level and retention of the unusual base J in trypanosome DNA. *Molecular microbiology*, **46**, 37-47.

34. Cliffe, L.J., Kieft, R., Southern, T., Birkeland, S.R., Marshall, M., Sweeney, K. and Sabatini, R. (2009) JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res*, **37**, 1452-1462.

35. Genest, P.A., ter Riet, B., Dumas, C., Papadopoulou, B., van Luenen, H.G. and Borst, P. (2005) Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. *Nucleic Acids Res*, **33**, 1699-1709.

36. Ekanayake, D. and Sabatini, R. (2011) Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryotic cell*, **10**, 1465-1472.

37. Michishita, E., Nakabayashi, K., Suzuki, T., Kaul, S.C., Ogino, H., Fujii, M., Mitsui, Y. and Ayusawa, D. (1999) 5-Bromodeoxyuridine induces senescence-like phenomena in mammalian cells regardless of cell type or species. *J Biochem*, **126**, 1052-1059.

38. Miki, K., Shimizu, M., Fujii, M., Takayama, S., Hossain, M.N. and Ayusawa, D. (2010) 5-bromodeoxyuridine induces transcription of repressed genes with disruption of nucleosome positioning. *FEBS J*, **277**, 4539-4548.

39. Goldsworthy, T., Dunn, C. and Popp, J. (1992) Dose effects of bromodeoxyuridine (BRDU) on rodent hepatocyte proliferation measurements. *Toxicologist*, **12**, 265.

40. Duque, A. and Rakic, P. (2011) Different effects of bromodeoxyuridine and [3H]thymidine incorporation into DNA on cell proliferation, position, and fate. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **31**, 15205-15217.

41. DiPaolo, C., Kieft, R., Cross, M. and Sabatini, R. (2005) Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Molecular cell*, **17**, 441-451.

42. Kapler, G.M., Coburn, C.M. and Beverley, S.M. (1990) Stable transfection of the human parasite Leishmania major delineates a 30-kilobase region sufficient for extrachromosomal replication and expression. *Mol Cell Biol*, **10**, 1084-1094.

43. Bernards, A., Van der Ploeg, L.H., Frasch, A.C., Borst, P., Boothroyd, J.C., Coleman, S. and Cross, G.A. (1981) Activation of trypanosome surface glycoprotein genes involves a duplication-transposition leading to an altered 3' end. *Cell*, **27**, 497-505.

44. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357-359.

45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

46. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**, 621-628.

47. Abuin, G., Colli, W., de Souza, W. and Alves, M.J. (1989) A surface antigen of Trypanosoma cruzi involved in cell invasion (Tc-85) is heterogeneous in expression and molecular constitution. *Molecular and biochemical parasitology*, **35**, 229-237.

48. Roditi, I., Schwarz, H., Pearson, T.W., Beecroft, R.P., Liu, M.K., Richardson, J.P., Bühring, H.J., Pleiss, J., Bülow, R. and Williams, R.O. (1989) Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. *J. Cell. Biol.*, **108**, 737-746.

49. Al Husini, N., Kudla, P. and Ansari, A. (2013) A role for CF1A 3' end processing complex in promoter-associated transcription. *PLoS genetics*, **9**, e1003722.

50. Fujii, M., Ito, H., Hasegawa, T., Suzuki, T., Adachi, N. and Ayusawa, D. (2002) 5-Bromo-2'-deoxyuridine efficiently suppresses division potential of the yeast Saccharomyces cerevisiae. *Biosci Biotechnol Biochem*, **66**, 906-909.

51. Tschudi, C., Fau, S.H. and Ullu, E. (2012) Small interfering RNA-producing loci in the ancient parasitic eukaryote Trypanosoma brucei. *BMC Genomics*, **13**.

52. Suzuki, T., Michishita, E., Ogino, H., Fujii, M. and Ayusawa, D. (2002) Synergistic induction of the senescence-associated genes by 5-bromodeoxyuridine and AT-binding ligands in HeLa cells. *Experimental cell research*, **276**, 174-184.

53. Suzuki, T., Yaginuma, M., Oishi, T., Michishita, E., Ogino, H., Fujii, M. and Ayusawa, D. (2001) 5-Bromodeoxyuridine suppresses position effect variegation of transgenes in HeLa cells. *Experimental cell research*, **266**, 53-63.

54. Hobson, D.J., Wei, W., Steinmetz, L.M. and Svejstrup, J.Q. (2012) RNA polymerase II collision interrupts convergent transcription. *Molecular cell*, **48**, 365-374.

55. Zheng, L.L., Wen, Y.Z., Yang, J.H., Liao, J.Y., Shao, P., Xu, H., Zhou, H., Wen, J.Z., Lun, Z.R., Ayala, F.J. *et al.* (2013) Comparative transcriptome analysis of small noncoding RNAs in different stages of Trypanosoma brucei. *RNA.*, **19**, 863-875.

56. Ngô, H., Tschudi, C., Gull, K. and Ullu, E. (1998) Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14687-14692.

57. Kelly, S., Kramer, S., Schwede, A., Maini, P.K., Gull, K. and Carrington, M. (2012) Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. *Open biology*, **2**, 120033.

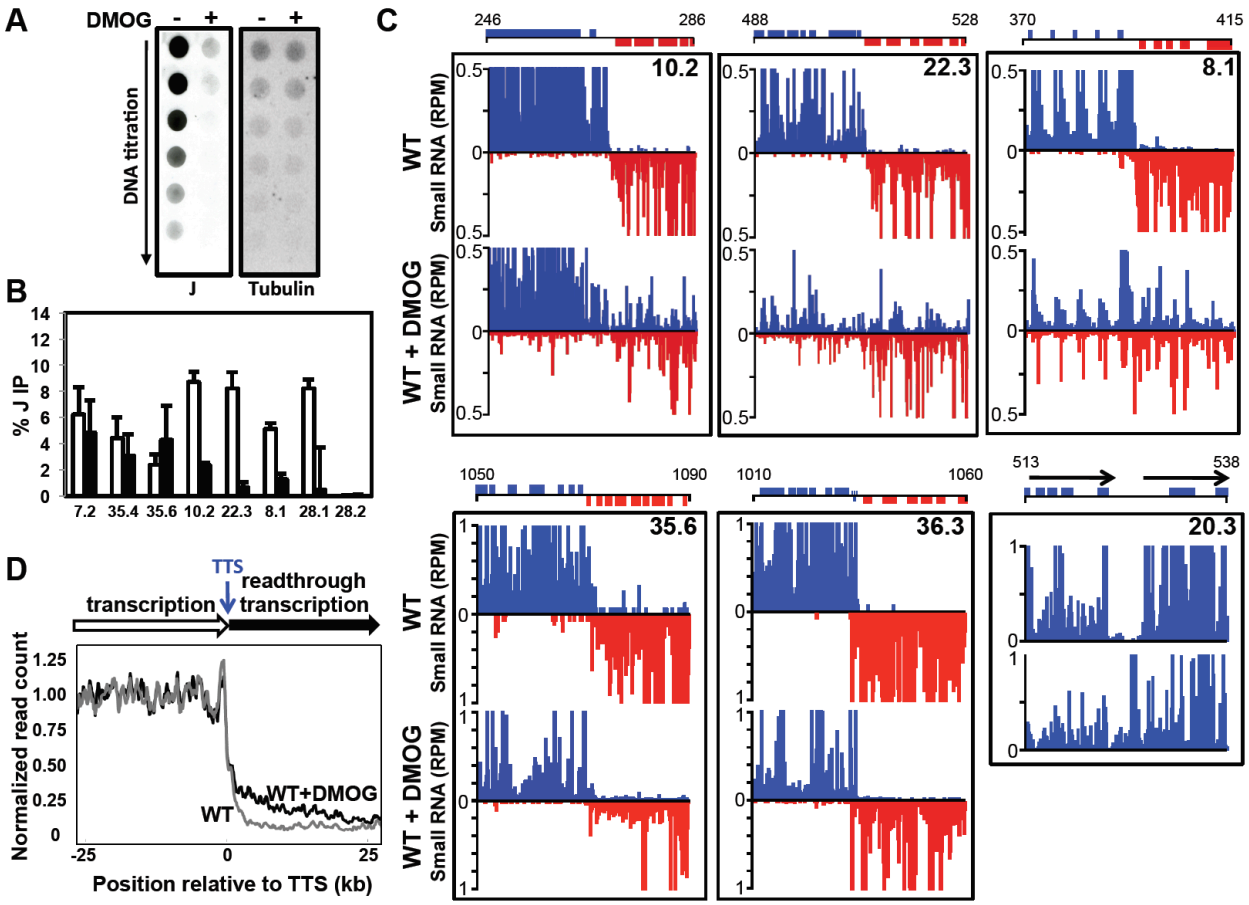


Figure 3.1 Loss of base J results in readthrough transcription and the production of antisense RNAs in *L. major*.

(A) Anti-base J dot blot analysis of WT *L. major*. A two-fold serial dilution of genomic DNA was spotted onto a membrane and incubated with anti-base J antisera. -,DMSO; +, 5mM DMOG. The membrane was stripped and probed using a radiolabeled beta tubulin probe as a loading control.

(B) Anti-base J IP qPCR analysis. The average of three independent IPs is plotted as the percent IP relative to the total input material. All IPs were background subtracted using a no antibody control. White bars, DMSO; black bars, DMOG. Seven cSSRs enriched for base J are shown and one previously identified J negative cSSR (28.2) as a negative control (see Supplemental Table 1 for genomic location). Error bars represent the standard deviation.

(C) Upper panels, small RNA sequencing reads for three cSSRs where J loss led to readthrough transcription are shown. Small RNA reads are plotted as reads per million reads mapped (rpm). Upper graphs, DMSO; lower graphs, DMOG. ORFs and the genomic location (kb) are shown

above the graphs. Blue, top strand; red, bottom strand. Lower panels, cSSR 35.6 illustrates a region where J was not reduced by DMOG (see Figure 3.1B) and there was no readthrough defect; cSSR 36.3 shows a site containing tRNA genes on both DNA strands where J was reduced by DMOG (see Figure 3.S1C), but did not result in a readthrough defect; and HT site 20.3 shows a non-cSSR termination site where J loss (see Figure 3.S1C) resulted in a termination defect.

(D) A metaplot summarizing the readthrough defect at cSSRs (n=36, 3 discarded) aligned by their TTS, shown as position 0 on the x-axis. Meta-coverage of each sample was normalized by the mean meta-coverage of upstream of TTS (See methods).

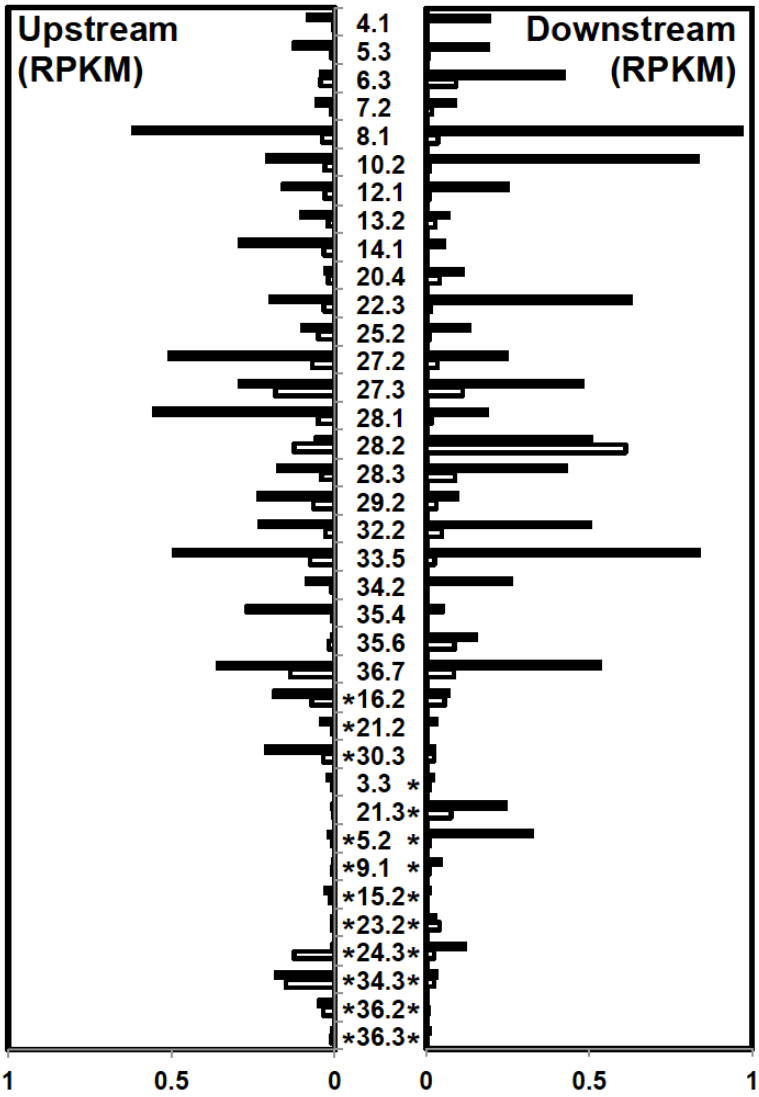


Figure 3.2 Quantification of readthrough transcription at individual cSSRs in *L. major*. The antisense small RNA reads per kb per million reads mapped (RPKM) within a 5kb window downstream of the TTS is plotted for each cSSR. cSSRs are labeled in the center. Bars to the left represent readthrough transcription on the bottom strand (readthrough transcription from right to left) and bars to the right represent readthrough transcription on the top strand (readthrough transcription from left to right). White bars, DMSO; black bars, DMOG. cSSRs containing tRNA genes are indicated by an asterisk; cSSRs with tRNAs on the top strand only (* on the right side of the cSSR number, bottom strand only (* on the left side) or both strands. See Supplemental Table 1 for the genomic location of each cSSR. cSSR 9.2 and 22.2 were excluded here because the downstream PTU was less than 5kb.

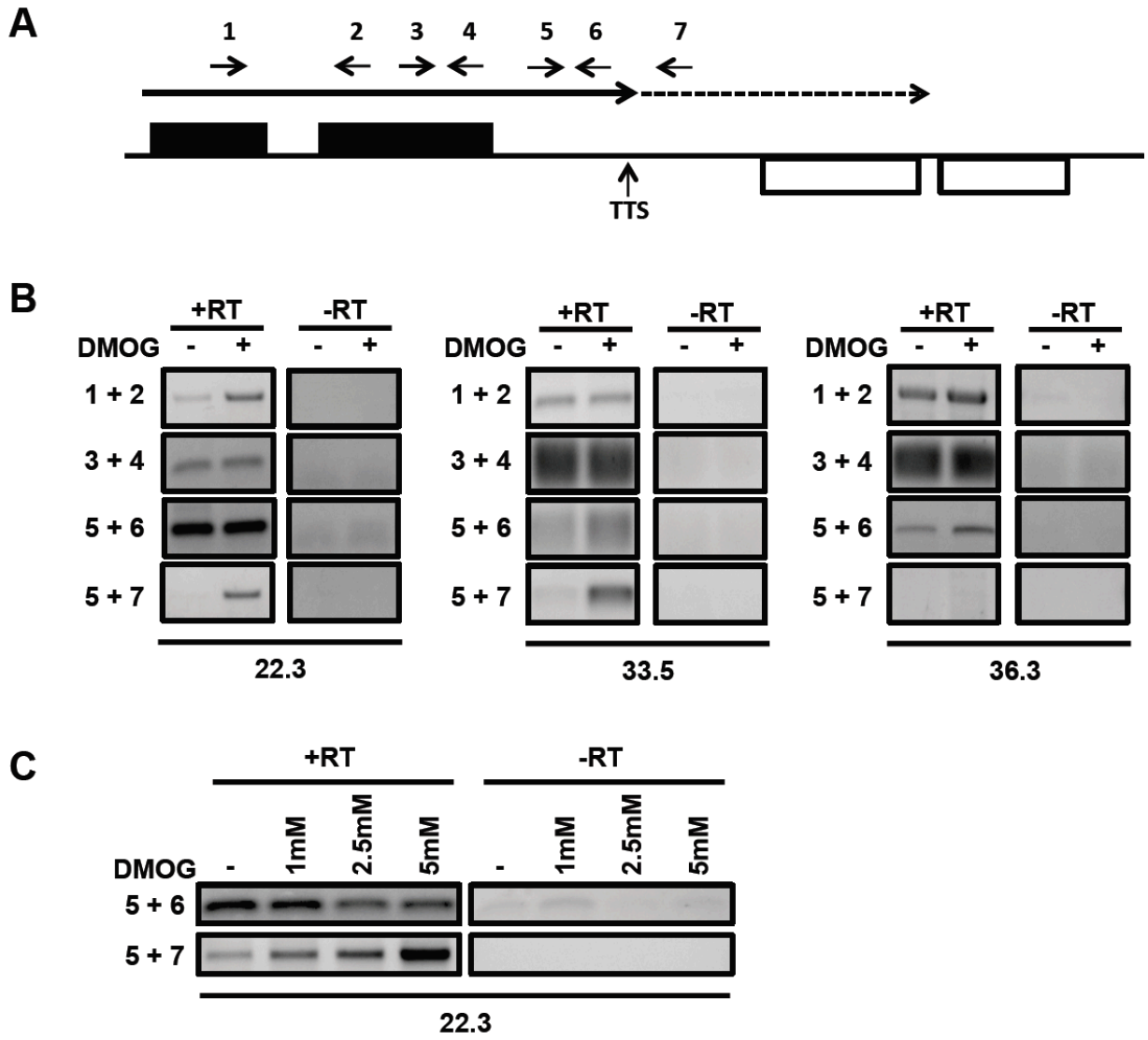


Figure 3.3 RNAP II fails to terminate following reduction of base J.

(A) Schematic representation of primer location and direction at a cSSR. The dashed arrow indicates readthrough transcription past the TTS.

(B) Single strand RT-PCR analysis. cDNA was synthesized using the reverse primers 2, 4, 6, and 7. PCR was performed using the same reverse primer used to make the cDNA plus a forward primer, as indicated. Three cSSRs were analyzed, two with readthrough upon J loss (22.3 and 33.5) and one without readthrough (36.3), as determined by small RNA-seq analysis. Plus RT and minus RT controls are shown. -, DMSO; +, DMOG (5mM).

(C) The amount of readthrough transcription correlates with the extent of J loss. Cells treated with 0, 1, 2.5, and 5mM DMOG were analyzed by single-strand RT PCR as described above.

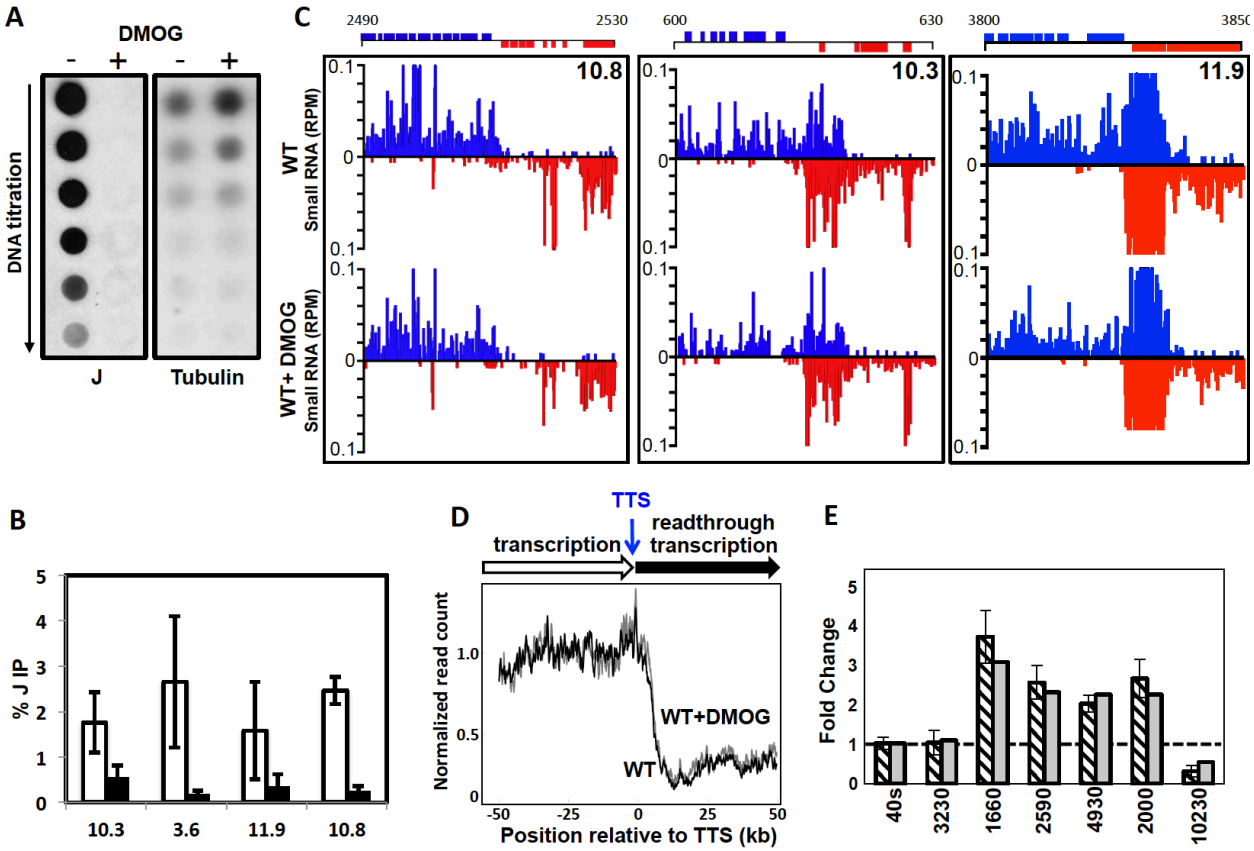


Figure 3.4 Loss of base J does not lead to readthrough transcription in *T. brucei* at cSSRs.

(A) Anti-base J dot blot analysis of WT *T. brucei* was performed as detailed for figure 1A, but *T. brucei* genomic DNA was isolated from cells treated with either 1mM DMOG or DMSO for 5 days.

(B) Anti-base J IP qPCR analysis. Same as shown in Figure 1B, where white bars show %IP for DMSO and black bars show % J IP from DMOG treated *T. brucei* DNA. Four cSSRs enriched for base J are shown. Error bars represent the standard deviation.

(C) Small RNA-seq analysis of three cSSRs (10.8, 10.3, and 11.9) is shown with reads plotted as reads per million reads mapped (rpm). Top graphs, DMSO treated WT; bottom graphs, DMOG treated WT. Reads mapped to the top strand are shown in blue and reads mapped to the bottom strand in red. ORFs and their chromosomal location in kb are shown above.

(D) A metaplot summarizing the readthrough defect at cSSRs (n=87, 11 discarded) aligned by their TTS, shown as position 0 on the x-axis. Meta-coverage of each sample was normalized by the mean meta-coverage of upstream of TTS (See methods).

(E). Confirmation of total RNA-seq transcript changes in *T. brucei* by RT qPCR.

Fold change in transcript abundance, with DMSO treated WT set to 1. Striped bars, fold change based on RT qPCR analysis of DMOG treated WT *T. brucei*; grey bars, fold change in the RPKM of DMOG treated WT *T. brucei* based on total RNA-seq analysis. For RT qPCR analysis, transcripts were normalized against 40s rRNA. Transcripts analyzed included 3230 (Tb427.07.3230), which did not change in abundance after DMOG treatment; 1660 (Tb427.08.1660), 2590 (Tb427.03.2590), 4930 (Tb427tmp.160.4930), and 2000 (Tb427.07.2000), which were increased by at least two-fold after DMOG treatment; and 10230 (Tb427.10.10230), which was decreased about two-fold following DMOG treatment. Error bars represent the standard deviation of three independent biological replicates.

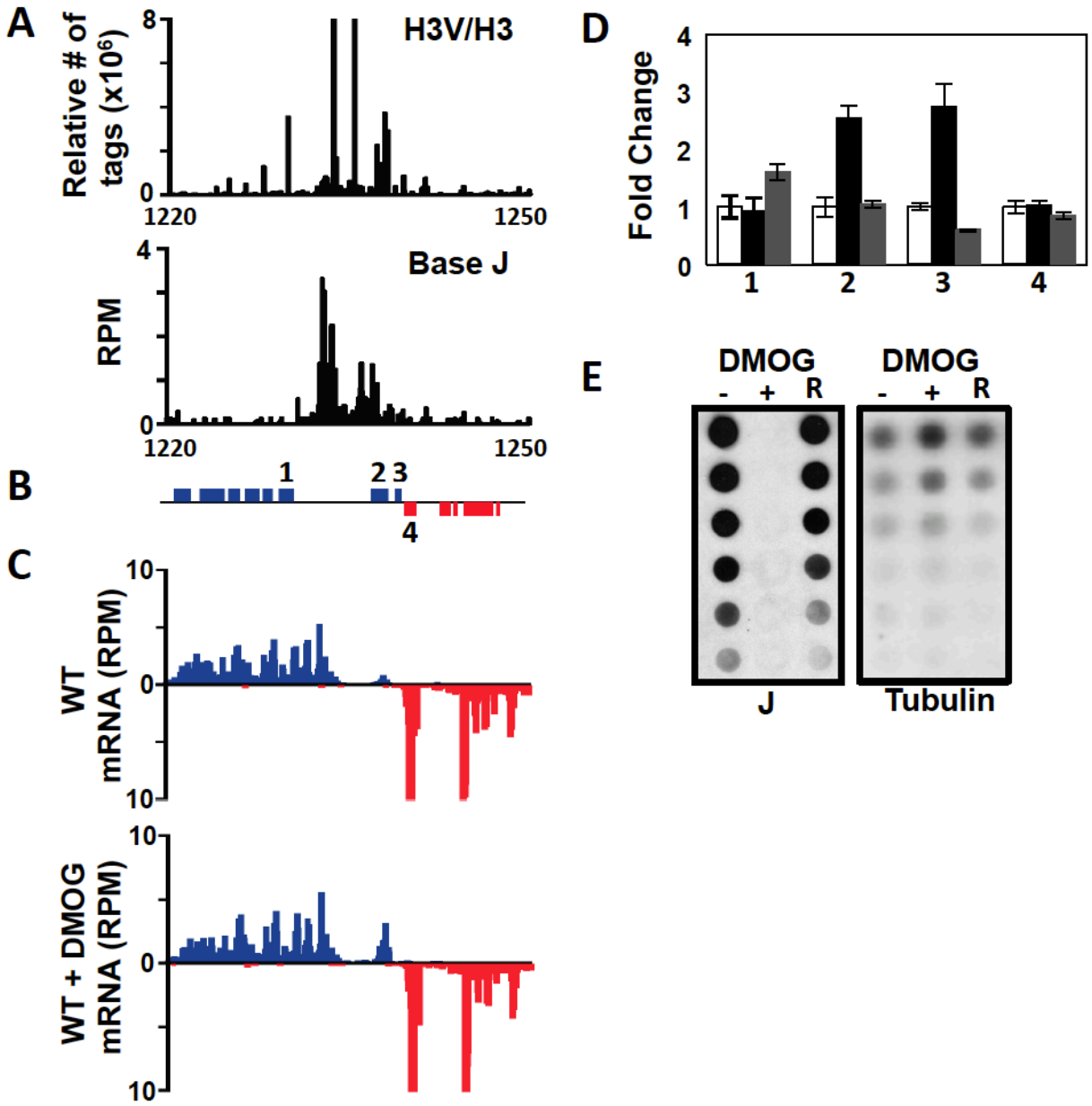


Figure 3.5 Base J regulates gene expression at the level of RNAP II transcription.

A region on chromosome 5 is shown where base J regulates RNAP II transcription.

(A) Base J and H3V co-localize at sites of RNAP II termination. H3V ChIP-seq reads are plotted as the relative number of sequence tags normalized to H3 ChIP-seq reads, as previously described (5) and base J IP-seq reads are plotted as reads per million mapped reads (RPM).

(B) ORFs are shown with the top strand in blue and the bottom strand in red. Genes analyzed by RT qPCR in (D) are labeled 1-4; 1, Tb427.05.3980 (hypothetical protein, conserved); 2,

Tb427.05.3990 (variant surface glycoprotein, atypical, putative); 3, Tb427.05.4000 (hypothetical protein); and 4, Tb427.05.4010 (hypothetical protein).
(C) Total RNA-seq reads plotted as RPM. Mapped reads from DMSO treated WT *T. brucei* are shown above and those from DMOG treated WT *T. brucei* are shown below. Reads that mapped to the top strand are shown in blue and reads that mapped to the bottom strand in red.
(D) RT qPCR confirmation of transcript changes. The transcript fold change is shown for the four genes (1-4) shown in (A). White bars, DMSO treated WT, set to 1; black bars, DMOG treated WT; and grey bars, J rescue, where cells were grown for 10 days in medium without DMOG, allowing J to be re-synthesized. Error bars represent the standard deviation of three independent biological replicates.
(E) Anti-base J dot blot analysis. -, DMSO treated WT; +, DMOG treated WT; and R, J rescue.

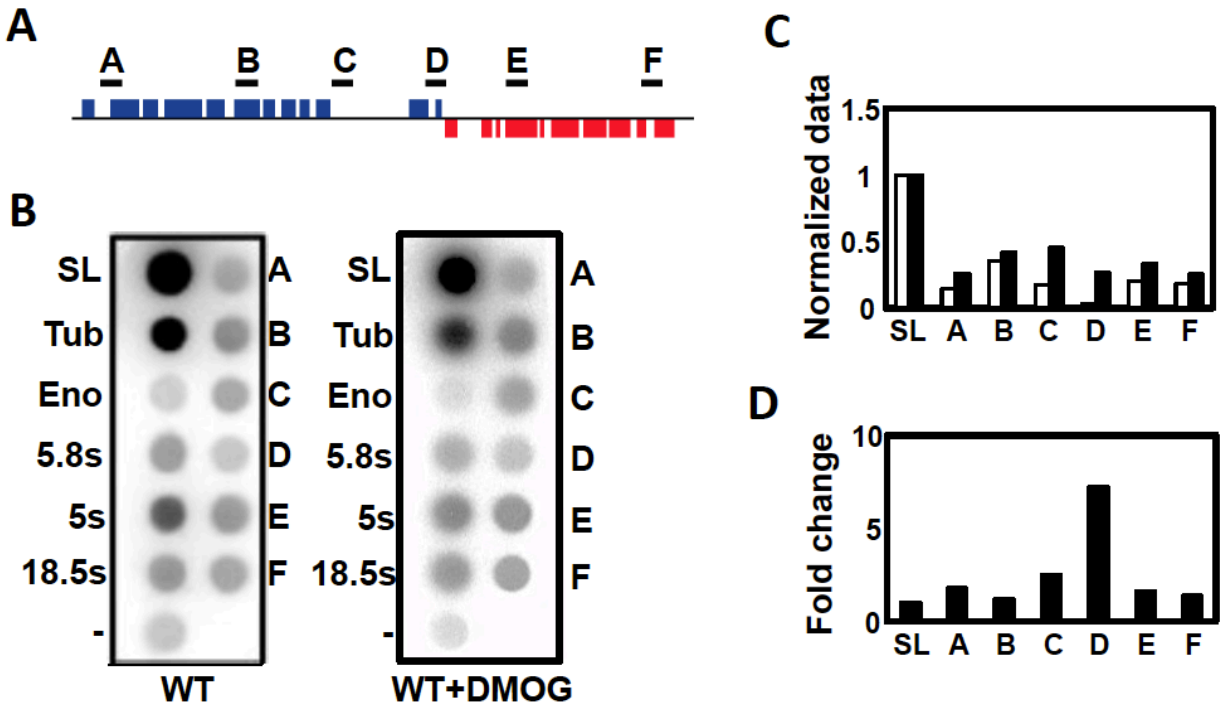


Figure 3.6 Nuclear run-on analysis of the region shown in Figure 5.

(A) The location of probes A-F is shown at the top.

(B) Results from a representative experiment are shown, with DMSO treated WT on the left and DMOG treated WT on the right. The probes used include SL, spliced leader; Tub, tubulin; Eno, enolase; 5.8s, 5.8s rRNA; 5s, 5s rRNA; 18.5s, 18.5s rRNA, -, empty pCR2.1 vector; and probes A-F illustrated above. The radioactive signal was measured using a phosphoimager and normalized using the SL signal. The empty pCR2.1 vector was used to subtract background signal.

(C) The normalized signal, white bars DMSO, black bars DMOG.

(D) Fold change in the DMOG treated WT signal, with the DMSO treated WT signal set to 1.

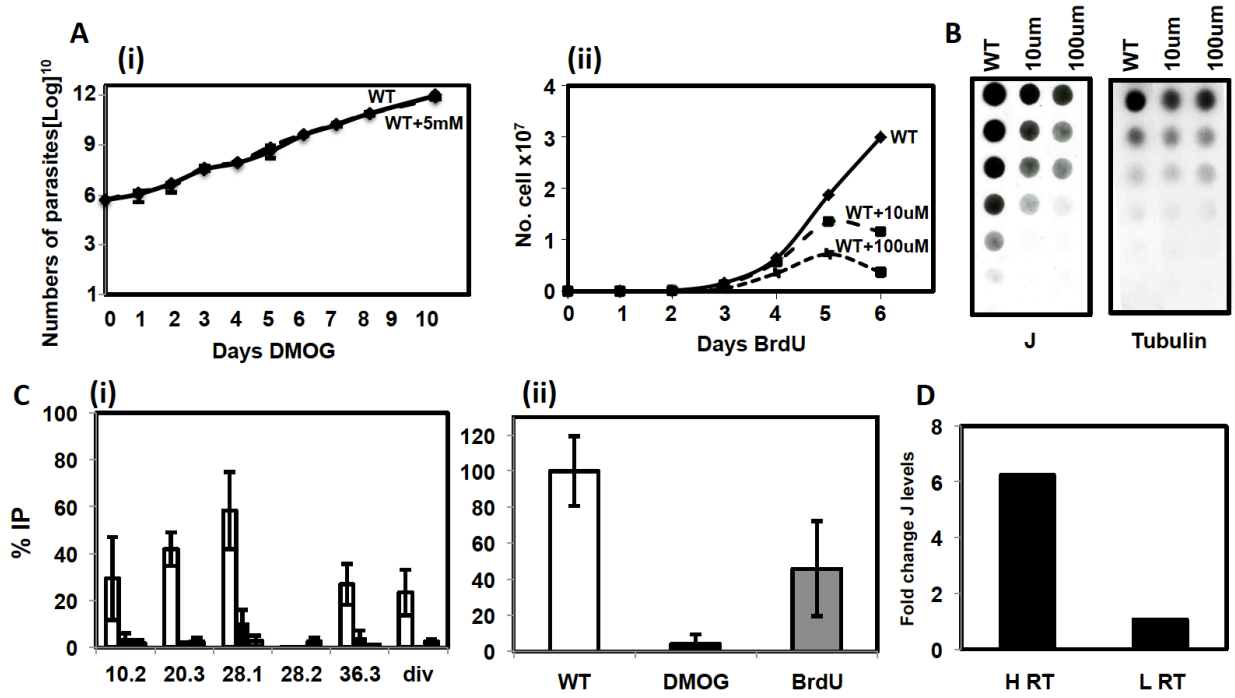


Figure 3.S1 DMOG reduces base J and does not result in cell death in *L. major*.

(A) Growth curves of DMSO and 5mM DMOG treated WT *L. major* (i) and 10μM and 100μM BrdU treated WT *L. major* (ii)

(B) Anti-base J dot blot analysis of total J levels in WT and BrdU treated cells, as indicated. DNA was isolated from cells on day 6.

(C) J IP qPCR analysis. White bars, DMSO; black bars, DMOG treated cells; and grey bars, BrdU treated cells. (i) cSSRs analyzed are indicated below the graph. Div indicates a divergent strand switch region (RNAP II initiation site). (ii) qPCR analysis of the telomeric repeats. Error bars represent the standard deviation of three independent IPs.

(D) The average fold reduction in J levels at cSSRs with high readthrough (HRT) and cSSRs with low readthrough (LRT). HRT cSSRs include 10.2, 22.3, 8.1, and 28.1. LRT cSSRs include 7.2, 35.4, and 35.6.

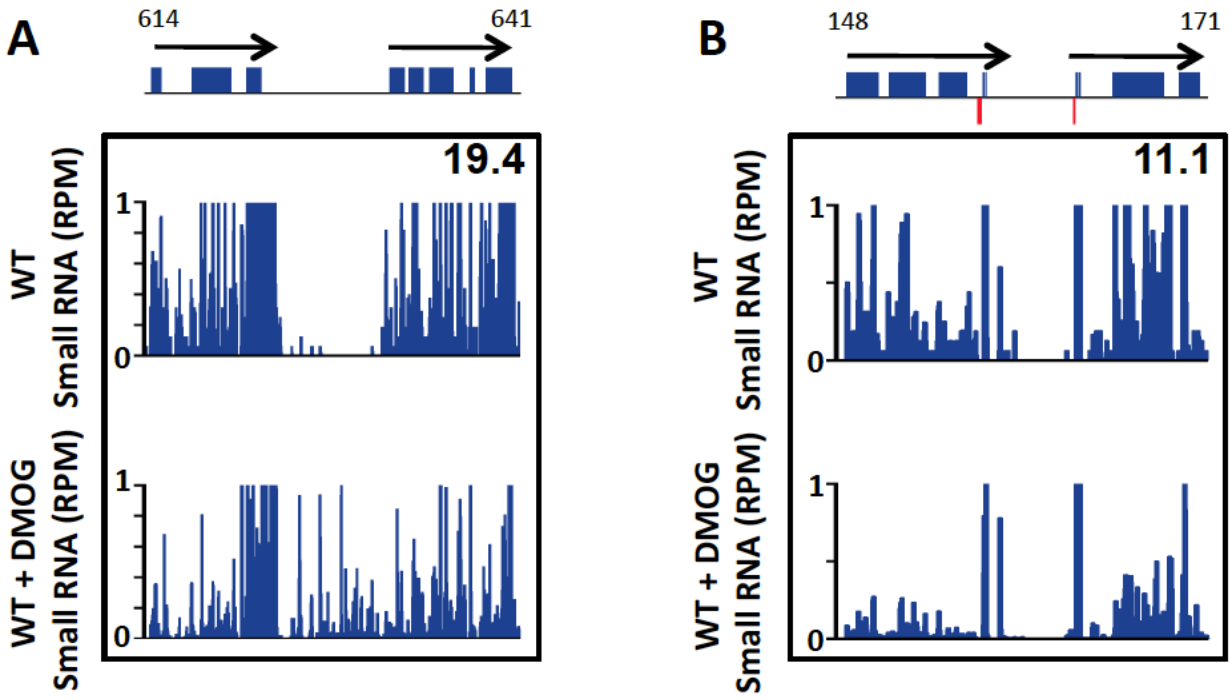


Figure 3.S2 RNAP III genes at HT sites prevent RNAP II transcriptional readthrough upon J reduction in *L. major*.

(A) A HT site on chromosome 19 from position 614-641kb is shown. Top, the location of ORFs. Genes on the top strand are shown in blue (no annotated genes present on the bottom strand). Arrows indicate the direction of RNAP II transcription. Small RNA-seq reads were mapped and are shown as reads per million reads mapped (rpm) for both DMSO treated WT *L. major* and DMOG treated WT *L. major*. Only reads mapped to the top strand are shown.

(B) Same as (A), but a HT site containing tRNA genes on chromosome 11 from position 148-171kb is shown. tRNA genes are indicated by thin lines in the center. Genes on the top strand are shown in blue, bottom strand in red.

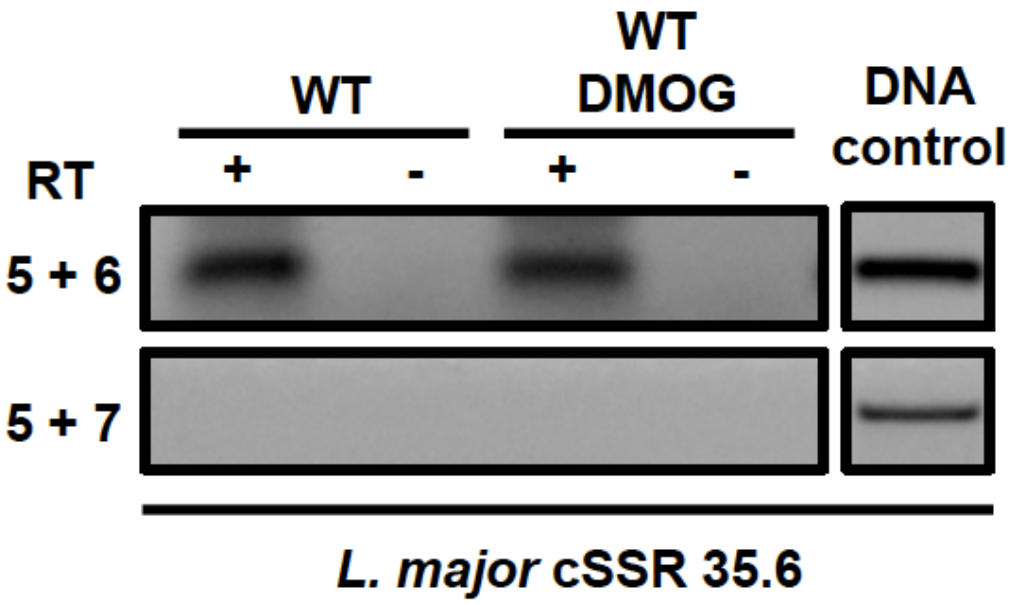


Figure 3.S3 Analysis of readthrough transcription in *L. major*.

L. major cSSR 35.6, where DMOG treatment did not significantly reduce base J, was analyzed by single-strand RT PCR. Primers correspond to the schematic shown in Figure 3A. Plus RT and minus RT controls are shown. WT cells were treated with DMSO only. Genomic DNA was used as a PCR positive control.

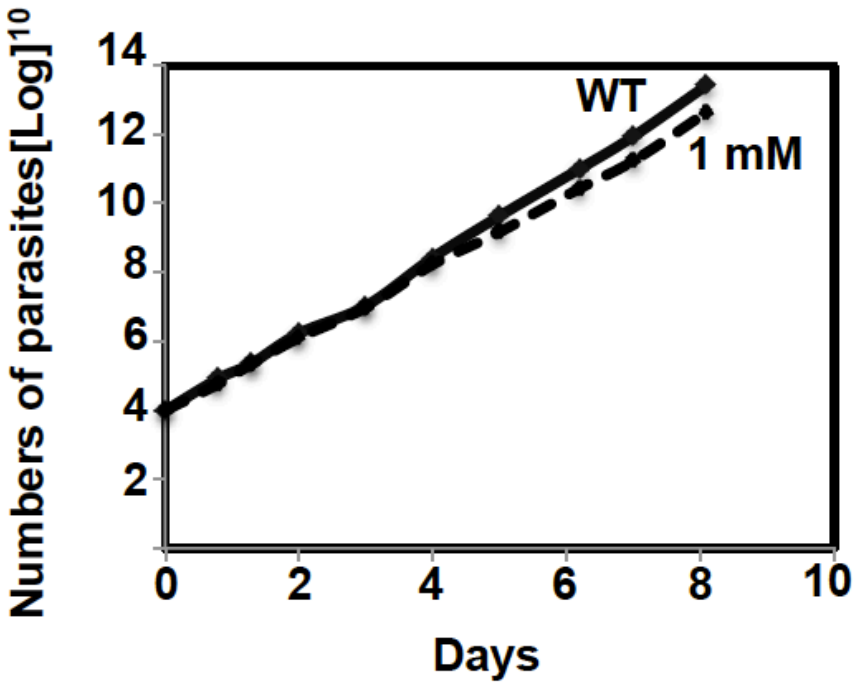


Figure 3.S4 WT *T. brucei* treated with 1mM DMOG does not result in a growth phenotype.

Cell number is plotted on a $\log_{10}$ scale. WT DMSO and WT+1mM DMOG are plotted.

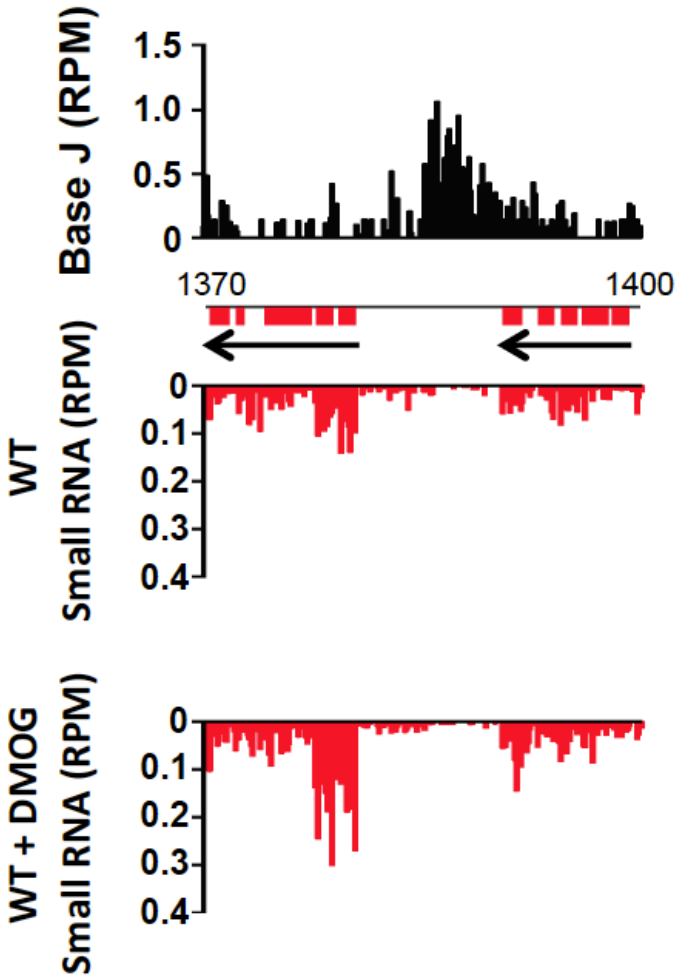


Figure 3.S5 Loss of base J from head-tail sites in *T. brucei* does not lead to readthrough transcription.

A HT site on chromosome 9 from position 1370-1400kb is shown. Mapped reads from base J IP-seq are shown at the top, plotted as reads per million reads mapped (rpm). ORFs are illustrated below. Small RNA-seq reads from DMSO treated WT and DMOG treated WT are plotted as rpm. Only reads mapped to the bottom strand are shown.

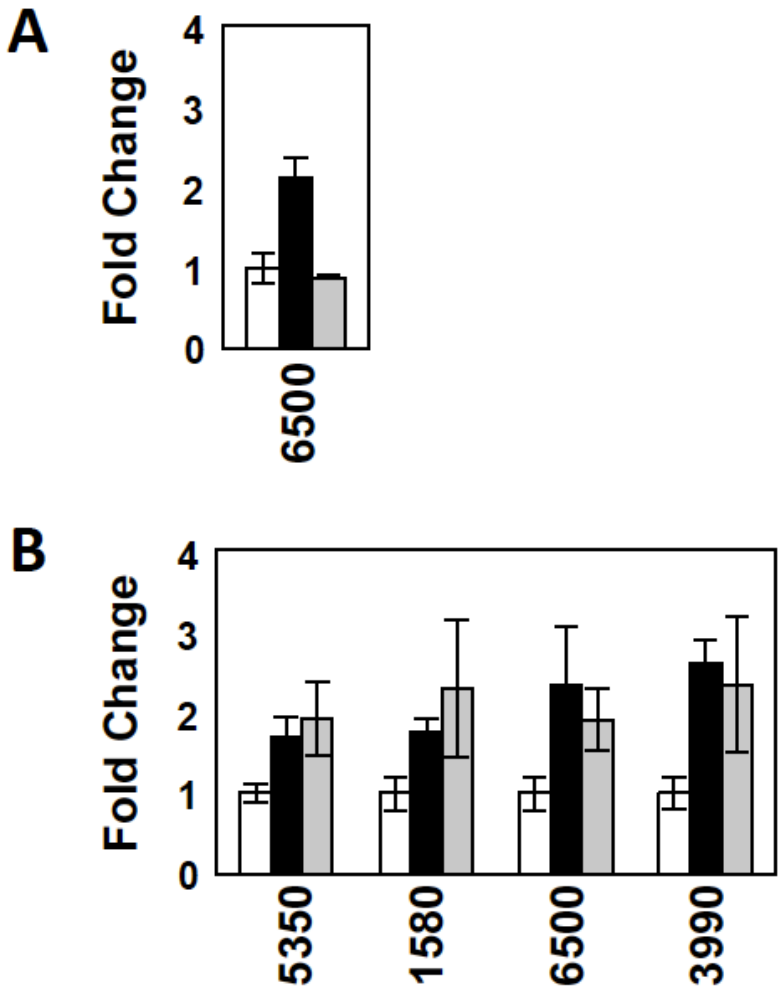


Figure 3.S6 Transcript increases in DMOG treated WT *T. brucei* can be rescued and are also found increased in JBP nulls (JBP1 and JBP2 KO).

(A RT qPCR analysis of transcripts from Tb427.07.6500, 6500. White bar, DMSO treated; black bar, DMOG. The grey bar represents a J rescue, where DMOG treated cells were grown in the absence of DMOG for 10 days. Error bars represent the standard deviation of three independent biological replicates.

(B)) RT qPCR analysis of genes found up-regulated by total RNA-seq following J loss. White bars, DMSO treated WT; black bars, DMOG treated WT; grey bars, JBP Null. DMSO treated WT was set to 1. Error bars represent the standard deviation of three independent biological replicates. Genes analyzed were 5350, Tb427tmp.160.5350; 1580, Tb427tmp.02.1580; 6500, Tb427.07.6500; and 3990, Tb427.05.3990.

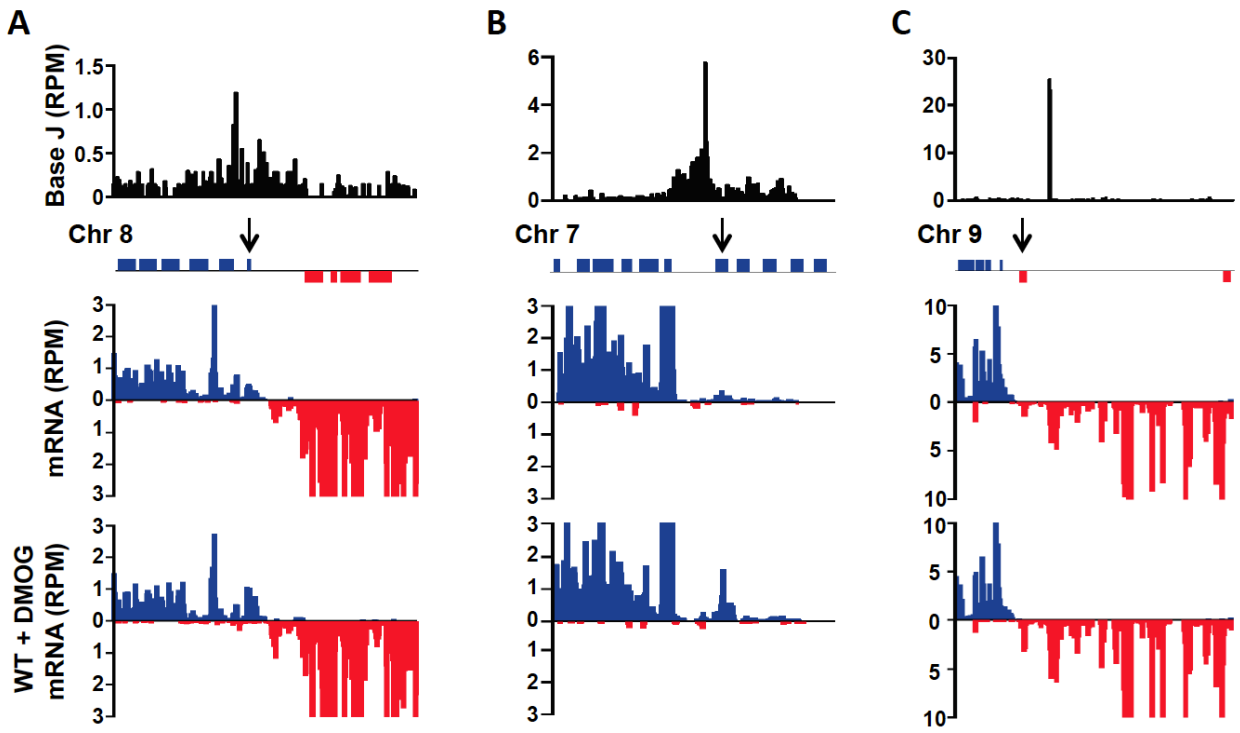


Figure 3.S7 Loss of base J results in up-regulated gene expression of downstream genes in *T. brucei*.

(A) A region on chromosome 8 from 540kb-570kb. Mapped reads from base J IP-seq are shown at the top, plotted as reads per million reads mapped (rpm). ORFs are illustrated below. Blue, genes on the top strand; red, genes on the bottom strand. Arrow indicates the gene found up-regulated in DMOG treated WT *T. brucei* by total RNA-seq (Tb427.08.1660). Total RNA-seq reads from DMSO treated WT and DMOG treated WT are plotted as rpm. Reads mapped to the top strand are shown in blue and reads mapped to the bottom strand in red.

(B) Same as in (A), but chromosome 7 from 1750kb-1780kb is shown. Arrow indicates gene Tb427.07.6500.

(C) Same as in (A), but chromosome 9 from 1120kb-1180kb is shown. Arrow indicates gene Tb427tmp.160.5350.

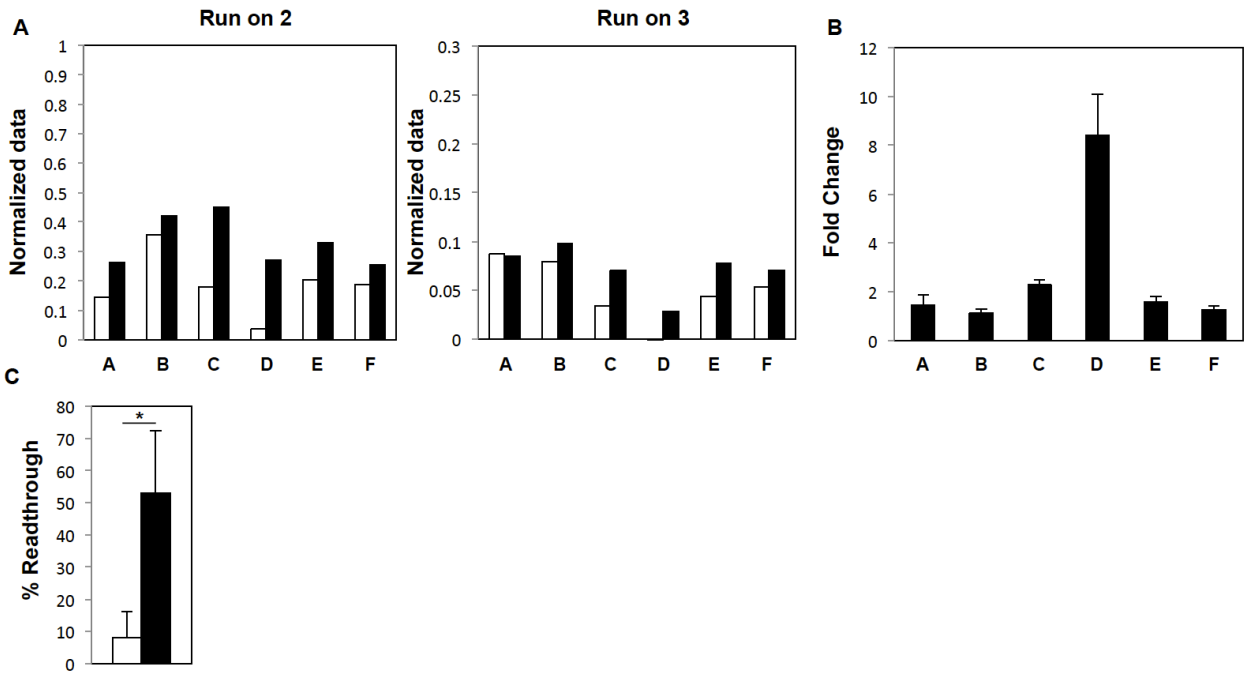


Figure 3.S8 Nuclear run on analysis.

(A) Normalized data from two additional independent run on experiments. White bars, WT *T. brucei*; Black bars, DMOG treated WT *T. brucei*. The probes used were the same as those shown in Figure 6A. Probe signal was normalized to SL, which was set to 1.

(B) The average fold change (DMOG treated WT normalized signal divided by WT *T. brucei* normalized signal) for three independent run on experiments. Error bars represent the fold change standard deviation of the three run on experiments with the exception of probe D. Fold change could not be determined for run on 3 probe D, shown in (A), given that the WT normalized signal was zero.

(C) The percent readthrough transcription for each experiment was determined by dividing the normalized signal for probe D by the upstream average normalized signal of probes A-C. White bar, WT *T. brucei*; Black bar, DMOG treated WT *T. brucei*. Error bars represent the standard deviation of three independent run on experiments. Asterisk indicates p-value < 0.05 by one-tailed Student's t-Test.

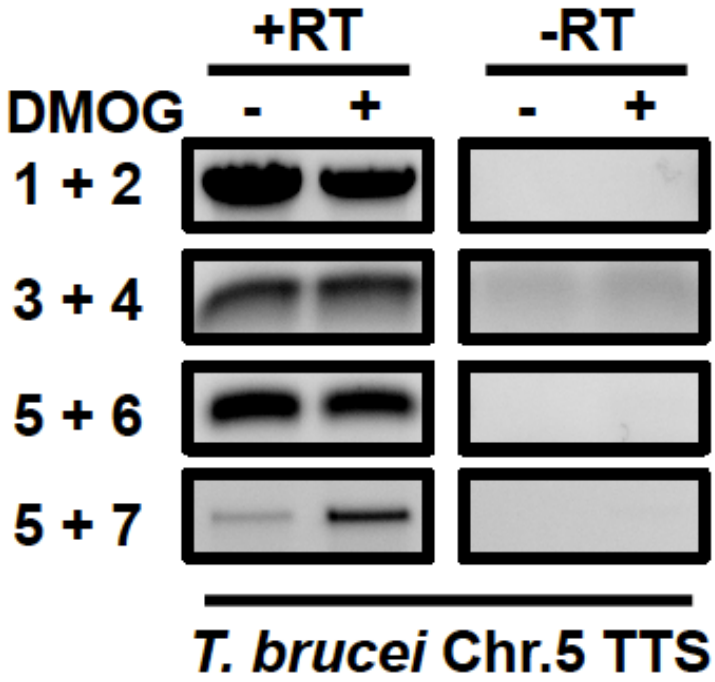


Figure 3.S9 Readthrough transcription in *T. brucei*.

WT and WT DMOG treated *T. brucei* cells were analyzed by single-strand RT PCR. The region upstream of gene Tb427.05.3990, which was found increased upon J loss, was analyzed. Primers correspond to the schematic shown in Figure 3A. Plus RT and minus RT controls are shown. -, DMSO; +, 1mM DMOG.

Table 3.1 *T. brucei* gene expression changes following J loss (2-fold or greater)

Up-regulated genes

Gene	Description	WT RPKM	WT DMOG RPKM	Fold Up-regulation[1]	Fits Model?[2]
Tb427.05.3990	variant surface glycoprotein (VSG, atypical), putative	2.61	12.33	4.72	Yes
Tb427.07.6500	variant surface glycoprotein (VSG), putative	1.24	4.5	3.64	Yes
Tb427tmp.v2.0170	variant surface glycoprotein (VSG, pseudogene), putative	0.41	1.32	3.24	Yes
Tb427.02.5384	retrotransposon hot spot protein (RHS, pseudogene), putative	0.85	2.72	3.21	Yes
Tb427tmp.160.3400	glucose transporter (pseudogene), putative	1.83	5.51	3.02	Yes
Tb427.08.1660	procyclin-associated gene (pseudogene), putative	6.59	18.76	2.85	Yes
Tb427tmp.160.4950	hypothetical protein, conserved	1.81	5.09	2.81	No
Tb427.04.140	hypothetical protein	1.25	3.45	2.76	No
Tb427tmp.02.1580	leucine-rich repeat protein (LRRP), putative	2.05	5.65	2.75	Yes[3]
Tb427.06.110	hypothetical protein	2.28	6.26	2.74	No
Tb427tmp.160.5350	variant surface glycoprotein (VSG)-related, putative	4.54	11.79	2.6	Yes
Tb427.02.720	retrotransposon hot spot protein (RHS, pseudogene), putative	2.24	5.77	2.58	Yes
Tb427.01.520	variant surface glycoprotein (VSG, pseudogene), putative	0.97	2.43	2.51	Yes
Tb427tmp.244.2020	nucleoside transporter 1, putative	0.38	0.95	2.48	Yes
Tb427tmp.02.1565	hypothetical protein	1.32	3.16	2.4	Yes[3]
Tb427.02.6220	adenosine transporter 2, putative (TbNT4)	1.43	3.42	2.39	No
Tb427.04.290	retrotransposon hot spot (RHS) protein, putative	0.4	0.95	2.38	Yes
Tb427tmp.244.0980	variant surface glycoprotein (VSG, pseudogene), putative	2.81	6.47	2.3	Yes
Tb427.07.1960	retrotransposon hot spot (RHS) protein, putative	0.3	0.69	2.26	Yes
Tb427.02.910	expression site-associated gene (ESAG, pseudogene), putative	14.19	32	2.26	No
Tb427.02.1260	expression site-associated gene (ESAG, pseudogene), putative	2.74	5.9	2.15	No
Tb427.03.2590	hypothetical protein	6.66	14.21	2.13	No
Tb427tmp.160.0010	variant surface glycoprotein (VSG, pseudogene), putative	0.87	1.84	2.12	Yes
Tb427.BES59.12	variant surface glycoprotein (VSG) (VSG 427-13)	0.47	0.98	2.08	N/A[4]
Tb427.07.2000	retrotransposon hot spot (RHS) protein, putative	0.45	0.93	2.08	N/A[4]
Tb427tmp.160.4930	hypothetical protein, conserved	2.07	4.3	2.07	Yes
Tb427.03.1490	leucine-rich repeat protein (LRRP), putative	0.36	0.75	2.07	No
Tb427.04.130	receptor-type adenylate cyclase GRESAG 4, pseudogene, putative	1.58	3.24	2.05	Yes
Tb427.BES5.7	unspecified product	1.31	2.65	2.02	No
H25N7.11	ESAG4, pseudogene	2.72	5.47	2.01	N/A[4]

1. Fold up-regulation calculated by dividing WT DMOG RPKM value by the WT RPKM value
2. Yes if gene is located within 10kb downstream of a J peak and within a gene cluster
   No if gene is either not within 10kb downstream of a J peak or if upstream J is found at a dSSR
3. J identified upstream of the gene by SMRT-seq analysis (unpublished data)
4. Not examined

Down-regulated genes

Gene	Description	WT RPKM	WT DMOG RPKM	Fold Down-regulation[5]
Tb427.01.4630	cyclin-like F-box protein (CFB1E)	1.66	0.69	2.41
Tb427.BES126.2	unspecified product	61.67	27.7	2.23
Tb427.07.6650	hypothetical protein, conserved	1.34	0.64	2.09
Tb427.03.3900	carnitine O-palmitoyltransferase II, putative (CPT II)	5.06	2.42	2.09
Tb427.BES29.3	unspecified product	6.94	3.44	2.02
Tb427.10.10230	procyclin-associated gene 5 (PAG5) protein (PAG5)	40.93	20.81	1.97

5. Fold down-regulation calculated by dividing WT RPKM value by WT DMOG RPKM value

Table 3.S1 Genomic coordinates of cSSRs in the *L. major* genome

TTS	Chromosome	Left	Right	RNAs
3.3	3	257,889	260,160	T
4.1	4	125,231	134,259	
5.2	5	358,712	368,098	T
5.3	5	413,600	416,349	
6.3	6	496,623	499,407	
7.2	7	56,646	61,002	
8.1	8	391,471	395,193	
9.1	9	271,624	278,097	T&B
9.2	9	410,541	420,123	T&B
10.2	10	266,993	271,075	
12.1	12	175,105	177,453	
13.2	13	234,147	235,431	
14.1	14	157,511	166,986	
15.2	15	322,199	330,364	T&B
16.2	16	455,043	455,990	B
20.4	20	655,704	657,200	
21.2	21	165,335	167,157	B
21.3	21	448,056	449,719	T
22.2	22	10,320	11,184	
22.3	22	508,165	509,101	
23.2	23	224,857	228,092	T&B
24.3	24	619,452	623,933	T&B
25.2	25	421,375	423,905	
27.2	27	371,632	385,084	
27.3	27	715,532	718,771	
28.1	28	111,902	112,522	
28.2	28	587,431	595,609	
28.3	28	1,038,359	1,039,013	
29.2	29	651,387	656,655	
30.3	30	784,998	785,934	B
32.2	32	536,618	540,807	
33.5	33	806,005	806,444	
34.2	34	298,242	304,334	
34.3	34	469,757	471,821	T&B
35.4	35	640,557	646,831	
35.6	35	1,069,989	1,070,801	
36.2	36	489,603	495,890	T&B
36.3	36	1,032,191	1,035,959	T&B
36.7	36	1,885,279	1,886,719	

Table 3.S2 High-throughput sequencing generated in this study

Library #	1	2	3	4	5	6
Species	T. brucei	T. brucei	T. brucei	T. brucei	L. major	L. major
treatment	DMSO	DMOG	DMSO	DMOG	DMSO	DMOG
Type of RNA sequenced	small RNA	small RNA	polyA enriched RNA (mRNA)	polyA enriched RNA (mRNA)	small RNA	small RNA
Genome used for alignment	T. brucei 427 v6.0	T. brucei 427 v6.0	T. brucei 427 v6.0	T. brucei 427 v6.0	L. major v4.2	L. major v4.2
Minimum read length considered for alignment (nt)	18	18	n/a	n/a	18	18
Total reads (millions)	43.0	37.6	30.2	26.9	21.2	65.2
Overall (unique and non unique) alignment rate %	97.12	97.49	95.87	95.40	97.72	97.99

CHAPTER 4

HISTONE H3 VARIANT REGULATES RNA POLYMERASE II TRANSCRIPTION TERMINATION AND DUAL STRAND TRANSCRIPTION OF SIRNA LOCI IN TRYPANOSOMA BRUCEI[1]

---

[1]**Reynolds D.**, Hofmeister B., Cliffe L., Alabady M., Siegel T. N., Schmitz R. J., and Sabatini R. 2016. *PLoS Genetics* 12(1): e1005758

ABSTRACT

Base J, β-D-glucosyl-hydroxymethyluracil, is a chromatin modification of thymine in the nuclear DNA of flagellated protozoa of the order Kinetoplastida. In *Trypanosoma brucei*, J is enriched, along with histone H3 variant (H3.V), at sites involved in RNA Polymerase (RNAP) II termination and telomeric sites involved in regulating variant surface glycoprotein gene (*VSG*) transcription by RNAP I. Reduction of J in *T. brucei* indicated a role of J in the regulation of RNAP II termination, where the loss of J at specific sites within polycistronic gene clusters led to read-through transcription and increased expression of downstream genes. We now demonstrate that the loss of H3.V leads to similar defects in RNAP II termination within gene clusters and increased expression of downstream genes. Gene derepression is intensified upon the subsequent loss of J in the *H3.V* knockout. mRNA-seq indicates gene derepression includes *VSG* genes within the silent RNAP I transcribed telomeric gene clusters, suggesting an important role for H3.V in telomeric gene repression and antigenic variation. Furthermore, the loss of H3.V at regions of overlapping transcription at the end of convergent gene clusters leads to increased nascent RNA and siRNA production. Our results suggest base J and H3.V can act independently as well as synergistically to regulate transcription termination and expression of coding and non-coding RNAs in *T. brucei,* depending on chromatin context (and transcribing polymerase). As such these studies provide the first direct evidence for histone H3.V negatively influencing transcription elongation to promote termination.

Author summary

Trypanosoma brucei is an early-diverged parasitic protozoan that causes African sleeping sickness in humans. The genome of *T. brucei* is organized into polycistronic gene

clusters that contain multiple genes that are co-transcribed from a single promoter. Because of this genome arrangement, it is thought that all gene regulation in *T. brucei* occurs after transcription at the level of RNA (processing, stability, and translation). We have recently described the presence of a modified DNA base J and variant of histone H3 (H3.V) at transcription termination sites within gene clusters where the loss of base J leads to read-through transcription and the expression of downstream genes. We now find that H3.V also promotes termination prior to the end of gene clusters, thus regulating the transcription of specific genes. Additionally, H3.V inhibits transcription of siRNA producing loci. Our data suggest H3.V and base J are utilized for regulating gene expression via terminating transcription within polycistronic gene arrays and regulating the synthesis of siRNAs in trypanosomes. These findings significantly expand our understanding of epigenetic regulatory mechanisms underlying transcription termination in eukaryotes, including divergent organisms that utilize polycistronic transcription, providing the first example of a histone variant that promotes transcription termination.

INTRODUCTION

Kinetoplastids are early-diverged protozoa that include the human parasites *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, which cause African sleeping sickness, Chagas disease, and leishmaniasis, respectively. The genomes of kinetoplastids are arranged into long gene clusters, or polycistronic transcription units (PTUs), which are transcribed by RNA polymerase (RNAP) II [1-3]. RNAP II transcription initiation and termination occurs at regions flanking PTUs called divergent strand switch regions (dSSRs) and convergent strand switch regions (cSSRs), respectively [4]. Pre-messenger RNAs (mRNA) are

processed to mature mRNA with the addition of a 5' spliced leader sequence through *trans*-splicing, followed by 3' polyadenylation [5-10]. The arrangement of genes into PTUs has led to the assumption that transcription is an unregulated process in these eukaryotes and a model in which gene regulation occurs strictly post-transcriptionally [11, 12]. However, specific chromatin marks have been characterized at sites of transcription initiation and termination, including histone variants and modified DNA base J, which could function to regulate polycistronic transcription and gene expression [13-17].

Base J, β-D-glucosyl-hydroxymethyluracil, is a modified DNA base consisting of O-linked glycosylation of thymine in the genome of kinetoplastids and closely related unicellular flagellates [18, 19]. Whilst J is largely a telomeric modification, it is also found internally within chromosomes at RNAP II transcription initiation and termination sites [13, 20-25]. As reviewed in Borst and Sabatini (2008), analysis of RNAP I transcribed telomeric polycistronic units in *T. brucei* led to the discovery of base J [20, 26]. Regulation of the ~15 telomeric variant surface glycoprotein expression sites (*VSG* ESs) allows the parasite to evade the host immune system in a process called antigenic variation [27, 28]. Although the genome of *T. brucei* has over 1,000 *VSG* genes, only one *VSG* is expressed at a given time. This is achieved through regulated transcription of the telomeric ESs, only one of which is productively transcribed at any time. The association of the modified base with silent ESs in the bloodstream life-cycle stage of the parasite has led to the hypothesis that J plays a role in the regulation of antigenic variation. However, no direct evidence has been provided.

Base J is synthesized in a two-step pathway in which a thymidine hydroxylase, JBP1 or JBP2, hydroxylates thymidine residues at specific positions in DNA to form hydroxymethyluracil, followed by the transfer of glucose to hydroxymethyluracil by the

glucosyltransferase, JGT [26, 29, 30]. JBP1 and JBP2 belong to the TET/JBP subfamily of dioxygenases, which require Fe^{2+} and 2-oxoglutarate for activity [31-34]. The synthesis of base J can be inhibited by competitive inhibition of the thymidine hydroxylase domain of JBP1 and JBP2 by dimethyloxalylglycine (DMOG), a structural analog of 2-oxoglutarate [31, 35]. Removal of both JBP1 and JBP2 or the JGT also results in *T. brucei* cells devoid of base J [29-31, 36].

The co-localization of base J with modified and variant histones at dSSRs and cSSRs suggested a functional role of modified DNA in the regulation of RNAP II transcription [13]. Our work in *T. cruzi* described a unique role of J in regulating RNAP II transcription initiation, where the loss of base J resulted in the formation of more active chromatin, increased RNAP II recruitment and increased PTU transcription rate [24, 37]. Recent studies have described a role for base J regulating RNAP II termination in *T. brucei* and *Leishmania*. van Luenen et al. (2012) found that reduction of base J in *L. tarentolae* is associated with the generation of RNAs downstream of the cSSR that are antisense to the genes on the opposing gene cluster [25]. Reduction of base J in *L. major* resulted in similar defects [35]. Strand-specific RT-PCR detection of the nascent transcript confirmed that the J-dependent generation of RNAs downstream of the cSSR is due to read-through transcription at cSSR termination sites. In contrast, loss of J in *T. brucei* failed to indicate any defect in termination at cSSRs [35]. However, we localized base J at sites within PTUs where the loss of J led to read-through transcription and upregulated expression of downstream genes. Therefore, base J is required for RNAP II termination in both *Leishmania* and *T. brucei*, but to different degrees and at different locations. In *L. major*, J regulates termination at the end of each PTU to prevent read-through transcription and the generation of RNAs antisense to the genes on the opposing PTU. In

contrast, although termination occurs at the end of each PTU in *T. brucei* in a J-independent manner, J-dependent termination within a PTU allows developmentally regulated expression of downstream genes.

The core histones H2A, H2B, H3 and H4, package DNA into nucleosomes and represent a critical component of higher order chromatin. All core histones have variant counterparts. Although histone post-translational modifications (PTMs) and their impact on transcription have been well documented, less is known about the role of histone variants in the regulation of transcription [38]. The most understood are variants of H2A and H3. Several variants of H2A exist, including H2A.Z, H2A.B, and macroH2A. Both H2A.Z and H2A.B are associated with transcriptional activation [39-41]. Knockdown of H2A.Z inhibits transcriptional activation [42-44]. Consistent with this, and perhaps the most direct evidence of a transcriptional role of a histone variant, H2A.Z positively correlates with rates of RNAP II elongation, such that the reduction of H2A.Z increases RNAP II stalling [40]. Presumably, the nucleosome destabilizing affect of H2A.Z [45] leads to more accessible DNA at promoter regions for transcription factor binding, as well as promoting RNAP II elongation through gene bodies. Like H2A.Z, H2A.B is enriched at promoter regions and its reduction largely results in the downregulation of gene expression [39, 46]. In contrast, macroH2A is enriched at transcriptionally repressed regions [47] and its reduction results in increased gene expression in an unknown mechanism [48]. Several H3 variants have also been characterized, including H3.3 and CENP-A, both of which are found in most eukaryotes including plants, mammals, and yeast. H3.3 differs from canonical H3 by only 4-5 amino acids and is found predominately at actively transcribed genes, forming more accessible nucleosomes [49-51]. H3.3 also provides a genome stabilization function at repetitive regions such as telomeres and centromeres [49, 52-55]. Recent studies have

implicated H3.3 in the maintenance of a repressed chromatin structure [56-58]. Evidence in mouse embryonic stem cells indicates H3.3 is enriched at lowly transcribed developmentally regulated genes where it promotes polycomb repressive complex 2 activity, which catalyzes the formation of the repressive modification H3K27me3 [57, 58]. These findings suggest H3.3 maintains the promoters of developmentally regulated genes in a repressed, but transcriptionally "poised" state important for proper differentiation. H3.3 has also been implicated in the maintenance of H3K9me3 at endogenous retroviral elements in mouse embryonic stem cells [56]. Presence of H3.3 (and H3K9me3) at endogenous retroviral elements repressed retrotransposition and expression of adjacent genes [56]. The centromeric specific histone variant has a well-characterized role in kinetochore formation, but its role in the regulation of transcription, if any, remains unknown [59]. Overall, although much progress has been achieved in the characterization of histone variants, few studies have revealed a direct link between histone variant function and transcriptional regulation.

The J-independent nature of termination at cSSRs in *T. brucei* led us to characterize the role of H3.V in regulating RNAP II termination. H3.V and base J co-localize at RNAP II termination sites in *T. brucei*, including cSSRs and PTU internal termination sites [13, 14]. H3.V and base J also co-localize at telomeric repeats involved in regulating RNAP I transcription of the *VSG* expression sites [14, 60]. *T. brucei* H3.V shares 45% sequence identity with canonical H3, much of the sequence divergence lying within the N-terminus, outside of the histone fold domain. H3.V appears to be unique to kinetoplastids [14, 60] and aside from its localization to termination sites and telomeres, very little is known about H3.V and its potential role in the regulation of transcription termination. We demonstrate here that, similar to phenotypes associated with the loss of J, loss of H3.V leads to defects in RNAP II termination

within gene clusters and increased expression of downstream genes. Interestingly, many of the gene expression changes in the *H3.V* knockout (KO) are further increased upon the subsequent loss of base J, suggesting that J and H3.V have independent but overlapping roles in regulating transcription termination in *T. brucei*. Although the loss of H3.V from cSSRs did not indicate any termination defects leading to transcription of the opposing strand of the adjacent convergent gene cluster, it does lead to increased generation of small interfering RNAs (siRNAs) that map to regions of overlapping transcription. Analysis of nascent RNA suggests this is due to increased transcription of the dual strand siRNA loci at cSSRs. We also detect increased expression of *VSGs* from silent *VSG* ESs in the *H3.V* KO, indicating H3.V can act independently in regulating telomeric repression and antigenic variation. Overall these findings provide the first known example of a histone H3 variant that functions as a repressive chromatin mark to promote transcription termination, in this case repressing both mRNAs and non-coding RNAs.

RESULTS

H3.V regulates RNAP II transcription at cSSRs

The co-localization of base J and H3.V at RNAP II termination sites in *T. brucei* prompted us to examine the role of H3.V in transcription termination. High-throughput sequencing of small RNAs has been shown previously to reveal transcription termination sites in trypanosomatids, as reflected in RNA degradation products [25, 35]. The reduction of base J in *L. major* by treatment with DMOG resulted in the production of antisense small RNAs corresponding to genes in the opposing PTU due to read-through transcription at cSSRs [35]. In contrast, we found no evidence of termination defects in *T. brucei* at cSSRs following DMOG

treatment and the complete loss of J. Antisense small RNAs, indicative of read-through transcription at cSSRs into the downstream PTU, were not increased following the loss of J [35]. We now show that the loss of H3.V also does not result in read-through transcription at cSSRs. No significant changes in antisense small RNAs corresponding to read-through transcription at cSSRs into the downstream PTU were detected by small RNA-seq in the *H3.V* KO compared to wild type (WT) *T. brucei* (Fig. 4.1A) (small RNA sequencing data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE70229). Subsequent loss of base J in the *H3.V* KO parasites, via DMOG treatment, also failed to uncover any defect. These results suggest H3.V is not required to prevent RNAP II read-through transcription at the end of convergent gene arrays in *T. brucei*.

Unique to *T. brucei*, large peaks of sense and antisense small RNAs map to regions of overlapping transcription at the ends of convergent PTUs at cSSRs ([35]; Fig. 4.1A). These presumably represent the previously characterized Dicer 2-dependent Argonaute-associated siRNAs derived from cSSRs [61]. Consistent with this, we show that the RNAs that map to these regions correspond to the previously characterized siRNA size range in bloodstream form *T. brucei* (21-27nt) [61-63] and exhibit characteristic phasing of siRNAs biogenesis, as indicated by the mapping of siRNA sequences in phased intervals at the target loci [64-66] (Fig. 4.1B and 4.S1). This pattern suggests that these siRNAs were enzymatically processed, most likely by the DCR complex [67], as opposed to random RNA degradation. siRNAs that map to cSSRs are 21-27nt compared to the 18-30nt range of small RNAs genome-wide (Fig. 4.1B and C). Although the loss of base J in WT *T. brucei* parasites has no effect on siRNA levels [35], the level of siRNAs derived from the cSSRs significantly increased in the *H3.V* KO (Fig. 4.1A). To

confirm these changes, we repeated the small RNA-seq analysis of WT and *H3.V* KO in triplicate. Quantitation of the small RNAs associated with cSSRs genome-wide indicates a statistically significant increase in 21-27nt RNAs at 30 out of 72 cSSRs in the *H3.V* KO (Fig. 4.1B and 4.S2 Table). This can be visualized by specifically mapping the siRNA size range of small RNAs (4.S1A Fig.). The increase in siRNAs is not restricted to cSSRs, but occurs at all previously characterized siRNA generating regions of the *T. brucei* genome [61, 62]; including the SLACS and ingi retrotransposable elements, CIR147 centromeric repeats, and inverted repeats (4.S2 Fig.). These regions are also enriched with H3.V [14, 68].

Small RNA-seq suggests that the siRNA mapping to dual strand transcription regions at cSSRs is due to continued RNAP II transcription of the convergent PTU resulting in overlapping transcription. An example that illustrates this is shown in Fig. 4.2A-D, where H3.V (and J) is found enriched at cSSR 2.5. Presumably, H3.V interferes with RNAP II elongation in these dual strand transcription regions. Removal of H3.V would then lead to increased nascent RNA corresponding to these regions (siRNA precursor) that is then processed, resulting in increased siRNA levels that map to both strands (Fig. 4.1, 4.2D, 4.S1A, and 4.S2 Table). We have previously utilized strand-specific RT-PCR to follow read-through transcription and increased nascent RNA production in *L. major* and *T. brucei* [35]. Here, primers were designed that span the poly(A) site of the final gene in the PTU, allowing us to specifically detect nascent, unprocessed RNA within cSSR 2.5 (Fig. 4.2E). Strand-specific RT-PCR results indicate an increase in nascent RNA in the *H3.V* KO at dual strand transcribed loci at cSSR 2.5 as well as cSSR 1.4 (Fig. 4.2F and G). Although no significant increase in siRNAs was observed in DMOG treated WT *T. brucei* [35], we do detect a slight increase in nascent RNA following the loss of base J at cSSR 2.5 (Fig. 4.2F and G). Nascent RNA is further increased following the

loss of both H3.V and J at the two cSSRs analyzed. Overall these results indicate that H3.V, and to a lesser extent J, attenuate transcription elongation within dual strand transcribed loci at cSSRs, potentially enabling regulated expression of siRNAs derived from these sites. Loss of H3.V does not lead to read-through transcription downstream of the dual strand transcribed loci however.

H3.V inhibits RNAP II elongation within PTUs

We have recently shown that base J is present along with H3.V at termination sites within a PTU where J loss results in read-through transcription and increased expression of downstream genes in *T. brucei* [35]. Because RNAP II elongation and gene expression is inhibited prior to the end of these PTUs, we refer to this as PTU internal termination. To explore the impact of H3.V on RNAP II elongation at these PTU internal termination sites we analyzed the downstream genes by RT-qPCR in the *H3.V* KO cell line. At three representative PTU internal termination sites we detect increased expression of downstream genes in the *H3.V* KO (Figs. 4.3A-D and 4.S3A-D). In each of these cases gene derepression is enhanced upon the subsequent loss of J in the *H3.V* KO following DMOG treatment (Figs. 4.3D and 4.S3D). Importantly, derepression is limited to genes downstream or within the peak of H3.V/base J. Strand specific RT-PCR using oligos flanking the termination site (based on mRNA-seq and base J localization [13, 35]) detects increased RNA in the *H3.V* KO (Fig. 4.3E-F, and 4.S3F). This is consistent with an increase in nascent read-through RNA resulting from continued transcription elongation at PTU internal termination sites marked by base J and H3.V. Consistent with the gene expression changes, read-through is visibly enhanced at region 7.3 upon the subsequent loss of J in the *H3.V* KO following DMOG treatment (Fig. 4.3F). These

results, combined with our previous study of J regulation of termination [35], suggest J and H3.V have independent and overlapping roles in regulating RNAP II termination through the inhibition of transcription elongation and the expression of downstream genes in *T. brucei*.

To further explore the role of H3.V in the regulation of termination and whether H3.V functions similarly across the genome, we performed mRNA-seq to compare the expression profiles of WT, WT+DMOG, *H3.V* KO and *H3.V* KO+DMOG cells (GEO accession number GSE69929). This led to the detection of 153 mRNAs that are increased at least 2-fold in one or more of the treatments (4.S1 Table). Many of the gene expression changes have been confirmed by RT-qPCR (4.S4 Fig. and below). Consistent with our previous mRNA-seq results, in WT cells treated with DMOG we observe similar increases in the expression of genes downstream of base J (and H3.V), which we previously demonstrated is caused by an RNAP II transcription termination defect within a PTU [35]. However, we now see that a significant number of genes downstream of J/H3.V within other PTUs are upregulated following the loss of H3.V, and that many are further increased following the subsequent loss of J (Fig. 4.4, 4.S1 and 4.S4 Table). In the *H3.V* KO we identified 71 genes that are upregulated (Fig. 4.5A and 4.S1 Table). Although many of these genes are not increased by 2-fold in the WT+DMOG condition, some respond at least slightly to the loss of J in WT cells: 28 of the 71 genes upregulated in the *H3.V* KO are also increased at least 1.3-fold in the WT+DMOG condition, suggesting J and H3.V have overlapping functions in regulating termination at these sites, where H3.V plays a dominant role. Consistent with this, 42 of the 71 H3.V regulated genes are upregulated even further upon subsequent loss of base J in the *H3.V* KO. This trend is evident in the heatmap shown in Fig. 4.5A. A specific example is shown in Fig. 4.5B-F where a cluster of genes is affected by the loss of H3.V, and to a lesser extent base J. Both chromatin marks are enriched upstream of and

within this gene cluster, which consists of genes annotated as *VSG* pseudogenes, an atypical *VSG*, and a hypothetical protein. mRNA-seq indicates gene upregulation is largest in the absence of H3.V and J, which we confirmed by RT-qPCR (Fig. 4.5E and F). An additional example is shown in 4.S5 Fig. Chromosome maps indicating the genomic location of upregulated genes upon the loss of H3.V and J are shown in 4.S6 Fig. These results also confirm our initial analyses of nascent and steady-state RNA indicating a role for H3.V (and J) regulating termination and expression of downstream genes (Fig. 4.3 and 4.S3).

80% of the affected genes are found within 10 kb of H3.V and base J (see 4.S1 Table and 4.S7 Fig.) and thus fit a model where derepression occurs as a result of deregulated transcription elongation/termination within a PTU following the loss of the two chromatin marks. In support of this model, strand-specific RT-PCR analysis links gene derepression with increased nascent RNA production downstream of J/H3.V marks within the PTU (Figs. 4.3E-F and 4.S3F). In a few cases, genes that are further than 10 kb downstream of base J and H3.V are still within a regulated cluster, explaining why some genes are indicated as not adjacent to J/H3.V but can still be regulated by these epigenetic marks (e.g. genes Tb427.07.6730-Tb427.07.6780, 4.S1 and 4.S4 Table). Although clusters of genes downstream of a J and H3.V enriched region are similarly upregulated in many cases, the 153 upregulated genes localize to 91 different PTUs (4.S4 Table). Therefore, H3.V and J repress gene expression genome-wide, often repressing the expression of a single gene, usually the last gene, within a PTU. Overall, the data indicate that base J and H3.V have independent but overlapping roles in regulating RNAP II transcription termination within specific PTUs and enable regulated expression of downstream genes.

Although the majority of the differentially expressed genes are upregulated, we also see some downregulated (4.S1 Table). 35 genes are downregulated by at least 2-fold in the *H3.V* KO cells. Unlike the upregulated genes, where many were increased furthest following the combined loss of J and H3.V, only 7 of the 35 downregulated genes decreased more in the *H3.V* KO+DMOG condition. However, 33 of the 35 downregulated genes are found proximal (within 10kb) to H3.V and J. The effect is also locus specific and mainly restricted to arrays of genes transcribed from the same strand. These points indicate a strong link between H3.V/J localization and gene expression changes observed. A few examples are indicated in Fig. 4.S8. Among the genes with most significant downregulation upon loss of H3.V were procyclins and procyclin associated genes (PAGs), which include surface protein encoding genes most highly expressed during the (procyclic) insect stage of the parasite [69]. Although PAGs are located in multiple copies in the genome, within RNAP II transcribed arrays or within RNAP I transcribed procyclin arrays [69, 70], the PAGs downregulated following the loss of H3.V (PAG1, PAG2, PAG4 and PAG5) are specifically arranged in an RNAP I transcribed array. For example, there are two PAG2 genes located on chromosome 10, one in an RNAP II transcribed PTU and the other in the RNAP I transcribed procyclin locus. The only gene that is significantly downregulated when H3.V is deleted is the one within the RNAP I procyclin locus. A similar locus specific alteration of PAG expression was seen upon the depletion of histone H1 [71]. Interestingly, the PAGs within this locus undergo overlapping RNAP II and RNAP I transcription, i.e. continued RNAP II transcription of the upstream opposing PTU produces antisense PAG RNAs [72]. This suggests a possible mechanism of PAG (and procyclin) downregulation resulting from increased formation of dsRNAs upon the loss of H3.V (see discussion). Similarly, other downregulated genes are arranged in opposing transcriptional

genes pairs where extended RNAP II transcription would result in dsRNA for each mRNA (4.S8 Fig.). We also identified 25 genes that are downregulated specifically in the *H3.V* KO+DMOG condition, though 22 of these genes are not located near H3.V or J. We therefore assume many of these changes are an indirect effect of genes that are upregulated in this cell line. For example, we have demonstrated that the genome-wide increase in RNAP II transcription in *T. cruzi* results in a global increase in gene expression that includes proteins that degrade specific mRNAs [24].

H3.V regulates expression of RNAP I transcribed *VSG* genes.

In addition to RNAP II termination sites, H3.V co-localizes with base J at telomeres, which in the *T. brucei* genome contain the RNAP I transcribed polycistronic units involved in antigenic variation (so called *VSG* expression sites, ESs) (Fig. 4.6A) [27, 28]. mRNA-seq results indicate that the deletion of H3.V leads to increased expression of *VSG* genes from silent ESs, which we have confirmed by RT-qPCR (4.S1 Table and Fig. 4.6B). Although *VSG*s within expression sites are not affected by the loss of J in WT cells, five of the *VSG*s are further upregulated upon the loss of J in the *H3.V* KO (4.S1 Table and Fig. 4.6B).

mRNA-seq indicates several expression site associated genes (*ESAGs*) within silent ESs with differential expression. Because the repetitive nature of *ESAGs* complicates read alignment of mRNA-seq data, we analyzed several genes by RT-qPCR using primers that are specific to ES *ESAGs* (4.S9 Fig.). Consistent with mRNA-seq results, with the apparent exception of *ESAG8*, most of the *ESAGs* are not upregulated in the *H3.V* KO (4.S9 Fig.). The lack of significant change in the majority of *ESAGs* within the ESs suggests derepression of telomere-proximal *VSG* genes after H3.V depletion is not due to transcriptional activation of silent

promoters. Repression of silent ESs is mediated in part by the inhibition of RNAP I elongation within the ES preventing the production of *VSG* mRNA from the silent ESs [73]. Similar to its inhibition of RNAP II transcription elongation at termination sites within PTUs genome-wide, H3.V may function at telomeric regions to attenuate transcription elongation within the silent ESs, thereby preventing transcription of silent *VSGs*. To further investigate this possibility and determine how telomere-proximal and telomere distal genes are affected after loss of H3.V, we used RT-qPCR to compare the derepression of a unique gene in the silent ES15 that is located 10 kb (*VSG* pseudogene 11, Tb427.BES126.13) and 1 kb (*VSG11*) upstream of the telomere (Fig. 4.6; [74]). Although *VSG11* was upregulated in the *H3.V* KO, we found that the *VSG* pseudogene is not significantly derepressed in the *H3.V* KO or upon the loss of H3.V and J (Fig. 4.6B). These results overall suggest H3.V is involved in the repression of *VSGs* in silent ESs, but does not significantly impact silencing of the entire ES, similar to the role of telomere localized TbRAP1 [75]. *ESAG8* derepression is consistent with ES derepression from the promoter, however upregulation of multiple *ESAGs* would be expected if derepression of the silent ESs occurs from the promoter, which was observed after ISWI depletion [76]. Although we cannot rule out possible alterations in VSG switching in the *H3.V* KO (see discussion), the data here suggest H3.V is involved in repressing RNAP I transcription of *VSGs* within the silent ESs as well as RNAP II transcription of *VSG* genes within genome internal PTUs.

DISCUSSION

H3.V is a kinetoplastid-specific H3 variant and appears to be the only H3 variant found in these early-diverged eukaryotes. The *T. brucei* H3.V shares only 45% sequence identity with the canonical H3 [60]. Although H3.V localizes to centromeres, it is not essential for viability

and does not contain sequence variations common to all identified centromeric H3 variants [14, 60, 68]. Aside from its localization to RNAP II termination sites and telomeres, the functional significance of H3.V and its potential role in the regulation of RNAP II termination has been unexplored. We have demonstrated that H3.V negatively regulates transcription elongation and promotes RNAP II termination in *T. brucei*. Several lines of evidence support this conclusion. At many of the same sites within gene clusters where we have previously shown J regulates transcription termination and expression of downstream genes, we detect similar increases in the expression of genes downstream of the termination site following the loss of H3.V. Also similar to J loss, detection of increased nascent, unprocessed RNA by strand-specific RT-PCR at these sites supports the conclusion that the loss of H3.V leads to read-through transcription. Furthermore, the loss of H3.V from regions where dual strand transcription naturally occurs at cSSRs leads to increased levels of siRNAs, and strand-specific RT-PCR indicates this is due to increased transcription of the cSSR. These findings overall suggest H3.V imparts a repressive chromatin structure that is refractory to transcription elongation, or potentially recruits other repressive factors or transcription termination factors at RNAP II termination sites. To our knowledge, this is the first example of a histone H3 variant that has been shown to repress the expression of mRNAs and non-coding RNAs by promoting transcription termination.

These results extend our previous findings that base J functions to prevent read-through transcription at termination sites within gene clusters in *T. brucei*, revealing an overlapping role of J and H3.V in the regulation of transcription termination. However, H3.V appears to have a broader and in many cases a more dominant role. In this study mRNA-seq indicated 71 genes were upregulated by 2-fold or more following the loss of H3.V. 39% of those genes were also affected by at least 1.3-fold following the loss of J alone, and 59% were further increased upon

the subsequent loss of base J in the *H3.V* KO. Thus H3.V and J appear to function similarly, but independently in the regulation of transcription termination. Although we cannot exclude a potential role of H3.V (and J) in the regulation of RNA processing, the increase in both unprocessed (nascent) and processed RNAs (mRNAs and siRNAs) strongly suggests H3.V regulates RNA abundance at the level of transcription and that the defects we observe are not simply due to an alteration of RNA processing. We propose a model in which both H3.V and base J inhibit RNAP II elongation, and therefore stimulate termination at sites within gene clusters (4.S10 Fig.). According to this model, the loss of J within a gene cluster results in read-through transcription and expression of genes that were previously silent, but the presence of H3.V within the downstream cSSR prevents increased dual strand transcription and thus siRNAs are not significantly increased. Similarly, the loss of H3.V leads to read-through transcription at termination sites within a gene cluster and subsequent gene derepression. H3.V loss also results in increased transcription of dual strand transcribed loci at cSSRs, giving rise to more siRNAs. Therefore, base J is a chromatin modification that specifically regulates the expression of a subset of genes in the bloodstream form of *T. brucei* parasites, whereas H3.V regulates a similar, but larger subset of genes, in addition to the generation of siRNAs.

Recent description of the trypanosome stress response has indicated that the location of a gene within a PTU can impact its expression, presumably via regulated transcription elongation [77]. Here we provide evidence that regulated transcription and expression of genes within PTUs can be achieved through their spatial organization and position relative to H3.V and base J. However, the biological significance of the gene expression changes we describe following the loss of these chromatin marks, remains unclear. It does not help that the majority of the regulated genes are annotated as hypothetical proteins of unknown function. Many of the H3.V

and base J regulated genes include *VSGs*, *ESAGs*, *RHS* proteins, and pseudogenes that are normally lowly expressed (or not at all) in wild type *T. brucei*. Interestingly, consistent with base J synthesis, *VSGs* and *ESAGs* are developmentally regulated, typically exclusively expressed in bloodstream form trypanosomes from the telomeric PTU (*VSG* ESs). Monoallelic expression of a *VSG* ES leads to the expression of a single VSG on the surface of the parasite, a key aspect of trypanosome antigenic variation. Therefore the repression of silent *VSGs* by H3.V/J allows the parasite to maintain this monoallelic expression. Another important aspect of antigenic variation is the periodic switching of the VSG protein expressed on the surface allowing the parasite to remain a step ahead of the host immune response. DNA recombination (i.e. gene conversion events) of silent *VSG* genes into the active ES is the dominant driver of trypanosome antigenic variation. Transcription of a donor DNA sequence has been shown to increase its use during gene conversion events in human cells [78]. It has also been demonstrated that active transcription in *T. brucei* stimulates DNA recombination [79]. Therefore, the regulation of transcription of silent *VSGs* by H3.V/J, including *VSGs* in silent telomeric ESs and *VSG* pseudogenes at the end of genomic internal PTUs, could play a role in gene conversion events. It is well characterized that during late phases of mammalian infection, trypanosomes predominately express mosaic *VSGs* comprised of multiple *VSGs* and pseudogenes [80]. These findings thus raise the possibility that regulated transcription of silent *VSGs* by H3.V/J, in particular *VSG* pseudogenes at internal PTUs, contributes to gene conversion events that result in the formation of these mosaic *VSGs*.

If H3.V and base J are utilized to effect specific gene expression changes, then mechanisms likely exist to overcome the silencing effects of these modifications, i.e. regulated addition and/or removal. Histone chaperones and chromatin remodeling complexes incorporate

histone variants at specific chromatin locations. Thus, regulation of (unidentified) histone chaperones that incorporate H3.V could enable regulated gene expression. Chromatin remodeling proteins could also be involved in the removal of H3.V. The first step of J synthesis consists of thymidine oxidation by JBP1 and 2, which utilize oxygen and 2-oxoglutarate and require Fe2+ as a cofactor. Changes in oxygen concentrations or metabolic changes could thus impact J synthesis and effect gene expression changes. We previously demonstrated oxygen regulation of JBP1/2 and J synthesis, which led to changes in gene expression and pathogenesis of *T. cruzi* [24, 31]. JBP1/2 have been shown to have differential chromatin substrates for de novo J synthesis in vivo [13]. Therefore, regulation of JBP1/2, or associated factors, could provide differential regulation of J synthesis at specific loci. Reiterative oxidations of thymidine residues by JBP1/2 [29], similar to TET mediated oxidation of cytosines [81], may also contribute to regulated J synthesis at specific loci.

Loss of H3.V also led to derepression of *VSG* genes within the silent telomeric ESs. We hypothesize, similar to its effect on RNAP II termination within PTUs genome-wide, H3.V localized to telomeric repeats limits basal levels of RNAP I transcription elongation within silent ESs [73]. A similar telomeric VSG derepression effect was observed following the loss of RAP1 in *T. brucei* [75]. The lack of significant *ESAG* gene derepression suggests loss of H3.V does not result in derepression from the promoters of silent ESs, though *ESAG8* upregulation is consistent with this possibility. Therefore, we acknowledge that further detailed analysis is required, including the use of tagged silent ESs, to fully understand the role of H3.V on ES transcription. Because of our inability to effectively measure VSG switching rates in our *H3.V* KO cell line, we also cannot exclude the possibility that H3.V restricts VSG switching, though a recent study has indicated that switching frequency does not appear to change significantly in

the *H3.V* KO or upon the loss of H3.V and J (Schulz, Papavasiliou, and Kim, personal communication). While base J has no apparent independent role in telomeric repression or VSG switching, the additional derepression of *VSG* genes observed in the *H3.V* KO upon loss of J suggests the novel modified base can act synergistically with H3.V in telomeric silencing and antigenic variation. This function is consistent with the distinct localization of base J in the silent ESs, with J density highest close to the telomeres [82].

We also found a specific RNAP I transcribed procyclin gene cluster that was downregulated following the loss of J and even more so by the loss of H3.V. As mentioned above, this locus has been shown to undergo overlapping RNAP II and RNAP I transcription in the procyclic form *T. brucei* [72]. In the procyclic form, transcription from the opposite strand was detectable from GU2 to EP1 (S8 Fig). Increased antisense transcription from the RNAP II PTU on the opposite strand led to antisense RNA and co-transcriptional silencing of PAG genes [72]. In contrast, there was little transcription from the opposing strand in bloodstream forms. The presence of H3.V, and to some extent J, may inhibit this dual strand transcription in bloodstream forms, thus reducing the formation of dsRNA and/or transcriptional interference by RNAP II that could interfere with the expression of the procyclin locus. However, analysis of this locus in procyclic cells has indicated that the loss of Argonaute did not appear to alter the expression of the procyclin genes [83], thus the role of dsRNAs at this locus, if any, remains unknown. Interestingly, several of the other genes that are downregulated in the *H3.V* KO are arranged in opposing transcriptional gene pairs where extended RNAP II transcription would result in dsRNA for each mRNA (4.S8 Fig.). Additional studies are needed to characterize the role of H3.V in regulating transcription at these loci. As we described for the effect of base J in *T. cruzi* [24], it is also possible that some transcripts are downregulated in the *H3.V* KO due to

secondary effects of derepressed genes, which could include regulatory proteins (destabilizing specific mRNAs).

Future studies are necessary to further elucidate the mechanisms by which H3.V regulates transcription, including what proteins interact with H3.V, the impact of H3.V on nucleosome structure/stability, whether H3.V undergoes any post-translational modifications, and how chromatin structure is affected by the loss of H3.V. It is not clear how sequence variation in H3.V confers its specific localization to transcription termination sites or its function. H3 PTMs have not been well characterized in *T. brucei*, however those identified by mass spectrometry analyses include S1 and K23 acetylation and K4, K32, and K76 methylation [84]. In comparison to the canonical H3, the *T. brucei* H3.V differs in that it contains an A1 and R23, and thus lacks the corresponding acetylation. However, differences in PTMs between H3.V and H3 have not been investigated.

Interestingly, the loss of H3.V and base J in *T. brucei* did not result in read-through transcription that extends into the downstream gene cluster encoded on the opposite strand and generation of antisense RNAs. In *L. major* the loss of J alone led to such read-through transcription at a majority of the cSSRs sites in the genome [35] whereas the loss of H3.V had no effect [85]. Overall these findings indicate that the function of epigenetic modifications in kinetoplastid parasites is not necessarily conserved. One obvious difference between *T. brucei* and *L. major* is the absence of a complete RNAi pathway in *L. major* [86, 87]. It would therefore be interesting to investigate the role of H3.V in regulating siRNAs in a *Leishmania* species with an intact RNAi pathway.

In addition to its role in the regulation of dual strand transcription at cSSRs and the generation of siRNAs, we also find H3.V regulates other characterized siRNA generating loci,

including the SLACS and ingi retrotransposable elements, CIR147 centromeric repeats, and inverted repeats. Aside from the SLACS and ingi derived siRNAs, the function of siRNAs in *T. brucei* is unclear. Mature SLACS and ingi transcripts are present at low levels in WT *T. brucei* due to the presence of a functional RNAi pathway [88]. Surprisingly, despite the increase in SLACS and ingi siRNAs following the loss of H3.V, we also observe a modest increase in SLACS (*Tb427tmp.211.5010*) and ingi transcripts by mRNA-seq (4.S1 Table). Although the relative increase in steady-state level of siRNAs is greater than that of the mature mRNA transcript, presumably the increased transcription of these loci in the absence of H3.V increases both RNA species. Dicer 2 is responsible for the formation of siRNAs derived from cSSRs, and its removal (and corresponding decrease in siRNAs) did not have a significant effect on the expression of genes located at cSSRs that coincide with the siRNA peak [61], suggesting that cSSR derived siRNAs do not regulate mRNA abundance. Consistent with this, at cSSRs where we detect increased siRNAs we do not observe significant decreases in mRNAs from genes that overlap the siRNA peak. Therefore the function, if any, of cSSR derived siRNAs remains unknown.

In summary, we have provided evidence for the connection between a histone H3 variant and transcription termination for the first time. These findings highlight the importance of chromatin modifications in the regulation of transcription termination, particularly in early-diverged eukaryotes with unique polycistronic transcription. These findings also have direct implications for a strictly post-transcriptional model of gene expression regulation in kinetoplastids.

MATERIALS AND METHODS

Parasite cell culture

WT and *H3.V* KO bloodstream form *T. brucei* 221a cell lines of strain 427 were cultured in HMI-9 medium as described previously [89]. The bloodstream form *T. brucei H3.V* KO cell line, generated by deleting both H3.V alleles by homologous recombination [60], was provided by George Cross. DMOG treatment of cells was performed by supplementing media with 1mM DMOG for 5 days as described previously [35].

Strand-specific RNA-seq library construction

Small RNA-sequencing was performed using two different methods. The analysis of WT, *H3.V* KO, and *H3.V* KO+DMOG (Fig. 4.1A, 4.2D, and 4.S2) was performed as previously described [35]. Briefly, small RNAs were isolated from *T. brucei* ($5x10^7$ cells) using a Qiagen miRNeasy kit according to the manufacturer's instructions. The small RNA-seq libraries were prepared using approximately 250ng small RNA by Vertis Biotechnology AG, Germany. The small RNA sample was poly(A)-tailed using poly(A) polymerase. Then, the 5'PPP and cap structures were removed using tobacco acid pyrophosphatase (TAP, Epicentre). Afterwards, an RNA adapter was ligated to the 5'-monophosphate of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. The resulting cDNAs were PCR-amplified to about 10-20 ng/µL using a high fidelity DNA polymerase. The cDNAs were purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics). Quality and concentration of all libraries was determined by capillary electrophoresis and high throughput sequencing was performed on a HiSeq2000 (Illumina). Sequencing reads were mapped to the *T. brucei* reference genome using Bowtie2 version 2.2.3

with local sensitive mode, all other parameters default, [90] and further processed using Samtools 1.2 [91]. Reads shorter than 18 bp were discarded before mapping. Genome and gene annotations of strain 427 version 6.0 were downloaded from EuPathDB [92] and used as the reference in all small RNA-seq analyses. RPM were calculated using a window size of 101 bp and a step size of 101 bp. Total sequence reads and overall alignment rate for all RNA-seq libraries discussed in this publications are listed in 4.S3 Table.

Small RNA-sequencing of the triplicate analysis of WT and *H3.V* KO (Figs. 4.1B, 4.1C, 4.S1A, and 4.S2 Table) was performed in a similar manner. Briefly, total RNA was isolated from log phase *T. brucei* cultures ($5x10^7$ cells) using Trizol according to the manufacturer's instructions. The small RNA-seq libraries were prepared using approximately 250ng total RNA using the Illumina-compatible NEBNext small RNA library preparation kit following the manufacturer protocol (New England Biolabs). Quality and concentration of all libraries was determined using a Bioanalyzer 2100 (Agilent). Libraries were pooled using equi-molar amounts and sequenced on a NextSeq500 (Illumina). Both library construction and sequencing were done at the Georgia Genomics Facility (GFF). Small RNA reads were quality and adapter trimmed using Cutadapt [93] and reads shorter than 18 nucleotides were discarded. Reads were mapped to the *T. brucei* reference genome using Bowtie2 version 2.2.3 with the following parameters "-a -D 10 -R 5 -N 1 -L 15 -i S,1,0.50" [90] and further processed using Samtools 1.2 [91], BEDTools [94], and custom scripts. RPM shown in Fig. 4.1B and 4.1C were calculated by dividing the total number of reads in each size class by the total million reads mapped. For Fig. 4.1B, only reads that mapped to the dual strand transcribed region on cSSR 11.9 (3826-3835 kb) were included, whereas Fig. 4.1C includes all mapped reads. Differential expression analysis on small RNA-seq read count data (WT versus *H3.V* KO) was performed using EdgeR (4.S2

Table). Significance testing was pairwise using Fisher's Exact test. Significance was assessed in both the total small RNA-seq reads and in the 21-27nt reads.

For the mRNA-seq, total RNA was isolated from log phase *T. brucei* cultures ($5x10^7$ cells) using Trizol. 12 mRNA-seq libraries were constructed (triplicate WT, WT+DMOG, *H3.V* KO, and *H3.V* KO+DMOG) using Illumina TruSeq Stranded RNA LT Kit following the manufacturer's instructions with limited modifications. The starting quantity of total RNA was adjusted to 1.3 µg, and all volumes were reduced to a third of the described quantity. High throughput sequencing was performed at the Georgia Genomics Facility (GFF) on a NextSeq500 (Illumina). Raw reads from mRNA-seq were first trimmed using Trimmomatic version 0.32 [95]. The single-end reads were trimmed for TruSeq3 adapters; leading and trailing bases with quality less than 15 and reads with average quality less 20 were removed. Finally, any reads shorter than 50 base pair were discarded. Remaining reads were locally aligned to the *T. brucei* Lister 427 version 9.0 genome, from EuPathDB [92], using Bowtie2 version 2.2.3 [90]. All settings were default except specifying sensitive local and further processed with Samtools 1.2 [91]. Transcript abundances were computed using the Cufflinks suite version 2.2.1 [96]. For individual replicates, Cuffnorm was used with the library type fr-firststrand flag and the *T. brucei* Lister 427 version 9.0 annotation (downloaded from EuPathDB [92]). To estimate gene expression levels for a condition, replicates were used together and analyzed by Cuffdiff with the *T. brucei* Lister 427 version 9.0 annotation. Default parameters were used except specifying library type fr-firststrand. All p values reported here, determined by Cuffdiff, reflect the FDR-adjusted p value. Correlation coefficients for mRNA-seq replicates of WT, WT+DMOG, and *H3.V* KO were all greater than 0.96 and *H3.V* KO+DMOG replicates were greater than 0.91. To express the transcripts levels for individual mRNA encoding genes as

shown in 4.S1 Table, we determined the number of reads per kilobase per million reads (RPKM) [97]. Briefly, we counted the number of reads mapped to all annotated transcriptomic features (e.g. mRNA) on the same strand (i.e. sense) and opposite strand (i.e. antisense). Both the sense and antisense read numbers were normalized by length of the feature (in kilobase) and the total number of reads (in millions) mapped to non-structural RNAs in the corresponding library (i.e. number of mappable reads excluding rRNA and tRNA reads). mRNA-seq data shown in Fig. 4.2C, 4.3C, 4.5D, 4.S3C, and 4.S5C are from our previously published dataset [35] and are consistent with mRNA-seq performed in this study (see 4.S3 Table for the RNA-seq datasets used in each figure). Genes were considered adjacent to base J and/or H3.V if the gene, according to the *T. brucei* Lister 427 annotation, overlapped within 10,000 base pairs upstream or downstream of the modification. All J IP-seq and H3.V ChIP-seq data shown here are from previously published work [13, 14]. Fold changes for the heatmaps were computed as $(RPKM_{var} + pseudocount) / (RPKM_{wt} + pseudocount)$, where pseudocount = 0.5. Once all fold changes were computed, any fold change value above five was set equal to five to improve visualization.

Strand-specific RT-PCR analysis of read through transcription

Total RNA was isolated using the hot phenol method, as described previously [98]. To ensure complete removal of contaminating genomic DNA, purified RNA was treated with Turbo DNase, followed by phenol:chloroform extraction. RNA concentration was determined using a spectrophotometer. Strand specific RT-PCR was performed as previously described [99]. ThermoScript™ Reverse Transcriptase from Life Technologies was used for cDNA synthesis at 60-65°C. 1-2 μg of RNA were used to make cDNA using a reverse primer as

described in the Figure legends. PCR was performed using GoTaq DNA Polymerase from Promega. A minus-RT control was used to ensure no contaminating genomic DNA was amplified. Primer sequences used in the analysis are available upon request.

Reverse transcription quantitative PCR (RT qPCR)

Total RNA was obtained using Qiagen RNeasy kits according to manufacturers instructions. First-strand cDNA was synthesized from 1 μg of total RNA using an iScript cDNA synthesis kit (Bio-Rad Laboratories, Hercules, CA) per the manufacturer's instructions. Quantification of selected genes were performed on an iCycler with an iQ5 multicolor real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA). Primer sequences used in the analysis are available upon request. The reaction mixture contained 5 pmol forward and reverse primer, 2x iQ SYBR green super mix (Bio-Rad Laboratories, Hercules, CA), and 2 μl of template cDNA. Standard curves were prepared for each gene using 5-fold dilutions of known quantity (100 ng/μl) of WT DNA. The quantities were calculated using iQ5 optical detection system software.

Acknowledgments

We are grateful to Jessica Lopes da Rosa-Spiegler, Rudo Kieft, Whitney Bullard, and Piet Borst for critical reading of the manuscript. This paper is dedicated to the memory of Dr. Laura Cliffe.

REFERENCES

1. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science. 2005; 309:416-22.
2. Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, et al. The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. PLoS Negl Trop Dis. 2010; 4:e658.
3. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005; 309:404-9.
4. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell. 2003; 11:1291-9.
5. Boothroyd JC, Cross GA. Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. Gene. 1982; 20:281-9.
6. Van der Ploeg LH, Liu AY, Michels PA, De Lange TD, Borst P, Majumder HK, et al. RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes. Nucleic Acids Res. 1982; 10:3591-604.
7. De Lange T, Liu AY, Van der Ploeg LH, Borst P, Tromp MC, Van Boom JH. Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? Cell. 1983; 34:891-900.
8. Nelson RG, Parsons M, Barr PJ, Stuart K, Selkirk M, Agabian N. Sequences homologous to the variant antigen mRNA spliced leader are located in tandem repeats and variable orphons in Trypanosoma brucei. Cell. 1983; 34:901-9.

9. Sutton RE, Boothroyd JC. Evidence for Trans splicing in trypanosomes. Cell. 1986; 47:527-35.

10. Agabian N. Trans splicing of nuclear pre-mRNAs. Cell. 1990; 61:1157-60.

11. Clayton CE. Life without transcriptional control? From fly to man and back again. EMBO J. 2002; 21:1881-8.

12. Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: why be normal? Microbes and Infection. 2003; 5:1231-40.

13. Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. Nucleic Acids Res. 2010; 38:3923-35.

14. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. Genes Dev. 2009; 23:1063-76.

15. Respuela P, Ferella M, Rada-Iglesias A, Aslund L. Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. J Biol Chem. 2008; 283:15884-92.

16. Thomas S, Green A, Sturm NR, Campbell DA, Myler PJ. Histone acetylations mark origins of polycistronic transcription in Leishmania major. BMC Genomics. 2009; 10:152.

17. Wright JR, Siegel TN, Cross GA. Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in Trypanosoma brucei. Mol Biochem Parasitol. 2010; 172:141-4.

18. van Leeuwen F, Taylor MC, Mondragon A, Moreau H, Gibson W, Kieft R, et al. beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. Proc Natl Acad Sci U S A. 1998; 95:2366-71.

19. Dooijes D, Chaves I, Kieft R, Dirks-Mulder A, Martin W, Borst P. Base J originally found in kinetoplastid is also a minor constituent of nuclear DNA of Euglena gracilis. Nucleic Acids Res. 2000; 28:3017-21.

20. Gommers-Ampt JH, Van Leeuwen F, de Beer AL, Vliegenthart JF, Dizdaroglu M, Kowalak JA, et al. beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. Cell. 1993; 75:1129-36.

21. van Leeuwen F, Kieft R, Cross M, Borst P. Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosal-hydroxymethyluracil in *Trypanosoma brucei*. Mol Biochem Parasitol. 2000; 109:133-45.

22. van Leeuwen F, Wijsman ER, Kieft R, van der Marel GA, van Boom JH, Borst P. Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. Genes Dev. 1997; 11:3232-41.

23. van Leeuwen F, Wijsman ER, Kuyl-Yeheskiely E, van der Marel GA, van Boom JH, Borst P. The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands. Nucleic Acids Res. 1996; 24:2476-82.

24. Ekanayake DK, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, et al. Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. Mol Cell Biol. 2011; 31:1690-700.

25. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. Cell. 2012; 150:909-21.

26. Borst P, Sabatini R. Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol. 2008; 62:235-51.

27. Horn D. Antigenic variation in African trypanosomes. Mol Biochem Parasitol. 2014; 195:123-9.

28. Horn D, McCulloch R. Molecular mechanisms underlying the control of antigenic variation in African trypanosomes. Curr Opin Microbiol. 2010; 13:700-5.

29. Bullard W, Lopes da Rosa-Spiegler J, Liu S, Wang Y, Sabatini R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J Biol Chem. 2014; 289:20273-82.

30. Sekar A, Merritt C, Baugh L, Stuart K, Myler PJ. Tb927.10.6900 encodes the glucosyltransferase involved in synthesis of base J in Trypanosoma brucei. Mol Biochem Parasitol. 2014; 196:9-11.

31. Cliffe LJ, Hirsch G, Wang J, Ekanayake D, Bullard W, Hu M, et al. JBP1 and JBP2 Proteins Are Fe2+/2-Oxoglutarate-dependent Dioxygenases Regulating Hydroxylation of Thymidine Residues in Trypanosome DNA. J Biol Chem. 2012; 287:19886-95.

32. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009; 324:930-5.

33. Iyer LM, Tahiliani M, Rao A, Aravind L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle. 2009; 8:1698-710.

34. Reynolds D, Cliffe L, Sabatini R. 2-Oxoglutarate-dependent hydroxylases involved in DNA base J (β-D-Glucopyranosyloxymethyluracil) synthesis. In: Schofield CJ, Hausinger RP, editors. 2-Oxoglutarate-Dependent Oxygenases. 1. Cambridge, U.K: Royal Society of Chemistry; 2015.

35. Reynolds D, Cliffe L, Forstner KU, Hon CC, Siegel TN, Sabatini R. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. Nucleic Acids Res. 2014; 42:9717-29.

36. Cliffe LJ, Kieft R, Southern T, Birkeland SR, Marshall M, Sweeney K, et al. JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res. 2009; 37:1452-62.

37. Ekanayake D, Sabatini R. Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. Eukaryot Cell. 2011; 10:1465-72.

38. Weber CM, Henikoff S. Histone variants: dynamic punctuation in transcription. Genes Dev. 2014; 28:672-82.

39. Tolstorukov Michael Y, Goldman Joseph A, Gilbert C, Ogryzko V, Kingston Robert E, Park Peter J. Histone Variant H2A.Bbd Is Associated with Active Transcription and mRNA Processing in Human Cells. Mol Cell. 2012; 47:596-607.

40. Weber CM, Ramachandran S, Henikoff S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol Cell. 2014; 53:819-30.

41. Chen Y, Chen Q, McEachin RC, Cavalcoli JD, Yu X. H2A.B facilitates transcription elongation at methylated CpG loci. Genome Res. 2014; 24:570-9.

42. Draker R, Sarcinella E, Cheung P. USP10 deubiquitylates the histone variant H2A.Z and both are required for androgen receptor-mediated gene activation. Nucleic Acids Research. 2011; 39:3529-42.

43. Cuadrado A, Corrado N, Perdiguero E, Lafarga V, Muñoz-Canoves P, Nebreda AR. Essential role of p18Hamlet/SRCAP-mediated histone H2A.Z chromatin incorporation in muscle differentiation. The EMBO Journal. 2010; 29:2014-25.

44. Gévry N, Hardy S, Jacques P-É, Laflamme L, Svotelis A, Robert F, et al. Histone H2A.Z is essential for estrogen receptor signaling. Genes Dev. 2009; 23:1522-33.

45. Bönisch C, Hake SB. Histone H2A variants in nucleosomes and chromatin: more or less stable? Nucleic Acids Research. 2012; 40:10719-41.

46. Chen Y, Chen Q, McEachin R, Cavalcoli J, Yu X. H2A.B facilitates transcription elongation at methylated CpG loci. Genome Res. 2014.

47. Gamble MJ, Frizzell KM, Yang C, Krishnakumar R, Kraus WL. The histone variant macroH2A1 marks repressed autosomal chromatin, but protects a subset of its target genes from silencing. Genes Dev. 2010; 24:21-32.

48. Buschbeck M, Uribesalgo I, Wibowo I, Rue P, Martin D, Gutierrez A, et al. The histone variant macroH2A is an epigenetic regulator of key developmental genes. Nat Struct Mol Biol. 2009; 16:1074-9.

49. Mito Y, Henikoff JG, Henikoff S. Genome-scale profiling of histone H3.3 replacement patterns. Nat Genet. 2005; 37:1090-7.

50. Delbarre E, Jacobsen BM, Reiner AH, Sørensen AL, Küntziger T, Collas P. Chromatin Environment of Histone Variant H3.3 Revealed by Quantitative Imaging and Genome-scale Chromatin and DNA Immunoprecipitation. Mol Biol Cell. 2010; 21:1872-84.

51. Ooi SL, Henikoff JG, Henikoff S. A native chromatin purification system for epigenomic profiling in Caenorhabditis elegans. Nucleic Acids Research. 2010; 38:e26-e.

52. Wong LH, Ren H, Williams E, McGhie J, Ahn S, Sim M, et al. Histone H3.3 incorporation provides a unique and functionally essential telomeric chromatin in embryonic stem cells. Genome Res. 2009; 19:404-14.

53. Santenard A, Ziegler-Birling C, Koch M, Tora L, Bannister AJ, Torres-Padilla M-E. Heterochromatin formation in the mouse embryo requires critical residues of the histone variant H3.3. Nat Cell Biol. 2010; 12:853-62.

54. Drané P, Ouararhni K, Depaux A, Shuaib M, Hamiche A. The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. Genes Dev. 2010; 24:1253-65.

55. Lewis PW, Elsaesser SJ, Noh K-M, Stadler SC, Allis CD. Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. Proceedings of the National Academy of Sciences. 2010; 107:14075-80.

56. Elsasser SJ, Noh KM, Diaz N, Allis CD, Banaszynski LA. Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. Nature. 2015; 522:240-4.

57. Banaszynski Laura A, Wen D, Dewell S, Whitcomb Sarah J, Lin M, Diaz N, et al. Hira-Dependent Histone H3.3 Deposition Facilitates PRC2 Recruitment at Developmental Loci in ES Cells. Cell. 2013; 155:107-20.

58. Goldberg AD, Banaszynski LA, Noh K-M, Lewis PW, Elsaesser SJ, Stadler S, et al. Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions. Cell. 2010; 140:678-91.

59. Law C, Cheung P. Histone Variants and Transcription Regulation. In: Kundu TK, editor. Epigenetics: Development and Disease. Subcellular Biochemistry. 61: Springer Netherlands; 2013. p. 319-41.

60. Lowell JE, Cross GA. A variant histone H3 is enriched at telomeres in Trypanosoma brucei. J Cell Sci. 2004; 117:5937-47.

61. Tschudi C, Fau SH, Ullu E. Small interfering RNA-producing loci in the ancient parasitic eukaryote Trypanosoma brucei. BMC Genomics. 2012; 13.

62. Zheng LL, Wen YZ, Yang JH, Liao JY, Shao P, Xu H, et al. Comparative transcriptome analysis of small noncoding RNAs in different stages of Trypanosoma brucei. RNA. 2013; 19:863-75.

63. Wen Y-Z, Zheng L-L, Liao J-Y, Wang M-H, Wei Y, Guo X-M, et al. Pseudogene-derived small interference RNAs regulate gene expression in African Trypanosoma brucei. Proceedings of the National Academy of Sciences. 2011; 108:8345-50.

64. Zamore PD, Tuschl T, Sharp PA, Bartel DP. RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals. Cell. 2000; 101:25-33.

65. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 2001; 15:188-200.

66. Allen E, Xie Z, Gustafson AM, Carrington JC. microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants. Cell. 2005; 121:207-21.

67. MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, et al. Structural Basis for Double-Stranded RNA Processing by Dicer. Science. 2006; 311:195-8.

68. Akiyoshi B, Gull K. Discovery of unconventional kinetochores in kinetoplastids. Cell. 2014; 156:1247-58.

69. Haenni S, Renggli CK, Fragoso CM, Oberle M, Roditi I. The procyclin-associated genes of Trypanosoma brucei are not essential for cyclical transmission by tsetse. Mol Biochem Parasitol. 2006; 150:144-56.

70. Kim HS, Park SH, Gunzl A, Cross GA. MCM-BP is required for repression of life-cycle specific genes transcribed by RNA polymerase I in the mammalian infectious form of Trypanosoma brucei. PLoS One. 2013; 8:e57001.

71. Pena AC, Pimentel MR, Manso H, Vaz-Drago R, Pinto-Neves D, Aresta-Branco F, et al. Trypanosoma brucei histone H1 inhibits RNA polymerase I transcription and is important for parasite fitness in vivo. Mol Microbiol. 2014; 93:645-63.

72. Liniger M, Bodenmüller K, Pays E, Gallati S, Roditi I. Overlapping sense and antisense transcription units in Trypanosoma brucei. Mol Microbiol. 2001; 40:869-78.

73. Vanhamme L, Poelvoorde P, Pays A, Tebabi P, Van Xong H, Pays E. Differential RNA elongation controls the variant surface glycoprotein gene expression sites of Trypanosoma brucei. Mol Microbiol. 2000; 36:328-40.

74. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, et al. Telomeric expression sites are highly conserved in Trypanosoma brucei. PLoS One. 2008; 3:e3527.

75. Yang X, Figueiredo LM, Espinal A, Okubo E, Li B. RAP1 is essential for silencing telomeric variant surface glycoprotein genes in Trypanosoma brucei. Cell. 2009; 137:99-109.

76. Stanne TM, Kushwaha M, Wand M, Taylor JE, Rudenko G. TbISWI regulates multiple polymerase I (Pol I)-transcribed loci and is present at Pol II transcription boundaries in Trypanosoma brucei. Eukaryot Cell. 2011; 10:964-76.

77. Kelly S, Kramer S, Schwede A, Maini PK, Gull K, Carrington M. Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. Open Biol. 2012; 2:120033.

78. Schildkraut E, Miller CA, Nickoloff JA. Transcription of a Donor Enhances Its Use during Double-Strand Break-Induced Gene Conversion in Human Cells. Mol Cell Biol. 2006; 26:3098-105.

79. Alsford S, Horn D. RNA polymerase I transcription stimulates homologous recombination in T. brucei. Mol Biochem Parasitol. 2007; 153:10.1016/j.molbiopara.2007.01.013.

80. Hall JPJ, Wang H, Barry JD. Mosaic VSGs and the Scale of Trypanosoma brucei Antigenic Variation. PLoS Pathog. 2013; 9:e1003502.

81. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science. 2011; 333:1300-3.

82. Bernards A, van Harten-Loosbroek N, Borst P. Modification of telomeric DNA in Trypanosoma brucei; a role in antigenic variation? Nucleic Acids Research. 1984; 12:4153-70.

83. Haenni S, Studer E, Burkard GS, Roditi I. Bidirectional silencing of RNA polymerase I transcription by a strand switch region in Trypanosoma brucei. Nucleic Acids Res. 2009; 37:5007-18.

84. Mandava V, Fernandez JP, Deng H, Janzen CJ, Hake SB, Cross GA. Histone modifications in Trypanosoma brucei. Mol Biochem Parasitol. 2007; 156:41-50.

85. Anderson BA, Wong IL, Baugh L, Ramasamy G, Myler PJ, Beverley SM. Kinetoplastid-specific histone variant functions are conserved in Leishmania major. Mol Biochem Parasitol. 2013; 191:53-7.

86. Lye LF, Owens K, Shi H, Murta SM, Vieira AC, Turco SJ, et al. Retention and loss of RNA interference pathways in trypanosomatid protozoans. PLoS Pathog. 2010; 6:e1001161.

87. Robinson KA, Beverley SM. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. Mol Biochem Parasitol. 2003; 128:217-28.

88. Djikeng A, Shi H, Tschudi C, Ullu E. RNA interference in Trypanosoma brucei: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs. RNA. 2001; 7:1522-30.

89. DiPaolo C, Kieft R, Cross M, Sabatini R. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. Mol Cell. 2005; 17:441-51.

90. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357-9.

91. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078-9.

92. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010; 38:D457-62.

93. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011; 17:10-2.

94. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841-2.

95. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114-20.

96. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotech. 2013; 31:46-53.

97. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621-8.

98. Roditi I, Schwarz H, Pearson TW, Beecroft RP, Liu MK, Richardson JP, et al. Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. J Cell Biol. 1989; 108:737-46.

99. Al Husini N, Kudla P, Ansari A. A role for CF1A 3' end processing complex in promoter-associated transcription. PLoS Genet. 2013; 9:e1003722.

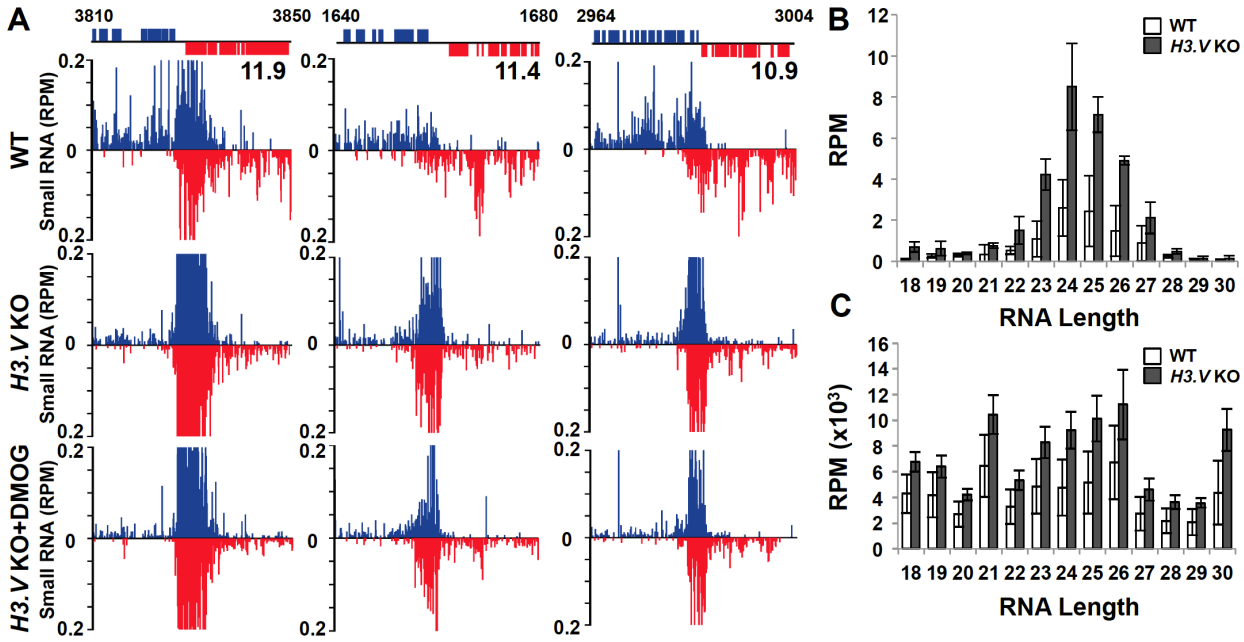


Fig. 4.1 Loss of H3.V stimulates the production of siRNAs in *T. brucei*. (A) Small RNA-sequencing reads for three representative cSSRs are shown (11.9, 11.4, and 10.9; where cSSR 11.9 refers to the ninth termination site on chromosome 11) where H3.V loss does not lead to read-through transcription, but does lead to increased siRNAs. Small RNA reads are plotted as reads per million reads mapped (RPM). ORFs and the genomic location (kb) are shown above the graphs. WT: wild type; KO: *H3.V* KO; KO+DMOG: *H3.V* KO + DMOG. Blue: top strand; red: bottom strand. (B and C) Length distribution of small RNAs. (B) Length distribution of small RNAs from cSSR 11.9. Shown is the RPM for each RNA length observed in cSSR 11.9, a site with a statistically significant increase in 21-27nt RNAs in the *H3.V* KO (4.S2 Table). The average of three independent small RNA-seq experiments is plotted. White bars, WT; black bars, *H3.V* KO. Error bars represent the standard deviation. (C) Length distribution of small RNAs genome-wide. The RPM for each RNA length observed in the entire small RNA-seq data set is shown. Data are plotted as in B.

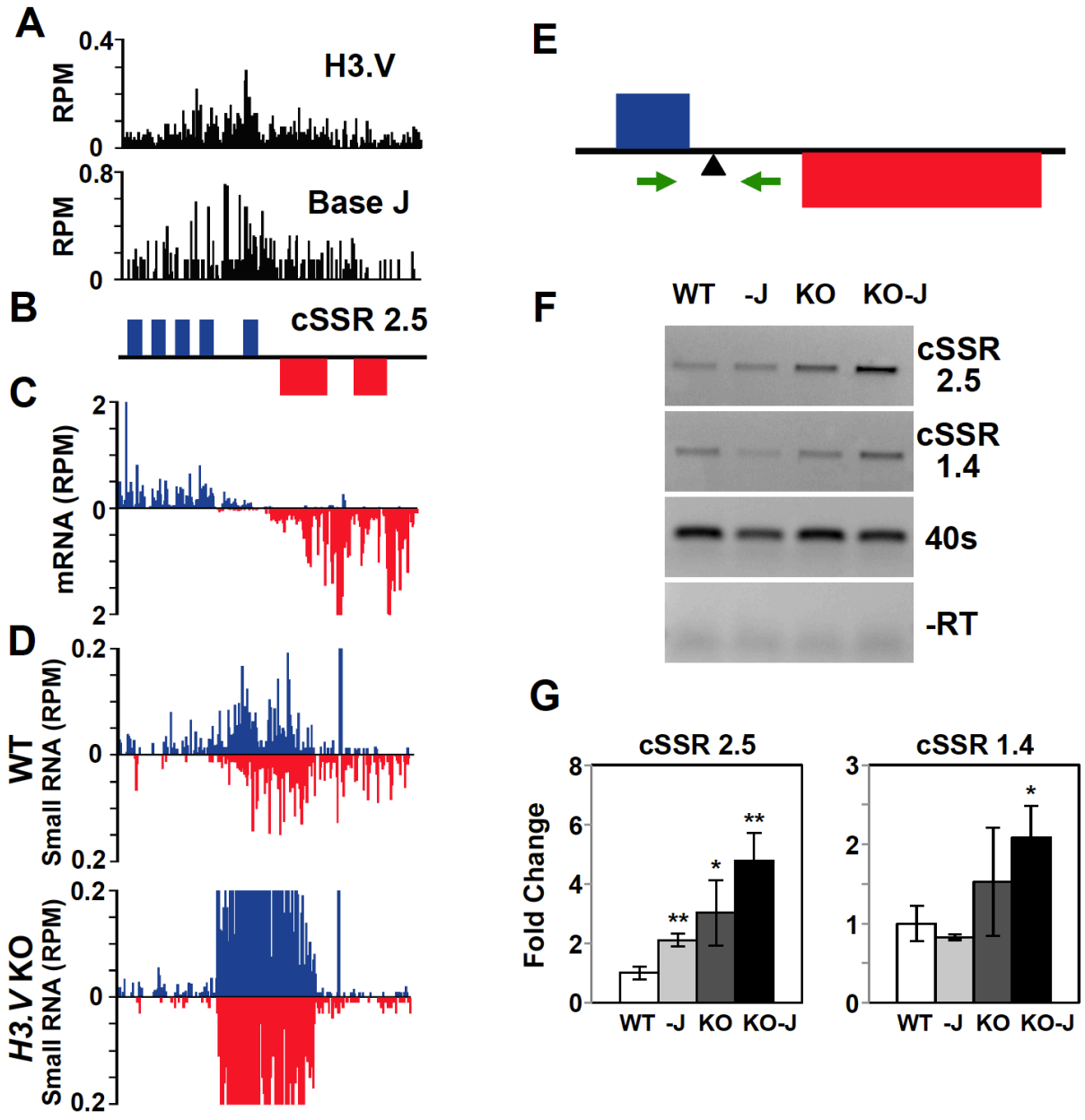


Fig. 4.2 Increased production of nascent RNA in cSSR following the loss of H3.V. A region on chromosome 2 (950-975 kb) representing cSSR 2.5 is shown where H3.V regulates transcription. (A) Base J and H3.V co-localize at sites of RNAP II termination within cSSRs [13, 14]. H3.V ChIP-seq reads and base J IP-seq reads are plotted as reads per million reads (RPM), as previously described [14, 35]. (B) ORFs are shown with the top strand in blue and the bottom strand in red. (C) mRNA-seq reads from wild type *T. brucei* are plotted as RPM.

Reads that mapped to the top strand are shown in blue and reads that mapped to the bottom strand in red. (D) Small RNA-seq reads from WT and *H3.V KO* are mapped as described in Fig. 4.1A. (E-G) Strand-specific RT-PCR analysis of nascent RNA in cSSRs. (E) Schematic representation (not to scale) of primer location and direction at cSSR 2.5 (primers shown as green arrows in this and all subsequent figures). The arrowhead below the line indicates the poly(A) processing site for the final gene in the PTU. (F) Strand-specific RT-PCR analysis. cDNA was synthesized using the reverse primer. PCR was performed using the same reverse primer to make the cDNA plus the forward primer, as indicated. Data is also presented for an additional cSSR on chromosome 1 (cSSR 1.4, 635-637 kb). Wild type: WT; Wild type+DMOG: -J; *H3.V* KO: KO; *H3.V* KO+DMOG: KO-J. 40S ribosomal protein S11 provides a positive control and minus RT (-RT) negative control is shown. (G) Nested qPCR. Primers were designed within the PCR reaction in F to use in subsequent qPCR analysis. White bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. All products were normalized to 40S ribosomal protein S11. The average of three independent strand-specific RT-PCR nested qPCR experiments is plotted. Error bars represent the standard deviation. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

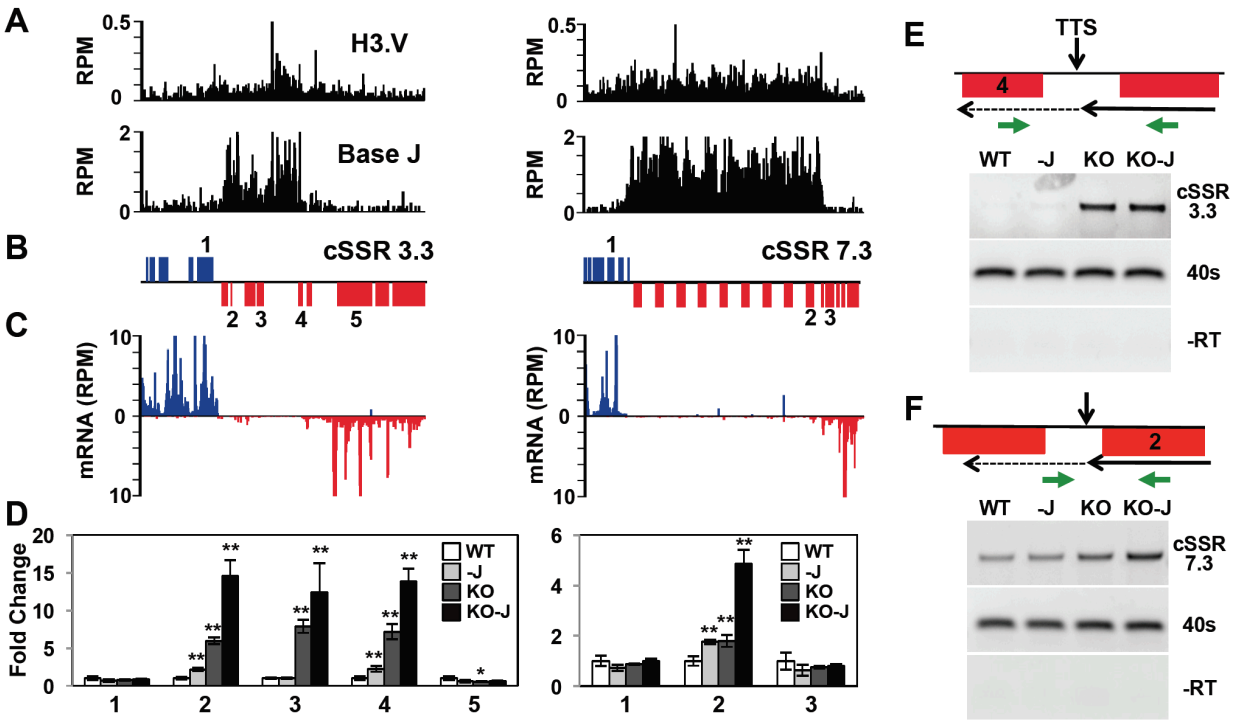


Fig. 4.3 Decreased efficiency of RNAP II termination and increased gene expression following the loss of histone H3.V. A region on chromosome 3 (617-670 kb) representing cSSR 3.3 and chromosome 7 (453-525 kb) representing cSSR 7.3 is shown where H3.V regulates transcription of a cluster of genes. (A-C) Base J and H3.V co-localize at sites of RNAP II termination within a PTU. H3.V ChIP-seq reads and base J IP-seq reads, ORFs, and mRNA-seq reads from wild type *T. brucei* are plotted for cSSR 3.3 (left) and cSSR 7.3 (right) as described in Fig. 4.2. (D) RT-qPCR analysis of genes numbered according to the ORF maps above in panel B. As described in Fig. 4.2G, white bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. Transcripts were

normalized against 40S ribosomal protein S11, and are plotted as the average and standard deviation of three replicates. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01. The silent gene cluster at cSSR 7.3 consists of nine highly similar retrotransposon hot spot protein genes, therefore the primers used to analyze gene 2 also amplify the additional upstream genes. (E and F) Strand-specific RT-PCR analysis of read-through transcription of the two cSSRs analyzed in A-D. Above each panel is a schematic representation (not to scale) of primer location and direction at a transcription termination site (TTS). The vertical arrow indicates the proposed TTS as described in the text [35]. The long solid arrow indicates the direction of transcription and the dashed arrow indicates read-through transcription past the TTS. cDNA was synthesized using the reverse primer (relative to transcription). PCR was performed using the same reverse primer to make the cDNA plus the forward primer, as indicated. 40S ribosomal protein S11 provides a positive control and a minus RT (-RT) negative control is shown.

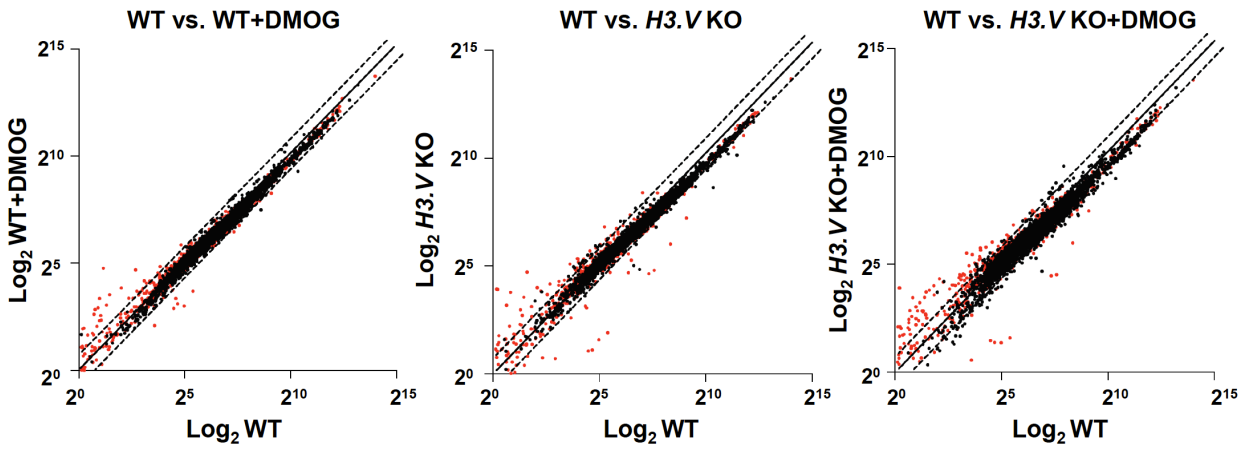


Fig. 4.4 Gene expression changes in the *H3.V* KO. The average reads per kilobase per million reads mapped (RPKM) of triplicate mRNA-seq libraries is plotted on a $\log_2$ scale. Genes differentially expressed by 2-fold or more in *T. brucei* following the loss of base J and/or H3.V fall above or below the dotted lines. Red dots indicate genes that are adjacent to H3.V (4.S1 Table). Only mRNAs with an RPKM≥1 are included.

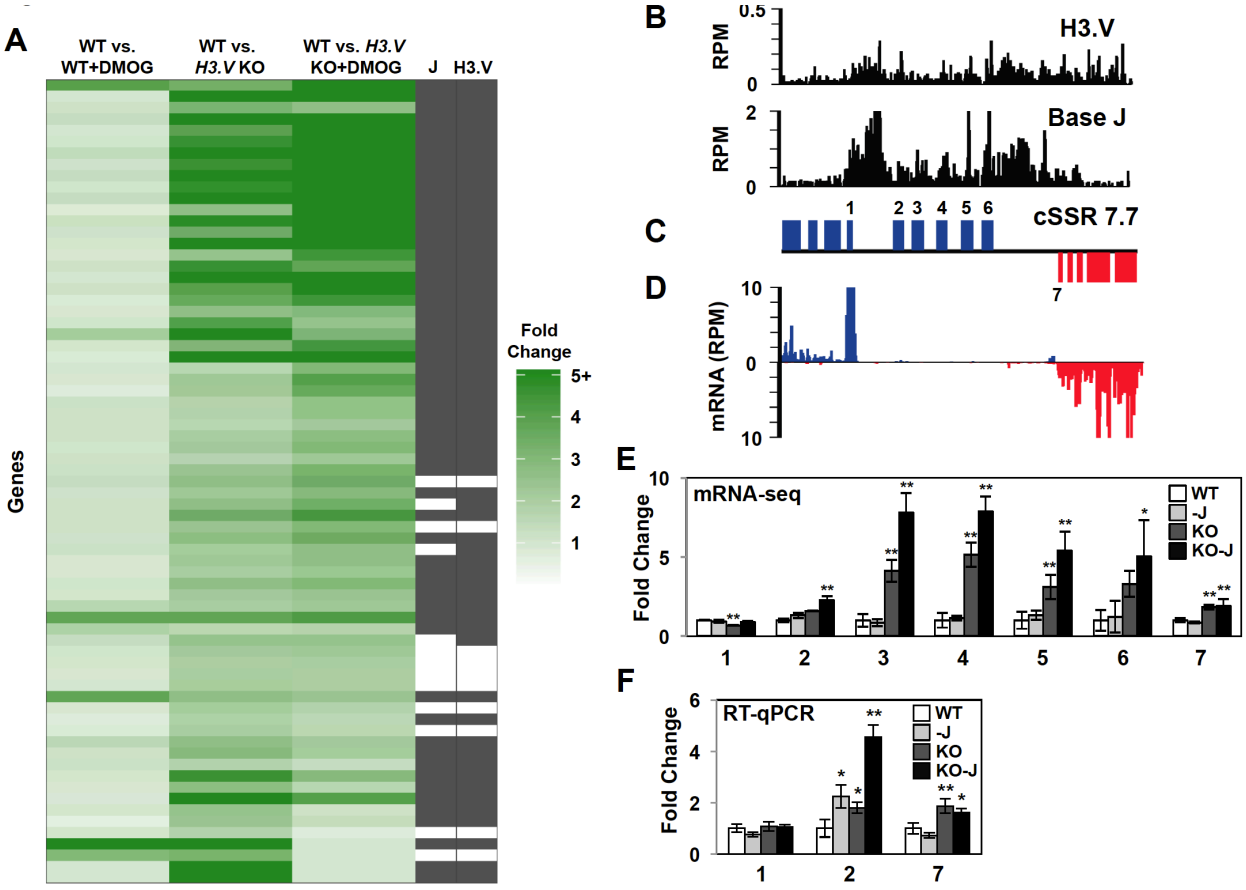


Fig. 4.5 H3.V and base J have independent yet additive roles in regulating termination and gene expression. (A) Heatmap of genes upregulated in the *H3.V* KO. For the list of genes represented on the heatmap see 4.S1 Table. Clustering of genes at the top indicate those that are further upregulated upon loss of base J in the *H3.V* KO. J and H3.V columns indicate whether each gene is located within 10 kb of the modification (filled black box), as described in the Materials and Methods section. (B-D) H3.V/J localization, gene map, and mRNA-seq reads plotted for a gene cluster on chromosome 7 at cSSR 7.7 (position 1750-1800 kb shown) is illustrated as described in Fig. 4.3. (E) Plot of the mRNA-seq data for the genes indicated (numbered) in the ORF map. The average RPKM of triplicate mRNA-seq libraries was used to determine fold changes, with wild type set to 1. Error bars indicate the standard deviation

between mRNA-seq replicates and p values, determined in Cuffdiff, are indicated by asterisks: *, p value ≤ 0.05; **, p value ≤ 0.01. (F) RT-qPCR analysis of gene expression for the indicated genes (according to the ORF map) as described in Fig. 4.3D. White bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

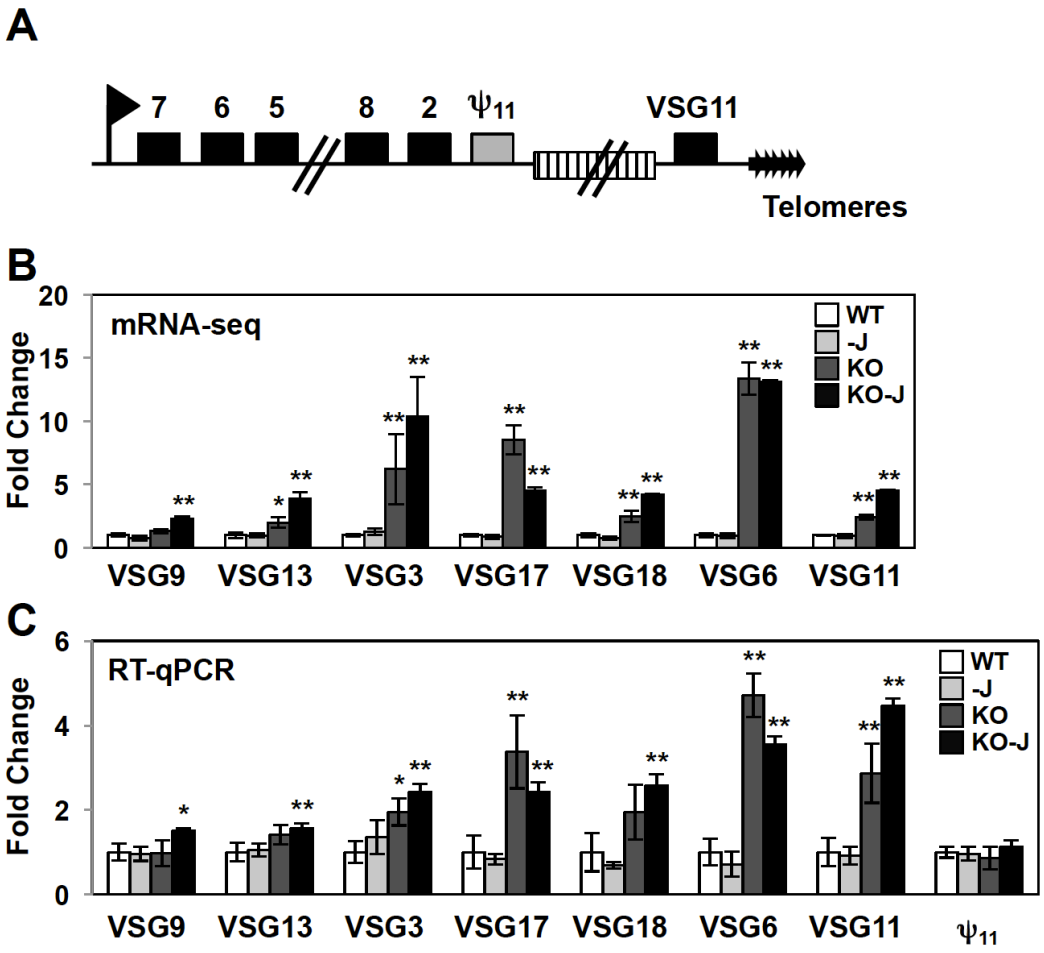


Fig. 4.6 H3.V regulates *VSG* gene expression from silent telomeric bloodstream expression sites. (A) A schematic diagram of the silent ES15 (not to scale). The box with stripes represents the 70 bp repeats. Numbers indicate *ESAG* genes. Grey box represents the *VSG* pseudogene 11 (Tb427.BES126.13). (B-C) mRNA-seq and RT-qPCR analysis of the indicated *VSG* genes in silent expression sites. As described in Fig. 4.3D and 4.5E, white bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. For mRNA-seq analysis, p values determined by Cuffdiff are indicated by asterisks: *, p value ≤ 0.05; **, p value ≤ 0.01. For RT-qPCR analysis, p values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

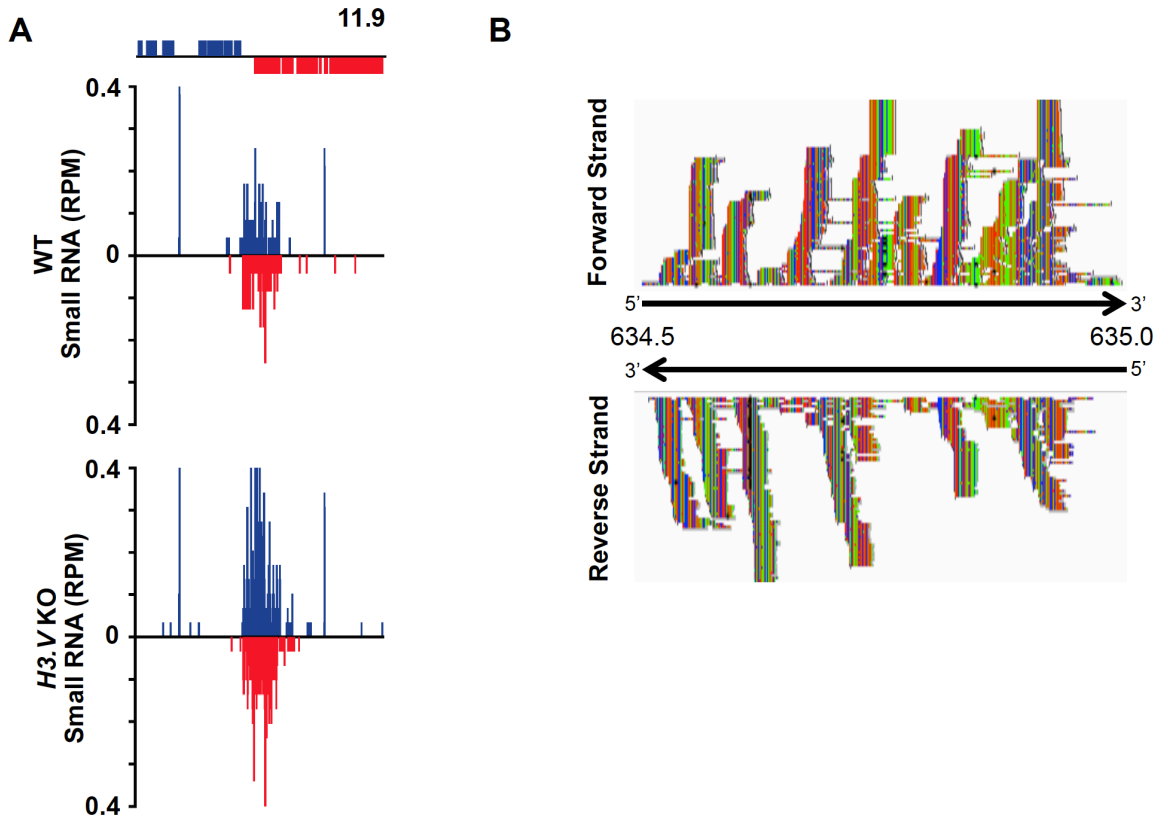


Fig. 4.S1 Regulation of siRNAs by H3.V. (A) Mapping of 23-26nt small RNAs to cSSR 11.9 in WT and *H3.V* KO. (B) Phasing of siRNAs mapping to a cSSR on chromosome 5 in WT cells. Position is indicated in kb. Colors indicate nucleotide: green, A; red, T; blue, C; and orange, G.

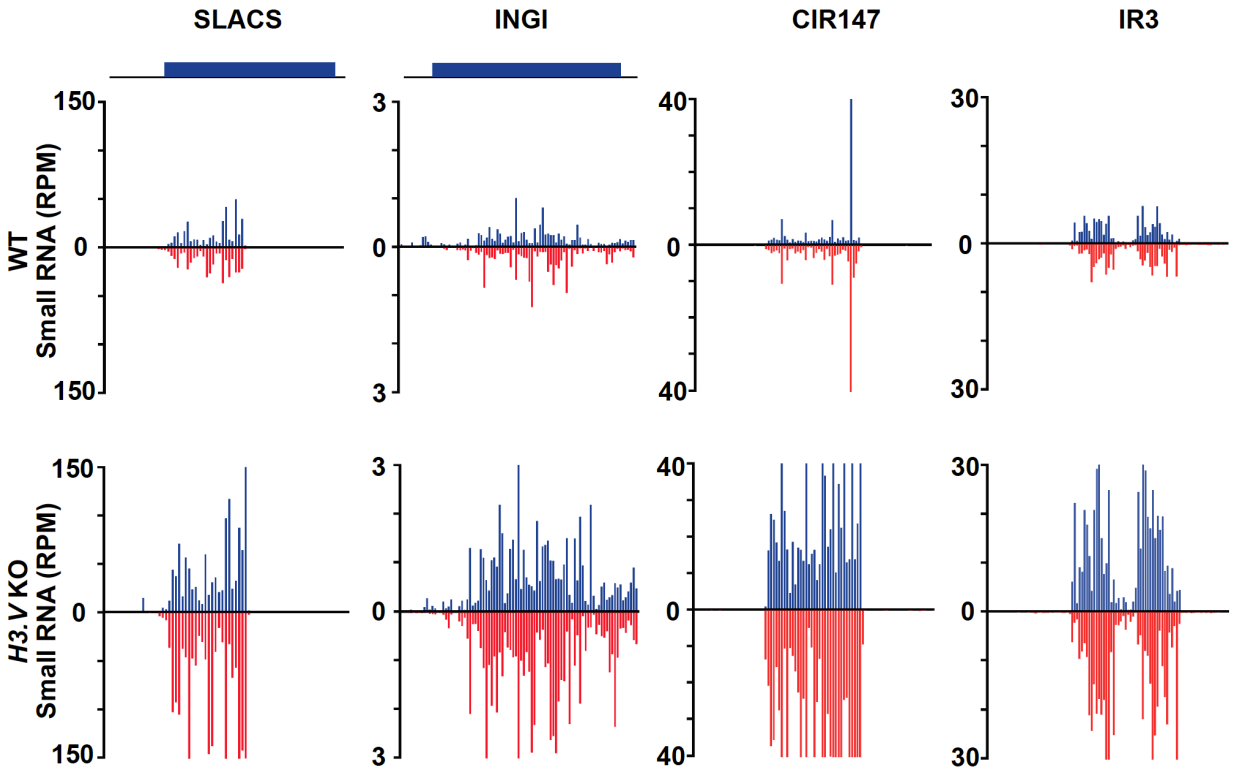


Fig. 4.S2 Regulation of siRNAs by H3.V at other loci. Mapping of small RNAs to SLACS, INGI, CIR147 and IR3

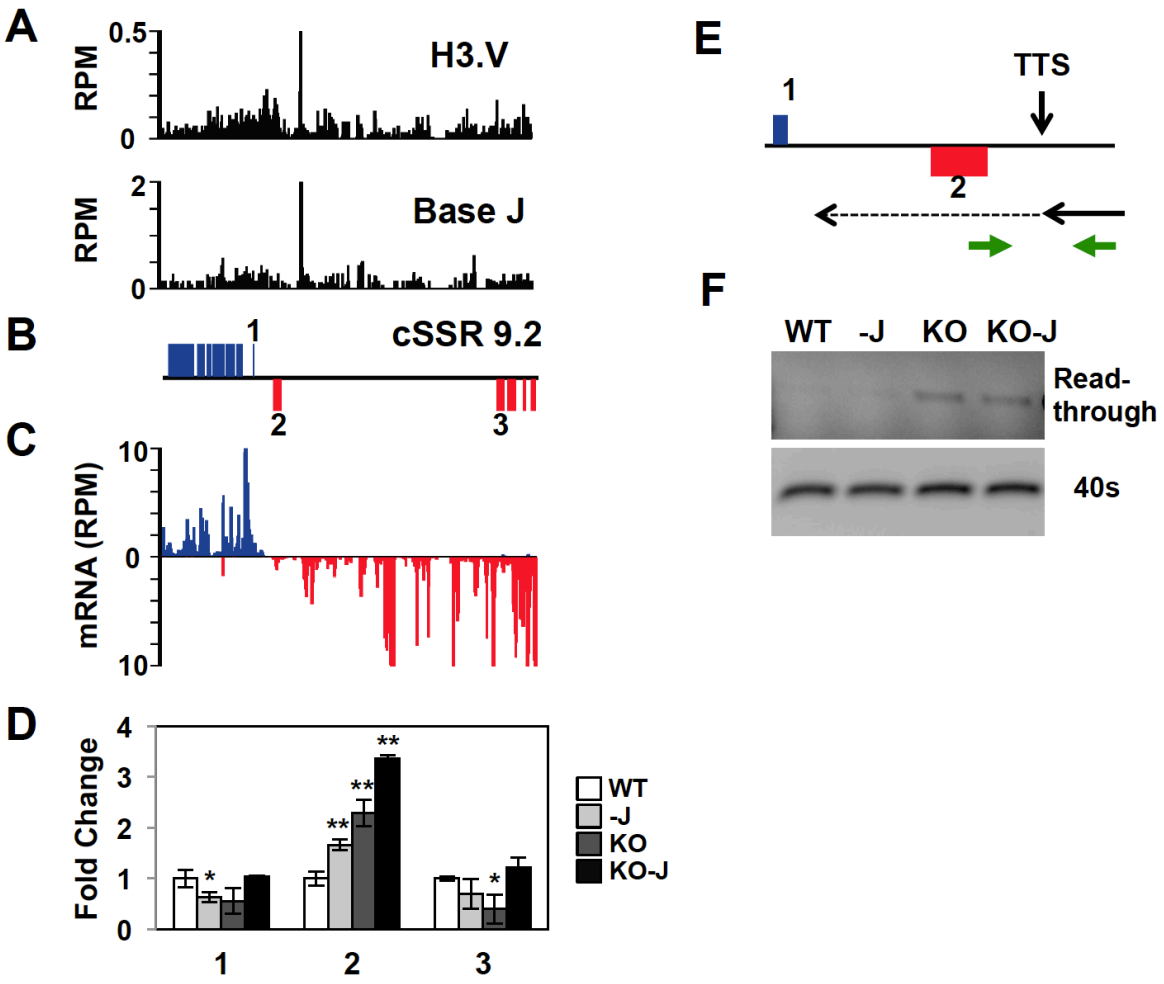


Fig. 4.S3 Regulation of termination and gene expression by H3.V. (A-C) Localization of H3.V, J, ORFs, and mRNA-seq reads from wild type *T. brucei* are plotted for cSSR 9.2 (position 1110-1190 kb is shown). (D-F) Gene expression changes and termination defects are analyzed as described in Fig. 4.3. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

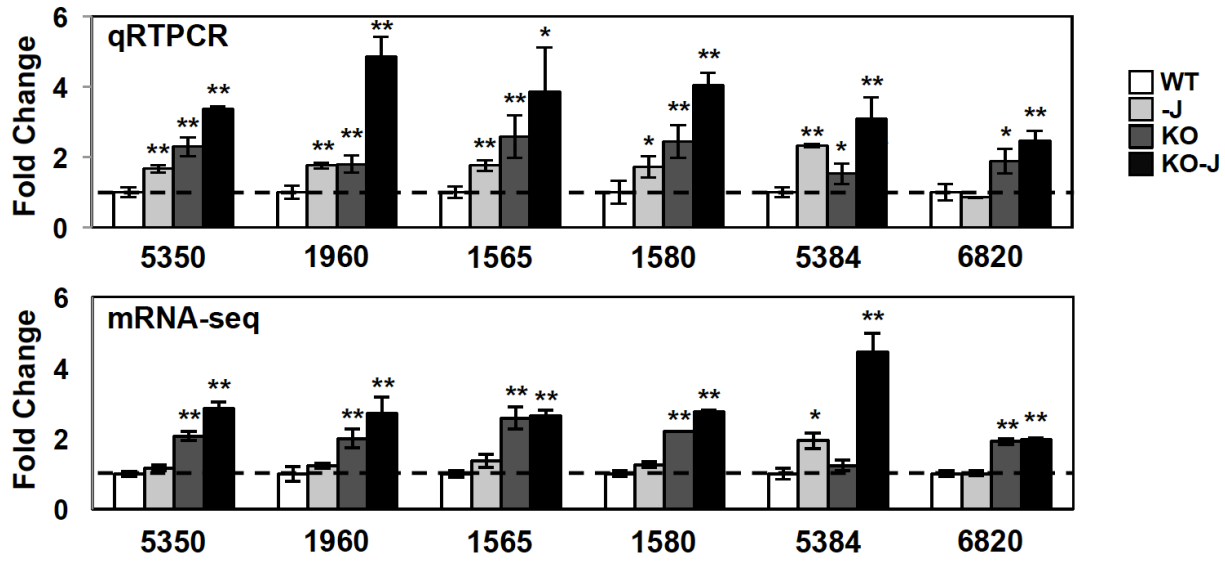


Fig. 4.S4 Confirmation of mRNA-seq transcript changes in *T. brucei* by RT-qPCR. RT-qPCR was performed as described in Fig. 4.3D. 5350: *Tb427tmp.160.5350*; 1960: *Tb427.07.1960*; 1565: *Tb427tmp.02.1565*; 1580: *Tb427tmp.02.1580*; 5384: *Tb427.02.5384*; and 6820: *Tb427.07.6820*. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

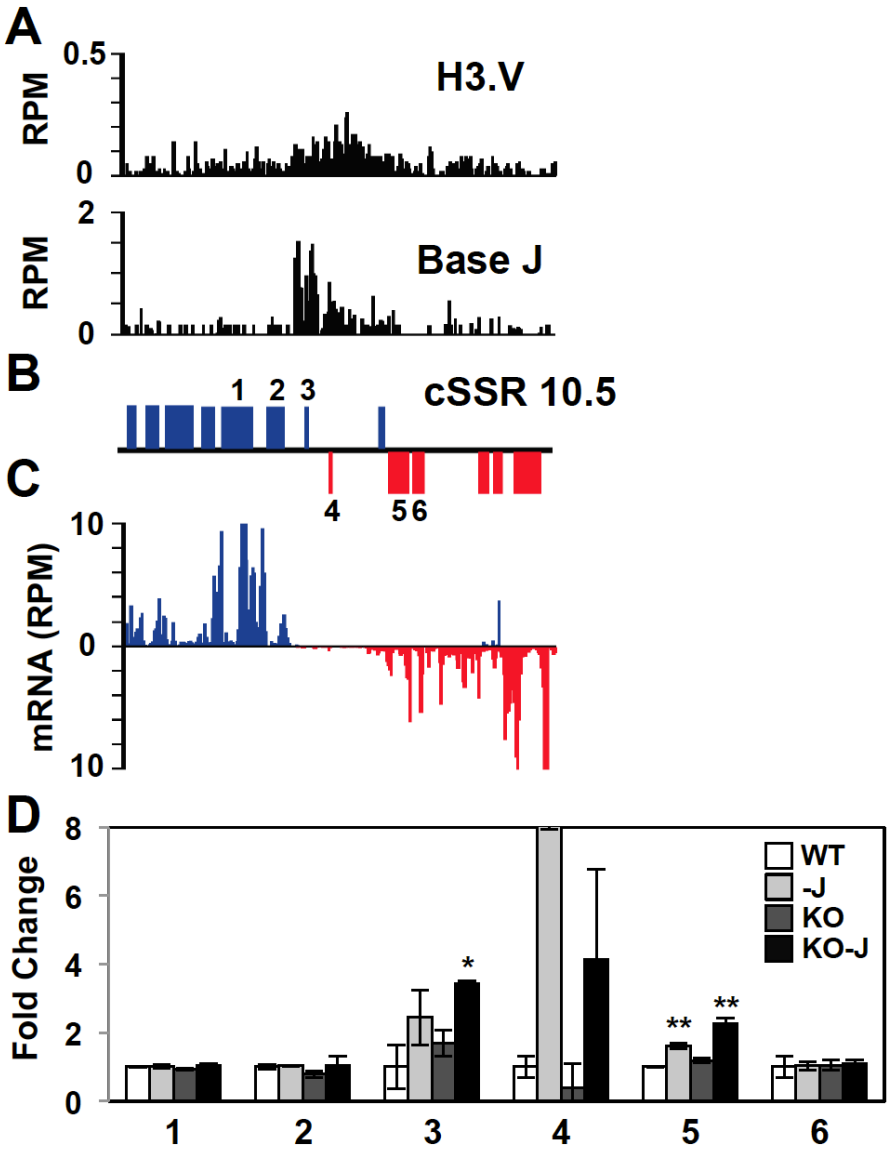
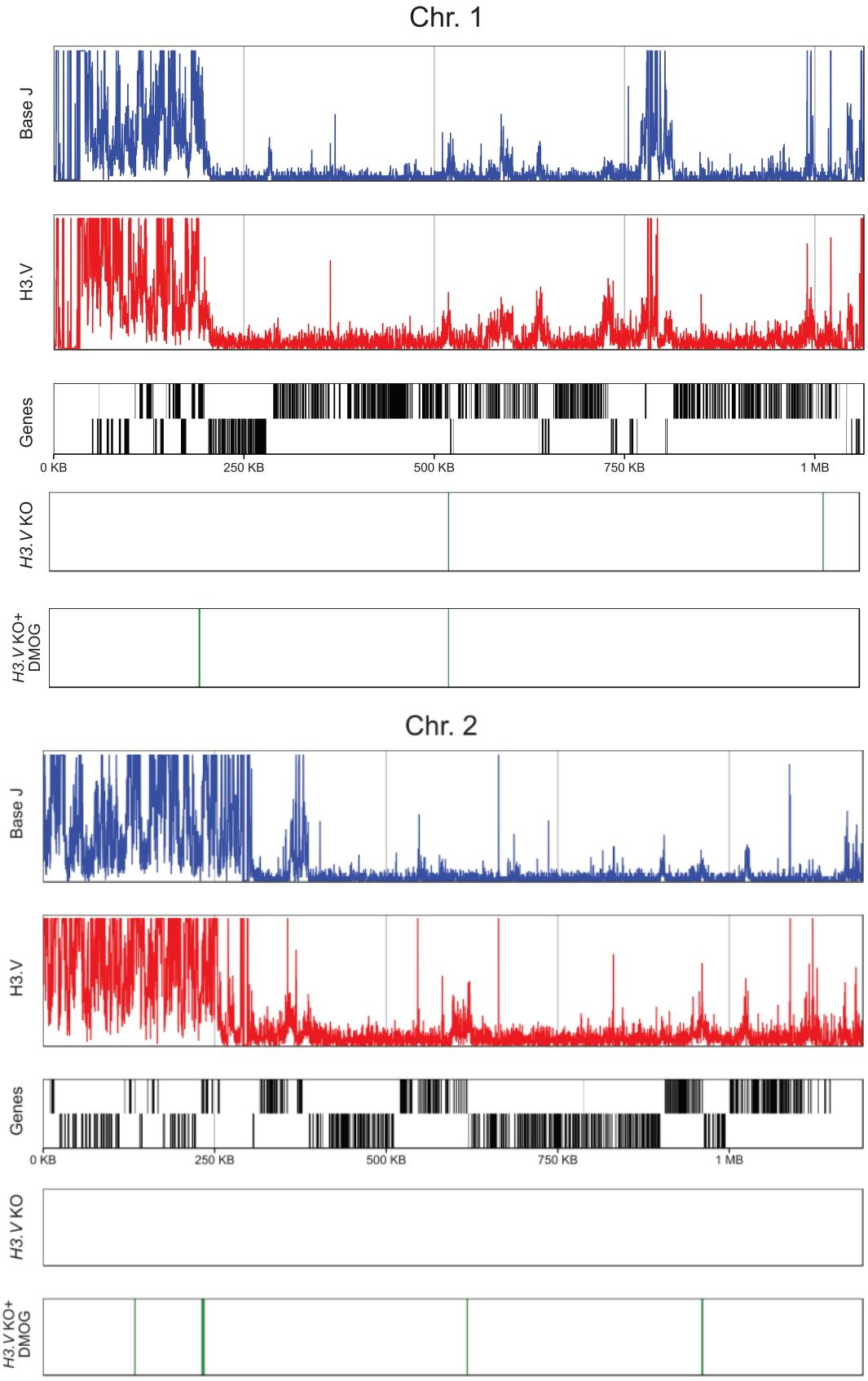
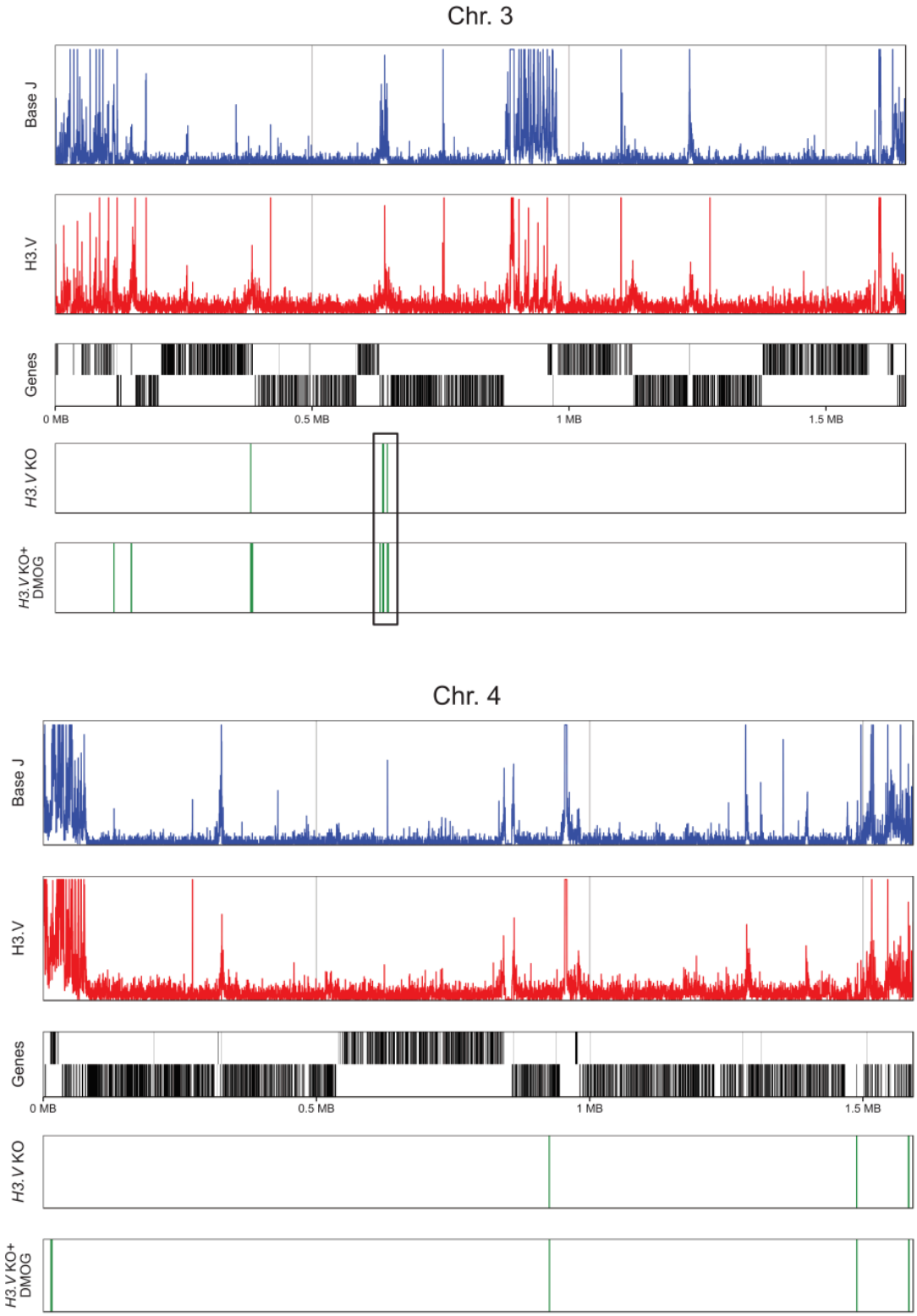
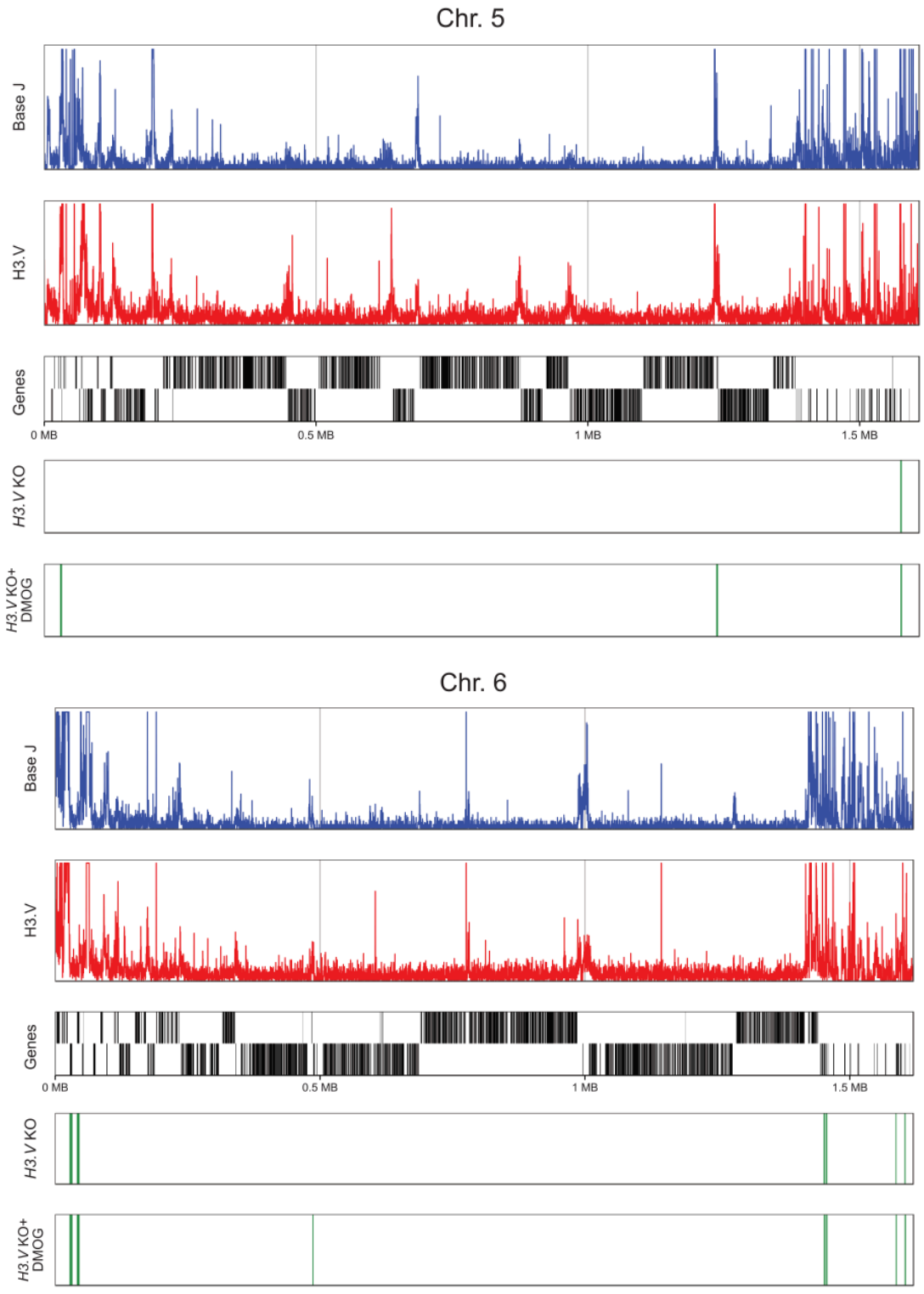
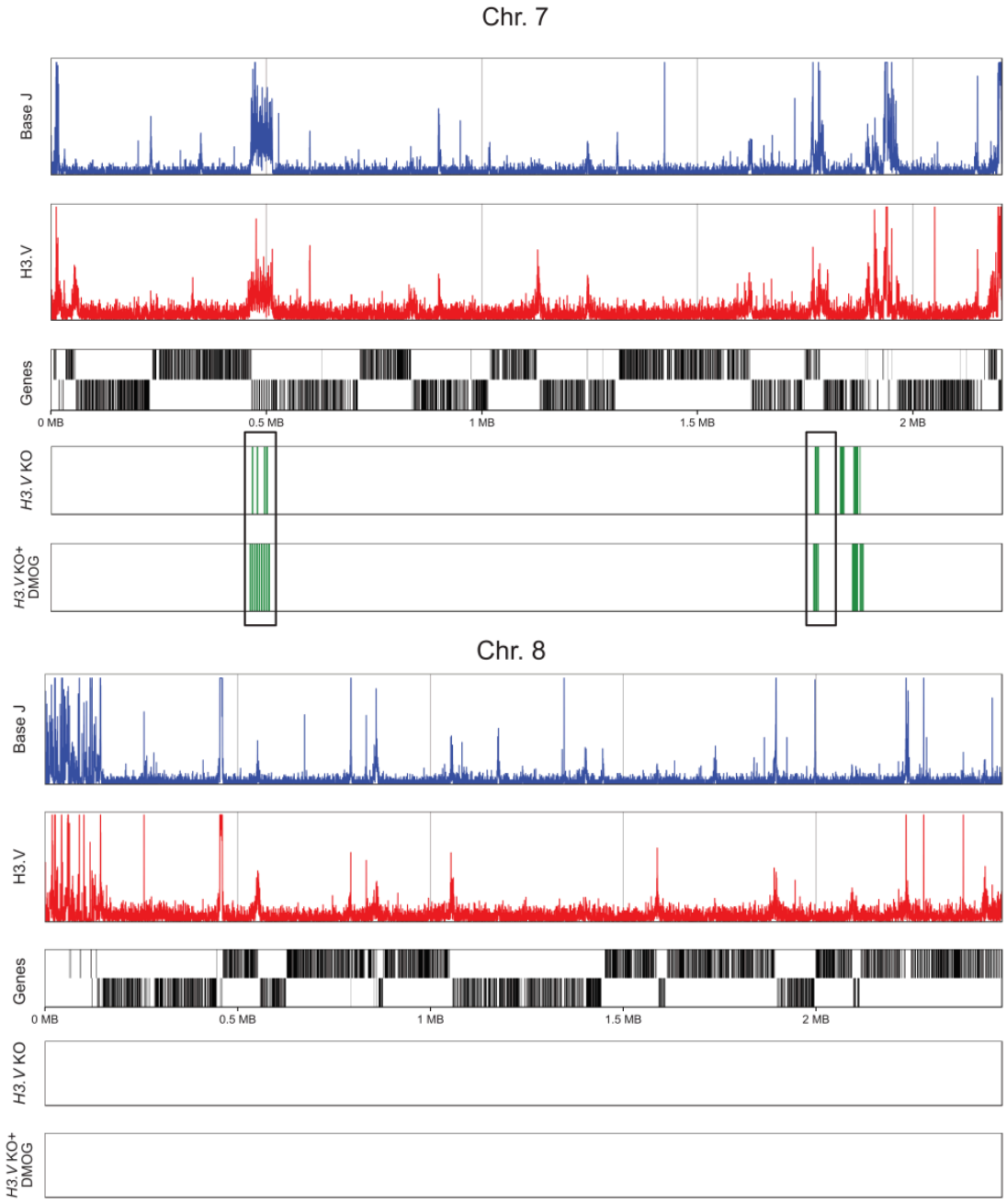


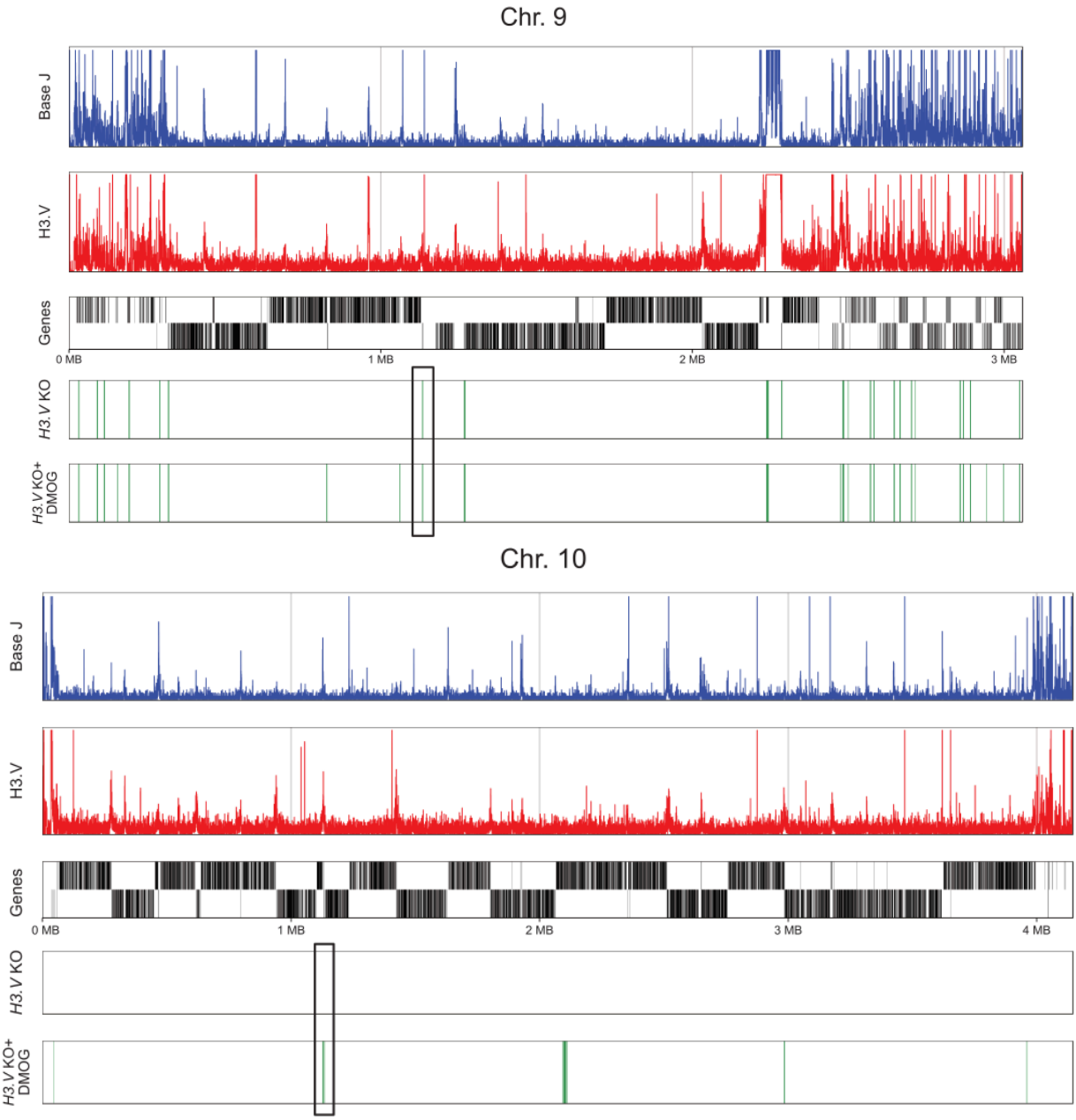
Fig. 4.S5 Regulation of termination and gene expression by H3.V. (A-C) Localization of H3.V, J, ORFs and mRNA-seq reads from wild type *T. brucei* are ploted for cSSR 10.5 (1120-1140). (D) mRNA-seq transcript fold changes of the genes indicated in the ORF map in B, as described in Fig. 4.5E. White bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. The fold change in the wild type+DMOG condition for gene 4 is 12.2, with a standard deviation of 4.3 and p value of 0.03.


Chr. 1
Base J
H3.V
Genes
0 KB
250 KB
500 KB
750 KB
1 MB
H3.V KO
H3.V KO+ DMOG
Chr. 2
Base J
H3.V
Genes
0 KB
250 KB
500 KB
750 KB
1 MB
H3.V KO
H3.V KO+ DMOG


Chr. 3
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
H3.V KO
H3.V KO+ DMOG
Chr. 4
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
H3.V KO
H3.V KO+ DMOG


Chr. 5
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
H3.V KO
H3.V KO+ DMOG
Chr. 6
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
H3.V KO
H3.V KO+ DMOG


Chr. 7
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
2 MB
H3.V KO
H3.V KO+ DMOG
Chr. 8
Base J
H3.V
Genes
0 MB
0.5 MB
1 MB
1.5 MB
2 MB
H3.V KO
H3.V KO+ DMOG


Chr. 9
Base J
H3.V
Genes
0 MB
1 MB
2 MB
3 MB
H3.V KO
H3.V KO+ DMOG
Chr. 10
Base J
H3.V
Genes
0 MB
1 MB
2 MB
3 MB
4 MB
H3.V KO
H3.V KO+ DMOG

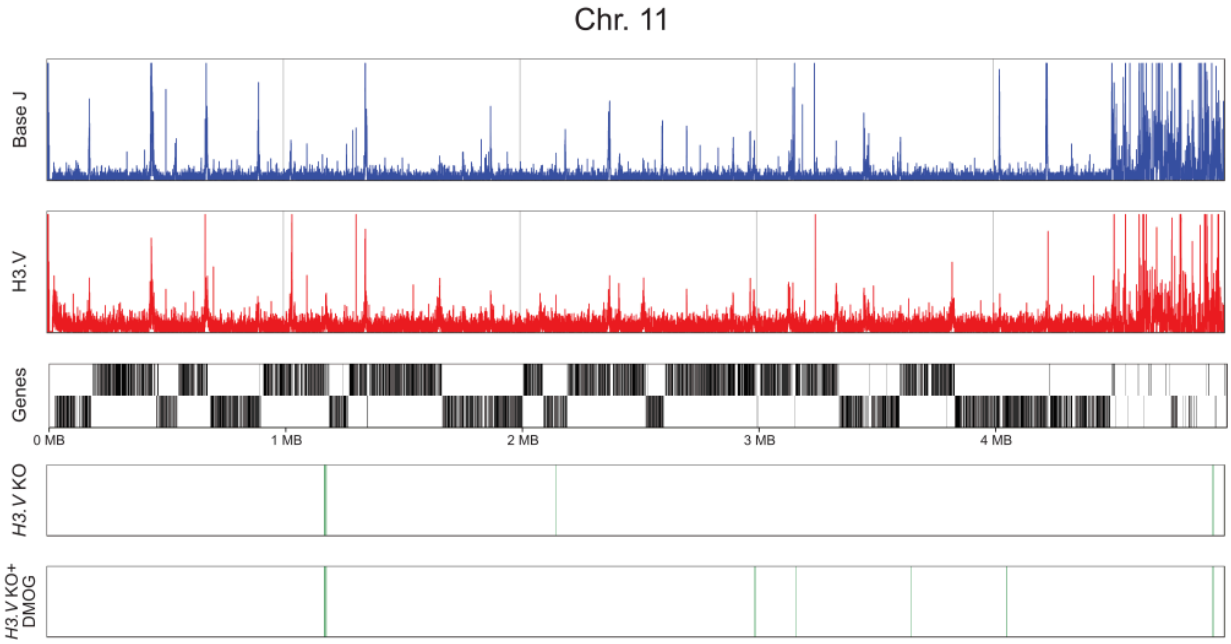


Fig. 4.S6 Chromosome maps. Whole chromosomes (Chr. 1-11, *T. brucei* Lister 427 version 9.0 genome) and the localization of base J (blue), H3.V (red), and mRNA coding genes (black lines; top strand is indicated by a line in the top half of the panel, bottom strand by a line in the bottom half) are shown. Genes on the top strand are transcribed from left to right and those on the bottom strand are transcribed from right to left. Position along each chromosome is indicated in kilobases (KB) or megabases (MB). Bottom two panels: mRNAs found upregulated by at least 2-fold or more in the *H3.V* KO (top) and *H3.V* KO+DMOG (bottom) relative to WT are indicated by a green line. Only mRNAs with an RPKM≥1 and significantly differentially expressed relative to wild type, as determined by Cuffdiff, are included. Boxes indicate sites examined in more detail in other figures. Genes are listed in 4.S1 and 4.S4 Tables.

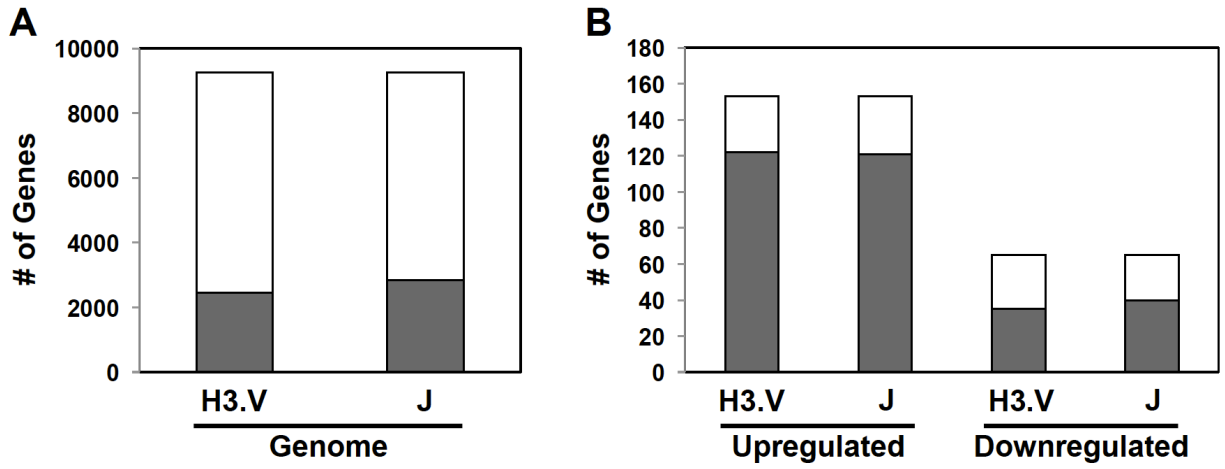


Fig. 4.S7 Enrichment of genes adjacent to H3.V and J following the loss of H3.V and/or J. (A) Genes were defined as adjacent to H3.V or J if located within 10 kb of an H3.V and J enriched region, respectively, and are indicated in grey. Genes not adjacent to H3.V or J are indicated in white [13, 14]. 2463 (27%) genes are adjacent to H3.V and 2837 (31%) are adjacent to J out of a total of 9266 annotated genes in the *T. brucei* genome. (B) 153 genes were upregulated in the absence of H3.V and/or J, 122 are adjacent to H3.V and 121 are adjacent to J (80%). 65 genes were downregulated in the absence of H3.V and/or J, 35 are adjacent to H3.V (54%) and 40 are adjacent to J (62%).

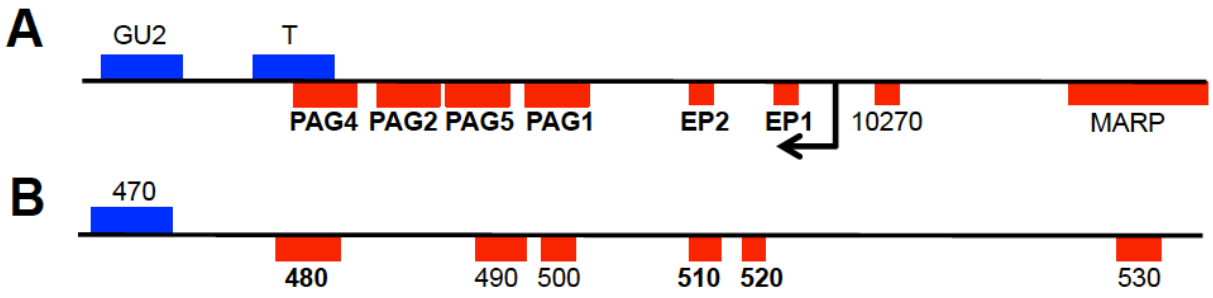


Fig. 4.S8 Genomic context of genes downregulated in the *H3.V* KO. (A) The EP/PAG1 loci. Small arrow indicates the RNAP I transcription start site in the promoter region. Genes in bold are downregulated in the *H3.V* KO. Genes in blue are transcribed by RNAP II on the top strand and EP and PAG genes in red are transcribed by RNAP I. MARP: microtubule-associated repetitive protein; EP1-2: procyclin; PAG; procyclin associated gene; T: 'T region' encoding transcripts containing small ORFs of <240 bp; GU2: gene of unknown function. The Fig. is drawn to scale. (B) Gene cluster on chromosome 6. Genes in bold are downregulated in the *H3.V* KO and identities are listed in 4.S1 Table.

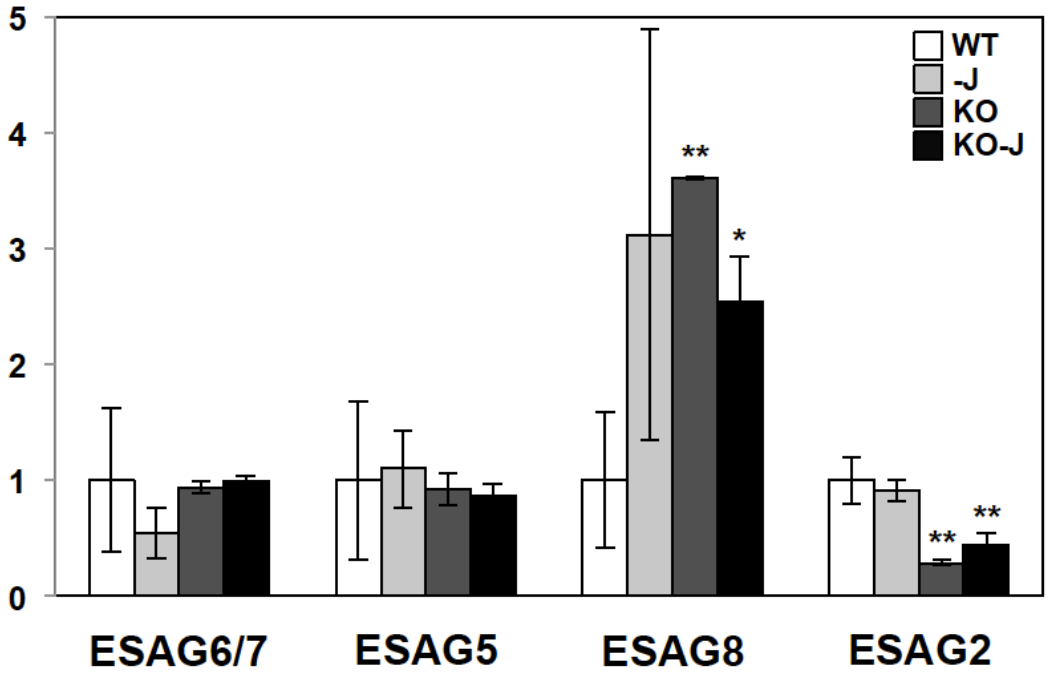


Fig. 4.S9 RT-qPCR analysis of ES associated *ESAGs*. RT-qPCR analysis of the indicated *ESAGs* was performed as described in Fig. 5.3D. P values were calculated using Student's t test. *, p value ≤ 0.05; **, p value ≤ 0.01.

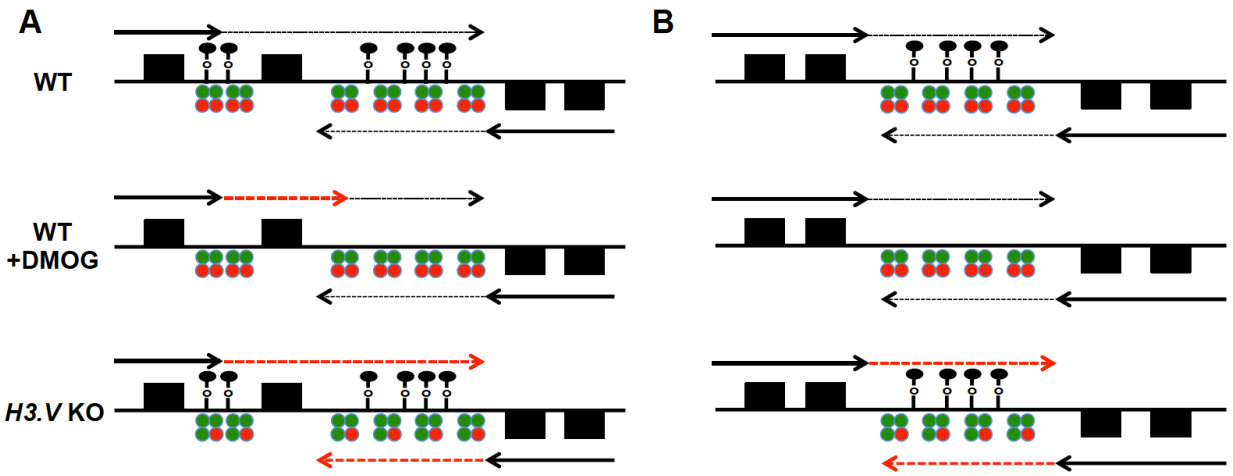


Fig. 4.S10 Working model for H3.V regulating RNAP II transcription and mRNA and siRNA expression. O-linked glycosylation of DNA (base J) is indicated by black line and dot. Nucleosomes are indicated by circles where green represents canonical histones and red represents histone H3 variant and an additional histone variant found at termination sites in *T. brucei* [14], histone H4 variant (H4.V). In the *H3.V* KO the H3.V is replaced with a canonical H3, with no change in nucleosome structure or H4.V, since it is currently unclear what happens to the nucleosome upon the loss of H3.V. According to the model, (A) the loss of base J leads to read-through transcription (indicated by the thicker red arrow) at internal termination sites within the cluster that is once again attenuated once it reaches H3.V within the cSSR. The loss of H3.V leads to read-through transcription at the internal site and continues into the cSSR, thus allowing increased dual strand transcription and generation of siRNAs. (B) At regions without an internal termination site, loss of base J has no affect on dual strand transcription. But, as described above, the loss of H3.V leads to increased transcription at cSSRs and generation of siRNAs.

Table 4.S1 *T. brucei* gene expression changes following H3.V and/or J loss. mRNAs found up or downregulated by 2-fold or more by mRNA-seq are listed, along with available gene descriptions, RPKM, fold change, and an indication of whether the genes are located within 10kb of H3.V and/or J. P values determined by Cuffdiff are listed for each condition compared to WT. For upregulated mRNAs, only those with an RPKM≥1 upon the loss of H3.V and/or J are included. For downregulated mRNAs, only those with an RPKM≥1 in WT cells are included.

	Upregulated Genes[1]		RPKM[3]					Fold Upregulation		Fold Upregulation			P value[6]		
	Gene	Description	WT	WT DMOG	*H3.V* KO	*H3.V* KO DMOG	WT DMOG/WT	*H3.V* KO/WT	*H3.V* KO DMOG/WT	*H3.V* KO DMOG/H3.V KO	H3.V?[4]	Base J?[5]	WT DMOG	*H3.V* KO	*H3.V* KO DMOG
H3.V KO[2]	*Tb427tmp.v2.0350*	variant surface glycoprotein (VSG, pseudogene), putative (VSG56)	0.0	1.5	1.2	3.1	N/A	N/A	N/A	2.6	Yes	Yes	1.00	1.00	0.39
	Tb427tmp.160.0220	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	4.8	4.1	N/A	N/A	N/A	0.9	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.v2.0300	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.1	1.1	0.8	N/A	N/A	N/A	0.7	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.354.0120	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.3	2.9	7.0	4.1	41.5	41.5	2.4	Yes	Yes	1.00	0.02	0.02
	Tb427tmp.160.0260	unspecified product	0.0	0.0	1.6	2.5	1.0	34.0	34.0	1.5	Yes	Yes	1.00	0.09	0.09
	Tb427tmp.244.0620	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.3	5.5	1.5	20.8	20.8	2.4	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.244.1740	variant surface glycoprotein (VSG), putative	0.7	1.2	15.6	30.5	1.8	23.8	23.8	1.9	Yes	Yes	0.11	0.00	0.00
	Tb427tmp.142.0040	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.3	4.7	1.8	21.4	21.4	2.1	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.11.0025	unspecified product	0.5	0.8	10.4	20.3	1.7	21.8	21.8	2.0	Yes	Yes	0.24	0.00	0.00
	Tb427tmp.142.0470	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.4	4.8	1.5	20.6	20.6	2.0	Yes	Yes	1.00	0.00	0.00
	Tb427.03.2540	variant surface glycoprotein (VSG)-related, putative	0.3	0.7	4.1	8.3	2.0	11.9	11.9	2.0	Yes	Yes	0.19	0.00	0.00
	Tb427.06.5700	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.0	1.2	2.6	0.0	9.6	9.6	2.2	Yes	Yes	1.00	0.48	0.48
	Tb427.03.2580	hypothetical protein, conserved	0.7	1.1	5.5	14.6	1.5	7.5	7.5	2.7	Yes	Yes	0.49	0.00	0.00
	Tb427tmp.244.0550	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.3	1.8	3.3	1.4	10.3	10.3	1.8	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.244.1410	variant surface glycoprotein (VSG), putative	0.7	0.8	8.5	10.5	1.1	12.7	12.7	1.2	Yes	Yes	0.86	0.00	0.00
	Tb427tmp.244.1480	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.1	1.2	2.5	0.7	7.3	7.3	2.1	Yes	Yes	1.00	0.00	0.00
	Tb427.05.5260	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.3	2.5	2.0	1.7	17.3	17.3	0.8	Yes	Yes	1.00	0.00	0.00
	Tb427.BES15.12	variant surface glycoprotein (VSG) (VSG 427-6)	1.1	1.1	15.1	14.8	1.0	13.4	13.4	1.0	No	No	0.98	0.00	0.00
	Tb427.BES65.13	variant surface glycoprotein (VSG) (VSG 427-3)	0.7	0.9	4.3	7.1	1.3	6.2	6.2	1.7	No	No	0.66	0.00	0.00
	Tb427.06.5260	variant surface glycoprotein (VSG), putative	0.3	0.2	2.5	3.2	0.5	7.9	7.9	1.3	Yes	Yes	1.00	0.00	0.00
	Tb427.04.5760	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.1	1.3	1.4	0.8	7.9	7.9	1.2	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.142.0130	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.1	1.1	1.5	15.8	15.8	0.5	Yes	Yes	1.00	0.00	0.01
	Tb427tmp.142.0135	expression site-associated gene (ESAG, pseudogene), putative	0.2	1.0	6.9	1.7	5.1	34.4	34.4	0.2	Yes	Yes	0.64	0.48	0.62
	Tb427.07.6520	variant surface glycoprotein (VSG, pseudogene), putative	0.5	0.6	2.5	3.9	1.1	5.1	5.1	1.5	Yes	Yes	0.86	0.00	0.00
	Tb427tmp.v2.0080	variant surface glycoprotein (VSG, pseudogene), putative	1.8	1.4	13.6	10.8	0.8	7.7	7.7	0.8	Yes	Yes	0.56	0.00	0.00
	Tb427.BES59.12	variant surface glycoprotein (VSG) (VSG 427-13)	1.1	1.1	2.2	4.3	1.0	2.0	2.0	2.0	No	No	0.98	0.02	0.00
	Tb427.BES126.15	variant surface glycoprotein (VSG) (VSG 427-11)	3.5	3.3	8.6	16.0	0.9	2.4	2.4	1.9	No	No	0.87	0.00	0.00
	Tb427.BES98.12	variant surface glycoprotein (VSG) (VSG 427-18)	2.6	2.0	6.5	11.0	0.7	2.5	2.5	1.7	No	No	0.37	0.00	0.00
	Tb427.07.1970	retrotransposon hot spot (RHS) protein, putative	0.9	1.2	2.0	3.1	1.4	2.2	2.2	1.6	Yes	Yes	0.40	0.01	0.00
	Tb427.07.2010	retrotransposon hot spot (RHS) protein, putative	2.4	3.0	4.9	7.1	1.2	2.0	2.0	1.5	Yes	Yes	0.51	0.00	0.00
	Tb427.07.1950	retrotransposon hot spot (RHS) protein, putative	1.0	1.3	2.0	2.7	1.3	2.0	2.0	1.4	Yes	Yes	0.47	0.01	0.00
	Tb427tmp.160.5350	variant surface glycoprotein (VSG)-related, putative	19.3	22.6	39.9	55.1	1.2	2.1	2.1	1.4	Yes	Yes	0.32	0.00	0.00
	Tb427.03.1480	hypothetical protein, conserved	16.2	17.9	36.3	50.1	1.1	2.2	2.2	1.4	Yes	Yes	0.72	0.00	0.00
	Tb427.07.1960	retrotransposon hot spot (RHS) protein, putative	1.5	1.9	3.1	4.1	1.2	2.0	2.0	1.4	Yes	Yes	0.59	0.00	0.00
	Tb427tmp.244.2010	expression site-associated gene 4 (ESAG4, pseudogene), putative	16.1	20.8	40.0	53.4	1.3	2.5	2.5	1.3	Yes	Yes	0.37	0.00	0.00
	Tb427tmp.211.0010	unspecified product	4.9	6.7	13.9	18.3	1.4	2.8	2.8	1.3	No	No	0.30	0.00	0.00
	Tb427.07.2000	retrotransposon hot spot (RHS) protein, putative	1.7	2.0	4.6	6.0	1.2	2.7	2.7	1.3	Yes	Yes	0.66	0.00	0.00
	Tb427tmp.02.1564	leucine-rich repeat protein (LRRP), putative	12.3	14.7	34.0	43.3	1.2	2.8	2.8	1.3	Yes	No	0.40	0.00	0.00
	Tb427.06.180	receptor-type adenylate cyclase GRESAG 4, putative	4.2	4.9	15.8	20.0	1.2	3.7	3.7	1.3	Yes	Yes	0.46	0.00	0.00
	Tb427tmp.211.0020	hypothetical protein	3.9	4.8	10.3	13.0	1.2	2.6	2.6	1.3	No	No	0.52	0.00	0.00
	Tb427.06.170	receptor-type adenylate cyclase GRESAG 4, putative	4.8	5.5	14.3	17.9	1.2	3.0	3.0	1.3	Yes	Yes	0.48	0.00	0.00
	Tb427tmp.02.1580	leucine-rich repeat protein (LRRP), putative	23.6	29.6	51.7	64.9	1.3	2.2	2.2	1.3	Yes	No	0.09	0.00	0.00
	Tb427tmp.160.0280	variant surface glycoprotein (VSG, atypical), putative	1.2	1.1	3.4	4.0	0.9	2.8	2.8	1.2	Yes	Yes	0.86	0.00	0.00
	Tb427.03.2530	expression site-associated gene (ESAG, pseudogene), putative	0.8	0.7	2.2	2.6	0.9	2.7	2.7	1.2	Yes	Yes	0.86	0.01	0.00
	Tb427tmp.244.0650	variant surface glycoprotein (VSG, pseudogene), putative	0.6	0.7	2.3	2.7	1.2	4.2	4.2	1.1	Yes	Yes	0.82	0.00	0.00
	Tb427tmp.244.2020	nucleoside transporter 1, putative	8.6	9.1	19.4	21.9	1.1	2.3	2.3	1.1	Yes	Yes	0.84	0.00	0.00
	Tb427tmp.244.1970	leucine-rich repeat protein (pseudogene), putative	5.9	10.4	12.7	14.3	1.8	2.2	2.2	1.1	Yes	Yes	0.13	0.02	0.00
	Tb427.02.960	hypothetical protein	2.7	11.6	11.7	13.0	4.2	4.3	4.3	1.1	Yes	Yes	0.14	0.14	0.12
	Tb427.06.250	hypothetical protein	0.7	1.9	1.6	1.7	2.5	2.1	2.1	1.1	Yes	Yes	0.77	0.77	0.78
	Tb427tmp.02.1565	hypothetical protein	15.6	21.3	40.2	41.1	1.4	2.6	2.6	1.0	Yes	No	0.13	0.00	0.00
	Tb427.07.6760	hypothetical protein, conserved	13.8	15.2	31.1	31.8	1.1	2.3	2.3	1.0	No	No	0.54	0.00	0.00
	Tb427.07.6750	hypothetical protein, conserved	22.0	23.6	45.1	45.9	1.1	2.0	2.0	1.0	No	No	0.71	0.00	0.00
	Tb427.07.6740	hypothetical protein, conserved	14.4	13.8	30.0	29.5	1.0	2.1	2.1	1.0	No	No	0.85	0.00	0.00
	Tb427.04.3640	protein phosphatase 1, putative	63.9	64.3	134.6	128.5	1.0	2.1	2.1	1.0	No	No	0.98	0.00	0.00
	Tb427.02.660	expression site-associated gene (ESAG, pseudogene), putative	6.4	25.8	17.8	16.3	4.0	2.8	2.8	0.9	Yes	Yes	0.61	0.78	0.78
	Tb427.07.6780	hypothetical protein	28.2	26.2	61.0	51.0	0.9	2.2	2.2	0.8	No	No	0.81	0.00	0.00
	Tb427.01.2070	expression site-associated gene (ESAG, pseudogene), putative	8.5	6.6	17.1	14.1	0.8	2.0	2.0	0.8	Yes	Yes	0.68	0.06	0.23
	Tb427.07.6670	hypothetical protein, conserved	16.6	13.8	34.5	27.9	0.8	2.1	2.1	0.8	No	No	0.10	0.00	0.00
	Tb427.04.5440	unspecified product	0.4	1.0	2.0	1.5	2.3	4.6	4.6	0.8	Yes	Yes	0.12	0.00	0.01
	Tb427.01.2060	expression site-associated gene (ESAG, pseudogene), putative	38.0	43.6	111.5	82.0	1.1	2.9	2.9	0.7	Yes	Yes	0.90	0.05	0.18
	Tb427tmp.57.0088	variant surface glycoprotein (VSG, pseudogene), putative	1.2	1.7	2.7	1.9	1.4	2.2	2.2	0.7	Yes	Yes	0.62	0.09	0.49
	Tb427tmp.211.5010	SLACS retrotransposable element (part), putative	1.6	1.6	9.0	5.4	1.1	5.7	5.7	0.6	Yes	Yes	0.90	0.00	0.00
	Tb427tmp.244.2890	SLACS reverse transcriptase, putative	2.4	2.2	7.8	4.6	0.9	3.2	3.2	0.6	Yes	Yes	0.81	0.00	0.01
	Tb427.BES56.13	variant surface glycoprotein (VSG) (VSG 427-17)	3.1	2.6	26.0	13.9	0.9	8.5	8.5	0.5	No	No	0.72	0.00	0.00
	Tb427.01.5090	expression site-associated gene (ESAG, pseudogene), putative	131.1	138.3	335.5	159.2	1.1	2.6	2.6	0.5	Yes	Yes	0.96	0.02	0.78
	Tb427tmp.211.5015	SLACS reverse transcriptase, putative	0.5	0.1	1.9	0.9	0.2	4.0	4.0	0.5	Yes	Yes	0.64	0.30	0.80
	Tb427tmp.02.5690	hypothetical protein	1.5	1.6	3.0	0.9	1.1	2.1	2.1	0.3	No	No	0.92	0.03	0.35
	Tb427.08.270	variant surface glycoprotein (VSG, pseudogene), putative	0.0	4.4	2.5	0.0	N/A	N/A	N/A	0.0	Yes	Yes	1.00	1.00	1.00
	Tb427tmp.02.0370	hypothetical protein, conserved	0.0	1.0	1.1	0.0	N/A	N/A	N/A	0.0	No	No	1.00	1.00	1.00
	Tb427tmp.11.0009	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	22.5	0.0	N/A	N/A	N/A	0.0	Yes	Yes	1.00	1.00	1.00
	Tb427tmp.14.0027	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	3.7	0.0	N/A	N/A	N/A	0.0	Yes	Yes	1.00	1.00	1.00
WT DMOG	*Tb427.03.600*	hypothetical protein	0.0	40.9	0.0	52.6	N/A	N/A	N/A		Yes	Yes	1.00	1.00	0.41
	Tb427.06.5220	variant surface glycoprotein (VSG, pseudogene), putative	0.0	2.2	0.0	0.0	N/A	N/A	N/A		Yes	Yes	1.00	1.00	1.00
	Tb427.10.4350	hypothetical protein	2.2	27.2	0.9	9.2	12.2	0.4	0.4		Yes	Yes	0.03	0.66	0.14
	Tb427.10.14060	hypothetical protein	0.4	3.9	0.4	2.7	9.7	1.1	1.1		Yes	Yes	0.48	1.00	0.48
	Tb427.08.5930	hypothetical protein	0.4	2.3	0.6	0.6	5.6	1.5	1.5		No	Yes	0.04	0.69	0.77
	Tb427tmp.244.1960	hypothetical protein	2.0	10.2	2.7	8.1	5.2	1.4	1.4		Yes	Yes	0.33	0.88	0.38
	Tb427tmp.12.0018	hypothetical protein	1.7	7.2	0.7	2.3	4.2	0.4	0.4		Yes	Yes	0.00	0.12	0.63
	Tb427.10.12570	hypothetical protein	1.9	7.6	1.5	2.4	4.0	0.8	0.8		Yes	Yes	0.02	0.81	0.82
	Tb427.05.3990	variant surface glycoprotein (VSG, atypical), putative	1.6	6.1	1.1	3.4	3.7	0.7	0.7		Yes	Yes	0.00	0.30	0.01
	Tb427.10.10270	hypothetical protein	4.0	14.3	3.1	6.5	3.6	0.8	0.8		Yes	Yes	0.00	0.68	0.26
	Tb427.10.7170	unspecified product	1.1	3.9	0.4	1.5	3.4	0.3	0.3		Yes	Yes	0.08	0.38	0.78
	Tb427.08.1660	procyclin-associated gene (pseudogene), putative	2.5	8.3	2.9	6.9	3.3	1.2	1.2		Yes	Yes	0.14	0.92	0.21
	Tb427.07.6010	hypothetical protein	1.2	3.9	0.7	2.2	3.3	0.6	0.6		Yes	Yes	0.00	0.52	0.22
	Tb427.08.5920	hypothetical protein	1.1	3.2	0.4	0.6	2.9	0.4	0.4		No	Yes	0.01	0.17	0.32
	Tb427tmp.01.2860	hypothetical protein	2.0	5.9	1.0	2.5	2.9	0.5	0.5		Yes	Yes	0.00	0.11	0.67
	Tb427tmp.01.3220	hypothetical protein	2.0	5.0	0.8	2.1	2.5	0.4	0.4		Yes	Yes	0.01	0.05	0.92
	Tb427tmp.01.3250	hypothetical protein	0.9	2.3	0.3	1.1	2.5	0.3	0.3		Yes	Yes	0.75	1.00	1.00
	Tb427.10.1220	hypothetical protein	0.8	2.1	0.8	2.4	2.5	1.0	1.0		Yes	Yes	0.16	0.98	0.09
	Tb427.10.4340	hypothetical protein	2.1	5.1	3.5	7.1	2.4	1.7	1.7		Yes	Yes	0.14	0.49	0.04
	Tb427.07.6030	hypothetical protein	1.3	3.1	0.7	2.5	2.3	0.5	0.5		Yes	Yes	0.04	0.31	0.17
	Tb427.04.5090	hypothetical protein	1.8	4.1	1.0	1.5	2.3	0.6	0.6		Yes	Yes	0.02	0.29	0.83
	Tb427.04.130	receptor-type adenylate cyclase GRESAG 4, pseudogene, putative	8.5	18.1	9.3	19.4	2.1	1.1	1.1		Yes	Yes	0.00	0.64	0.00
	Tb427tmp.v1.0970	hypothetical protein	1.6	3.3	1.8	4.8	2.1	1.1	1.1		Yes	Yes	0.08	0.86	0.01
	Tb427.06.110	hypothetical protein	28.5	57.8	19.3	49.1	2.0	0.7	0.7		Yes	Yes	0.02	0.38	0.10
	Tb427.04.140	hypothetical protein	5.1	10.4	3.5	6.4	2.0	0.7	0.7		Yes	Yes	0.15	0.64	0.79
	Tb427.02.1260	expression site-associated gene (ESAG, pseudogene), putative	14.4	28.8	15.6	29.5	2.0	1.1	1.1		Yes	Yes	0.00	0.61	0.00

***H3.V* KO DMOG**	*Tb427tmp.12.0019*	unspecified product	0.0	0.1	0.2	1.8	N/A	N/A	N/A	No	No	1.00	1.00	0.48
	Tb427tmp.55.0028	expression site-associated gene (ESAG) protein, putative	0.0	0.0	0.0	32.2	N/A	N/A	N/A	Yes	Yes	1.00	1.00	1.00
	Tb427tmp.v2.0450	hypothetical protein, conserved, frameshifted fragment	0.0	0.0	0.0	1.5	N/A	N/A	N/A	Yes	Yes	1.00	1.00	0.02
	Tb427.03.5890	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	0.6	1.0	0.0	26.3	26.3	Yes	Yes	1.00	0.56	0.56
	Tb427.06.5780	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	0.8	1.1	1.7	7.9	7.9	Yes	Yes	1.00	0.06	0.03
	Tb427tmp.244.1260	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.1	0.8	1.0	0.8	6.6	6.6	Yes	Yes	1.00	0.01	0.00
	Tb427.06.5240	variant surface glycoprotein (VSG, atypical), putative	0.2	0.1	1.0	1.4	0.7	5.3	5.3	Yes	Yes	1.00	0.00	0.00
	Tb427tmp.244.1700	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.2	0.9	1.5	1.3	5.0	5.0	Yes	Yes	1.00	0.01	0.00
	Tb427.07.6510	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.2	0.9	1.8	0.8	4.1	4.1	Yes	Yes	1.00	0.00	0.00
	Tb427.07.6530	variant surface glycoprotein (VSG, pseudogene), putative	0.3	0.4	1.0	1.7	1.3	3.1	3.1	Yes	Yes	1.00	0.01	0.00
	Tb427.07.6880	hypothetical protein, conserved	0.4	0.9	0.9	1.3	2.3	2.1	2.1	Yes	Yes	0.18	0.23	0.05
	Tb427.07.6730	hypothetical protein, conserved	17.3	16.4	34.2	36.1	0.9	2.0	2.0	No	No	0.79	0.00	0.00
	Tb427.07.6810	hypothetical protein, conserved	14.5	15.7	28.5	31.8	1.1	2.0	2.0	Yes	Yes	0.62	0.00	0.00
	Tb427.03.2590	hypothetical protein	10.2	18.0	20.1	45.6	1.8	2.0	2.0	Yes	Yes	0.00	0.00	0.00
	Tb427.07.6790	hypothetical protein, conserved	18.3	16.2	35.6	36.9	0.9	2.0	2.0	No	No	0.51	0.00	0.00
	Tb427tmp.01.5580	hypothetical protein, conserved	32.1	32.1	62.1	66.4	1.0	1.9	1.9	No	No	1.00	0.00	0.00
	Tb427.07.6800	hypothetical protein, conserved	15.4	17.6	29.6	36.8	1.1	1.9	1.9	Yes	Yes	0.57	0.00	0.00
	Tb427.07.3260	expression site-associated gene (ESAG) protein, putative	86.8	127.4	165.1	179.0	1.5	1.9	1.9	Yes	Yes	0.76	0.51	0.44
	Tb427.01.520	variant surface glycoprotein (VSG, pseudogene), putative	2.5	3.3	4.8	7.1	1.3	1.9	1.9	Yes	Yes	0.49	0.02	0.00
	Tb427.07.2020	retrotransposon hot spot (RHS) protein, putative	2.0	2.1	3.6	5.5	1.1	1.8	1.8	Yes	Yes	0.92	0.01	0.00
	Tb427tmp.244.2060	expression site-associated gene (ESAG) protein, putative	46.9	61.0	84.5	113.3	1.3	1.8	1.8	Yes	Yes	0.01	0.00	0.00
	Tb427.03.1490	leucine-rich repeat protein (LRRP), putative	9.5	9.4	17.1	22.5	1.0	1.8	1.8	Yes	Yes	0.98	0.00	0.00
	Tb427.07.1980	retrotransposon hot spot (RHS) protein, putative	1.2	1.6	2.0	3.2	1.4	1.8	1.8	Yes	Yes	0.35	0.03	0.00
	Tb427.10.90	hypothetical protein	3.6	4.4	6.3	11.9	1.2	1.8	1.8	Yes	Yes	0.78	0.26	0.01
	Tb427tmp.160.4950	hypothetical protein, conserved	11.7	16.9	20.3	24.9	1.4	1.7	1.7	Yes	Yes	0.05	0.00	0.00
	Tb427.05.4640	unspecified product	0.2	0.5	0.4	1.2	2.2	1.7	1.7	Yes	Yes	0.39	1.00	0.06
	Tb427.02.3330	hypothetical protein	8.2	14.2	13.7	22.7	1.7	1.7	1.7	Yes	No	0.03	0.05	0.00
	Tb427.03.2500	hypothetical protein	1.6	1.3	2.6	3.9	0.8	1.6	1.6	Yes	Yes	0.67	0.12	0.00
	Tb427.07.1990	retrotransposon hot spot protein (RHS, pseudogene), putative	1.3	1.4	2.2	3.4	1.1	1.6	1.6	Yes	Yes	0.91	0.06	0.00
	Tb427.03.590	adenosine transporter, putative	7.9	10.3	12.6	17.1	1.3	1.6	1.6	Yes	Yes	0.18	0.00	0.00
	Tb427.07.6500	variant surface glycoprotein (VSG), putative	2.0	2.6	3.1	4.5	1.3	1.6	1.6	Yes	Yes	0.43	0.09	0.00
	Tb427tmp.244.1950	hypothetical protein	0.3	0.9	0.5	1.3	2.6	1.6	1.6	Yes	Yes	0.54	0.81	0.37
	Tb427tmp.01.3915	RNA-binding protein, putative (RBP5)	118.1	108.4	173.8	249.2	0.9	1.5	1.5	No	Yes	0.70	0.00	0.00
	Tb427.10.8500	glucose transporter, putative	80.5	127.0	111.5	164.8	1.6	1.4	1.4	No	No	0.00	0.00	0.00
	Tb427tmp.142.0300	hypothetical protein (pseudogene)	0.9	1.5	1.3	2.3	1.7	1.4	1.4	Yes	Yes	0.32	0.63	0.04
	Tb427.10.8460	glucose transporter, putative	168.8	291.0	227.8	369.0	1.7	1.3	1.3	No	No	0.00	0.01	0.00
	Tb427tmp.01.3240	trans-sialidase, putative	15.9	20.4	21.3	38.2	1.3	1.3	1.3	Yes	Yes	0.04	0.01	0.00
	Tb427.BES129.14	variant surface glycoprotein (VSG) (VSG 427-9)	2.7	2.1	3.6	6.3	0.8	1.3	1.3	No	No	0.51	0.33	0.00
	Tb427.03.510	hypothetical protein, conserved	0.7	1.2	1.0	2.6	1.6	1.3	1.3	Yes	Yes	0.44	0.71	0.00
	Tb427.10.8470	glucose transporter, putative	148.4	256.4	193.1	317.9	1.7	1.3	1.3	No	No	0.00	0.02	0.00
	Tb427.02.720	retrotransposon hot spot protein (RHS, pseudogene), putative	13.3	23.6	16.8	33.4	1.8	1.3	1.3	Yes	Yes	0.01	0.46	0.00
	Tb427.05.160	retrotransposon hot spot protein (RHS, pseudogene), putative	9.1	16.7	11.5	23.5	1.8	1.3	1.3	Yes	Yes	0.00	0.46	0.00
	Tb427.10.8440	glucose transporter 1B (THT1-)	157.5	262.3	194.8	321.3	1.7	1.2	1.2	No	No	0.00	0.09	0.00
	Tb427.02.5384	retrotransposon hot spot protein (RHS, pseudogene), putative	1.6	3.1	2.0	7.2	1.9	1.2	1.2	Yes	Yes	0.02	0.64	0.00
	Tb427.10.8480	glucose transporter, putative	155.4	270.0	190.4	313.0	1.7	1.2	1.2	No	No	0.00	0.11	0.00
	Tb427.05.170	hypothetical protein, conserved	6.7	12.9	8.2	16.4	1.9	1.2	1.2	Yes	Yes	0.03	0.68	0.00
	Tb427.07.1930	nucleoside diphosphatase, putative	13.3	15.9	16.1	26.7	1.2	1.2	1.2	Yes	Yes	0.34	0.30	0.00
	Tb427.10.4370	expression site-associated gene (ESAG) protein, putative	7.1	11.5	8.4	16.1	1.6	1.2	1.2	Yes	Yes	0.00	0.51	0.00
	Tb427.10.16100	peptidylprolyl isomerase-like protein, putative	243.0	185.3	275.7	748.3	0.8	1.1	1.1	No	No	0.02	0.45	0.00
	Tb427.10.12270	hypothetical protein, conserved	11.1	17.3	12.1	30.5	1.6	1.1	1.1	Yes	Yes	0.02	0.80	0.00
	Tb427tmp.57.0077	unspecified product	8.0	13.3	8.5	20.7	1.7	1.1	1.1	Yes	Yes	0.48	0.96	0.11
	Tb427tmp.01.7120	hypothetical protein, conserved	188.2	155.8	187.5	435.0	0.8	1.0	1.0	No	No	0.08	0.99	0.00
	Tb427.06.1300	hypothetical protein	2.7	3.8	2.5	6.4	1.4	0.9	0.9	Yes	Yes	0.56	0.95	0.05
	Tb427tmp.160.3400	glucose transporter (pseudogene), putative	1.3	2.2	0.9	3.3	1.7	0.7	0.7	Yes	Yes	0.07	0.37	0.00
	Tb427.10.12280	hypothetical protein	2.1	3.5	1.3	5.2	1.6	0.6	0.6	Yes	Yes	0.40	0.50	0.05
	Tb427.10.1760	hypothetical protein	2.8	3.4	1.3	5.6	1.2	0.5	0.5	Yes	Yes	0.74	0.14	0.05

	Downregulated Genes[7]		**RPKM**				**Fold Downregulation**					**P value**[6]		
	Gene	**Gene Description**	**WT**	**WT DMOG**	**H3.V KO**	**H3.V KO DMOG**	**WT/WT DMOG**	**WT/H3.V KO**	**WT/H3.V KO DMOG**	**H3.V?**[5]	**Base J?**[6]	**WT DMOG**	***H3.V* KO**	***H3.V* KO DMOG**
***H3.V* KO**	*Tb427.01.2780*	hypothetical protein	1.3	0.3	0.0	0.6	4.8	N/A	2.0	Yes	Yes	0.51	0.06	0.74
	Tb427.03.450	variant surface glycoprotein (VSG, pseudogene), putative	1.1	2.1	0.0	0.1	0.5	N/A	18.8	Yes	Yes	0.30	0.00	0.47
	Tb427.10.15350	H3V (h3vaR)	133.8	123.1	0.0	0.1	1.1	N/A	1672.0	No	No	0.67	0.00	0.47
	Tb427tmp.354.0100	variant surface glycoprotein (VSG, pseudogene), putative	1.1	1.0	0.0	0.0	1.1	N/A	N/A	Yes	Yes	0.95	0.00	0.00
	Tb427.01.5280	expression site-associated gene (ESAG, pseudogene), putative	10.4	8.7	0.1	0.1	1.2	74.5	208.6	Yes	Yes	0.56	0.01	0.58
	Tb427.01.5320	variant surface glycoprotein (VSG, pseudogene), putative	5.5	6.7	0.0	0.0	0.8	137.5	275.0	Yes	Yes	0.43	0.22	0.58
	Tb427tmp.24.0010	variant surface glycoprotein (VSG, pseudogene), putative	11.7	9.7	0.2	0.7	1.2	61.8	18.1	Yes	Yes	0.66	0.53	0.00
	Tb427.10.7160	procyclin-associated gene 1 (PAG1) protein, putative	12.0	4.3	0.8	1.5	2.8	15.0	8.2	Yes	Yes	0.00	0.00	0.00
	Tb427.10.10210	procyclin-associated gene 4 (PAG4) protein (PAG4)	25.4	8.6	2.1	2.6	3.0	11.9	9.8	Yes	Yes	0.00	0.00	0.00
	Tb427.10.10220	procyclin-associated gene 2 (PAG2) protein (PAG2)	41.9	12.9	3.7	3.0	3.2	11.3	14.0	Yes	Yes	0.00	0.00	0.00
	Tb427.10.10230	procyclin-associated gene 5 (PAG5) protein (PAG5)	31.5	8.0	3.0	2.6	3.9	10.6	12.3	Yes	Yes	0.00	0.00	0.00
	Tb427.10.10240	procyclin-associated gene 1 (PAG1) protein (PAG1)	22.3	7.4	2.1	2.7	3.0	10.8	8.1	Yes	Yes	0.00	0.00	0.00
	Tb427.06.520	EP3-2 procyclin	192.2	108.8	27.7	22.8	1.8	6.9	8.4	Yes	Yes	0.00	0.00	0.00
	Tb427.10.7150	hypothetical protein	4.3	0.0	0.6	0.0	N/A	7.0	N/A	Yes	Yes	0.14	0.46	0.14
	Tb427.10.10250	EP2 procyclin (EP2)	160.7	88.3	24.8	21.9	1.8	6.5	7.3	Yes	Yes	0.00	0.00	0.00
	Tb427tmp.160.0130	expression site-associated gene 3 (ESAG3, pseudogene), putative	1.4	1.6	0.2	0.2	0.9	6.8	9.5	Yes	Yes	0.88	0.00	0.01
	Tb427.10.10260	EP1 procyclin (EP1)	323.8	215.6	64.1	63.2	1.5	5.0	5.1	Yes	Yes	0.00	0.00	0.00
	Tb427.06.210	leucine-rich repeat protein (LRRP, pseudogene), putative	7.6	7.5	1.6	3.9	1.0	4.7	1.9	Yes	Yes	0.95	0.00	0.00
	Tb427.06.480	unspecified product	119.1	89.6	28.4	25.5	1.3	4.2	4.7	No	Yes	0.06	0.00	0.00
	Tb427.06.510	GPEET2 procyclin precursor	545.7	307.1	149.6	176.5	1.8	3.6	3.1	Yes	Yes	0.00	0.00	0.00
	Tb427tmp.211.0865	60S ribosomal protein L37, putative	1312.4	627.2	396.2	502.3	2.1	3.3	2.6	No	No	0.00	0.00	0.00
	Tb427.08.480	phosphatidic acid phosphatase protein, putative	82.4	54.8	25.6	50.1	1.5	3.2	1.6	Yes	Yes	0.00	0.00	0.00
	Tb427.06.450	unspecified product	98.3	62.1	32.3	38.5	1.6	3.0	2.6	No	Yes	0.00	0.00	0.00
	Tb427.08.510	hypothetical protein	5.0	5.8	1.7	4.4	0.9	3.0	1.2	Yes	Yes	0.71	0.00	0.74
	Tb427.10.7170	unspecified product	1.1	3.9	0.4	1.5	0.3	2.9	0.7	Yes	Yes	0.08	0.38	0.78
	Tb427tmp.01.3220	hypothetical protein	2.0	5.0	0.8	2.1	0.4	2.6	0.9	Yes	Yes	0.01	0.05	0.92
	Tb427.08.500	hypothetical protein	21.3	17.5	8.3	15.8	1.2	2.6	1.3	Yes	Yes	0.50	0.00	0.22
	Tb427.08.5920	hypothetical protein	1.1	3.2	0.4	0.6	0.3	2.5	1.9	No	Yes	0.01	0.17	0.32
	Tb427.10.4350	hypothetical protein	2.2	27.2	0.9	9.2	0.1	2.5	0.2	Yes	Yes	0.03	0.66	0.14
	Tb427tmp.02.0500	60S ribosomal protein L37, putative	2830.6	1845.8	1120.0	1428.6	1.5	2.5	2.0	No	No	0.00	0.00	0.00
	Tb427tmp.12.0018	hypothetical protein	1.7	7.2	0.7	2.3	0.2	2.5	0.8	Yes	Yes	0.00	0.12	0.63
	Tb427.08.490	hypothetical protein	91.1	66.4	38.3	45.1	1.4	2.4	2.0	Yes	Yes	0.00	0.00	0.00
	Tb427.08.520	hypothetical protein	20.3	15.4	8.7	16.1	1.3	2.3	1.3	Yes	Yes	0.29	0.00	0.41
	Tb427.10.5720	hypothetical protein	1.1	2.2	0.5	1.7	0.5	2.3	0.7	Yes	Yes	0.21	0.31	0.51
	Tb427.10.1760	hypothetical protein	2.8	3.4	1.3	5.6	0.8	2.2	0.5	Yes	Yes	0.74	0.14	0.05
	Tb427.02.1290	hypothetical protein	12.7	15.5	6.0	15.9	0.8	2.1	0.8	Yes	Yes	0.74	0.09	0.69
WT DMOG	*Tb427tmp.v2.0190*	expression site-associated gene 3 (ESAG3, pseudogene), putative	2.1	0.0	2.0	0.0	N/A	1.1	N/A	Yes	Yes	1.00	1.00	1.00
	Tb427.04.2090	hypothetical protein	1.5	0.5	1.1	0.9	3.3	1.3	1.7	No	Yes	0.11	0.69	0.40
	Tb427.10.3050	hypothetical protein	2.3	0.7	2.0	1.6	3.3	1.2	1.5	Yes	Yes	0.06	0.83	0.49
	Tb427.01.275	expression site-associated gene (ESAG, pseudogene), putative	24.0	10.2	29.6	37.3	2.3	0.8	0.6	Yes	Yes	0.78	0.84	0.80
	Tb427tmp.211.1800	hypothetical protein, conserved	390.0	180.5	282.2	153.3	2.2	1.4	2.5	No	No	0.00	0.03	0.00
***H3.V* KO DMOG**	*Tb427.10.3410*	hypothetical protein	6.8	5.0	6.5	2.4	1.3	1.0	2.8	No	No	0.63	0.96	0.05
	Tb427tmp.01.4390	leucine-rich repeat protein (LRRP), putative	8.3	7.4	7.4	3.2	1.1	1.1	2.6	No	No	0.91	0.91	0.62
	Tb427tmp.02.4710	hypothetical protein	20.2	17.1	12.1	7.9	1.2	1.7	2.6	No	No	0.56	0.02	0.00
	Tb427.06.1790	hypothetical protein	6.5	5.2	4.4	2.6	1.2	1.5	2.5	No	No	0.62	0.30	0.01
	Tb427tmp.02.4850	hypothetical protein, conserved	383.9	362.2	273.4	157.8	1.1	1.4	2.4	No	No	0.93	0.45	0.02
	Tb427.10.3400	hypothetical protein, conserved	7.2	4.6	5.6	3.2	1.6	1.3	2.3	No	No	0.04	0.31	0.00
	Tb427.10.3420	NADH-ubiquinone oxidoreductase c, putative	2.9	3.1	3.3	1.3	0.9	0.9	2.3	No	No	0.91	0.84	0.14
	Tb427.08.4300	hypothetical protein	252.9	172.6	193.3	114.7	1.5	1.3	2.2	No	No	0.23	0.46	0.01
	Tb427.10.12770	unspecified product	4.3	3.7	3.9	2.0	1.1	1.1	2.2	No	No	0.86	0.91	0.16
	Tb427tmp.01.5820	hypothetical protein, conserved	21.7	15.7	21.8	10.0	1.4	1.0	2.2	No	No	0.01	0.98	0.00
	Tb427.08.4820	eukaryotic translation initation factor 4 gamma, putative	166.8	101.7	123.5	78.2	1.6	1.4	2.1	No	No	0.00	0.01	0.00
	Tb427tmp.160.5510	unspecified product	83.5	68.4	67.6	39.4	1.2	1.2	2.1	No	No	0.23	0.19	0.00
	Tb427.04.2310	asparaginyl-tRNA synthetase, putative	249.7	155.7	205.5	116.4	1.6	1.2	2.1	No	No	0.00	0.28	0.00
	Tb427.04.4870	amino acid transporter, putative	428.4	298.1	323.4	207.2	1.4	1.3	2.1	No	No	0.01	0.06	0.00
	Tb427.10.12660	pumilio/PUF RNA binding protein 2 (PUF2)	9.0	6.7	8.1	4.4	1.4	1.1	2.1	No	No	0.05	0.59	0.00
	Tb427tmp.42.0003	T-complex protein 1, beta subunit, putative	639.8	470.4	420.8	311.6	1.4	1.5	2.1	No	Yes	0.09	0.01	0.00
	Tb427tmp.160.3820	nucleolar RNA binding protein, putative	222.6	150.5	161.6	109.9	1.5	1.4	2.0	No	No	0.00	0.01	0.00
	Tb427.10.2370	RNA-binding protein, putative (LA)	372.4	239.9	286.6	183.9	1.6	1.3	2.0	Yes	Yes	0.00	0.03	0.00
	Tb427.05.1480	hypothetical protein	9.6	8.9	7.6	4.7	1.1	1.3	2.1	No	Yes	0.87	0.53	0.03
	Tb427.04.1910	hypothetical protein, conserved	32.0	24.0	28.0	15.6	1.3	1.1	2.1	No	No	0.08	0.55	0.00
	Tb427tmp.02.1120	adenylosuccinate synthetase, putative	312.1	247.8	206.3	156.6	1.3	1.5	2.0	Yes	No	0.17	0.00	0.00
	Tb427tmp.160.4240	nucleosome assembly protein-like protein	290.4	197.9	214.2	146.4	1.5	1.4	2.0	No	No	0.00	0.01	0.00
	Tb427.08.730	nucleolar RNA-binding protein, putative	866.1	591.9	648.9	429.9	1.5	1.3	2.0	No	No	0.00	0.06	0.00
	Tb427.10.6060	universal minicircle sequence binding protein (UMSBP), putative	675.4	546.3	514.5	341.0	1.2	1.3	2.0	No	No	0.14	0.03	0.00
	Tb427tmp.01.3940	RNA-binding protein, putative (RBP9)	114.8	84.5	93.7	57.9	1.4	1.2	2.0	No	No	0.00	0.06	0.00

[1]Fold upregulation calculated by dividing the *H3.V* KO and/or DMOG treated RPKM by WT RPKM

[2]Genes upregulated in the *H3.V* KO are listed in the same order as the heatmap shown in Fig. 5A

[3]The average RPKM value of triplicate mRNA-seq libraries is listed

[4]Yes: if gene is located within 10 kb of H3.V

[5]Yes: if gene is located within 10 kb of base J

[6]P value determined by Cuffdiff

[7]Fold downregulation calculated by dividing WT RPKM by the *H3.V* KO and/or DMOG treated RPKM

Table 4.S2 Small RNA-seq RPM and statistical significance at cSSRs. The average small RNA-seq RPM of triplicate libraries at cSSRs is listed. Chromosome number and the 5' and 3' position of regions quantified are included. Statistical significance was assessed using pairwise Fisher's Exact test on both total reads and specifically 21-27bp reads. Yellow highlight indicates significance at a p value ≤ 0.05.

			Small RNA RPM[1]		Total Reads[2]	21-27bp Reads[3]
Chromosome	5' end	3' end	WT	H3.V KO	P Value	P Value
1	120,619	136,017	3658	9165	2.53E-01	8.93E-04
1	159,590	174,513	6441	14884	9.38E-01	6.19E-01
1	190,101	204,264	3258	7502	9.77E-01	4.62E-01
1	517,884	520,878	4	10	4.76E-01	1.59E-01
1	636,387	642,042	69	148	1.74E-01	7.02E-02
1	723,975	731,389	7	20	3.33E-03	6.99E-04
1	993,994	999,602	4	9	3.07E-01	5.62E-01
2	20,587	24,801	701	1615	9.69E-01	6.24E-01
2	132,165	145,338	464	968	2.39E-02	2.77E-04
2	166,620	174,460	106	240	7.05E-01	9.55E-01
2	355,961	367,695	3279	7561	9.59E-01	5.27E-01
2	378,789	386,362	50	126	2.71E-01	8.84E-02
2	616,771	622,424	3	12	4.77E-05	6.18E-04
2	958,385	967,079	9	32	8.37E-06	2.85E-07
3	149,748	153,368	2	8	1.50E-03	2.53E-04
3	381,684	389,146	2	5	9.38E-01	3.26E-01
3	1,122,114	1,136,077	15	39	1.56E-01	1.72E-01
4	24,266	38,655	154	361	9.36E-01	8.19E-05
4	320,899	323,164	58	110	9.19E-03	2.66E-03
4	978,180	983,793	16	29	1.05E-02	1.78E-01
5	27,660	35,994	107	239	5.65E-01	8.97E-01
5	68,899	71,485	8	30	4.73E-06	5.38E-06
5	94,691	104,893	16	29	8.47E-03	5.93E-03
5	125,081	131,045	16	31	4.36E-02	5.38E-01
5	441,989	458,585	33	57	2.38E-04	1.69E-01
5	632,018	645,237	333	619	2.43E-05	1.41E-11
5	964,085	968,719	4	9	7.88E-01	6.16E-01
5	1,232,496	1,234,311	5	9	1.51E-01	9.53E-01
6	16,199	28,277	123	342	2.38E-03	2.59E-06
6	90,293	98,790	114	250	2.39E-01	2.99E-02
6	114,773	121,065	119	257	1.36E-01	1.22E-02
6	167,386	179,608	45	103	6.62E-01	2.16E-01
6	339,037	344,823	17	34	3.30E-02	2.45E-01
6	987,390	992,887	3010	6927	9.75E-01	4.92E-01
6	995,924	1,008,942	5776	13374	9.12E-01	7.38E-01
7	54,845	61,519	13	31	1.00E+00	2.87E-01
7	461,538	514,339	60	133	3.58E-01	8.49E-01
7	831,874	841,946	8	18	5.34E-01	1.00E+00
7	1,126,324	1,135,240	569	1440	3.56E-02	1.43E-03
7	1,242,271	1,248,969	3	6	3.44E-01	5.98E-01
7	1,785,211	1,789,989	2517	5813	9.23E-01	6.10E-01
7	1,962,135	1,964,696	227	398	5.25E-04	5.00E-02

8	551,137	558,187	110	209	3.43E-03	5.89E-06
8	862,870	863,473	108	205	4.77E-03	1.49E-05
8	1,052,049	1,055,429	6	16	1.17E-01	3.60E-02
8	1,586,151	1,588,699	5	10	1.55E-01	8.92E-01
8	1,892,235	1,893,385	149	232	2.38E-04	6.18E-07
8	1,894,715	1,896,190	573	1326	8.91E-01	7.97E-01
8	2,090,232	2,097,712	11	27	8.08E-01	5.59E-01
8	2,233,445	2,235,800	3648	8564	9.25E-01	4.86E-01
8	2,236,077	2,236,492	3	5	1.46E-02	3.50E-03
9	823,112	833,729	11	25	9.55E-01	2.90E-01
9	1,126,658	1,135,397	1	8	8.32E-05	3.39E-06
9	2,446,192	2,453,293	114	345	8.90E-06	1.03E-12
10	615,666	623,300	88	177	3.33E-02	1.49E-02
10	937,355	943,346	6	13	4.77E-01	8.77E-01
10	1,127,056	1,128,837	8	24	8.73E-03	1.11E-05
10	1,422,953	1,427,810	5	12	5.74E-01	2.89E-01
10	2,980,784	2,987,816	5	13	4.16E-01	3.47E-01
10	3,177,865	3,186,377	12	25	1.77E-01	8.78E-01
11	443,663	454,995	3	9	2.59E-02	1.79E-03
11	669,438	679,992	3229	7353	8.76E-01	1.88E-01
11	1,180,869	1,185,840	1	5	2.27E-03	1.23E-03
11	1,342,791	1,347,652	9	20	4.46E-01	6.49E-01
11	1,651,056	1,661,888	3	10	1.32E-02	3.22E-03
11	2,085,053	2,092,052	5	13	9.61E-01	7.24E-02
11	2,517,606	2,523,771	7	14	1.86E-01	1.35E-01
11	2,985,656	2,997,876	6	14	8.61E-01	2.14E-01
11	3,133,413	3,152,201	16	33	3.87E-01	9.25E-01
11	3,333,649	3,344,869	19	39	1.26E-01	1.02E-01
11	3,826,110	3,835,331	11	34	3.30E-03	3.71E-04
11	4,222,111	4,229,999	42	69	7.35E-08	2.87E-08

[1]RPM represents the average of triplicate libraries (total mapped small RNA-seq reads)

[2]Pairwise significance tested using Fisher's Exact test on total mapped small RNA-seq reads

[3]Pairwise significance tested using Fisher's Exact test on 21-27bp mapped small RNA-seq reads

Table 4.S3 High-throughput sequencing information. Information about all sequencing experiments performed in this study is listed. Also indicates the figures in which the data are presented.

Species	Genotype	Treatment	Library #	Type of RNA sequenced	Genome used for alignment	Minimum read length considered for alignment (nt)	Total reads (millions)	Overall (unique and non unique) alignment rate %	Sequenced by	Data shown in
T. brucei	WT	DMSO	1	small RNA	T. brucei 427 v6.0	18	20.71	92.36	Vertis Biotechnology	Fig. 1A, 2D, S1B, S2
T. brucei	H3.V KO	DMSO	2	small RNA	T. brucei 427 v6.0	18	32.93	84.38	Vertis Biotechnology	Fig. 1A, 2D, S2
T. brucei	H3.V KO	DMOG	3	small RNA	T. brucei 427 v6.0	18	37.36	86.16	Vertis Biotechnology	Fig. 1A
T. brucei	WT	DMSO	4	small RNA	T. brucei 427 v6.0	18	7.53	82.68	Georgia Genomics Facility	Fig. 1B, 1C, S2 Table
T. brucei	WT	DMSO	5	small RNA	T. brucei 427 v6.0	18	9.34	80.52	Georgia Genomics Facility	Fig. 1B, 1C, S1A, S2 Table
T. brucei	WT	DMSO	6	small RNA	T. brucei 427 v6.0	18	6.70	83.76	Georgia Genomics Facility	Fig. 1B, 1C, S2 Table
T. brucei	WT	DMOG	7	small RNA	T. brucei 427 v6.0	18	7.33	82.39	Georgia Genomics Facility	not shown
T. brucei	WT	DMOG	8	small RNA	T. brucei 427 v6.0	18	9.87	82.65	Georgia Genomics Facility	not shown
T. brucei	WT	DMOG	9	small RNA	T. brucei 427 v6.0	18	13.24	81.19	Georgia Genomics Facility	not shown
T. brucei	H3.V KO	DMSO	10	small RNA	T. brucei 427 v6.0	18	14.07	82.02	Georgia Genomics Facility	Fig. 1B, 1C, S1A, S2 Table
T. brucei	H3.V KO	DMSO	11	small RNA	T. brucei 427 v6.0	18	6.51	80.08	Georgia Genomics Facility	Fig. 1B, 1C, S2 Table
T. brucei	H3.V KO	DMSO	12	small RNA	T. brucei 427 v6.0	18	8.77	81.69	Georgia Genomics Facility	Fig. 1B, 1C, S2 Table
T. brucei	H3.V KO	DMOG	13	small RNA	T. brucei 427 v6.0	18	6.45	84.21	Georgia Genomics Facility	not shown
T. brucei	H3.V KO	DMOG	14	small RNA	T. brucei 427 v6.0	18	8.25	81.64	Georgia Genomics Facility	not shown
T. brucei	H3.V KO	DMOG	15	small RNA	T. brucei 427 v6.0	18	6.87	84.66	Georgia Genomics Facility	not shown
T. brucei	WT	DMSO	16	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	57.72	98.60	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	WT	DMSO	17	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	31.22	98.75	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	WT	DMSO	18	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	36.30	98.83	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	WT	DMOG	19	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	48.54	98.60	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	WT	DMOG	20	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	34.62	98.81	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	WT	DMOG	21	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	26.66	98.59	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMSO	22	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	59.22	98.54	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMSO	23	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	29.88	98.48	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMSO	24	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	30.84	98.58	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMOG	25	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	60.88	98.85	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMOG	26	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	29.06	98.46	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4
T. brucei	H3.V KO	DMOG	27	polyA enriched RNA (mRNA)	T. brucei 427 v9.0	50	34.40	98.62	Georgia Genomics Facility	Fig. 4, 5A, 5E, 6B, S4, S5, Table 1 and 4

Table 4.S4 *T. brucei* upregulated genes following H3.V and/or J loss. Similar to 4.S1 Table, but upregulated genes are organized according to their location within PTUs. Genes sharing the same number are located in the same PTU and genes with different numbers are located in different PTUs.

	Upregulated Genes[1]		RPKM				Fold Upregulation				
PTU[2]	Gene	Description	WT	WT DMOG	H3.V KO	H3.V KO DMOG	WT DMOG/WT	H3.V KO/WT	H3.V KO DMOG/WT	H3.V?[3]	Base J?[4]
1	Tb427.01.520	variant surface glycoprotein (VSG, pseudogene), putative	2.5	3.3	4.8	7.1	1.3	1.9	1.9	Yes	Yes
2	Tb427.01.2060	expression site-associated gene (ESAG, pseudogene), putative	38.0	43.6	111.5	82.0	1.1	2.9	2.9	Yes	Yes
2	Tb427.01.2070	expression site-associated gene (ESAG, pseudogene), putative	8.5	6.6	17.1	14.1	0.8	2.0	2.0	Yes	Yes
3	Tb427.01.5090	expression site-associated gene (ESAG, pseudogene), putative	131.1	138.3	335.5	159.2	1.1	2.6	2.6	Yes	Yes
4	Tb427.02.660	expression site-associated gene (ESAG, pseudogene), putative	6.4	25.8	17.8	16.3	4.0	2.8	2.8	Yes	Yes
4	Tb427.02.720	retrotransposon hot spot protein (RHS, pseudogene), putative	13.3	23.6	16.8	33.4	1.8	1.3	1.3	Yes	Yes
5	Tb427.02.960	hypothetical protein	2.7	11.6	11.7	13.0	4.2	4.3	4.3	Yes	Yes
6	Tb427.02.1260	expression site-associated gene (ESAG, pseudogene), putative	14.4	28.8	15.6	29.5	2.0	1.1	1.1	Yes	Yes
7	Tb427.02.3330	hypothetical protein	8.2	14.2	13.7	22.7	1.7	1.7	1.7	Yes	No
8	Tb427.02.5384	retrotransposon hot spot protein (RHS, pseudogene), putative	1.6	3.1	2.0	7.2	1.9	1.2	1.2	Yes	Yes
9	Tb427.03.510	hypothetical protein, conserved	0.7	1.2	1.0	2.6	1.6	1.3	1.3	Yes	Yes
10	Tb427.03.590	adenosine transporter, putative	7.9	10.3	12.6	17.1	1.3	1.6	1.6	Yes	Yes
11	Tb427.03.600	hypothetical protein	0.0	40.9	0.0	52.6	N/A	N/A	N/A	Yes	Yes
12	Tb427.03.1480	hypothetical protein, conserved	16.2	17.9	36.3	50.1	1.1	2.2	2.2	Yes	Yes
12	Tb427.03.1490	leucine-rich repeat protein (LRRP), putative	9.5	9.4	17.1	22.5	1.0	1.8	1.8	Yes	Yes
13	Tb427.03.2500	hypothetical protein	1.6	1.3	2.6	3.9	0.8	1.6	1.6	Yes	Yes
13	Tb427.03.2530	expression site-associated gene (ESAG, pseudogene), putative	0.8	0.7	2.2	2.6	0.9	2.7	2.7	Yes	Yes
13	Tb427.03.2540	variant surface glycoprotein (VSG)-related, putative	0.3	0.7	4.1	8.3	2.0	11.9	11.9	Yes	Yes
13	Tb427.03.2580	hypothetical protein, conserved	0.7	1.1	5.5	14.6	1.5	7.5	7.5	Yes	Yes
13	Tb427.03.2590	hypothetical protein	10.2	18.0	20.1	45.6	1.8	2.0	2.0	Yes	Yes
14	Tb427.03.5890	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	0.6	1.0	0.0	26.3	26.3	Yes	Yes
15	Tb427.04.130	receptor-type adenylate cyclase GRESAG 4, pseudogene, putative	8.5	18.1	9.3	19.4	2.1	1.1	1.1	Yes	Yes
15	Tb427.04.140	hypothetical protein	5.1	10.4	3.5	6.4	2.0	0.7	0.7	Yes	Yes
16	Tb427.04.3640	protein phosphatase 1, putative	63.9	64.3	134.6	128.5	1.0	2.1	2.1	No	No
17	Tb427.04.5090	hypothetical protein	1.8	4.1	1.0	1.5	2.3	0.6	0.6	Yes	Yes
18	Tb427.04.5440	unspecified product	0.4	1.0	2.0	1.5	2.3	4.6	4.6	Yes	Yes
18	Tb427.04.5760	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.1	1.3	1.4	0.8	7.9	7.9	Yes	Yes
19	Tb427.05.160	retrotransposon hot spot protein (RHS, pseudogene), putative	9.1	16.7	11.5	23.5	1.8	1.3	1.3	Yes	Yes
20	Tb427.05.170	hypothetical protein, conserved	6.7	12.9	8.2	16.4	1.9	1.2	1.2	Yes	Yes
21	Tb427.05.3990	variant surface glycoprotein (VSG, atypical), putative	1.6	6.1	1.1	3.4	3.7	0.7	0.7	Yes	Yes
22	Tb427.05.4640	unspecified product	0.2	0.5	0.4	1.2	2.2	1.7	1.7	Yes	Yes
23	Tb427.05.5260	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.3	2.5	2.0	1.7	17.3	17.3	Yes	Yes
24	Tb427.06.110	hypothetical protein	28.5	57.8	19.3	49.1	2.0	0.7	0.7	Yes	Yes
25	Tb427.06.170	receptor-type adenylate cyclase GRESAG 4, putative	4.8	5.5	14.3	17.9	1.2	3.0	3.0	Yes	Yes
26	Tb427.06.180	receptor-type adenylate cyclase GRESAG 4, putative	4.2	4.9	15.8	20.0	1.2	3.7	3.7	Yes	Yes
27	Tb427.06.250	hypothetical protein	0.7	1.9	1.6	1.7	2.5	2.1	2.1	Yes	Yes
28	Tb427.06.1300	hypothetical protein	2.7	3.8	2.5	6.4	1.4	0.9	0.9	Yes	Yes
29	Tb427.06.5220	variant surface glycoprotein (VSG, pseudogene), putative	0.0	2.2	0.0	0.0	N/A	N/A	N/A	Yes	Yes
29	Tb427.06.5240	variant surface glycoprotein (VSG, atypical), putative	0.2	0.1	1.0	1.4	0.7	5.3	5.3	Yes	Yes
29	Tb427.06.5260	variant surface glycoprotein (VSG), putative	0.3	0.2	2.5	3.2	0.5	7.9	7.9	Yes	Yes
29	Tb427.06.5700	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.0	1.2	2.6	0.0	9.6	9.6	Yes	Yes
29	Tb427.06.5780	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	0.8	1.1	1.7	7.9	7.9	Yes	Yes
30	Tb427.07.1930	nucleoside diphosphatase, putative	13.3	15.9	16.1	26.7	1.2	1.2	1.2	Yes	Yes
31	Tb427.07.1950	retrotransposon hot spot (RHS) protein, putative	1.0	1.3	2.0	2.7	1.3	2.0	2.0	Yes	Yes
31	Tb427.07.1960	retrotransposon hot spot (RHS) protein, putative	1.5	1.9	3.1	4.1	1.2	2.0	2.0	Yes	Yes
31	Tb427.07.1970	retrotransposon hot spot (RHS) protein, putative	0.9	1.2	2.0	3.1	1.4	2.2	2.2	Yes	Yes
31	Tb427.07.1980	retrotransposon hot spot (RHS) protein, putative	1.2	1.6	2.0	3.2	1.4	1.8	1.8	Yes	Yes
31	Tb427.07.1990	retrotransposon hot spot protein (RHS, pseudogene), putative	1.3	1.4	2.2	3.4	1.1	1.6	1.6	Yes	Yes
31	Tb427.07.2000	retrotransposon hot spot (RHS) protein, putative	1.7	2.0	4.6	6.0	1.2	2.7	2.7	Yes	Yes
31	Tb427.07.2010	retrotransposon hot spot (RHS) protein, putative	2.4	3.0	4.9	7.1	1.2	2.0	2.0	Yes	Yes
31	Tb427.07.2020	retrotransposon hot spot (RHS) protein, putative	2.0	2.1	3.6	5.5	1.1	1.8	1.8	Yes	Yes
32	Tb427.07.3260	expression site-associated gene (ESAG) protein, putative	86.8	127.4	165.1	179.0	1.5	1.9	1.9	Yes	Yes
33	Tb427.07.6010	hypothetical protein	1.2	3.9	0.7	2.2	3.3	0.6	0.6	Yes	Yes
34	Tb427.07.6030	hypothetical protein	1.3	3.1	0.7	2.5	2.3	0.5	0.5	Yes	Yes
35	Tb427.07.6500	variant surface glycoprotein (VSG), putative	2.0	2.6	3.1	4.5	1.3	1.6	1.6	Yes	Yes
35	Tb427.07.6510	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.2	0.9	1.8	0.8	4.1	4.1	Yes	Yes
35	Tb427.07.6520	variant surface glycoprotein (VSG, pseudogene), putative	0.5	0.6	2.5	3.9	1.1	5.1	5.1	Yes	Yes
35	Tb427.07.6530	variant surface glycoprotein (VSG, pseudogene), putative	0.3	0.4	1.0	1.7	1.3	3.1	3.1	Yes	Yes
36	Tb427.07.6670	hypothetical protein, conserved	16.6	13.8	34.5	27.9	0.8	2.1	2.1	No	No
36	Tb427.07.6730	hypothetical protein, conserved	17.3	16.4	34.2	36.1	0.9	2.0	2.0	No	No
36	Tb427.07.6740	hypothetical protein, conserved	14.4	13.8	30.0	29.5	1.0	2.1	2.1	No	No
36	Tb427.07.6750	hypothetical protein, conserved	22.0	23.6	45.1	45.9	1.1	2.0	2.0	No	No
36	Tb427.07.6760	hypothetical protein, conserved	13.8	15.2	31.1	31.8	1.1	2.3	2.3	No	No
36	Tb427.07.6780	hypothetical protein	28.2	26.2	61.0	51.0	0.9	2.2	2.2	No	No
36	Tb427.07.6790	hypothetical protein, conserved	18.3	16.2	35.6	36.9	0.9	2.0	2.0	No	No
36	Tb427.07.6800	hypothetical protein, conserved	15.4	17.6	29.6	36.8	1.1	1.9	1.9	Yes	Yes
36	Tb427.07.6810	hypothetical protein, conserved	14.5	15.7	28.5	31.8	1.1	2.0	2.0	Yes	Yes
37	Tb427.07.6880	hypothetical protein, conserved	0.4	0.9	0.9	1.3	2.3	2.1	2.1	Yes	Yes
38	Tb427.08.270	variant surface glycoprotein (VSG, pseudogene), putative	0.0	4.4	2.5	0.0	N/A	N/A	N/A	Yes	Yes
39	Tb427.08.1660	procyclin-associated gene (pseudogene), putative	2.5	8.3	2.9	6.9	3.3	1.2	1.2	Yes	Yes
40	Tb427.08.5920	hypothetical protein	1.1	3.2	0.4	0.6	2.9	0.4	0.4	No	Yes
40	Tb427.08.5930	hypothetical protein	0.4	2.3	0.6	0.6	5.6	1.5	1.5	No	Yes

41	Tb427.10.90	hypothetical protein	3.6	4.4	6.3	11.9	1.2	1.8	1.8	Yes	Yes
42	Tb427.10.1220	hypothetical protein	0.8	2.1	0.8	2.4	2.5	1.0	1.0	Yes	Yes
43	Tb427.10.1760	hypothetical protein	2.8	3.4	1.3	5.6	1.2	0.5	0.5	Yes	Yes
44	Tb427.10.4340	hypothetical protein	2.1	5.1	3.5	7.1	2.4	1.7	1.7	Yes	Yes
44	Tb427.10.4350	hypothetical protein	2.2	27.2	0.9	9.2	12.2	0.4	0.4	Yes	Yes
45	Tb427.10.4370	expression site-associated gene (ESAG) protein, putative	7.1	11.5	8.4	16.1	1.6	1.2	1.2	Yes	Yes
46	Tb427.10.7170	unspecified product	1.1	3.9	0.4	1.5	3.4	0.3	0.3	Yes	Yes
47	Tb427.10.8440	glucose transporter 1B (THT1-)	157.5	262.3	194.8	321.3	1.7	1.2	1.2	No	No
47	Tb427.10.8460	glucose transporter, putative	168.8	291.0	227.8	369.0	1.7	1.3	1.3	No	No
47	Tb427.10.8470	glucose transporter, putative	148.4	256.4	193.1	317.9	1.7	1.3	1.3	No	No
47	Tb427.10.8480	glucose transporter, putative	155.4	270.0	190.4	313.0	1.7	1.2	1.2	No	No
47	Tb427.10.8500	glucose transporter, putative	80.5	127.0	111.5	164.8	1.6	1.4	1.4	No	No
48	Tb427.10.10270	hypothetical protein	4.0	14.3	3.1	6.5	3.6	0.8	0.8	Yes	Yes
49	Tb427.10.12270	hypothetical protein, conserved	11.1	17.3	12.1	30.5	1.6	1.1	1.1	Yes	Yes
49	Tb427.10.12280	hypothetical protein	2.1	3.5	1.3	5.2	1.6	0.6	0.6	Yes	Yes
50	Tb427.10.12570	hypothetical protein	1.9	7.6	1.5	2.4	4.0	0.8	0.8	Yes	Yes
51	Tb427.10.14060	hypothetical protein	0.4	3.9	0.4	2.7	9.7	1.1	1.1	Yes	Yes
52	Tb427.10.16100	peptidylprolyl isomerase-like protein, putative	243.0	185.3	275.7	748.3	0.8	1.1	1.1	No	No
N/A	Tb427.BES126.15	variant surface glycoprotein (VSG) (VSG 427-11)	3.5	3.3	8.6	16.0	0.9	2.4	2.4	No	No
N/A	Tb427.BES129.14	variant surface glycoprotein (VSG) (VSG 427-9)	2.7	2.1	3.6	6.3	0.8	1.3	1.3	No	No
N/A	Tb427.BES15.12	variant surface glycoprotein (VSG) (VSG 427-6)	1.1	1.1	15.1	14.8	1.0	13.4	13.4	No	No
N/A	Tb427.BES59.12	variant surface glycoprotein (VSG) (VSG 427-13)	1.1	1.1	2.2	4.3	1.0	2.0	2.0	No	No
N/A	Tb427.BES65.13	variant surface glycoprotein (VSG) (VSG 427-3)	0.7	0.9	4.3	7.1	1.3	6.2	6.2	No	No
N/A	Tb427.BES56.13	variant surfave glycoprotein (VSG) (VSG 427-17)	3.1	2.6	26.0	13.9	0.9	8.5	8.5	No	No
N/A	Tb427.BES98.12	variant surface glycoprotein (VSG) (VSG 427-18)	2.6	2.0	6.5	11.0	0.7	2.5	2.5	No	No
53	Tb427tmp.01.2860	hypothetical protein	2.0	5.9	1.0	2.5	2.9	0.5	0.5	Yes	Yes
54	Tb427tmp.01.3220	hypothetical protein	2.0	5.0	0.8	2.1	2.5	0.4	0.4	Yes	Yes
55	Tb427tmp.01.3240	trans-sialidase, putative	15.9	20.4	21.3	38.2	1.3	1.3	1.3	Yes	Yes
56	Tb427tmp.01.3250	hypothetical protein	0.9	2.3	0.3	1.1	2.5	0.3	0.3	Yes	Yes
57	Tb427tmp.01.3915	RNA-binding protein, putative (RBP5)	118.1	108.4	173.8	249.2	0.9	1.5	1.5	No	Yes
58	Tb427tmp.01.5580	hypothetical protein, conserved	32.1	32.1	62.1	66.4	1.0	1.9	1.9	No	No
59	Tb427tmp.01.7120	hypothetical protein, conserved	188.2	155.8	187.5	435.0	0.8	1.0	1.0	No	No
60	Tb427tmp.02.0370	hypothetical protein, conserved	0.0	1.0	1.1	0.0	N/A	N/A	N/A	No	No
61	Tb427tmp.02.1564	leucine-rich repeat protein (LRRP), putative	12.3	14.7	34.0	43.3	1.2	2.8	2.8	Yes	No
61	Tb427tmp.02.1565	hypothetical protein	15.6	21.3	40.2	41.1	1.4	2.6	2.6	Yes	No
61	Tb427tmp.02.1580	leucine-rich repeat protein (LRRP), putative	23.6	29.6	51.7	64.9	1.3	2.2	2.2	Yes	No
62	Tb427tmp.02.5690	hypothetical protein	1.5	1.6	3.0	0.9	1.1	2.1	2.1	No	No
63	Tb427tmp.11.0009	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	22.5	0.0	N/A	N/A	N/A	Yes	Yes
64	Tb427tmp.11.0025	unspecified product	0.5	0.8	10.4	20.3	1.7	21.8	21.8	Yes	Yes
65	Tb427tmp.12.0018	hypothetical protein	1.7	7.2	0.7	2.3	4.2	0.4	0.4	Yes	Yes
65	Tb427tmp.12.0019	unspecified product	0.0	0.1	0.2	1.8	N/A	N/A	N/A	No	No
66	Tb427tmp.14.0027	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	3.7	0.0	N/A	N/A	N/A	Yes	Yes
67	Tb427tmp.142.0040	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.3	4.7	1.8	21.4	21.4	Yes	Yes
67	Tb427tmp.142.0130	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.1	1.1	1.5	15.8	15.8	Yes	Yes
67	Tb427tmp.142.0135	expression site-associated gene (ESAG, pseudogene), putative	0.2	1.0	6.9	1.7	5.1	34.4	34.4	Yes	Yes
67	Tb427tmp.142.0300	hypothetical protein (pseudogene)	0.9	1.5	1.3	2.3	1.7	1.4	1.4	Yes	Yes
68	Tb427tmp.142.0470	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.4	4.8	1.5	20.6	20.6	Yes	Yes
69	Tb427tmp.160.0220	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.0	4.8	4.1	N/A	N/A	N/A	Yes	Yes
69	Tb427tmp.160.0260	unspecified product	0.0	0.0	1.6	2.5	1.0	34.0	34.0	Yes	Yes
70	Tb427tmp.160.0280	variant surface glycoprotein (VSG, atypical), putative	1.2	1.1	3.4	4.0	0.9	2.8	2.8	Yes	Yes
71	Tb427tmp.160.3400	glucose transporter (pseudogene), putative	1.3	2.2	0.9	3.3	1.7	0.7	0.7	Yes	Yes
72	Tb427tmp.160.4950	hypothetical protein, conserved	11.7	16.9	20.3	24.9	1.4	1.7	1.7	Yes	Yes
73	Tb427tmp.160.5350	variant surface glycoprotein (VSG)-related, putative	19.3	22.6	39.9	55.1	1.2	2.1	2.1	Yes	Yes
74	Tb427tmp.211.0010	unspecified product	4.9	6.7	13.9	18.3	1.4	2.8	2.8	No	No
74	Tb427tmp.211.0020	hypothetical protein	3.9	4.8	10.3	13.0	1.2	2.6	2.6	No	No
75	Tb427tmp.211.5010	SLACS retrotransposable element (part), putative	1.6	1.6	9.0	5.4	1.1	5.7	5.7	Yes	Yes
75	Tb427tmp.211.5015	SLACS reverse transcriptase, putative	0.5	0.1	1.9	0.9	0.2	4.0	4.0	Yes	Yes
76	Tb427tmp.244.0550	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.3	1.8	3.3	1.4	10.3	10.3	Yes	Yes
76	Tb427tmp.244.0620	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.2	2.3	5.5	1.5	20.8	20.8	Yes	Yes
76	Tb427tmp.244.0650	variant surface glycoprotein (VSG, pseudogene), putative	0.6	0.7	2.3	2.7	1.2	4.2	4.2	Yes	Yes
77	Tb427tmp.244.1260	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.1	0.8	1.0	0.8	6.6	6.6	Yes	Yes
78	Tb427tmp.244.1410	variant surface glycoprotein (VSG), putative	0.7	0.8	8.5	10.5	1.1	12.7	12.7	Yes	Yes
79	Tb427tmp.244.1480	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.1	1.2	2.5	0.7	7.3	7.3	Yes	Yes
80	Tb427tmp.244.1700	variant surface glycoprotein (VSG, pseudogene), putative	0.2	0.2	0.9	1.5	1.3	5.0	5.0	Yes	Yes
80	Tb427tmp.244.1740	variant surface glycoprotein (VSG), putative	0.7	1.2	15.6	30.5	1.8	23.8	23.8	Yes	Yes
80	Tb427tmp.244.1950	hypothetical protein	0.3	0.9	0.5	1.3	2.6	1.6	1.6	Yes	Yes
80	Tb427tmp.244.1960	hypothetical protein	2.0	10.2	2.7	8.1	5.2	1.4	1.4	Yes	Yes
80	Tb427tmp.244.1970	leucine-rich repeat protein (pseudogene), putative	5.9	10.4	12.7	14.3	1.8	2.2	2.2	Yes	Yes
81	Tb427tmp.244.2010	expression site-associated gene 4 (ESAG4, pseudogene), putative	16.1	20.8	40.0	53.4	1.3	2.5	2.5	Yes	Yes
81	Tb427tmp.244.2020	nucleoside transporter 1, putative	8.6	9.1	19.4	21.9	1.1	2.3	2.3	Yes	Yes
82	Tb427tmp.244.2060	expression site-associated gene (ESAG) protein, putative	46.9	61.0	84.5	113.3	1.3	1.8	1.8	Yes	Yes
83	Tb427tmp.244.2890	SLACS reverse transcriptase, putative	2.4	2.2	7.8	4.6	0.9	3.2	3.2	Yes	Yes
84	Tb427tmp.354.0120	variant surface glycoprotein (VSG, pseudogene), putative	0.1	0.3	2.9	7.0	4.1	41.5	41.5	Yes	Yes
85	Tb427tmp.55.0028	expression site-associated gene (ESAG) protein, putative	0.0	0.0	0.0	32.2	N/A	N/A	N/A	Yes	Yes
86	Tb427tmp.57.0077	unspecified product	8.0	13.3	8.5	20.7	1.7	1.1	1.1	Yes	Yes
87	Tb427tmp.57.0088	variant surface glycoprotein (VSG, pseudogene), putative	1.2	1.7	2.7	1.9	1.4	2.2	2.2	Yes	Yes
88	Tb427tmp.v1.0970	hypothetical protein	1.6	3.3	1.8	4.8	2.1	1.1	1.1	Yes	Yes
88	Tb427tmp.v2.0080	variant surface glycoprotein (VSG, pseudogene), putative	1.8	1.4	13.6	10.8	0.8	7.7	7.7	Yes	Yes
89	Tb427tmp.v2.0300	variant surface glycoprotein (VSG, pseudogene), putative	0.0	0.1	1.1	0.8	N/A	N/A	N/A	Yes	Yes
90	Tb427tmp.v2.0350	variant surface glycoprotein (VSG, pseudogene), putative (VSG56)	0.0	1.5	1.2	3.1	N/A	N/A	N/A	Yes	Yes
91	Tb427tmp.v2.0450	hypothetical protein, conserved, frameshifted fragment	0.0	0.0	0.0	1.5	N/A	N/A	N/A	Yes	Yes

[1]Fold upregulation calculated by dividing the *H3.V* KO and/or DMOG treated RPKM by WT RPKM

[2]Upregulated genes within the same PTU have the same number. A different number indicates the genes are in separate PTUs

[3]Yes: if gene is located within 10 kb of H3.V

[4]Yes: if gene is located within 10 kb of base J

CHAPTER 5

BASE J REPRESSES GENES AT THE END OF POLYCISTRONIC GENE CLUSTERS IN *LEISHMANIA MAJOR* BY PROMOTING RNAP II TERMINATION[1]

[1]**Reynolds D.**, Hofmeister B. T., Cliffe L., Siegel T. N., Anderson B. A., Beverley S. M., Schmitz R. J., and Sabatini R. 2016. Submitted to *Molecular Microbiology*.

ABSTRACT

The genomes of kinetoplastids are organized into polycistronic gene clusters that are flanked by the modified DNA base J. Previous work has established a role of base J in promoting RNA polymerase II termination in *Leishmania* spp. where the loss of J leads to termination defects and transcription into adjacent gene clusters. It remains unclear whether these termination defects affect gene expression and whether read through transcription is detrimental to cell growth, thus explaining the essential nature of J. We now demonstrate that reduction of base J at specific sites within polycistronic gene clusters in *L. major* leads to read through transcription and increased expression of downstream genes in the cluster. Interestingly, subsequent transcription into the opposing polycistronic gene cluster does not lead to downregulation of sense mRNAs. These findings indicate a conserved role for J regulating transcription termination and expression of genes within polycistronic gene clusters in trypanosomatids. In contrast to the expectations often attributed to opposing transcription, the essential nature of J in *Leishmania* spp. is related to its role in gene repression rather than preventing transcriptional interference resulting from read through and dual strand transcription.

INTRODUCTION

Kinetoplastids are a group of early-diverged eukaryotes collectively responsible for multiple diseases including African sleeping sickness, Chagas disease, and leishmaniasis. Kinetoplastid parasites, which include *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, progress through life stages by cycling between an insect vector and a mammalian host. The parasites must therefore adapt to changes in their environment, such as changes in pH, temperature, oxygen concentrations, and host immune responses, among others.

Unlike most eukaryotes, the entire genome of kinetoplastids is arranged into gene clusters (1-3) where RNA Polymerase (RNAP) II initiates transcription at sites called divergent strand switch regions (dSSRs), transcribes polycistronically (4-6) through tens to hundreds of genes, and terminates at convergent strand switch regions (cSSRs) (7). Gene clusters are also adjacently arranged on the same DNA strand in what are called a head-tail (HT) arrangement, such that transcription of an upstream cluster terminates and initiation of an independent gene cluster occurs downstream (8-10). Pre-mRNAs are processed through trans-splicing with the addition of a 39 nucleotide spliced leader (SL) sequence to the 5' end of all mRNAs (11-16) (reviewed in (17)), which is coupled to the 3' polyadenylation of the upstream transcript (18).

Relatively little is known about transcriptional regulation in kinetoplastids, in large part because the polycistronic nature of the genome has led to the belief that gene expression regulation does not occur at the transcriptional level and instead predominately occurs post-transcriptionally through mechanisms such as mRNA processing, mRNA stability, and translation efficiency (19,20). Evidence has begun to emerge however that has challenged the notion that kinetoplastids lack transcriptional regulation. Multiple chromatin modifications are enriched at sites of RNAP II initiation and termination, including histone methylation, acetylation, histone variants, and the DNA modification base J that could be utilized by the parasites to regulate transcription (8,9,21,22). Of these chromatin modifications, base J has been the most studied and demonstrated to play a functional role in the regulation of transcription in all kinetoplastids examined thus far (23-27). J consists of a glucosylated thymidine (28) and has been found only in the nuclear DNA of kinetoplastids, *Diplonema*, and *Euglena* (29,30). The DNA modification is synthesized through the hydroxylation of thymidine by either JBP1 or JBP2 (31), forming hydroxymethyluridine, followed by the transfer of a glucose by the

glucosyltransferase enzyme JGT (32,33) (reviewed in (34,35)). Analogous to the TET proteins in mammals (36,37), the JBPs utilize 2-oxoglutarate, oxygen, and Fe^{2+} in the hydroxylation reaction (31). The central roles of JBP1/2 and JGT in synthesizing J have been established in *T. brucei*, the only kinetoplastid species tested thus far in which genetic deletions of *JBP1/2* and *JGT* are viable (32,38). Attempts to generate *T. cruzi* and *Leishmania* spp. *JBP1/2* KO cells have been unsuccessful, suggesting an essential role of J in these kinetoplastids. Addition of the 2-oxoglutarate structural analog dimethyloxalylglycine (DMOG) to the growth medium or limiting oxygen concentrations inhibit hydroxylase activity and thus enable J reduction in cells without genetic modification (31).

Investigations in *T. cruzi* have revealed a function of J in the repression of RNAP II initiation, such that J loss increases active chromatin marks and transcription initiation, resulting in global gene expression changes (23,24). No defect in RNAP II termination was identified upon J loss in *T. cruzi*, however in *T. brucei* and *Leishmania* spp. J has been found to promote RNAP II termination (25-27). In *T. brucei,* J located prior to the end of gene clusters promotes termination, repressing downstream genes in the same cluster (26,27). Loss of J from within gene clusters results in read through and expression of the downstream genes, and we therefore refer to these sites as gene cluster internal termination sites. H3.V co-localizes with J at RNAP II termination sites in *T. brucei* (22) and has recently been shown to play a similar role in regulating termination (27,39). J and H3.V independently function to promote termination within gene clusters, such that the combined loss of J and H3.V results in a synergistic increase in read through transcription at gene cluster internal termination sites and derepression of downstream genes (27). Genes regulated by this process include many that are developmentally regulated in *T. brucei* and code for proteins involved in optimal growth and immune evasion

during infection of the mammalian host (the specific trypanosome life stage where J is synthesized) (27,39).

While J and H3.V also co-localize at termination sites at the end of gene clusters (i.e. cSSRs) in *T. brucei*, they are not required for terminating transcription and preventing elongation of RNAP II into the opposing gene cluster (27). However, in *L. major* and *L. tarentolae* J does function to prevent read through transcription at cSSRs and the formation of antisense RNAs (25,26). For example, reduction of base J levels in *L. major* following DMOG treatment results in transcription of the antisense strand of the adjacent gene cluster genome-wide (26). It has been proposed that dual strand transcription resulting from read through at cSSRs disrupts expression of sense mRNAs via transcriptional interference and is therefore detrimental to *Leishmania* spp. cell growth (25). Whether J loss and subsequent read through at cSSRs impact mRNA transcript abundance in *Leishmania* spp. is not yet clear however. It is also not known if J functions to promote gene cluster internal termination in *Leishmania* spp., and if so what role this process has in parasite growth and explaining the apparent essential nature of J. The role of H3.V is also unclear. Removal of H3.V in *L. major* did not reveal defects in RNAP II termination (40), however it is not known if H3.V localizes to termination sites in *L. major*, as is observed in *T. brucei*.

The recently identified role of H3.V and base J in promoting RNAP II termination within gene clusters in *T. brucei* led us to further investigate the function of these epigenetic marks in *L. major*. We found that like *T. brucei*, J is localized within several gene clusters in *L. major* where the acute J loss induced by the J synthesis inhibitor DMOG results in defects of RNAP II termination within the cluster and increased expression of downstream genes. We also demonstrate here that, similar to *T. brucei*, H3.V co-localizes with base J at termination sites in

L. major. Surprisingly, however, H3.V apparently regulates J synthesis in *L. major* as the loss of H3.V reduces the level of J at termination sites with no effects on RNAP II termination and minimal gene expression changes. Further reduction of J at termination sites in the *H3.V* knockout (KO) using DMOG revealed greater termination defects, more significant gene expression changes, and greatly reduced cell growth, compared with wild type (WT) cells treated with DMOG. Whilst read through defects in *L. major* include the extension of RNAP II onto the adjacent opposing gene cluster and dual strand transcription, we saw no evidence of transcription interference resulting in significant downregulation of mRNAs on the opposing gene cluster in either WT or *H3.V* KO cells treated with DMOG. These results indicate a conserved role for J regulating RNAP II transcription termination and expression of genes within polycistronic gene clusters in trypanosomatids and suggest that the essential nature of J in *Leishmania* spp. is related to its role in repressing specific genes at the end of gene clusters and not the prevention of dual strand transcription.

MATERIAL AND METHODS

Enzymes and chemicals

Enhanced chemiluminescence and Hybond-N+ were purchased from Amersham. Goat anti-rabbit HRP (horseradish peroxidase) and goat anti-mouse HRP antibodies were purchased from Southern Biotec Inc. DMOG was purchased from Frontier Scientific, Inc. All other chemicals were purchased from Sigma-Aldrich.

Parasite cell culture

L. major Friedlin strain promastigotes, wild type and the *H3.V* KO (40), were grown at 26°C in M199 media supplemented with 10% FBS as described (41). DMOG treatment of cells was performed by supplementing media with 5mM DMOG for 2-10 days. Control cells were treated with equal amounts of vehicle (dimethyl sulfoxide).

Chromatin immunoprecipitation-quantitative PCR (ChIP-qPCR)

To determine histone H3 and H3.V enrichment in the *L. major* genome, ChIP-qPCR was performed using $2x10^8$ cells resuspended in 20mL of phosphate buffered saline. Cells were cross-linked in 1% formaldehyde for 20 minutes. Subsequent ChIP steps were performed as previously described (23). *L. major* histone H3 and H3.V specific antibodies used in this study were previously described (40). Input DNA was used as a positive control for qPCR (10% of the IP). Quantification of selected regions of the genome was performed on an iCycler with an iQ5 multicolor real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA). All primers used in this study were ordered from Integrated DNA Technologies, Inc. or Invitrogen. Primer sequences used in the analysis are available upon request. The reaction mixture contained 5 pmol forward and reverse primer, 2x iQ SYBR green super mix (Bio-Rad Laboratories, Hercules, CA), and 2 µl of template DNA. Standard curves were prepared for each gene using 5-fold dilutions of known quantity (100 ng/µl) of WT DNA. The quantities were calculated using iQ5 optical detection system software.

Determination of the genomic level of J

To quantify the levels of base J at specific regions of the *L. major* genome, genomic DNA was sonicated and anti J immunoprecipitation was performed as described (22). Immunoprecipitated J containing DNA was used for qPCR analysis, as described above for ChIP-qPCR analysis.

Western blots

Proteins from 10^7 cell equivalents were separated by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS page 8% gel), transferred to nitrocellulose and probed with anti-H3 and H3.V as described (40). Anti-alpha tubulin (Sigma-Aldrich) was used as a loading control. Bound antibodies were detected by goat anti-rabbit or goat anti-mouse secondary antibodies conjugated to HRP and visualized by enhanced chemiluminescence. Equal loading was also assessed by Coomassie Brilliant Blue staining.

Strand-specific RNA-seq library construction and data analysis

Small RNAs were isolated from *L. major* using a Qiagen miRNeasy kit according to the manufacturer's instructions. $5x10^7$ cells were used per sample, and isolated at the log phase of parasite growth. Two small RNA-seq libraries (*L. major H3.V* KO and *H3.V* KO+DMOG) were prepared by Vertis Biotechnology AG, Germany, as previously described (27). All WT and WT+DMOG small RNA-seq data shown here are from our previous study (26). Libraries were sequenced using Illumina's HiSeq2000. Reads were mapped to the *L. major* Friedlin strain genome version 4.2 as previously described (26). Information about all sequencing data generated in this study is listed in Supplementary Table 2.

For strand-specific mRNA-seq, total RNA was isolated using Trizol from log phase *L. major* WT and *H3.V* KO cells ($5x10^7$ cells) treated with and without 5mM DMOG for two days. Eight mRNA-seq libraries were constructed (two replicates each; WT, WT+DMOG, *H3.V* KO, and *H3.V* KO+DMOG) using Illumina's TruSeq Stranded RNA LT Kit, sequenced by the Georgia Genomics Facility using Illumina's NextSeq500, and resulting reads were processed and mapped to the *L. major* Friedlin strain genome version 9.0 as previously described (27). Determination of transcript abundances was performed as previously described (27) using the Cufflinks suite version 2.2.1 (42). To estimate gene expression levels for a condition, replicates were analyzed by Cuffdiff with the *L. major* Friedlin strain genome version 9.0 annotation. Default parameters were used except specifying library type fr-firststrand. All p-values reported here, determined by Cuffdiff, reflect the FDR-adjusted p-value. Correlation coefficients for mRNA-seq replicates of WT, WT+DMOG, *H3.V* KO and *H3.V* KO +DMOG were all greater than 0.95. All J IP-seq data shown here are from previously published work (25).

Strand-specific reverse transcription-PCR (RT-PCR) analysis of read through transcription

Total RNA was isolated using the hot phenol method, as described previously (43). To ensure complete removal of contaminating genomic DNA, purified RNA was treated with Turbo DNaseI (Invitrogen), followed by phenol:chloroform extraction. Strand-specific RT-PCR was performed as previously described (26,27). Briefly, ThermoScript Reverse Transcriptase from Life Technologies was used for cDNA synthesis at 60-65°C. 1-2 μg of RNA were used to make cDNA using a reverse primer as described in the Figure legends. PCR was performed using GoTaq DNA Polymerase from Promega. A minus-RT control was used to ensure no

contaminating genomic DNA was amplified. Primer sequences used in the analysis are available upon request.

Reverse transcription-quantitative PCR (RT-qPCR)

Total RNA was obtained using Qiagen RNeasy kits according to manufacturers instructions. First-strand cDNA was synthesized from 1 μg of total RNA using an iScript cDNA synthesis kit (Bio-Rad Laboratories, Hercules, CA) per the manufacturer's instructions. qPCR was performed using 1-2 μl of template cDNA, and as described above.

RESULTS

L. major H3.V co-localizes with base J at termination sites and regulates J synthesis

We have found that in *T. brucei*, H3.V and base J co-localize at RNAP II termination sites and both epigenetic marks work together to promote RNAP II termination (27,39). However, while base J has clearly been shown to regulate RNAP II termination in *L. major* (26) the deletion of H3.V from this parasite did not indicate defects in termination (40). To take a closer look, we first determined whether H3.V localizes with base J at termination sites in *L. major*. Using chromatin immunoprecipitation-qPCR (ChIP-qPCR), and primers that spanned a cSSR (Figure 5.1A), we found that while histone H3 is present at similar levels across the cSSR, H3.V is enriched at the RNAP II termination site (Figure 5.1B and Supplementary Figure 5.S1A and B). In contrast, we did not detect enrichment of H3.V at dSSRs, where RNAP II initiation takes place (Figure 5.1D and E, Supplementary Figure 5.S1A and B). Lack of H3.V ChIP-qPCR signal in *H3.V* KO cells confirms the specificity of the *L. major* H3.V antibody (Supplementary Figure 5.S1D). Consistent with previous findings (26), by J IP-qPCR we observed a peak

enrichment of J at the RNAP II termination site, confirming that J and H3.V co-localize at RNAP II termination sites in *L. major* (Figure 5.1B and C). As expected, we also detect a peak of J within dSSRs (Figure 5.1F).

While H3.V and base J co-localize and cooperatively regulate RNAP II termination in *T. brucei*, H3.V does not influence base J synthesis (22). In contrast, the deletion of *H3.V* in *L. major* led to significantly reduced levels of base J in cSSRs (Figure 5.1C and Supplementary Figure S1C). This loss of J was not restricted to termination sites however, as we also observed reduced J at a dSSR in the *H3.V* KO, despite the lack of enrichment of H3.V at these sites (Figure 5.1E and F). This suggests that, for at least dSSRs, the loss of H3.V indirectly impacts J synthesis. To determine whether J plays a role in histone/nucleosome localization, we analyzed the levels of H3 and H3.V at cSSRs of cells with reduced J following treatment with the J synthesis inhibitor DMOG. We have previously demonstrated that DMOG treatment of *L. major* leads to significant reductions in base J levels, and corresponding read through transcription, at cSSRs genome-wide (26). We find here that reduction of base J did not have a significant effect on the abundance of H3.V or canonical histone H3 within cSSRs (Supplementary Figure 5.S2A and B). There was a slight trend towards increased H3.V and H3 following J loss, however the increase was not statistically significant at any of the sites analyzed. Thus, the loss of H3.V results in reduced J levels at both RNAP II termination and initiation sites, presumably through an indirect effect. However, J is not required for H3.V or nucleosome localization at RNAP II termination sites.

To begin to investigate how the loss of H3.V decreases J levels at RNAP II termination sites, we analyzed the effect of H3.V loss on nucleosome abundance by H3 ChIP-qPCR. Using primers located near the termination site we found a trend of reduced H3 in cSSRs of the *H3.V*

KO cell line (Supplementary Figure 5.S2C). Total protein levels of H3 were similar in the WT and *H3.V* KO cell lines (Supplementary Figure 5.S3) and the loss of H3.V did not affect transcript levels for any of the histones or enzymes involved in J synthesis (see below). Therefore, loss of H3.V leads to reduced assembly of H3 containing nucleosomes at cSSRS. These findings implicate chromatin structure, specifically nucleosome abundance, in the regulation of J synthesis at termination sites, but it remains unclear how the loss of H3.V negatively affects J synthesis at dSSRs.

L. major H3.V is not required for RNAP II termination

Given that we have previously shown that reduction of J in *L. major* results in read through transcription at RNAP II termination sites, regardless of how J is reduced, we would expect a defect in RNAP II termination in the *H3.V* KO cells. Anderson et al. have recently found however, that the loss of H3.V does not lead to a detectable termination defect in *L. major*, as determined by SL RNA-seq (40). While both SL RNA-seq and small RNA-seq are capable of revealing termination defects following J reduction, small RNA-seq is more sensitive (25) and thus we used this method to further analyze the *H3.V* KO for termination defects genome-wide by small RNA-seq. As we have previously demonstrated, the reduction of base J in *L. major* by treatment with DMOG resulted in the production of antisense small RNAs due to read through transcription at cSSRs and into the opposing gene cluster (Figure 5.2A, Supplementary Figure S4A and Supplementary Figure 5.S5) (26). Read through transcription following J loss is confirmed by strand-specific RT-PCR, where an RNA species spanning the termination site is detected only following J loss (Figure 5.2B and Supplementary Figure 5.S4B), and in agreement with the SL RNA-seq data of Anderson et al., we did not find

evidence of read through transcription in the *H3.V* KO by small RNA-seq or strand-specific RT-PCR (Figure 5.2A and B, Supplementary Figure S4A and B, and Supplementary Figure 5.S5). Only upon further loss of J in the *H3.V* KO following DMOG treatment is read through transcription detected (Figure 5.2, Supplementary Figure 5.S4 and Supplementary Figure 5.S5). Interestingly, DMOG treatment of the *H3.V* KO cells results in a greater decrease in J and greater read through transcription than the similar DMOG treatment of WT cells (Figure 5.2 and Supplementary Figure 5.S4). The increased read through transcription in the *H3.V* KO+DMOG is observed by small RNA-seq (several cSSRs are shown in Figure 5.2A, Supplementary Figure 5.S4A, and Supplementary Figure 5.S5 and quantitation across all cSSRs is shown in Supplementary Figure 5.S6), and also confirmed by semi-quantitative strand-specific RT-PCR and quantitative strand-specific RT-qPCR analysis of several termination sites (Figure 5.2B, Supplementary Figure 5.S4B, and Supplementary Figure 5.S7). This nicely reflects the direct relationship of J levels and degree of read through transcription previously characterized in *L. major* (26). These data support the conclusion that, as opposed to *T. brucei*, H3.V does not directly regulate RNAP II termination in *L. major*.

The fact that reduction of base J in the *H3.V* KO does not result in termination defects, unless J is further reduced by DMOG, suggests there is a threshold of J required for RNAP II termination. To address this, we performed a DMOG titration in WT cells to follow the effect of varying levels of J on termination at two cSSRs (Figure 5.3). We found that read through transcription does not occur until approximately 75% of the initial J is lost at the termination site. In the DMOG titration we see that ~40% (Figure 5.3A) and ~55% (Figure 5.3B) reduction of J in WT cells is not sufficient for a termination defect. ~75% (or greater) reduction of J in WT cells, with increasing DMOG concentration, leads to read through transcription. Deletion of

H3.V resulted in ~60% reduction in base J, thus explaining the lack of a detectable defect in termination. However, subsequent treatment of the *H3.V* KO with DMOG decreased J by 90-95% relative to WT cells, and resulted in a termination defect (Figure 5.3). As described above, similar treatment of WT and *H3.V* KO with DMOG resulted in significantly more J loss in the *H3.V* KO and a stronger defect in termination. In some cases the loss of J in the WT+DMOG (or *H3.V* KO) is not enough for significant read through transcription at the cSSR and only when J is reduced further in the *H3.V* KO+DMOG is there strong termination defect (Supplementary Figure 5.S4 and Supplementary Figure 5.S5, cSSR 7.2). We conclude that a threshold of J between 25-40% of WT levels is sufficient to prevent read through transcription, thus explaining the lack of read through in the *H3.V* KO cells. Once below this threshold the degree of termination defect is directly related to the decrease in J, explaining why we detect greater read through in the *H3.V* KO+DMOG cell line than WT+DMOG. Taken together these data support the conclusion that regulation of termination in *L. major* is all about base J. H3.V does not regulate termination, but does regulate J levels. However, the loss of H3.V does not decrease J levels below the threshold required for termination.

J promotes termination prior to the end of gene clusters, repressing gene expression

We have shown previously that J reduction and subsequent read through transcription in WT *L. major* cells treated with DMOG result in only a modest defect in cell growth *in vitro* (26). These findings contrast with those by van Luenen et al., who found that further reduction of J in a *JBP2* KO cell line by bromodeoxyuridine (BrdU) is lethal in *L. tarentolae* (25). We now find that the *H3.V* KO cell line is hypersensitive to DMOG compared to WT cells (Supplementary Figure 5.S8). While the loss of H3.V alone has no effect on cell growth (40),

H3.V KO+DMOG cells grow significantly slower than WT+DMOG cells after 48 hours and continue to slowly divide thereafter. We conclude that the greater J loss and read through defects in the *H3.V* KO+DMOG cells are detrimental to cell growth and that base J is essential in *Leishmania* spp. due to its role in regulating RNAP II termination.

One possibility is that the essential nature of J in *Leishmania* spp. is indicated by the termination defects at cSSRs resulting in RNAP II transcription into opposing gene clusters genome-wide that occur following the loss of J (25,26). For example, read through transcription at convergent gene clusters may result in transcriptional interference where transcription of the opposing strand of the adjacent gene cluster exerts a direct negative impact on the transcription of the sense strand. In this collision model, active antisense transcription suppresses sense RNA transcription resulting in downregulation of mRNAs for (essential) genes located at the end of convergent gene clusters. To test this model we performed mRNA-seq in parasites treated with and without DMOG. For each condition (WT, WT+DMOG, *H3.V* KO, and *H3.V* KO+DMOG) two independent mRNA-seq libraries were sequenced. Replicates were strongly correlated with each other (with R^2 values >0.95 for all replicates). We focused on genes that were up or downregulated by twofold or more with an RPKM of at least one in each replicate. Using these cutoffs we identified very limited gene expression changes in WT *L. major* cells incubated with DMOG (Supplementary Table 1). Even fewer gene expression changes were observed in *H3.V* KO cells, consistent with previous SL RNA-seq (40) (Supplementary Table 1). mRNA-seq indicated only 22 transcripts were changed more than twofold following the decrease of J in WT cells, the majority of which (21 transcripts) had increased expression. We confirm many of these changes by RT-qPCR (Supplementary Figure 5.S9). The limited number of affected genes indicates read through transcription does not lead to significant downregulation of mRNAs. For

example, at two cSSRs that are known to have some of the highest levels of read through transcription following loss of J with DMOG (10.2 and 8.1) (26), none of the mRNAs for genes within 10kb of the termination site of the four gene clusters are significantly downregulated in either WT+DMOG or *H3.V* KO+DMOG (Figure 5.4). In fact, none of the genes on these chromosomes are significantly downregulated in any of the conditions examined here. Consistent with small RNA-seq analysis, read through transcription through the cSSR into the opposing gene cluster following the loss of J is apparent in the mRNA-seq data (Figure 5.4A and B). Therefore, in agreement with van Luenen et al., some fraction of read through (antisense) RNA is processed similarly to mRNAs (25). However, as seen previously, steady-state mRNA does not necessarily reflect the extent of read through transcription compared to small RNA-seq or analysis of nascent RNA by strand-specific RT-PCR. Nevertheless, consistent with small RNA-seq, plotting the mRNA-seq data supports the conclusion that transcriptional read through is highest in *H3.V* KO+DMOG. Interestingly, while none of genes localized at the end of gene clusters are significantly downregulated, even in the highest degree of read through transcription in the *H3.V* KO+DMOG, we do detect a slight trend of decreasing mRNA levels with increasing degree of read through (Figure 5.4C and D). Therefore, while there may be some transcriptional interference resulting from RNAP II extension into the adjacent convergent gene cluster, this does not result in significant repression of mRNA levels.

In contrast, we find many more genes upregulated upon stimulation of read through transcription (Supplementary Table 5.1). Similar to findings we previously described in *T. brucei* (26,27), all 21 of the upregulated genes in WT+DMOG *L. major* cells are located at the end of a gene cluster immediately downstream or within a J peak, where J regulation of transcription termination within a gene cluster attenuates transcription of downstream genes. For

example, on chromosome 9 only one gene is upregulated in WT+DMOG, and represents the last gene of a gene cluster at cSSR 9.1 (Figure 5.5A). The peak of base J extends from the cSSR into the open reading frame (ORF) of the last gene in the gene cluster that is lowly expressed in the presence of J (WT), but increases upon the loss of J following DMOG treatment (Figure 5.5B-E). mRNA levels are further increased upon the loss of J in the *H3.V* KO+DMOG, presumably due to increased read through transcription. As discussed above, no read through and minimal gene expression changes are detected in the *H3.V* KO. cSSR 9.1 (Figure 5.5) is an example where attenuation of transcription within an ORF is sufficient for gene repression. Termination of transcription prior to the 3' UTR would prevent the production of full-length precursor transcript and therefore, the production of processed steady-state mRNA. At this gene cluster internal termination site, decreased levels of base J leads to read through transcription on the top strand that results in the complete transcription of the final gene, which is processed to mature mRNA. Read through following the loss of J extends down to an RNAP III-transcribed gene in the cSSR, which are known to terminate RNAP II transcription independent of J in *Leishmania* spp. (25,26). The presence of RNAP III-transcribed genes (a 5S rRNA gene on the bottom strand and tRNA genes on the top strand) at this site terminates transcription on both strands in the absence of base J (25,26). Thus, while our previous genome-wide studies quantitating the production of antisense RNAs failed to detect a defect in RNAP II termination at this cSSR due to J-independent mechanisms preventing transcription into the opposing gene clusters, we now illustrate that base J, in fact, regulates termination on the sense strand of a gene cluster prior to the cSSR. No expression change is detected for the genes immediately upstream or the final two genes of the adjacent convergent gene cluster in either WT or *H3.V* KO cells treated with DMOG (Figure 5.5E).

Additional examples of J regulation of gene expression via termination include H-T regions 26.5 (Figure 5.6) and 23.1 (Supplementary Figure 5.S10A), and cSSR 32.2 (Supplementary Figure 5.S10B) where J is found upstream of the last gene within the gene cluster. The downstream (and final) gene is lowly expressed in the presence of J (WT), but increases upon the loss of J following DMOG treatment and further increases in the *H3.V* KO+DMOG. Adjacent genes, upstream of J or within the neighboring gene cluster, are not affected. Further evidence for read through transcription at a gene cluster internal termination site following J loss is provided by strand-specific RT-PCR analysis (Figure 5.6F). Following the decrease in base J in WT cells we detect an increase in a nascent transcript that extends downstream of the proposed termination site. The level of this RNA species increases along with decreasing levels of J in the *H3.V* KO+DMOG, directly linking the degree of read through with mRNA abundance of the downstream gene.

27 of the 43 (63%) upregulated genes in DMOG treated parasites (WT and/or *H3.V* KO cells) fit a model where derepression occurs as a result of deregulated transcription elongation/termination within a gene cluster following the loss of base J (Supplementary Table 5.1 and Supplementary Figure 5.S11). These 27 upregulated genes are located in 23 independent gene clusters, out of approximately 184 gene clusters in the *L. major* genome. In this way, J represses gene expression genome-wide in *L. major*, often repressing the expression of a single gene, usually the last gene, within a gene cluster. 14 of the 27 genes that fit this model have been identified as developmentally regulated in *L. major* (44) (see Discussion). These findings overall suggest a conserved role of J in the promotion of RNAP II termination prior to the end of gene clusters in kinetoplastids as a mechanism for regulated gene expression.

DISCUSSION

The arrangement of functionally unrelated genes into co-transcribed gene clusters and the apparent lack of sequence specific DNA transcription factors has made it difficult to envision regulated gene expression at the level of transcription in kinetoplastids. Evidence has begun to emerge however, that chromatin modifications, including the DNA modification base J and H3.V, can provide some measure of transcriptional control. In *T. brucei* both J and H3.V promote RNAP II termination prior to the end of gene clusters, thereby repressing downstream genes in the same cluster (27,39). We demonstrate here that this function of J is conserved between *T. brucei* and *L. major*. Whereas H3.V does not have a direct role, base J promotes RNAP II termination and regulation of genes at the end of gene clusters in *L. major*, including several that are developmentally regulated. We have therefore identified a conserved epigenetic mechanism of regulated gene expression via control of RNAP II transcription termination in kinetoplastids. Consistent with previous findings in *L. tarentolae* (25), we also show that a large decrease of base J is detrimental to *L. major* growth. The absence of major gene expression changes at sites where J loss leads to read through transcription into the adjacent gene cluster and the production of antisense RNAs suggests that defects in cell growth are not due to transcriptional interference, and instead implicates the conserved function of J in the repression genes at the ends of gene clusters in the maintenance of cell viability.

Function of H3.V in termination

In *T. brucei*, J and H3.V co-localize at RNAP II termination sites genome-wide and loss of either mark at specific sites within gene clusters leads to read through transcription and increases expression of downstream genes (26,27). These studies indicate that base J and H3.V

synergistically regulate transcription termination and gene expression. Importantly, the effect of H3.V is not simply due to base J since the levels of the modified base are not reduced in the *T. brucei H3.V* KO (22). Rather, H3.V appears to regulate transcription in *T. brucei* independent of base J. Regions of overlapping transcription in cSSRs of WT *T. brucei* lead to the production of siRNAs (45,46). Loss of H3.V, but not base J, leads to increased transcription within these cSSRs and increased production of siRNAs. *Variant surface glycoprotein* genes within the silent RNAP I transcribed telomeric gene clusters are also derepressed in the *H3.V* KO. These results indicate a specific role of H3.V, independent of base J, in telomeric gene repression and non-coding RNA expression in *T. brucei*. Here we now demonstrate that H3.V and base J co-localize at termination sites in *L. major*, and confirm that loss of H3.V does not lead to significant effects on RNAP II termination. Therefore, in contrast to its function in *T. brucei*, H3.V does not directly regulate termination in *L. major*. Rather, the loss of H3.V in *L. major* decreases levels of base J by an unknown mechanism, possibly due to subtle chromatin structural changes (Supplementary Figure 5.S2). When analyzing read through transcription via antisense RNA production by small RNA-seq (Figure 5.2A, Supplementary Figure 5.S4A, and Supplementary Figure 5.S5) or SL RNA-seq (40), no evidence of a termination defect was identified in the *H3.V* KO. Through strand-specific mRNA-seq we did detect a few gene expression changes in the *H3.V* KO due to read through at internal termination sites (a total of six genes at five different termination sites, see Supplementary Table 5.1). In general these genes are lowly expressed in both WT and *H3.V* KO cells, potentially explaining why these genes were previously missed. Therefore, at a majority of termination sites (179 out of ~184) enough J remains in the *H3.V* KO to efficiently terminate transcription, but at five sites J is apparently decreased enough to result in a small degree of read through and subsequently increased steady-

state mRNA for the downstream gene(s). Evidence that J regulates these genes is provided by the response to reducing J in WT and the *H3.V* KO. At each of the five internal termination sites, the genes are upregulated in WT cells where J is reduced via DMOG. Consistent with greater J loss and read through at termination sites upon DMOG treatment compared with deletion of *H3.V*, the upregulation is higher in WT+DMOG than *H3.V* KO. Additional loss of J in the *H3.V* KO, via DMOG, leads to further read through and corresponding increases in the expression of these six genes as well as genes downstream of J in other gene clusters (discussed below). Thus, although the co-localization of J and H3.V at termination sites is conserved between *T. brucei* and *L. major*, in contrast to *T. brucei*, H3.V does not function to promote termination in *L. major*, but it is involved (potentially indirectly) in the maintenance of J levels.

Essential nature of base J in *Leishmania* spp.

Whereas *T. brucei* can survive without base J (38), *Leishmania* spp. appear to require J because *JBP1* KO cells are not viable (47). *JBP2* KO *Leishmania* spp. cells are viable, retaining approximately 30% of WT J levels (25,48). When *L. tarentolae JBP2* KO cells are grown in BrdU, they lose more J and die (48). As an explanation for this J-less death, the Borst lab proposed that the resulting massive read through transcription into the adjacent gene cluster interferes with sense RNA synthesis (25). To test this hypothesis and avoid the toxicity and indirect effects of BrdU, we utilized DMOG to inhibit thymine hydroxylase activity of the JBP enzymes and reduce J synthesis *in vivo* (31). DMOG treatment of WT *L. major* leads to significant reduction of J and read through transcription with minimal effect on parasite growth (26). Although the deletion of *H3.V* results in a significant loss of J, the loss is not sufficient for read through transcription or any measurable defect in cell growth. Compared to WT cells,

DMOG treatment of the *H3.V* KO cells leads to the greatest reduction of J (Figure 5.2C, Figure 5.3, and Supplementary Figure 5.S4C), more severe read through (Figure 5.2A and B, Figure 5.3, Figure 5.6F, Supplementary Figure 5.S4A and B, Supplementary Figure 5.S5, Supplementary Figure 5.S6, and Supplementary Figure 5.S7), and a more severe growth phenotype (Supplementary Figure 5.S8). Therefore, independent approaches to reduce J (BrdU treatment of a *JBP2* KO and DMOG treatment of an *H3.V* KO) have led to the observation that the growth defect following J loss in *Leishmania* spp. is inversely correlated with J levels and directly correlated with the degree of read through, strongly suggesting that the essential nature of J is due to its role in preventing read through transcription.

To address the essential nature of J, we therefore took advantage of the *H3.V* KO+DMOG cells to study the consequence of increased read through on gene expression. In contrast to the hypothesis that read through transcription disrupts sense mRNAs, our mRNA-seq analysis indicates only 20 genes are downregulated more than twofold in the *H3.V* KO+DMOG. Of the 20 downregulated genes, none were reduced by more than 2.4 fold and only four represent the last gene of a convergent gene cluster, where the greatest amount of read through from the opposing gene cluster would be expected. In fact, of the 76 convergent gene clusters in the *L. major* genome, only six have a significantly downregulated gene within 10kb of the termination site in the *H3.V* KO+DMOG. Clearly, the expression of convergently arranged genes in *L. major* is not significantly negatively correlated with the production of antisense RNAs by RNAP II read through transcription. Examination of the genome-wide transcriptome data from WT and the *H3.V* KO cells with reduced J does reveal slightly downregulated genes near several cSSRs with high read through, suggesting that a transcriptional interference/collision model could explain some of these modest reductions in gene expression.

While the changes are much less than our twofold significance cutoff, the downregulation correlates with levels of base J and degree of read through transcription. The lack of significantly downregulated genes at the ends of the majority of gene clusters genome-wide indicates that the process of read through transcription at cSSRs and subsequent dual strand transcription does not negatively affect gene expression at the mRNA level to any significant extent.

Rather than observing downregulation caused by J loss and transcriptional interference, we instead found that the majority of genes affected by J loss were upregulated. 21 of the 22 genes affected in WT+DMOG cells were upregulated. Similarly, 42 out of the 62 genes affected in *H3.V* KO+DMOG cells were upregulated. 28 of the genes upregulated following J loss are located within at least 10kb of base J, most of which are immediately downstream of J or overlap regions of J enrichment (Figure 5.5, 5.6, and Supplementary Table 5.1), and therefore fit a model in which J promotes termination prior to the gene and represses its expression. Upon J loss, RNAP II elongation continues and gives rise to stabilized processed mRNA. Consistent with this model, of the 21 genes upregulated in WT cells treated with DMOG, 18 are further upregulated in the *H3.V* KO+DMOG, thus, expression is inversely correlating with J levels. By strand-specific RT-PCR we are also able to detect nascent RNA spanning the termination site and (antisense) RNA downstream of the termination site that is similarly inversely correlated with J levels and directly correlated with increased mRNA levels of the downstream gene. These findings reveal a conserved function of base J among kinetoplastids in the promotion of RNAP II termination prior to the end of gene clusters (26,27,39).

Regardless of how J loss directly impacts gene expression, 62 genes are differentially expressed after DMOG treating *H3.V* KO cells, and it is possible that both decreases and

increases in transcript abundance contribute to defects in cell growth following J loss. We also cannot exclude the possibility that antisense RNAs generated by transcriptional read through are somehow detrimental to cell viability. These RNAs appear to be processed similarly to sense mRNAs, and although they do not significantly decrease sense mRNA abundance, it is possible that other stages of gene expression are deregulated by antisense RNAs, for example mRNA nuclear export and translation. Future studies are underway to clarify the essential nature of J in *L. major*. However, given that the major impact of J loss on transcript levels (in terms of the number of genes affected and degree of expression change) is the increased expression of genes at the end of gene clusters, the increasing growth defect with decreased J in *L. major* promastigotes is more likely linked to the de-repression of genes as opposed to the downregulation of mRNAs or uncharacterized transcriptional interference processes. While the majority of the upregulated genes encode for products of unknown function, approximately 50% of the genes are known to be developmentally regulated in *L. major* (44). Included are putative amastin-like surface proteins that are upregulated upon differentiation to metacyclic promastigotes. De-repression of these genes following the loss of J may lead to growth defects in our *in vitro L. major* promastigote cultures. These findings raise the possibility that regulation of J levels during different life stages or in response to environmental signals allows parasites to regulate gene expression at the transcriptional level. Developmental regulation of global levels of base J in kinetoplastids is consistent with this possibility (29,49), though high resolution mapping of J across life stages will be necessary to more thoroughly test this hypothesis. Further investigation of the proteins involved in regulating J synthesis, including JBP associated proteins, will also provide insight to the biological significance of J regulation of gene expression in kinetoplastids.

This study has thus identified a conserved function of base J in kinetoplastids in the promotion of RNAP II termination prior to the end of gene clusters and strongly implicates gene repression as the essential role of J in maintaining *L. major* parasite viability.

Funding

Research reported in this publication was supported by the National Institutes of Health (grant number R01AI109108) to RS; funding from the Office of the Vice President for Research to RJS; funding by the Human Frontier Science Program to TNS; the National Institutes of Health (grant number R01129646) to SMB; a T32 GM007067 fellowship and the Washington University Berg-Morse and Schlesinger Graduate Fellowships to BAA; and the American Heart Association grant 15PRE23240002 to DLR. Funding for open access charge: National Institute of Health R01AI109108.

Accession number

All sequencing data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession numbers GSE77713 and GSE77632.

Supplementary data

Supplementary Data are available at NAR online: Supplementary Tables S1-S2 and Supplementary Figures S1-S11.

Acknowledgement

We are grateful for critical reading of the manuscript by Rudo Kieft, Jessica Lopes da Rosa-Spiegler, Whitney Bullard, and Piet Borst. We also thank Nick Rohr for preparing mRNA-seq libraries.

REFERENCES

1. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B. *et al.* (2005) The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416-422.
2. Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R. *et al.* (2005) The genome of the kinetoplastid parasite, Leishmania major. *Science*, **309**, 436-442.
3. El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G. *et al.* (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science*, **309**, 409-415.
4. Johnson, P.J., Kooter, J.M. and Borst, P. (1987) Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*, **51**, 273-281.
5. Mottram, J.C., Murphy, W.J. and Agabian, N. (1989) A transcriptional analysis of the Trypanosoma brucei hsp83 gene cluster. *Mol. Biochem. Parasitol.*, **37**, 115-127.
6. Martinez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K. and Myler, P.J. (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Mol. Cell*, **11**, 1291-1299.

7. Martinez-Calvillo, S., Nguyen, D., Stuart, K. and Myler, P.J. (2004) Transcription initiation and termination on Leishmania major chromosome 3. *Eukaryotic cell*, **3**, 506-517.
8. Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S. and Cross, G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev.*, **23**, 1063-1076.
9. Thomas, S., Green, A., Sturm, N.R., Campbell, D.A. and Myler, P.J. (2009) Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*, **10**, 152.
10. Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S. and Tschudi, C. (2010) The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog.*, **6**, e1001090.
11. Boothroyd, J.C. and Cross, G.A. (1982) Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. *Gene*, **20**, 281-289.
12. Van der Ploeg, L.H., Liu, A.Y., Michels, P.A., De Lange, T.D., Borst, P., Majumder, H.K., Weber, H., Veeneman, G.H. and Van Boom, J. (1982) RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes. *Nucleic Acids Res*, **10**, 3591-3604.
13. De Lange, T., Liu, A.Y., Van der Ploeg, L.H., Borst, P., Tromp, M.C. and Van Boom, J.H. (1983) Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? *Cell*, **34**, 891-900.
14. Nelson, R.G., Parsons, M., Barr, P.J., Stuart, K., Selkirk, M. and Agabian, N. (1983) Sequences homologous to the variant antigen mRNA spliced leader are located in tandem repeats and variable orphons in Trypanosoma brucei. *Cell*, **34**, 901-909.

15. Sutton, R.E. and Boothroyd, J.C. (1986) Evidence for Trans splicing in trypanosomes. *Cell*, **47**, 527-535.

16. Agabian, N. (1990) Trans splicing of nuclear pre-mRNAs. *Cell*, **61**, 1157-1160.

17. Borst, P. (1986) Discontinuous transcription and antigenic variation in trypanosomes. *Annu. Rev. Biochem.*, **55**, 701-732.

18. LeBowitz, J.H., Smith, H.Q., Rusche, L. and Beverley, S.M. (1993) Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev.*, **7**, 996-1007.

19. Clayton, C.E. (2002) Life without transcriptional control? From fly to man and back again. *EMBO J*, **21**, 1881-1888.

20. Campbell, D.A., Thomas, S. and Sturm, N.R. (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.*, **5**, 1231-1240.

21. Respuela, P., Ferella, M., Rada-Iglesias, A. and Aslund, L. (2008) Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. *J. Biol. Chem.*, **283**, 15884-15892.

22. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A. and Sabatini, R. (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res*, **38**, 3923-3935.

23. Ekanayake, D. and Sabatini, R. (2011) Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryotic cell*, **10**, 1465-1472.

24. Ekanayake, D.K., Minning, T., Weatherly, B., Gunasekera, K., Nilsson, D., Tarleton, R., Ochsenreiter, T. and Sabatini, R. (2011) Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. *Mol. Cell. Biol.*, **31**, 1690-1700.

25. van Luenen, H.G., Farris, C., Jan, S., Genest, P.A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G. *et al.* (2012) Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, **150**, 909-921.

26. Reynolds, D., Cliffe, L., Forstner, K.U., Hon, C.C., Siegel, T.N. and Sabatini, R. (2014) Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res*, **42**, 9717-9729.

27. Reynolds, D., Hofmeister, B.T., Cliffe, L., Alabady, M., Siegel, T.N., Schmitz, R.J. and Sabatini, R. (2016) Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005758.

28. Gommers-Ampt, J.H., Van Leeuwen, F., de Beer, A.L., Vliegenthart, J.F., Dizdaroglu, M., Kowalak, J.A., Crain, P.F. and Borst, P. (1993) beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. *Cell*, **75**, 1129-1136.

29. van Leeuwen, F., Taylor, M.C., Mondragon, A., Moreau, H., Gibson, W., Kieft, R. and Borst, P. (1998) beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. *Proc. Natl. Acad. Sci. U S A*, **95**, 2366-2371.

30. Dooijes, D., Chaves, I., Kieft, R., Dirks-Mulder, A., Martin, W. and Borst, P. (2000) Base J originally found in kinetoplastid is also a minor constituent of nuclear DNA of Euglena gracilis. *Nucleic Acids Res*, **28**, 3017-3021.
31. Cliffe, L.J., Hirsch, G., Wang, J., Ekanayake, D., Bullard, W., Hu, M., Wang, Y. and Sabatini, R. (2012) JBP1 and JBP2 Proteins Are Fe2+/2-Oxoglutarate-dependent Dioxygenases Regulating Hydroxylation of Thymidine Residues in Trypanosome DNA. *J. Biol. Chem.*, **287**, 19886-19895.
32. Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y. and Sabatini, R. (2014) Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. *J. Biol. Chem.*, **289**, 20273-20282.
33. Sekar, A., Merritt, C., Baugh, L., Stuart, K. and Myler, P.J. (2014) Tb927.10.6900 encodes the glucosyltransferase involved in synthesis of base J in Trypanosoma brucei. *Mol. Biochem. Parasitol.*, **196**, 9-11.
34. Borst, P. and Sabatini, R. (2008) Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol*, **62**, 235-251.
35. Reynolds, D., Cliffe, L. and Sabatini, R. (2015) *2-Oxoglutarate-dependent hydroxylases involved in DNA base J (β-D-Glucopyranosyloxymethyluracil) synthesis.* In Schofield, C. J. and Hausinger, R. P. (eds.), *2-Oxoglutarate-Dependent Oxygenases*. Royal Society of Chemistry, Cambridge, U.K, Vol. 1.
36. Iyer, L.M., Tahiliani, M., Rao, A. and Aravind, L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698-1710.

37. Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930-935.

38. Cliffe, L.J., Kieft, R., Southern, T., Birkeland, S.R., Marshall, M., Sweeney, K. and Sabatini, R. (2009) JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res*, **37**, 1452-1462.

39. Schulz, D., Zaringhalam, M., Papavasiliou, F.N. and Kim, H.-S. (2016) Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005762.

40. Anderson, B.A., Wong, I.L., Baugh, L., Ramasamy, G., Myler, P.J. and Beverley, S.M. (2013) Kinetoplastid-specific histone variant functions are conserved in Leishmania major. *Mol. Biochem. Parasitol.*, **191**, 53-57.

41. Kapler, G.M., Coburn, C.M. and Beverley, S.M. (1990) Stable transfection of the human parasite Leishmania major delineates a 30-kilobase region sufficient for extrachromosomal replication and expression. *Mol. Cell. Biol.*, **10**, 1084-1094.

42. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechol.*, **31**, 46-53.

43. Roditi, I., Schwarz, H., Pearson, T.W., Beecroft, R.P., Liu, M.K., Richardson, J.P., Bühring, H.J., Pleiss, J., Bülow, R. and Williams, R.O. (1989) Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. *J. Cell. Biol.*, **108**, 737-746.

44. Dillon, Laura A.L., Okrah, K., Hughitt, V.K., Suresh, R., Li, Y., Fernandes, M.C., Belew, A.T., Corrada Bravo, H., Mosser, D.M. and El-Sayed, N.M. (2015) Transcriptomic profiling

of gene expression and RNA processing during Leishmania major differentiation. *Nucleic Acids Res*, **43**, 6799-6813.

45. Tschudi, C., Fau, S.H. and Ullu, E. (2012) Small interfering RNA-producing loci in the ancient parasitic eukaryote Trypanosoma brucei. *BMC Genomics*, **13**.
46. Zheng, L.L., Wen, Y.Z., Yang, J.H., Liao, J.Y., Shao, P., Xu, H., Zhou, H., Wen, J.Z., Lun, Z.R., Ayala, F.J. *et al.* (2013) Comparative transcriptome analysis of small noncoding RNAs in different stages of Trypanosoma brucei. *RNA.*, **19**, 863-875.
47. Genest, P.A., ter Riet, B., Dumas, C., Papadopoulou, B., van Luenen, H.G. and Borst, P. (2005) Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. *Nucleic Acids Res*, **33**, 1699-1709.
48. Vainio, S., Genest, P.A., ter Riet, B., van Luenen, H. and Borst, P. (2009) Evidence that J-binding protein 2 is a thymidine hydroxylase catalyzing the first step in the biosynthesis of DNA base J. *Mol. Biochem. Parasitol.*, **164**, 157-161.
49. Ekanayake, D.K., Cipriano, M.J. and Sabatini, R. (2007) Telomeric co-localization of the modified base J and contingency genes in the protozoan parasite Trypanosoma cruzi. *Nucleic Acids Res*, **35**, 6367-6377.

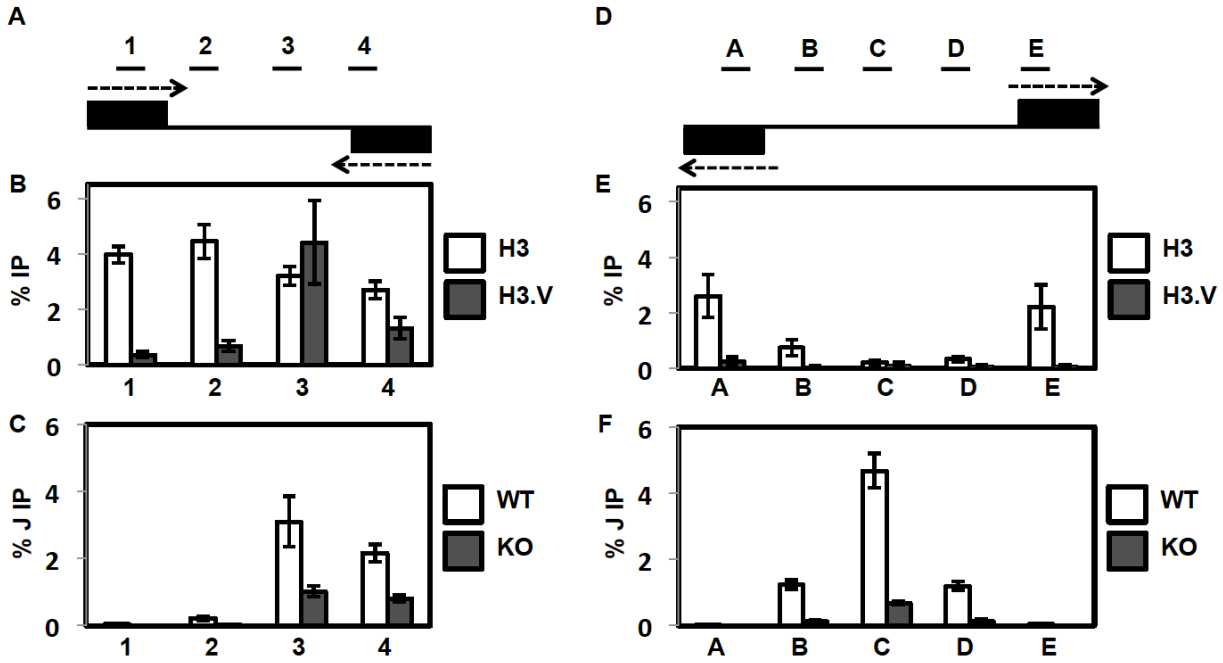


Figure 5.1 H3.V co-localizes with base J at cSSRs and regulates J synthesis.

(A) Map of cSSR 12.1 (12.1 indicates the first termination site on chromosome 12, following nomenclature established by van Luenen et al. (25). Genomic coordinates for all *L. major* cSSRs are listed in Reynolds et al. (26)). Black boxes represent the final gene in the opposing gene clusters. Arrows indicate the direction of transcription. Lines above the map indicate PCR amplified regions 1-4 utilized in the ChIP-qPCR analyses. Not to scale.

(B) Localization of H3 and H3.V across cSSR 12.1 in WT cells determined by ChIP-qPCR using H3 antisera and H3.V antisera. The average of three independent IPs is plotted as the percent IP relative to the total input material. All IPs were background subtracted using a no antibody control. White bars, H3 ChIP; dark grey bars, H3.V ChIP. Error bars represent the standard deviation.

(C) J localization across a cSSR in WT and *H3.V* KO cells. Anti-base J IP-qPCR analysis was performed for regions 1-4 within cSSR 12.1 of the indicated cell lines. The peak of J and the TTS have been shown to be within region 3 (25,26). The average of three independent IPs is plotted as the percent IP relative to the total input material. All IPs were background subtracted using a no antibody control. White bars, WT; dark grey bars, *H3.V* KO. Error bars represent the standard deviation.

(D) Map of dSSR 6.1 (6.1 indicates the first initiation site on chromosome 6, following nomenclature established by van Luenen et al. (25)). Black boxes represent the initial gene in the opposing gene clusters. Arrows indicate the direction of transcription. Lines above the map indicate PCR amplified regions A-E utilized in the ChIP-qPCR analyses. Not to scale.
(E) Localization of H3 and H3.V across dSSR 6.1 in WT cells. H3 and H3.V ChIP-qPCR were performed as described in B.
(F) J localization across a dSSR in WT and *H3.V* KO cells. Anti-base J IP-qPCR analysis was performed for regions A-E within dSSR as described in C.

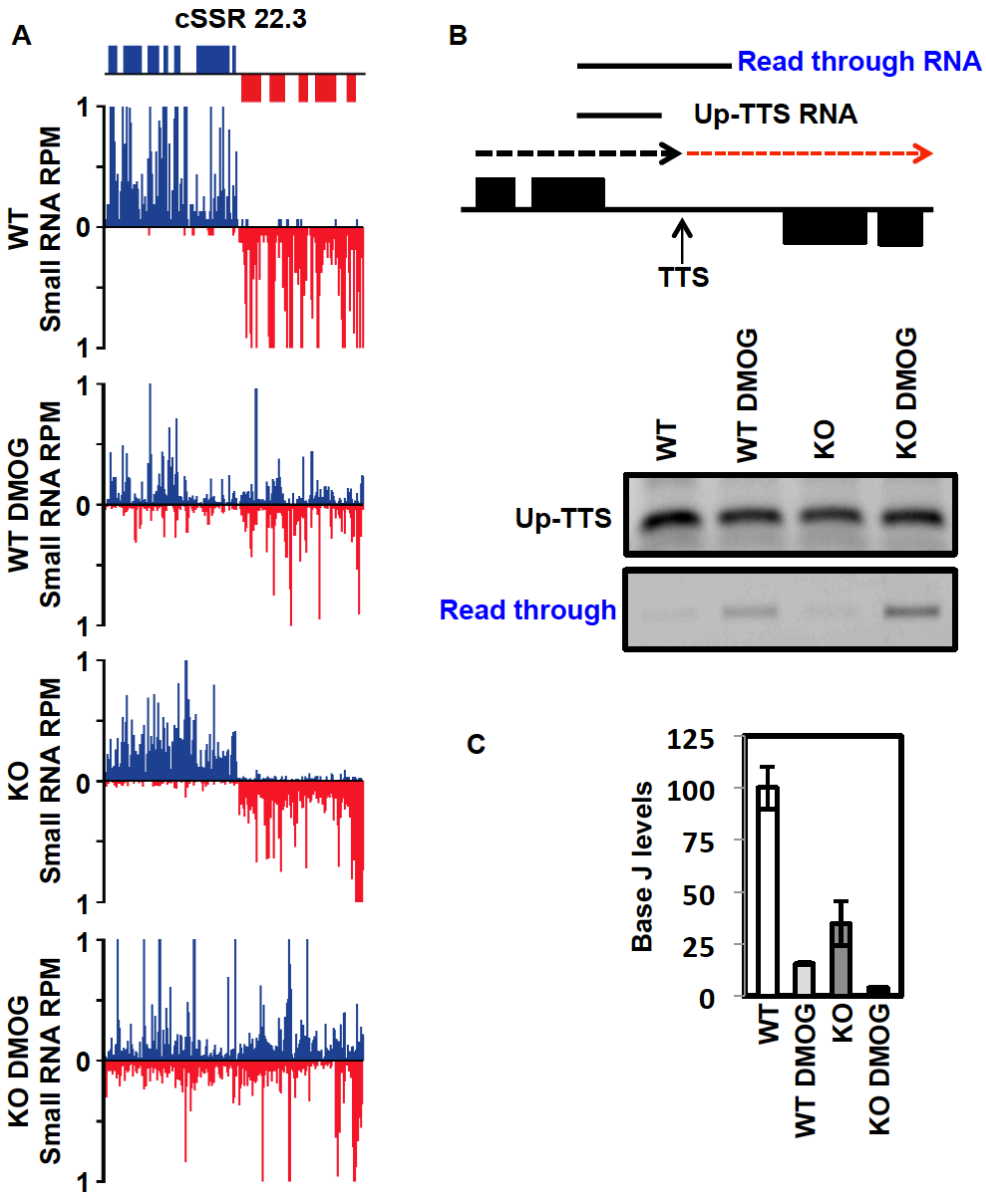


Figure 5.2 H3.V does not promote transcription termination in *L. major*.

(A) Small RNA-seq reads for a representative cSSR (22.3) are shown where J loss leads to read through transcription, but loss of H3.V does not. Small RNA reads are plotted as reads per million reads mapped (RPM). ORFs are shown above the graphs. The genomic location shown includes position 488-528kb on chromosome 22. WT: wild type; WT DMOG; WT+DMOG; KO: *H3.V* KO; KO DMOG: *H3.V* KO+DMOG. Blue: top strand; red: bottom strand.

(B) Strand-specific RT-PCR analysis. Above, schematic representation (not to scale) of cSSR 22.3 illustrating the nascent RNA species assayed by RT-PCR. The dashed red arrow indicates read through transcription past the transcription termination site (TTS). Up-TTS: RNA species

upstream of the TTS; Read through RNA: RNA species resulting from read through transcription. Below, cSSR 22.3 was analyzed in WT and *H3.V* KO cells in the absence and presence of 5mM DMOG.

(C) The amount of read through transcription correlates with the extent of J loss. The levels of J at the TTS in cSSR 22.3 was analyzed by anti-J IP-qPCR for the indicated cell lines as described in Figure 5.1C. WT level is set to 100%.

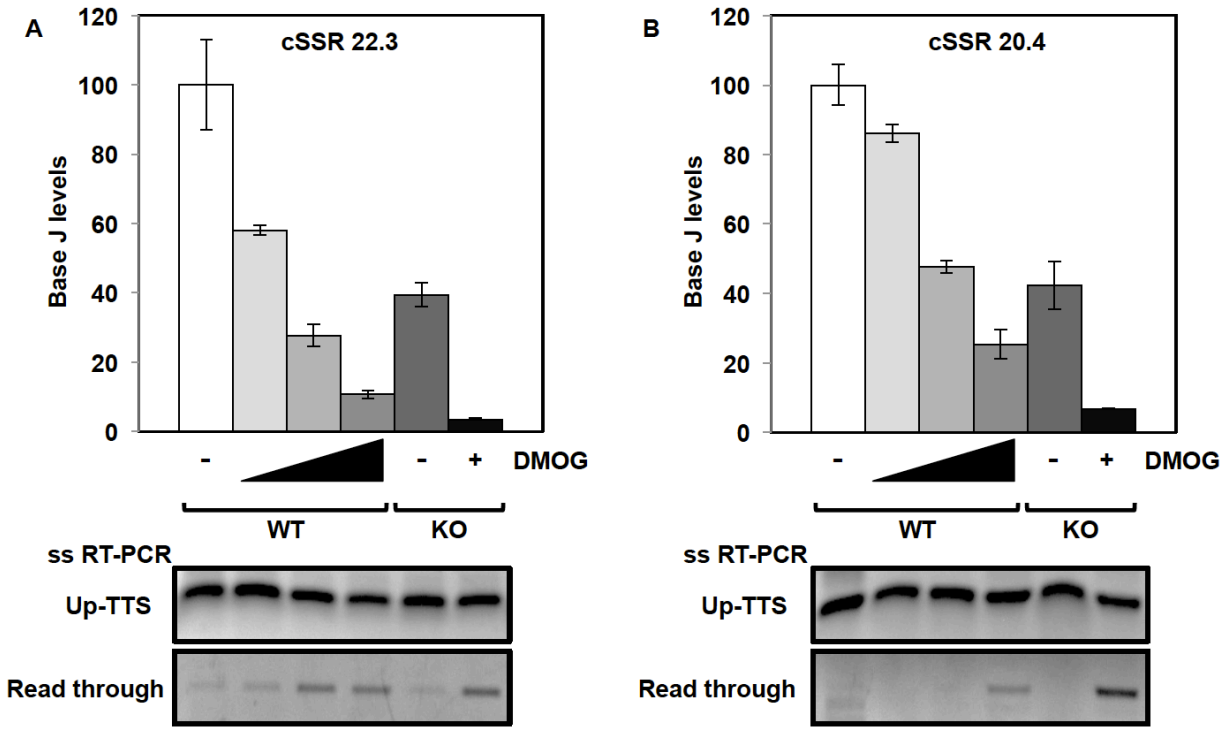


Figure 5.3 Levels of J remaining in the *H3.V* KO are sufficient for terminating transcription.

(A and B) The levels of J at the TTS of two representative cSSRs were determined by J IP-qPCR analysis in WT cells with increasing DMOG concentrations (0, 1, 2.5, and 5mM) and in *H3.V* KO cells in the absence and presence of 5mM DMOG. Below, strand-specific RT-PCR analysis of each of the cSSRs as described in Figure 5.2B.

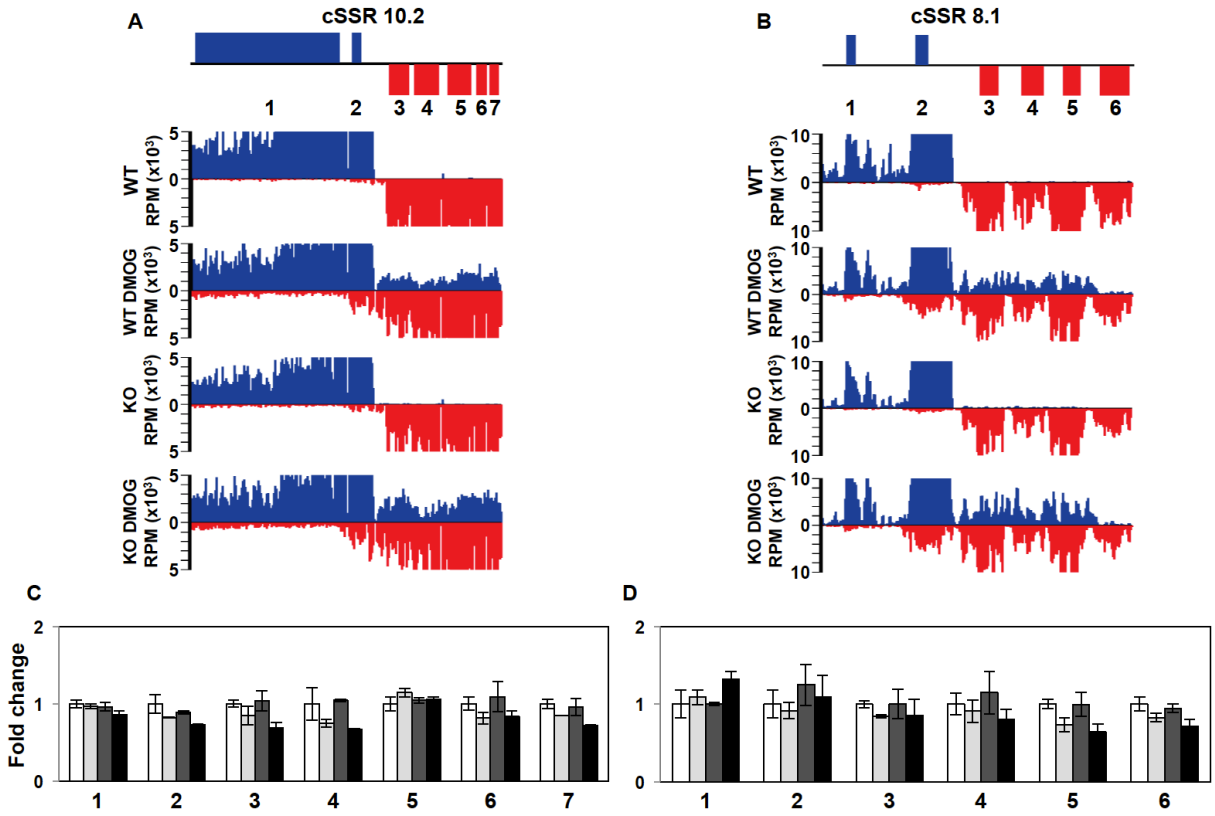


Figure 5.4 Read through transcription does not lead to transcriptional interference.

(A and B) Strand-specific mRNA-seq reads for two representative cSSRs where loss of J leads to high degree of read through transcription into the adjacent gene cluster are shown. The genomic region shown is from 243-287kb on chromosome 10 in (A) and from 384-406kb on chromosome 8 in (B).

(C and D) Plot of the mRNA-seq data for the genes indicated (numbered) in the ORF map for the corresponding regions above. The average RPKM of mRNA-seq replicate libraries was used to determine the fold changes, with WT set to one. Error bars indicate the standard deviation between mRNA-seq replicates. White bars: Wild type; grey bars: Wild type+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. None of the genes shown are significantly differentially expressed relative to WT except for gene 5 (*LmjF.08.0890*) in D in the *H3.V* KO+DMOG condition, as determined by Cuffdiff (p-value of 0.03), but gene 5 is only downregulated by 1.6 fold, which does not meet our twofold cutoff.

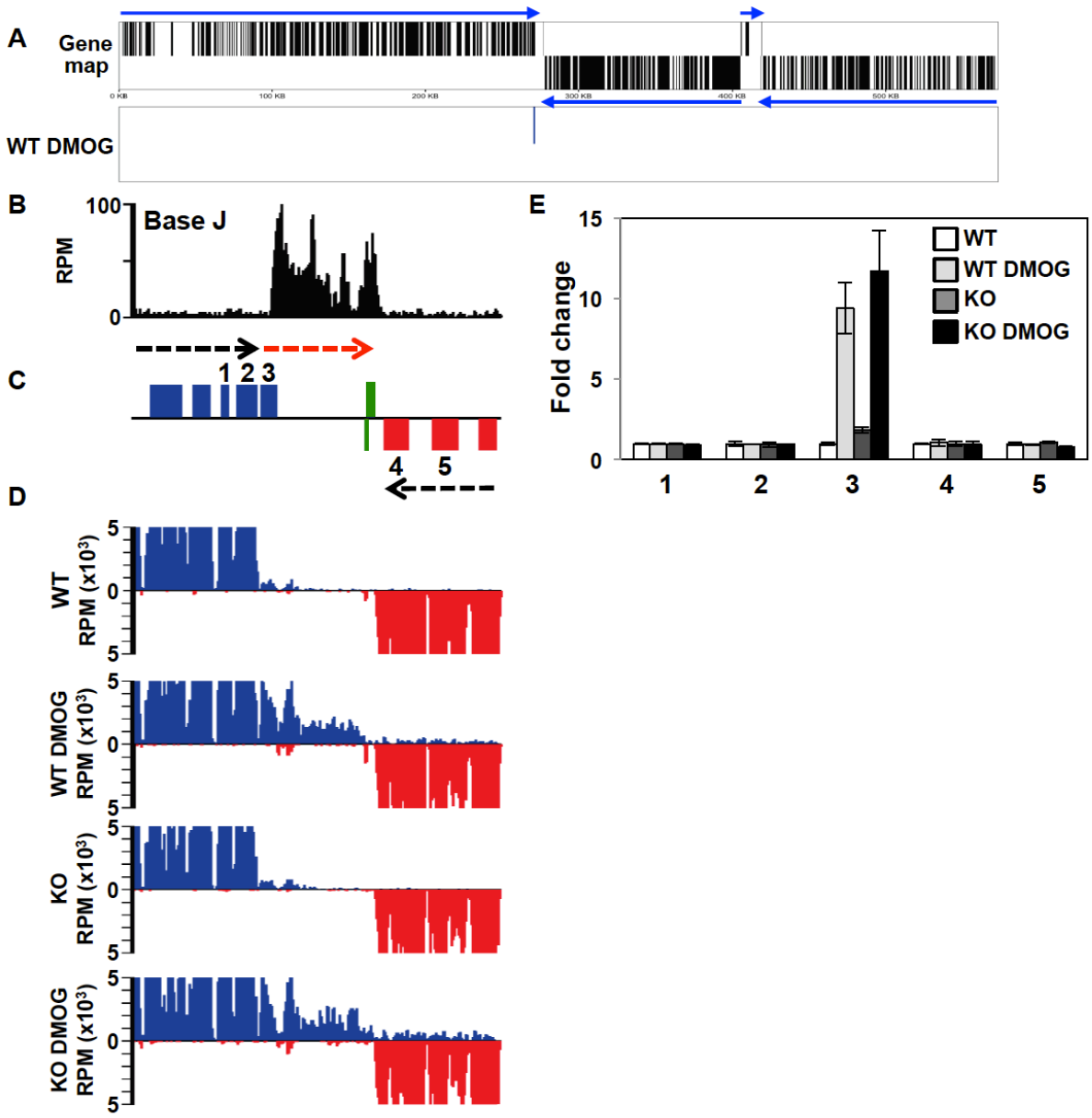


Figure 5.5 Decreased efficiency of RNAP II termination and increased gene expression following the loss of J.

(A) Gene map of chromosome 9 is shown. mRNA coding genes on the top strand are indicated by black lines in the top half of the panel, bottom strand by a line in the bottom half. Genes on the top strand are transcribed from left to right and those on the bottom strand are transcribed from right to left, indicated by blue arrows. Panel below (WT DMOG) indicates the location of

the single mRNA found upregulated by at least twofold or more in WT cells treated with DMOG relative to WT. No other expression changes (up or downregulated) were detected. (B-D) A region on chromosome 9 from 263-285kb where J regulates transcription termination and gene expression at a cSSR is shown. (B) Base J localizes at the site of RNAP II termination within the gene cluster (prior to the last gene before the cSSR). Base J IP-seq reads are plotted as reads per million (RPM), as previously described (27). (C) ORFs are shown with the top strand in blue and the bottom strand in red. The red arrow indicates read through transcription following the loss of J. Green boxes indicate RNAP III transcribed genes (tRNA and 5S rRNA). (D) mRNA-seq reads from the indicated cell lines are mapped as described in Figure 5.4A and B.

(E) Plot of the mRNA-seq data for the genes numbered in the ORF map in panel B, as described in Figure 5.4C and D. The upregulated gene, 3, is *LmjF.09.0690*.

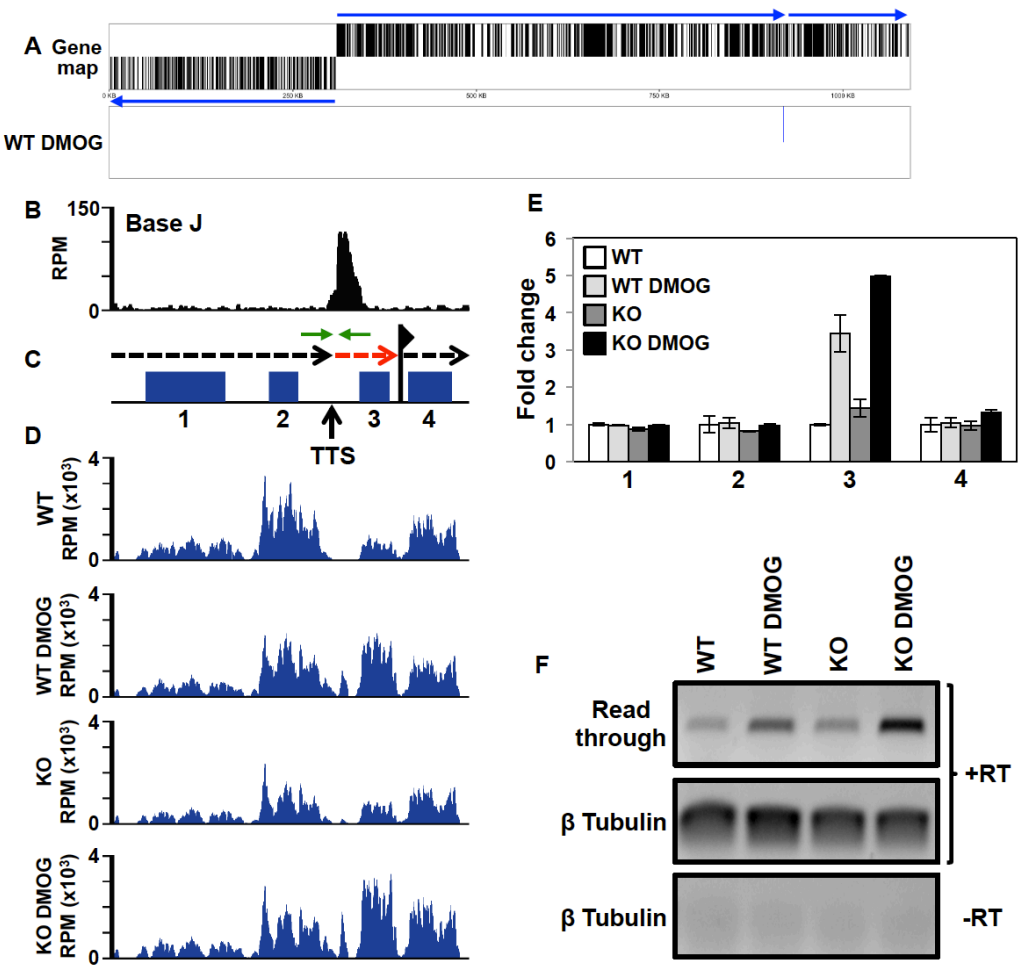


Figure 5.6 Base J regulates RNAP II termination and gene expression at head-tail regions within gene clusters.

(A) Gene map of chromosome 26 is shown where loss of J leads to upregulation of a single mRNA at the end of a gene cluster at a head-tail region. Labeling is as described in Figure 5.5A. (B-D) Base J IP-seq reads, ORFs, and mRNA-seq reads are plotted for the head-tail region on chromosome 26 from 912-922kb, as described in Figure 5.5. (B) Base J localizes at the transcription termination site (TTS). The vertical arrow indicates the proposed TTS as described in the text (26,27). The black dashed arrow above the map indicates the direction of transcription and the dashed red arrow indicates read through transcription past the TTS. The flag indicates the transcription start site for the downstream gene cluster as indicated by H3 acetylation localization (9). Green arrows flanking the TTS represent the PCR oligos utilized in strand-specific RT-PCR.

(E) Plot of the mRNA-seq data for the genes numbered in the ORF map in panel B, as described in Figure 5.4C and D. The upregulated gene, 3, is *LmjF.26.2280*.
(F) Strand-specific RT-PCR analysis of read through transcription at the TTS analyzed in B-E. Above the gene map in panel B is a schematic representation (not to scale) of primer location and direction at the TTS. cDNA was synthesized using the reverse primer (relative to transcription) and RNA from WT and *H3.V* KO cells treated with and without DMOG. Read through was detected by PCR using the same reverse primer used to make the cDNA plus the forward primer, as indicated. Tubulin provides a positive control and a minus RT (-RT) negative control is shown.

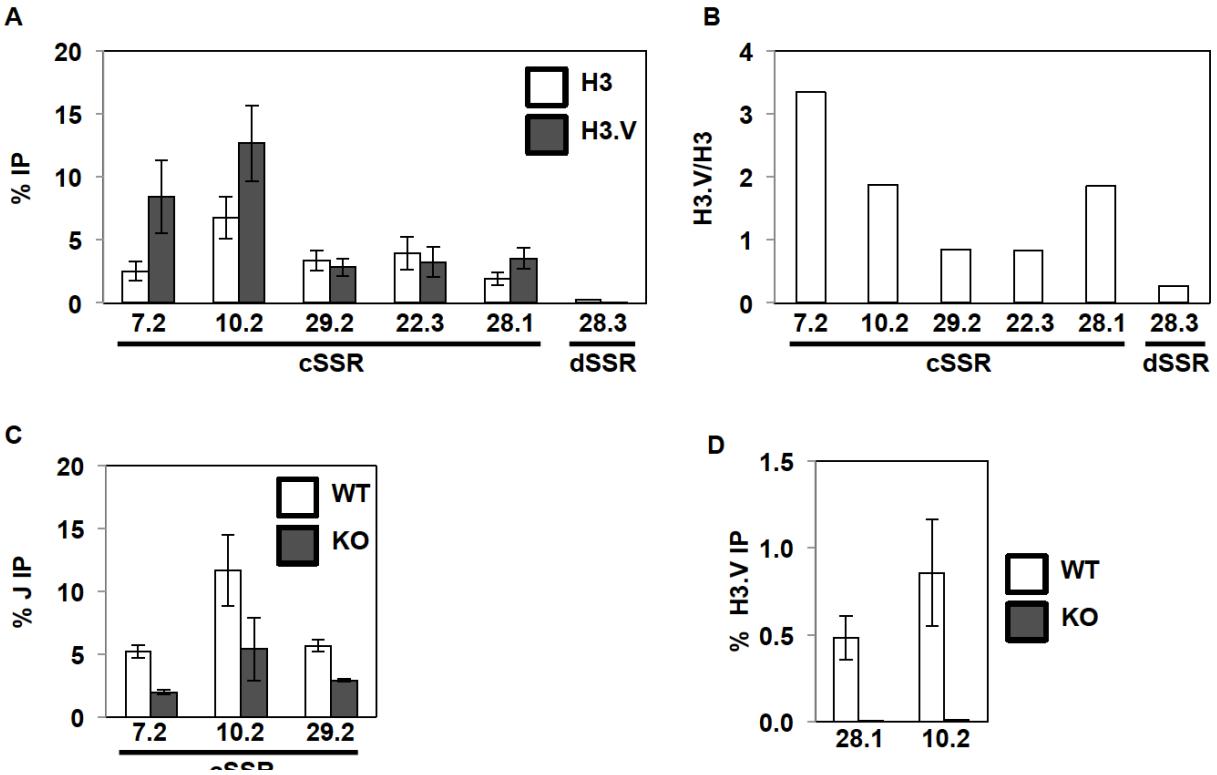


Figure 5.S1 H3.V is enriched at cSSRs and its loss reduces J levels.

(A) H3 (white bars) and H3.V (dark grey bars) ChIP-qPCR analysis of the indicated SSRs, as described in Figure 5.1B.

(B) ChIP-qPCR analysis in (A) was used to plot the ratio of H3.V:H3 at the indicated SSRs.

(C) J IP-qPCR analysis of the indicated cSSRs in WT and *H3.V* KO cells, as described in Figure 5.1C.

(D) H3.V ChIP-qPCR analysis of the indicated cSSRs in WT and *H3.V* KO cells, as described in Figure 5.1B. H3.V ChIP in the *H3.V* KO cells is barely detectable, indicating the specificity of the H3.V antibody.

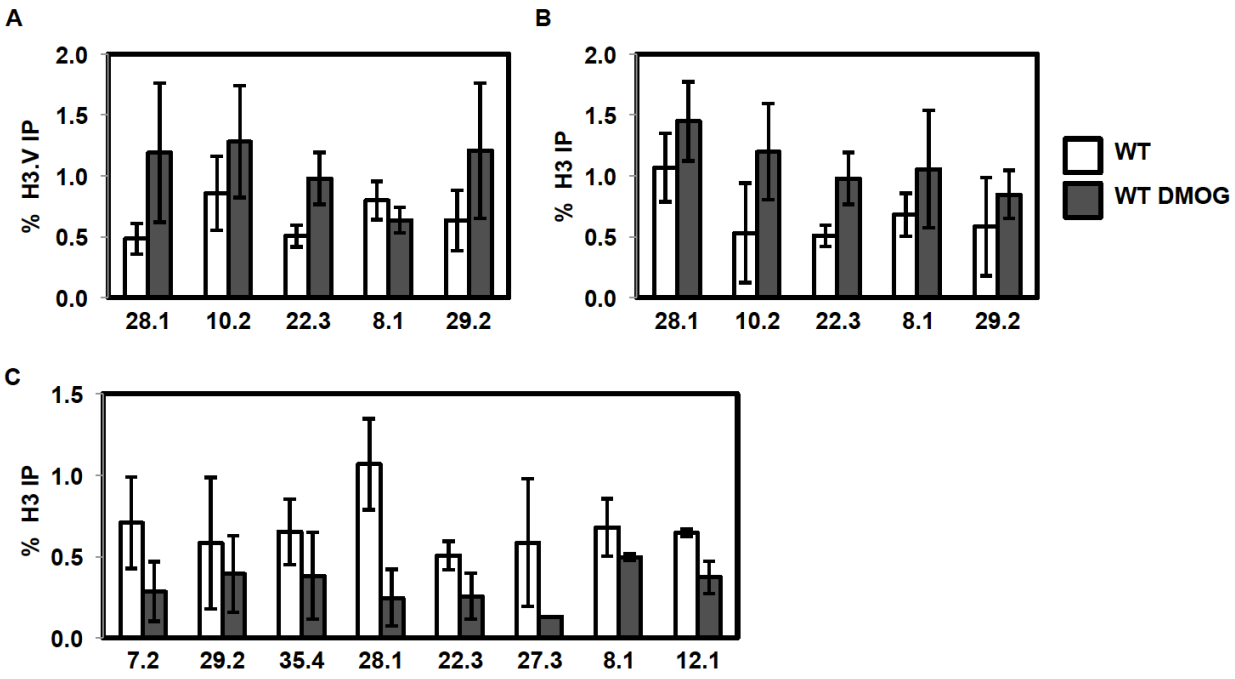


Figure 5.S2 H3 and H3.V levels in cSSRs upon the loss of J and H3.V.

(A and B) H3.V (A) and H3 (B) ChIP-qPCR analysis of the indicated cSSRs in WT cells treated with (dark grey bars) and without (white bars) DMOG. Data are shown as described in Figure 5.1B.

(C) H3 ChIP-qPCR analysis of the indicated cSSRs in WT and *H3.V* KO cells, as described in Figure 5.1B. The decrease in H3 is statistically significant at cSSR 28.1 and 12.1, with p-value<0.05, as determined by two-tailed Student's t-test.

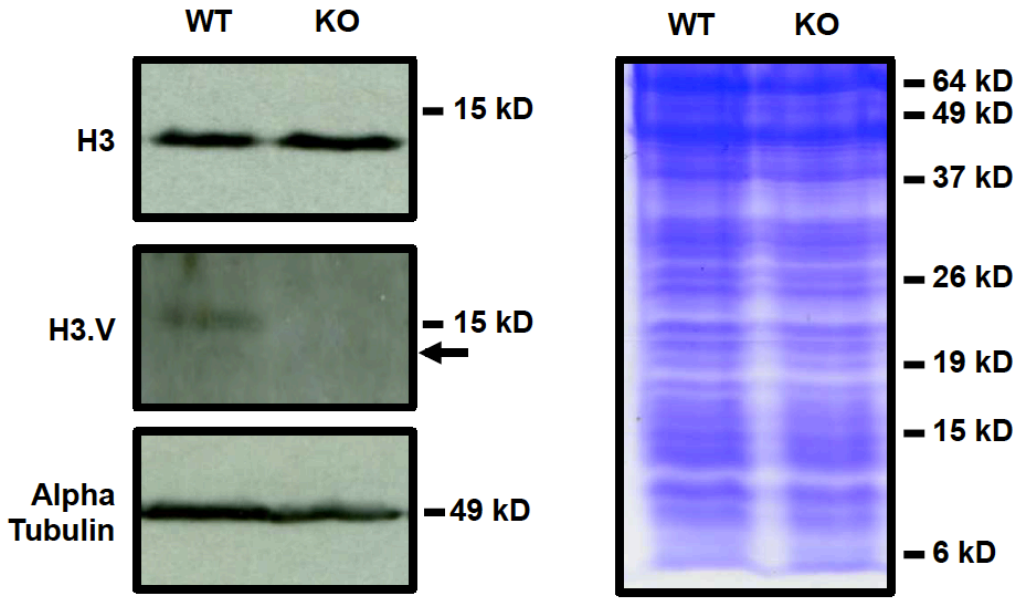


Figure 5.S3 H3 and H3.V protein levels in the *H3.V* KO.

Left; anti-H3, H3.V, and alpha tubulin Western blots using total protein lysates from WT and *H3.V* KO cells. Size markers are indicated on the right in kilodaltons (kD). Histone H3 runs just under 15kD, whereas H3.V runs at 15kD. The black arrow on the H3.V Western blot indicates the size where H3 runs. Right; Coomassie blue staining of WT and *H3.V* KO total protein lysates. Size markers are indicated on the right.

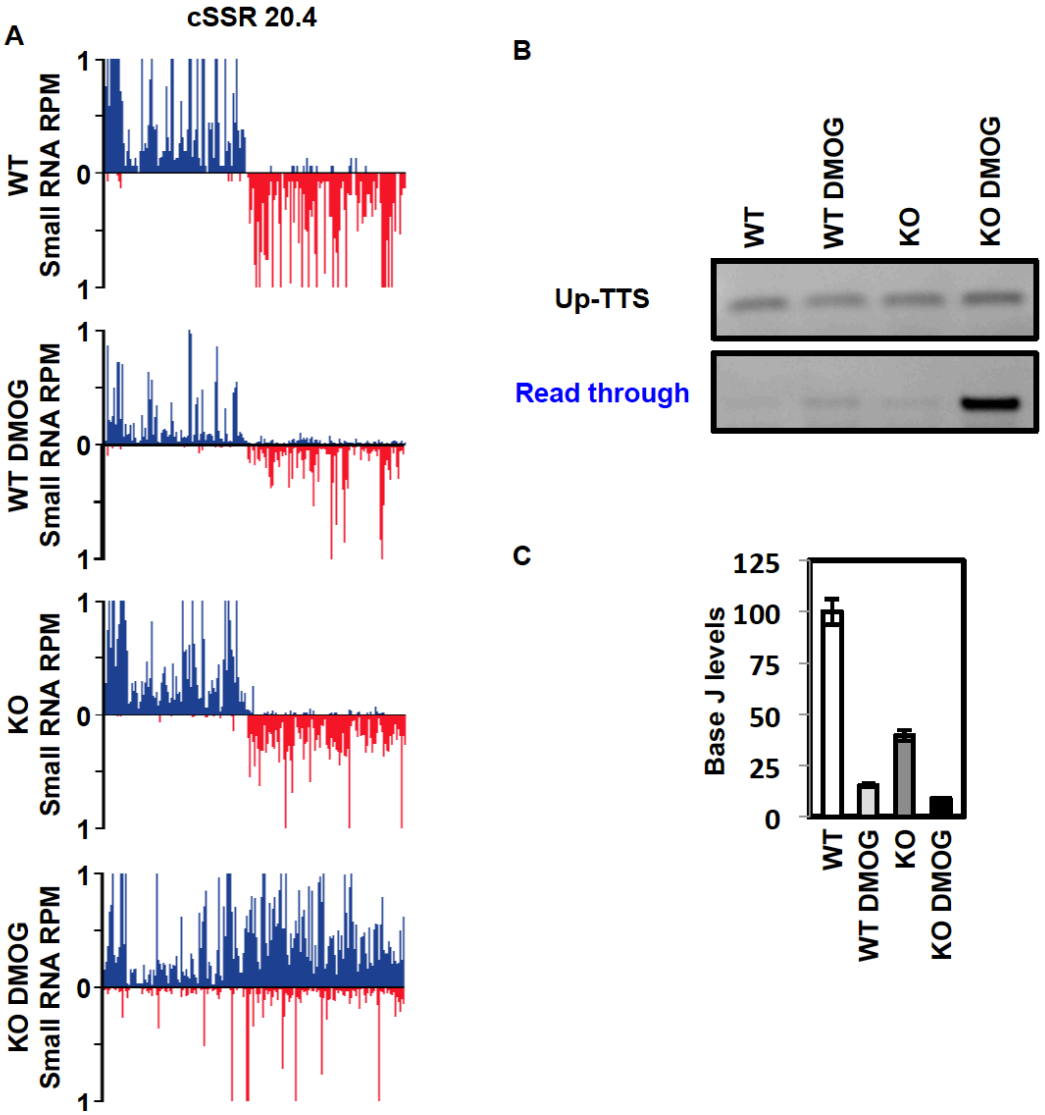


Figure 5.S4 H3.V does not regulate transcription termination.

(A) Small RNA-seq reads for cSSR 20.4 are shown where J loss leads to read through transcription, but loss of H3.V does not, as described in Figure 5.2. Small RNA reads are plotted as reads per million reads mapped (RPM). The genomic location shown includes position 637-677kb on chromosome 20. WT: wild type; WT DMOG; WT+DMOG; KO: *H3.V* KO; KO DMOG: *H3.V* KO+DMOG. Blue: top strand; red: bottom strand.

(B) Strand-specific RT-PCR analysis of read through transcription as described in Figure 5.2B.

(C) The levels of J at the TTS in the cSSR was analyzed by anti-J IP-qPCR for the indicated cell lines as described in Figure 5.1C.

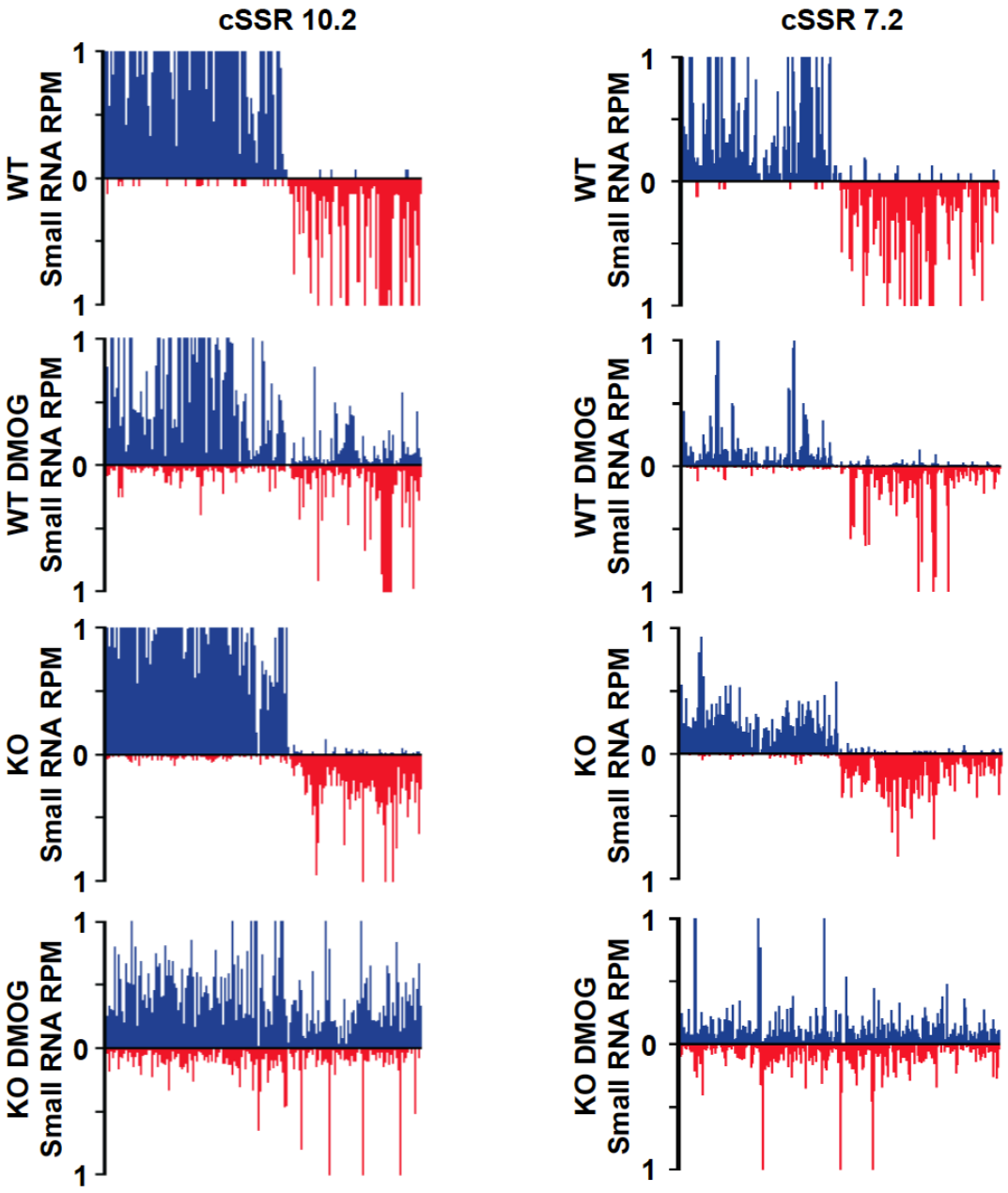


Figure 5.S5 H3.V does not promote transcription termination.

Small RNA-seq reads are shown for two additional cSSRs, as described in Figure 5.2. The region plotted for cSSR 10.2 (left) is from 246-286kb on chromosome 10 and cSSR 7.2 (right) is from 40-80kb on chromosome 7.

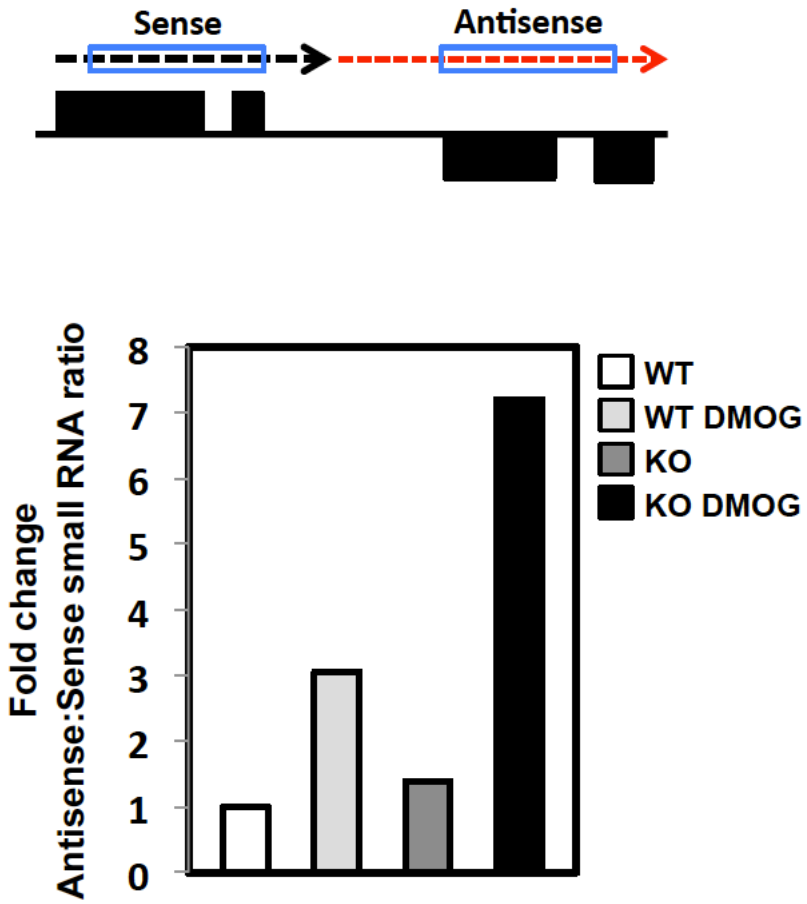


Figure 5.S6 Quantitation of read through at cSSRs by small RNA-seq.

Top: Diagram of a cSSR (not to scale), with genes indicated by black boxes. The black arrow indicates the direction of transcription on the top strand and the red arrow indicates read through transcription past the termination site. Blue boxes show the 5kb windows flanking the cSSR used to determine the antisense:sense small RNA-seq RPM ratio. 5kb windows were set using the genomic positions for all cSSRs in the *L. major* genome (coordinates are listed in (26)). Analysis of small RNA-seq reads on the bottom strand was performed similarly. Bottom: The antisense:sense small RNA-seq RPM ratio was determined at all cSSRs (excluding cSSR 9.2, in which case the flanking gene cluster was less than 5kb) and WT (white bar) was set to one. Fold change relative to WT in WT+DMOG (light grey), *H3.V* KO (dark grey), and *H3.V* KO+DMOG (black) cells is plotted.

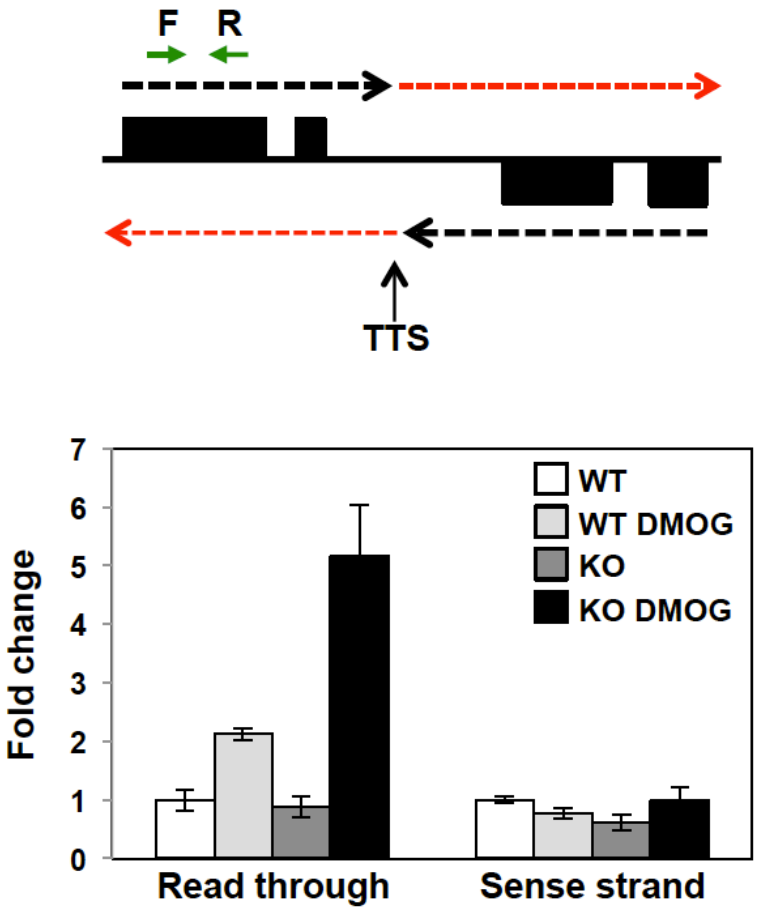


Figure 5.S7 Strand-specific RT-qPCR analysis of read through.

Top: a diagram of cSSR 22.3. Black boxes represent genes, black arrows indicate the direction of transcription, and red arrows indicate read through transcription past the transcription termination site (TTS), as identified by small RNA-seq. Green arrows indicate the location of primers utilized for strand-specific RT-qPCR analysis. Diagram not to scale. Bottom: Strand-specific RT-qPCR. Read through transcription on the bottom strand was quantitated by performing site-specific cDNA synthesis using primer F illustrated in the diagram above, followed by qPCR using primers F and R. Abundance was normalized using beta tubulin (a gene specific primer against beta tubulin was added to the same cDNA synthesis reaction with primer F, followed by qPCR using beta tubulin primers). White bars: WT; light grey bars: WT+DMOG; dark grey bars: *H3.V* KO; black bars: *H3.V* KO+DMOG. Data are plotted as fold change, with WT set to one. Strand-specific RT-qPCR was performed in triplicate, with error bars indicating standard deviation. RNA from the sense strand was similarly quantitated by generating cDNA using primer R, followed by qPCR with primers F and R.

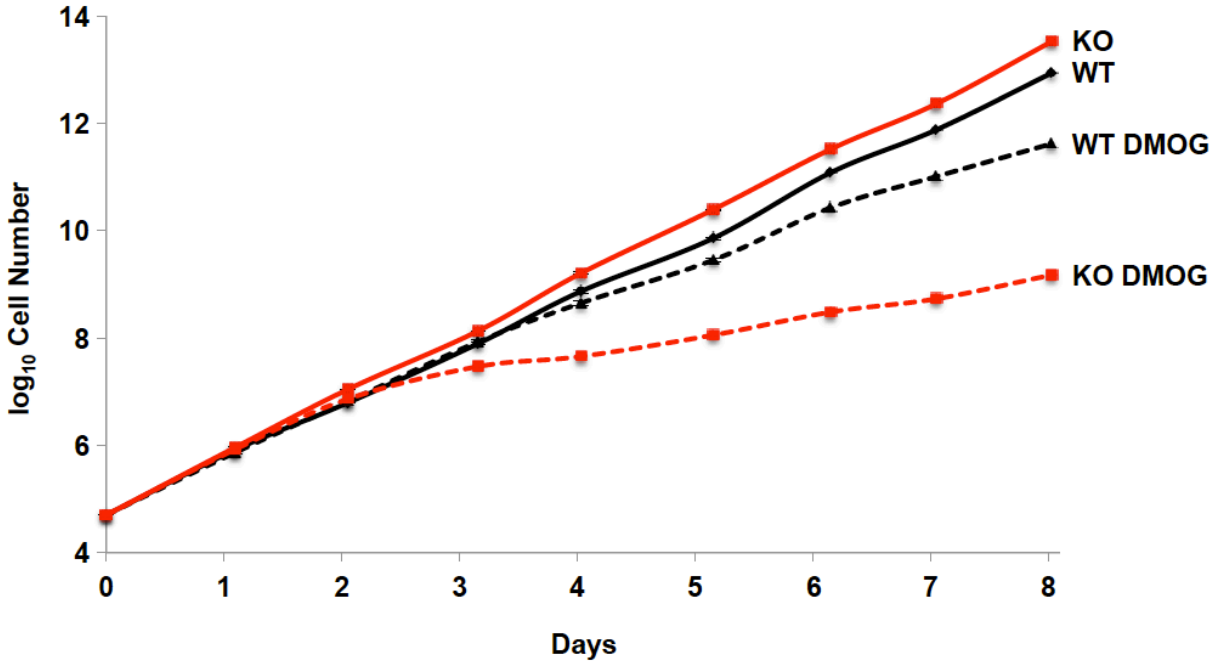


Figure 5.S8 Growth defect following the loss of base J in *L. major*.

Density of *L. major* WT and *H3.V* KO cells treated with and without 5mM DMOG was followed for 8 days. Error bars represent standard deviation of two biological replicates.

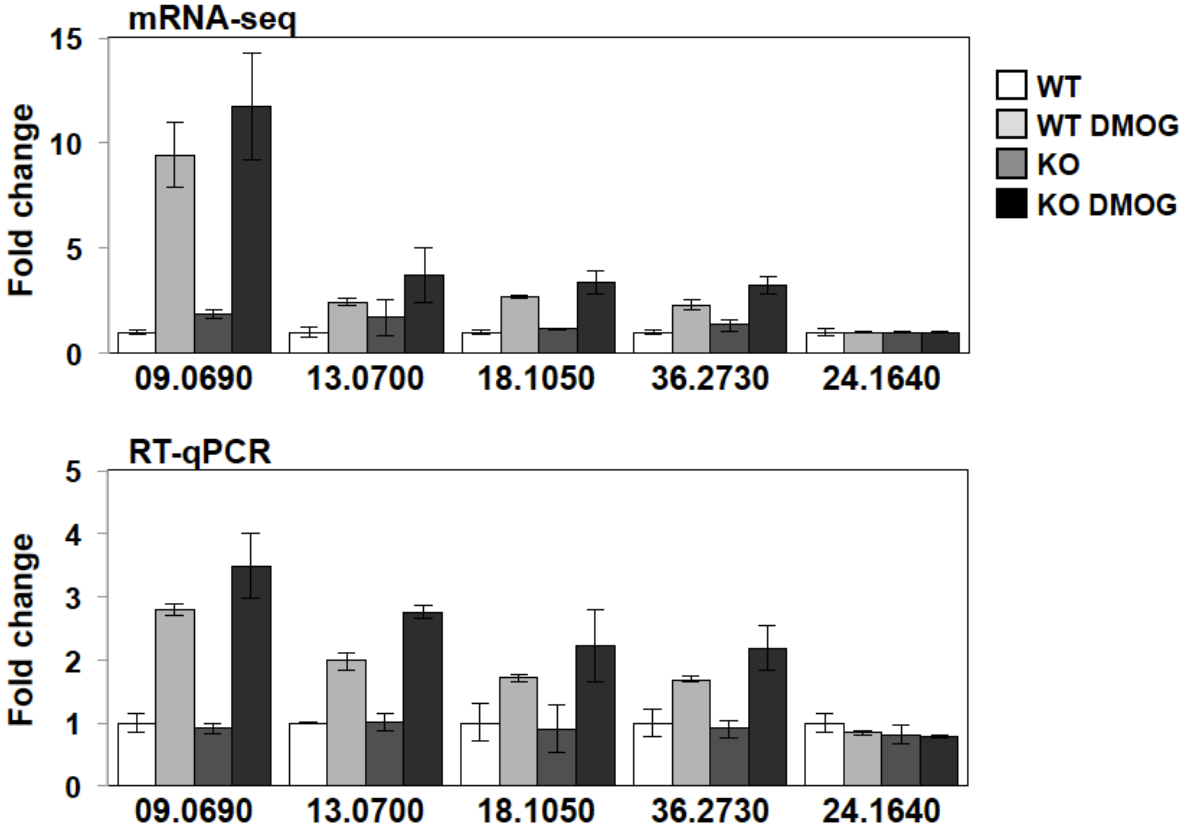


Figure 5.S9 Confirmation of mRNA-seq transcript changes by RT-qPCR.

Top; mRNA-seq data was plotted for the indicated genes, as described in Figure 5.4C and D. Bottom; RT-qPCR analysis was performed for the indicated genes as described in Material and Methods. Transcripts were normalized against beta tubulin mRNA, and are plotted as the average and standard deviation of three replicates. Gene 24.1640 represents a negative control.

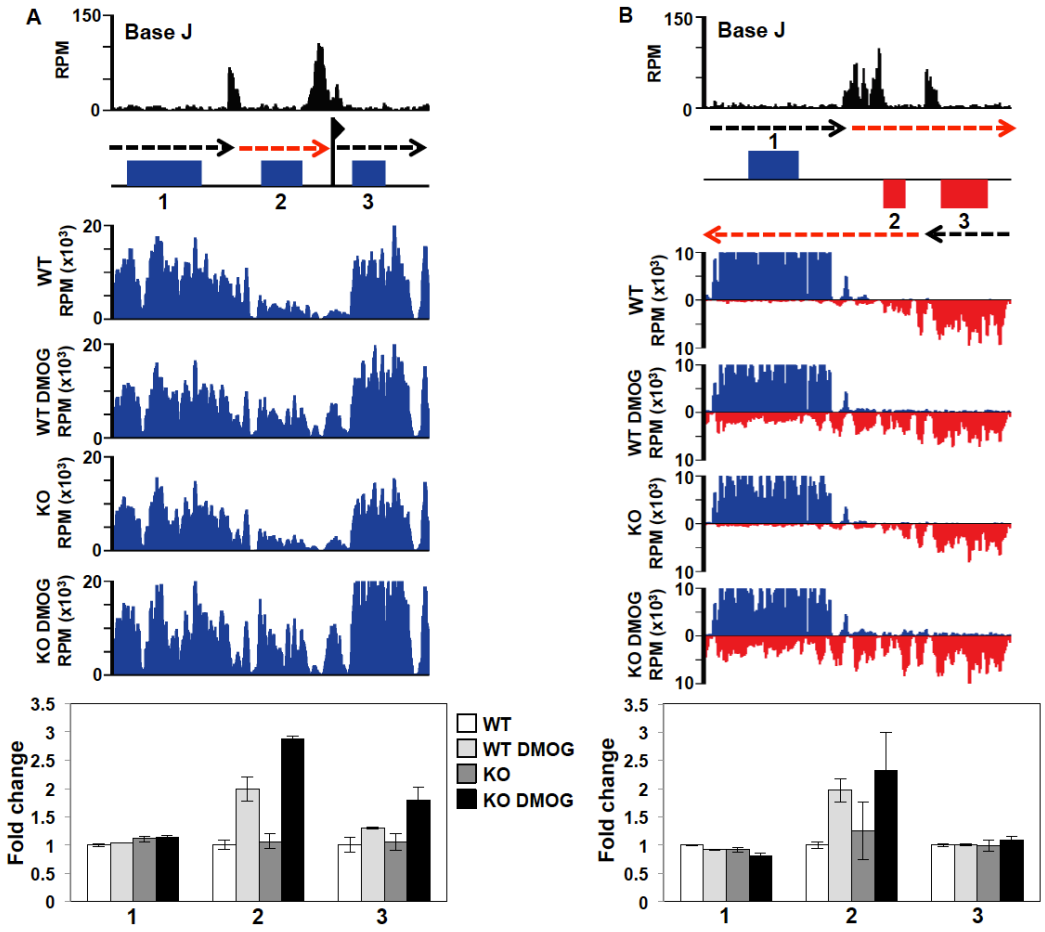


Figure 5.S10 Base J regulates termination and repression of a gene at the end of a gene cluster on chromosome 23 and 32.

(A) Base J IP-seq reads, ORFs, and mRNA-seq reads are plotted for the head-tail region on chromosome 23 as described in Figure 5.5. The region shown is from 208-219 on chromosome 23. Below, is the plot of the mRNA-seq data for the genes numbered in the ORF map, as described in Figure 5.4C and D. The upregulated gene, 2, is *LmjF.23.1590*.

(B) Base J IP-seq reads, ORFs, and mRNA-seq reads are plotted for the cSSR on chromosome 32 as described in Figure 5.5. Below, is the plot of the mRNA-seq data for the genes numbered in the ORF map above, as described in Figure 5.4C and D. The upregulated gene, 2, is *LmjF.32.1380*.

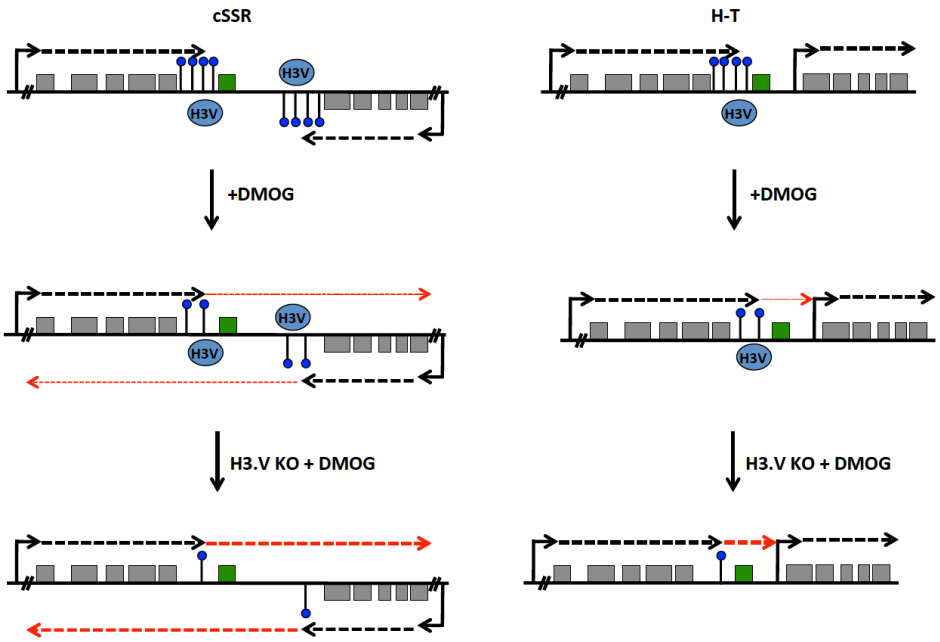


Figure 5.S11 Model of J regulation of RNAP II transcription termination and mRNA expression within gene clusters in *L. major*.

A cSSR and head-tail region (H-T) is illustrated with polycistronic transcription of genes (grey boxes) indicated by dashed black arrows. For several polycistronic gene clusters RNAP II transcription terminates prior to the final gene in the cluster (highlighted in green), where base J, shown as a black line and a blue dot, is enriched. J is also enriched within the cSSR. H3.V co-localizes with base J at RNAP II termination sites. The transcription start site of the downstream gene cluster at H-T regions is marked by acetylated H3 (not shown). Reduction of base J by DMOG treatment results in read through transcription (thin dashed red arrows), which increases the expression of the last gene on the top strand, but does not significantly affect the expression of other genes. At cSSRs, read through includes transcription into the adjacent opposing gene cluster that does not interfere with production of sense mRNAs. Loss of H3.V leads to some J reduction, but not enough to result in read through transcription (not shown). Loss of H3V in cells treated with DMOG leads to further decreases in J and corresponding increases in read through transcription (thick dashed red arrows) and mRNA levels of the downstream genes. The increased expression of silenced genes in *H3.V* KO+DMOG correlates with a strong defect in cell growth, suggesting that the essential nature of J in *L. major* is due to repression of specific genes via regulation of RNAP II termination.

Table 5.S1 *L. major* gene expression changes following H3.V and/or J loss

Upregulated Genes			RPKM[1]				Fold Change[2]			P-Value[3]						
	Gene	Description	WT	WT+DMOG	*H3.V* KO	*H3.V* KO+DMOG	WT+DMOG	*H3.V* KO	*H3.V* KO+DMOG	WT+DMOG	*H3.V* KO	*H3.V* KO+DMOG	Last Gene?[4]	J?[5]	Fit Model?[6]	Developmentally Regulated?[7]
All Conditions	*LmjF.35.2130*	hypothetical protein, unknown function	3.13	64.61	9.70	81.85	20.64	3.10	26.16	0.00	0.01	0.00	Yes	Yes	Yes	Yes
	LmjF.27.1740	hypothetical protein, unknown function	1.05	11.03	2.86	26.62	10.53	2.73	25.41	0.57	0.96	0.57	Yes	Yes	Yes	No
	LmjF.35.2600	hypothetical protein, unknown function	0.33	3.34	0.97	6.82	10.23	2.97	20.87	0.61	0.95	0.60	Yes	Yes	Yes	No
	LmjF.14.0470	hypothetical protein, conserved	2.43	18.78	7.44	18.62	7.73	3.06	7.67	0.00	0.04	0.00	Yes	Yes	Yes	Yes
	LmjF.24.1910	hypothetical protein, conserved	114.59	404.55	335.03	473.22	3.53	2.92	4.13	0.00	0.00	0.00	No	Yes	Yes	Yes
	LmjF.24.1905	hypothetical protein, conserved	33.83	85.82	77.15	108.45	2.54	2.28	3.21	0.00	0.00	0.00	Yes	Yes	Yes	Yes
WT and *H3.V* KO+DMOG	*LmjF.09.0690*	hypothetical protein, conserved	2.41	22.72	4.46	28.33	9.42	1.85	11.74	0.00	0.62	0.00	Yes	Yes	Yes	No
	LmjF.36.3080	lipoate protein ligase, putative	19.22	70.06	28.05	84.49	3.65	1.46	4.40	0.00	0.44	0.00	Yes	Yes	Yes	Yes
	LmjF.22.0020	hypothetical protein, unknown function	19.76	70.80	35.06	100.46	3.58	1.77	5.08	0.00	0.33	0.00	Yes	Yes	Yes	Yes
	LmjF.26.2280	nitrilase, putative	41.63	143.22	59.94	206.77	3.44	1.44	4.97	0.00	0.25	0.00	Yes	Yes	Yes	Yes
	LmjF.18.1050	hypothetical protein, conserved	6.48	17.43	7.33	21.83	2.69	1.13	3.37	0.00	1.00	0.00	Yes	Yes	Yes	No
	LmjF.20.0830	phosphopantetheinyl transferase-like protein	33.32	88.54	56.00	100.53	2.66	1.68	3.02	0.00	0.04	0.00	No	Yes	Yes	Yes
	LmjF.13.0700	kinesin, putative	1.68	4.08	2.85	6.25	2.42	1.69	3.71	0.00	0.27	0.00	Yes	Yes	Yes	No
	LmjF.35.2610	hypothetical protein, unknown function	1.10	2.65	1.44	4.43	2.42	1.31	4.03	0.53	1.00	0.16	No	Yes	Yes	No
	LmjF.33.1320	hypothetical protein, conserved	22.90	54.77	28.49	59.28	2.39	1.24	2.59	0.00	0.89	0.00	Yes	Yes	Yes	Yes
	LmjF.36.2730	D-tyrosyl-tRNA deacylase, putative	51.71	117.75	70.28	168.36	2.28	1.36	3.26	0.00	0.63	0.00	Yes	Yes	Yes	Yes
	LmjF.20.1175	hypothetical protein	9.66	21.53	13.59	21.02	2.23	1.41	2.18	0.67	1.00	0.70	Yes	Yes	Yes	No
	LmjF.20.0820	hypothetical protein, conserved	21.18	44.94	33.65	51.78	2.12	1.59	2.45	0.00	0.13	0.00	No	Yes	Yes	Yes
	LmjF.33.1760	hypothetical protein, unknown function	2.95	6.23	5.73	8.15	2.11	1.94	2.76	0.43	0.56	0.14	Yes	Yes	Yes	Yes
	LmjF.36.5370	tyrosine specific protein phosphatase, putative	12.51	25.50	14.77	33.11	2.04	1.18	2.65	0.00	0.98	0.00	Yes	Yes	Yes	No
WT+DMOG only	*LmjF.01.0820*	potassium channel subunit-like protein	6.84	13.79	3.82	11.40	2.01	0.56	1.67	0.00	0.06	0.05	No	Yes	Yes	No
***H3.V* KO+DMOG only**	*LmjF.23.1590*	oxidoreductase-like protein	15.66	31.13	16.67	45.09	1.99	1.06	2.88	0.00	1.00	0.00	Yes	Yes	Yes	No
	LmjF.12.0490	hypothetical protein, unknown function	1.04	1.28	1.66	2.92	1.23	1.59	2.80	1.00	0.98	0.38	No	No	No	Yes
	LmjF.17.0890	hypothetical protein, conserved (META1)	316.46	454.79	598.35	839.58	1.44	1.76	2.65	0.16	0.00	0.00	No	No	No	Yes
	LmjF.31.0350	amino acid transporter aATP11, putative (AAT1.4)	407.76	486.66	729.20	1077.87	1.19	1.79	2.64	0.89	0.02	0.00	No	No	No	No
	LmjF.12.1090	surface antigen protein, putative	404.99	639.63	587.39	947.19	1.58	1.45	2.34	0.11	0.26	0.00	No	No	No	No
	LmjF.32.3400	hypothetical protein, conserved	83.56	121.91	111.71	194.61	1.46	1.34	2.33	0.11	0.33	0.00	No	No	No	Yes
	LmjF.32.1380	hypothetical protein, conserved	10.37	20.46	12.99	24.09	1.97	1.25	2.32	0.05	0.94	0.01	Yes	Yes	Yes	Yes
	LmjF.35.4900	hypothetical protein, conserved	78.57	95.58	106.38	174.73	1.22	1.35	2.22	0.88	0.51	0.00	No	No	No	No
	LmjF.35.5380	hypothetical protein, conserved	32.67	54.15	51.90	71.99	1.66	1.59	2.20	0.07	0.11	0.00	Yes	Yes	Yes	No
	LmjF.28.0980	P27 protein, putative (P27)	30.51	34.75	52.89	64.89	1.14	1.73	2.13	1.00	0.05	0.00	No	No	No	Yes
	LmjF.26.2710	glutamate 5-kinase, putative	52.61	81.41	69.91	110.40	1.55	1.33	2.10	0.07	0.49	0.00	Yes	Yes	Yes	No
	LmjF.08.1270	hypothetical protein, unknown function	39.61	59.47	50.35	82.29	1.50	1.27	2.08	0.14	0.69	0.00	No	Yes	No	Yes
	LmjF.34.1760	amastin-like surface protein, putative	8.66	10.26	9.30	17.81	1.18	1.07	2.06	1.00	1.00	0.19	No	No	No	Yes
	LmjF.36.4890	hypothetical protein, conserved	7.22	10.41	10.46	14.81	1.44	1.45	2.05	0.68	0.66	0.06	Yes	Yes	Yes	No
	LmjF.34.1960	amastin-like surface protein, putative	61.94	85.90	84.13	127.03	1.39	1.36	2.05	0.38	0.46	0.00	No	No	No	Yes
	LmjF.36.5365	hypothetical protein, conserved	35.98	52.38	38.17	73.69	1.46	1.06	2.05	0.13	1.00	0.00	No	Yes	Yes	Yes
	LmjF.34.1840	amastin-like surface protein, putative	102.33	122.37	133.76	209.45	1.20	1.31	2.05	0.85	0.49	0.00	No	No	No	Yes
	LmjF.34.1800	amastin-like surface protein, putative	9.18	9.76	11.85	18.65	1.06	1.29	2.03	1.00	0.99	0.21	No	No	No	Yes
	LmjF.24.1770	inhibitor of cysteine peptidase (ICP)	279.73	403.06	432.45	566.80	1.44	1.55	2.03	0.16	0.06	0.00	No	No	No	No
	LmjF.34.1780	amastin-like surface protein, putative	9.47	11.11	11.66	19.15	1.17	1.23	2.02	1.00	1.00	0.23	No	No	No	Yes
	LmjF.34.1880	amastin-like surface protein, putative	74.57	92.35	101.15	149.85	1.24	1.36	2.01	0.76	0.42	0.00	No	No	No	Yes
	LmjF.34.1600	amastin-like surface protein, putative	44.62	58.26	60.38	89.21	1.31	1.35	2.00	0.72	0.61	0.01	No	No	No	Yes
***H3.V* KO only**	*LmjF.09.0158*	ATG8/AUT7/APG8/PAZ2, putative (ATG8C.5)	6.74	13.09	13.72	5.93	1.94	2.04	0.88	0.66	0.60	1.00	No	No	No	No
	LmjF.09.0152	ATG8/AUT7/APG8/PAZ2, putative (ATG8C.2)	7.90	9.01	15.79	10.79	1.14	2.00	1.37	1.00	0.64	1.00	No	No	No	No
Downregulated Genes																
WT and *H3.V* KO+DMOG	*LmjF.33.1790*	hypothetical protein, conserved	78.37	35.13	64.27	32.26	2.23	1.22	2.43	0.00	0.78	0.00	No	Yes		Yes
***H3.V* KO and *H3.V* KO+DMOG**	*LmjF.11.0900*	60S ribosomal protein L24, putative	243.84	219.40	121.79	110.64	1.11	2.00	2.20	1.00	0.00	0.00	No	No		Yes
***H3.V* KO+DMOG only**	*LmjF.16.1130*	tyrosyl or methionyl-tRNA synthetase-like protein	172.83	142.57	103.57	59.65	1.21	1.67	2.90	0.79	0.02	0.00	Yes	Yes		No
	LmjF.16.1110	hypothetical protein, conserved	98.36	88.69	59.34	42.53	1.11	1.66	2.31	1.00	0.01	0.00	No	Yes		No
	LmjF.16.1310	cytochrome c, putative	804.22	711.04	515.23	372.20	1.13	1.56	2.16	0.98	0.04	0.00	No	No		No
	LmjF.16.1320	cytochrome c, putative	939.26	850.09	645.25	441.26	1.10	1.46	2.13	1.00	0.12	0.00	No	No		No
	LmjF.04.0850	serine peptidase, Clan S-, family S54, putative,rhomboid-like protein	161.03	145.28	98.90	76.02	1.11	1.63	2.12	1.00	0.02	0.00	No	No		Yes
	LmjF.11.0630	aminopeptidase, putative,metallo-peptidase, Clan MF, Family M17	426.81	389.28	313.45	201.69	1.10	1.36	2.12	1.00	0.42	0.00	No	No		Yes
	LmjF.16.1100	hypothetical protein, conserved	52.24	48.01	35.15	24.82	1.09	1.49	2.10	1.00	0.09	0.00	No	No		Yes
	LmjF.11.0470	pumilio protein 10, putative,pumilio-repeat, RNA-binding protein, putative (PUF10)	123.10	113.75	67.57	58.80	1.08	1.82	2.09	1.00	0.00	0.00	No	Yes		No
	LmjF.04.0370	hypothetical protein, conserved	103.06	94.85	65.49	49.52	1.09	1.57	2.08	1.00	0.04	0.00	No	Yes		No
	LmjF.16.1370	hypothetical protein, conserved	244.41	201.71	145.47	119.04	1.21	1.68	2.05	0.87	0.05	0.00	No	Yes		Yes
	LmjF.16.0530	dihydroorotate dehydrogenase (fumarate) (DHODH)	124.28	113.59	70.50	60.91	1.09	1.76	2.04	1.00	0.00	0.00	No	No		No
	LmjF.12.0400	3'-nucleotidase/nuclease, putative	213.85	153.44	212.70	105.13	1.39	1.01	2.03	0.20	1.00	0.00	Yes	Yes		Yes
	LmjF.16.0420	hypothetical protein, conserved	17.25	18.06	10.80	8.49	0.96	1.60	2.03	1.00	0.14	0.01	No	No		Yes
	LmjF.28.2700	hypothetical protein, unknown function	15.44	12.33	16.54	7.61	1.25	0.93	2.03	0.86	1.00	0.03	Yes	Yes		No
	LmjF.16.1120	hypothetical protein, conserved	52.26	49.97	33.18	25.78	1.05	1.57	2.03	1.00	0.12	0.00	No	Yes		Yes
	LmjF.16.1140	hypothetical protein, conserved	30.09	27.67	19.45	15.01	1.09	1.55	2.01	1.00	0.07	0.00	Yes	Yes		No
	LmjF.11.0620	aminopeptidase, putative,metallo-peptidase, Clan MF, Family M17	374.34	342.18	265.87	187.29	1.09	1.41	2.00	1.00	0.20	0.00	No	No		No
	LmjF.16.1230	hypothetical protein, conserved	95.84	93.60	64.06	48.38	1.02	1.50	1.98	1.00	0.12	0.00	No	No		No
***H3.V* KO only**	*LmjF.02.0400*	hypothetical protein, unknown function	47.51	39.83	16.90	49.02	1.19	2.81	0.97	1.00	0.75	1.00	No	No		No

[1]The average RPKM value of duplicate mRNA-seq libraries is listed

[2]Fold change was determined by dividing the experimental sample RPKM value by the WT RPKM value for fold upregulation or by dividing the WT RPKM value by the experimental sample RPKM value for fold downregulation

[3]P-value determined by Cuffdiff

[4]Yes: if gene is the final gene in a gene cluster

[5]Yes: if gene is located within 10 kb of base J

[6]Genes fit our model of J regulated transcription termination if the gene is located near the end of a gene cluster, downstream or overlapping a J enriched region

[7]Developmental regulation determined using the Dillon et al., 2015 dataset

Table 5.S2 High-throughput sequencing information

Species	Genotype	Treatment	Library #	Type of RNA sequenced	Genome used for alignment	Minimum read length considered for alignment (nt)	Total reads (millions)	Overall (unique and non unique) alignment rate %	Sequenced by	Data shown in
L. major	H3.V KO	DMSO	1	small RNA	*L. major* Friedlin v4.2	18	28.7	54.1	Vertis Biotechnology	Figure 2A, Supplemental Figure S4A and S7
L. major	H3.V KO	DMOG	2	small RNA	*L. major* Friedlin v4.2	18	31.9	43.5	Vertis Biotechnology	Figure 2A, Supplemental Figure S4A and S7
L. major	WT	DMSO	3	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	62.3	98.8	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	WT	DMSO	4	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	23.8	98.0	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	WT	DMOG	5	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	57.6	98.7	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	WT	DMOG	6	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	29.1	98.5	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	H3.V KO	DMSO	7	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	52.6	98.9	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	H3.V KO	DMSO	8	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	25.4	98.8	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	H3.V KO	DMOG	9	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	59.0	98.7	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8
L. major	H3.V KO	DMOG	10	polyA enriched RNA (mRNA)	*L. major* Friedlin v9.0	50	18.5	98.1	Georgia Genomics Facility	Figure 4, 5, 6, Supplemental Figure S6, S8

CHAPTER 6

CONCLUSIONS AND DISCUSSION

Several unusual features of the early-diverged kinetoplastid species, most notably the arrangement of functionally unrelated genes into polycistronically transcribed gene clusters, have led to post-transcriptional mechanisms as the primary focus of gene expression regulation in kinetoplastids (1,2). The exception to this focus has been in the study of antigenic variation in *T. brucei*, where growing evidence implicates transcriptional regulation, including epigenetic mechanisms, in the regulation of antigenic variation and RNA Polymerase (RNAP) I transcription of subtelomeric expression sites (3). Despite these findings, RNAP II transcription of chromosome internal gene clusters is still largely assumed to be a constitutive, unregulated process.

A major challenge to this dogma arose when multiple chromatin modifications enriched at RNAP II initiation and termination sites were identified, suggesting that perhaps epigenetic mechanisms of transcriptional regulation were overlooked (4-7). Indeed, studies in *T. cruzi* and *L. tarentolae* have revealed that the DNA modification base J does affect RNAP II transcription (8-10); however, the use of genetic deletions to reduce J, and the resulting global gene expression changes, has made it difficult to link specific gene expression changes directly to J reduction. The functional characterization of J and histone H3.V in *T. brucei* and *L. major* (described in Chapters 3-5) has provided further evidence that chromatin modifications impact the process of RNAP II transcription in kinetoplastids. Importantly, specific gene expression changes at the end of gene clusters following the loss of J and H3.V implicate these

modifications in directly regulating the process of RNAP II elongation/termination and repressing the expression of genes near the end of gene clusters.

In *T. brucei*, J and H3.V independently promote termination prior to the end of gene clusters and repress expression of downstream genes (11-13). Both modifications are also involved in the maintenance of monoallelic expression of *variant surface glycoprotein* (*VSG*) genes from expression sites, indicating that similar mechanisms are utilized to affect RNAP I and II transcription. H3.V also has a specific role at convergent strand switch regions (cSSRs), where transcription of two gene clusters converges, leading to dual strand transcription and the formation of siRNAs. The presence of H3.V within cSSRs suppresses dual strand transcription and therefore the production of siRNAs (12).

In *L. major*, base J has a more extensive role in promoting termination within cSSRs, preventing read through transcription and the formation of antisense RNAs (13). Interestingly, the production of these antisense RNAs does not significantly negatively affect the abundance of the corresponding sense mRNAs (Chapter 5). We instead find that J promotes termination prior to the end of some gene clusters similar to its function in *T. brucei*, repressing the expression of genes at the end of the cluster. These findings strongly implicate the repression of genes at the end of gene clusters as the essential role of J in *Leishmania* spp. In contrast, H3.V does not directly impact RNAP II transcription in *L. major* and instead alters J levels at cSSRs, where the loss of H3.V leads to reduction of base J, but, surprisingly, to a level that does not result in a termination defect. Thus, H3.V localization is conserved between *T. brucei* and *L. major*, but its role in promoting termination is not. We present evidence that the lack of a termination defect in *H3.V* KO *L. major* cells, despite J reduction, is due to a low level of J that is sufficient to prevent read through transcription. Beyond this threshold level of around 25-40%

of wild type J abundance, read through transcription is inversely correlated with further J loss (Chapter 5).

Overall these findings reveal that chromatin modifications do impact the process of RNAP II transcription of gene clusters in kinetoplastids, specifically through inhibiting transcription elongation, leading to termination. Although much work remains to more fully elucidate the extent of transcriptional regulation in kinetoplastids, this work highlights the importance of chromatin modifications in the promotion of termination and its role in repressing gene expression. Here, potential mechanisms and the biological significance of J and H3.V regulation of RNAP II transcription termination are discussed.

MECHANISM OF J AND H3.V INHIBITION OF RNAP II ELONGATION

The mechanism through which J promotes termination in *T. brucei* and *Leishmania* spp. is currently unknown. Given that JBP1 is the only protein identified thus far in kinetoplastids that binds to J modified DNA, the mechanism of J termination likely does not involve recruitment of J-binding proteins. The abundance of J-modified sites far outnumbers the amount of JBP1 protein in a given cell (14), though a role of JBP1 or other unidentified J-binding proteins in promoting termination cannot be excluded. Our current hypothesis is that the glucose group of base J, which projects into the major groove of DNA, acts as a steric block to elongating RNAP II, thereby promoting pausing and release of RNAP II from the DNA. DNA modifications have been shown to inhibit both yeast and mammalian RNAP II elongation in vitro (15). Additional, non-mutually exclusive, possibilities of J's impact on termination include alteration to the physical properties of DNA, blocking of DNA binding proteins, or impacts on chromatin structure. None of these possibilities have been thoroughly examined.

We similarly hypothesize that transcription elongation is inhibited by H3.V in *T. brucei*, promoting termination both within RNAP I transcribed silent expression site gene clusters and within RNAP II transcribed gene clusters. This hypothesis of J and H3.V inhibition of RNAP I and II elongation is consistent with observations that low levels of RNAP I initiation takes place at the promoters of silent expression sites in *T. brucei*, but transcription terminates prior to the *VSG* gene at the end of silent expression sites (16). Fully productive elongation occurs only at the active expression site, which appears to be depleted of both J and H3.V (5,17), thereby maintaining monoallelic *VSG* expression. Very little is known about the molecular mechanism by which H3.V represses transcription, such as whether it affects chromatin structure, recruits other proteins, and how it is deposited into chromatin. Nonetheless, the finding that H3.V directly promotes termination in *T. brucei* but does not in *L. major* demonstrates that the functions of chromatin modifications are not necessarily conserved among kinetoplastids.

BIOLOGICAL SIGNIFICANCE OF J AND H3.V REGULATION OF GENE EXPRESSION

These findings indicate that both J and H3.V in *T. brucei* and J in *L. major* promote termination, repressing genes at the end of gene clusters. A major question is whether these modifications are dynamically added and removed to effect gene expression changes, and if so, what is the biological significance of such regulation? It is possible that these modifications constitutively repress RNAP II transcription at the end of gene clusters and therefore do not regulate gene expression in the sense of modulating expression in response to environmental signals or developmental cues. To more fully address the role of J and H3.V in regulating gene

expression it will be important to understand how these modifications are established and maintained and whether mechanisms exist to remove them.

Establishment, maintenance, and removal of J and H3.V

Much is now known about the synthesis of base J (Chapter 2), however it remains unclear how J is established at specific regions of kinetoplastid genomes. Efforts to identify proteins associated with JBP1/2 will likely lead to insight in this area. Preliminary data suggest that JBP1/2 and JGT do not interact. Thus, whether the activities of JBP1/2 and JGT are coordinated is not known. An exciting possibility is that JBP1/2 reiteratively oxidize thymidine, analogous to the TET proteins in mammalian cells (18), forming formyluridine and carboxyuridine in addition to hydroxymethyluridine. Non-J modified hydroxymethyluridine and formyluridine have been identified in *T. brucei* DNA by mass spectrometry (19). Functional elucidation of oxidized DNA bases is only now beginning in higher eukaryotes, where oxidized forms of cytosine have been linked to alternative splicing in mammalian cells (20). Although potential functions of further oxidized forms of thymidine remain to be explored in kinetoplastids, reiterative oxidation by JBP1/2 without glucosylation presents a mechanism to modulate J levels. In addition to JBP1/2, regulation of JGT could provide a mechanism to alter J levels, however incorporation of hydroxymethyluridine randomly in the genome in *T. brucei* leads to random J synthesis, suggesting that the JGT is sequence non-specific in vivo (21). In vitro analyses support the sequence non-specific nature of JGT activity (22). Active removal of base J, as opposed to passive loss through DNA replication following inhibition of J synthesis, remains a possibility, though no enzymes capable of removing J have been identified yet.

Much less is known about the establishment, maintenance, and removal of H3.V. In other eukaryotes histone variants are specifically incorporated outside of the DNA synthesis phase of the cell cycle, in contrast to canonical histones (23). Whether H3.V incorporation occurs independently of the cell cycle in kinetoplastids is unknown. Histone chaperones and chromatin remodelers involved in H3.V incorporation or removal are also unknown. While base J, H3.V, and H4.V co-localize at termination sites in *T. brucei* (5,7), it does not appear the modifications are dependent on each other for their specific localization. Loss of H3.V does not lead to J reduction in *T. brucei* (unpublished data) and the observation that the combined loss of H3.V and J (DMOG treatment of *H3.V* KO *T. brucei* cells) leads to greater derepression of gene expression compared to the individual loss of J or H3.V suggests that J loss alone does not reduce H3.V enrichment at termination sites. Furthermore, preliminary data reveal that loss of H4.V does not increase transcript levels downstream of gene cluster internal termination sites similarly to H3.V removal, suggesting that loss of H4.V does not result in H3.V loss at termination sites. In *L. major* loss of H3.V does result in reduction of J at termination sites, but the loss of J also at transcription initiation sites where H3.V is not enriched suggests the relationship between H3.V and J is indirect. Conversely, reduction of J from termination sites does not significantly alter H3.V enrichment (Chapter 5). It is not clear whether H4.V is present in *L. major*.

To investigate the possibility that J and H3.V are utilized to effect changes in gene expression, it will be important to compare the levels of each modification across different life stages of kinetoplastid parasites and whether they are affected by relevant environmental stimuli such as heat shock, pH stress, or nutrient availability. Electron microscopy has revealed large-scale chromatin structural differences between insect and mammalian stages of *T. brucei* cells

(24), but higher resolution analysis of chromatin modifications, e.g. chromatin immunoprecipitation followed by high throughput sequencing, across life stages has only been performed in a few instances in kinetoplastid species. Base J is developmentally regulated in *T. brucei*; the modification is only present during the mammalian stage (25). Interestingly, the genes upregulated following J loss in the mammalian stage are not upregulated in the insect stage (26). Additional mechanisms therefore likely exist to repress gene expression during the insect stage.

In *Leishmania* spp. it is not known whether J levels or localization differ across life stages. Multiple genes with increased expression following J loss in *L. major* are developmentally regulated however (Chapter 5 and (27)). Specifically, we identified 13 genes that are both developmentally regulated and are located near the end of a gene cluster downstream of or overlapping a J enriched region, thus fitting our model of J inhibition of RNAP II elongation to repress gene expression (Chapter 5). Efforts are currently underway to measure J levels at these sites across *L. major* life stages, directly testing the hypothesis that J is utilized to effect gene expression changes. We predict an inverse correlation between gene expression and J levels.

What genes do J and H3.V repress?

Although dynamic modulation of J and H3.V levels to achieve biologically relevant changes in gene expression in kinetoplastids is an exciting prospect, regardless of whether this is the case, the question of why some genes are specifically repressed by base J and H3.V remains. In *T. brucei* the majority of the genes affected by J and H3.V loss are annotated as hypothetical or unknown genes, however about a fifth are annotated as *VSG* genes or pseudogenes (12).

Monoallelic expression of these genes is important during mammalian infections in the process of antigenic variation. In liquid culture in vitro, J and H3.V loss does not negatively affect *T. brucei* cell growth, however the effect of J and H3.V loss in vivo has not been examined. Therefore, it will be important to compare mammalian infections of wild type *T. brucei* and cells without J and H3.V to assess the biological significance of *VSG* repression by these chromatin modifications.

In *L. major* the most significant effect of J reduction on mRNA abundance is an increase in transcripts from genes located near the end of gene clusters, most of which are annotated as hypothetical or unknown. In contrast to *T. brucei* however, the large reduction of J in *H3.V* KO *L. major* cells treated with dimethyloxalylglycine (DMOG) is detrimental to cell growth in vitro. *L. major H3.V* KO cells alone grow similarly to wild type cells and experience minimal gene expression changes. Thus, genes near the end of gene clusters that are derepressed following J loss in *H3.V* KO+DMOG cells likely includes genes whose repression by J is essential for cell viability. Future efforts should therefore investigate the function of these genes, particularly whether their overexpression in *L. major* is detrimental to cell growth. One possibility is that genes repressed through gene cluster internal termination include those important for proper developmental progression. As mentioned above, 13 of the J repressed genes are developmentally regulated in their expression. The studies described here were performed in *L. major* promoastigotes, which represent the insect stage of the parasite. During their developmental progression, promastigotes differentiate into non-dividing metacyclic promastigotes in the salivary glands of sand flies, which then enter mammalian hosts through an insect bite. Very little is understood about the molecular underpinnings of developmental progression in *Leishmania* spp. We are currently investigating whether J reduction stimulates

differentiation to metacyclic promastigotes. If so, this could explain reduced cell growth (due to differentiation to a non-dividing stage) and increased expression of developmentally regulated surface proteins that do not appear to be directly regulated by J at gene cluster internal termination sites, and thus are secondary effects of J loss. The results of these studies could also reveal genes involved in directly promoting differentiation, though it remains possible that derepression of these genes generates an unknown cell stress that triggers differentiation. Regardless of whether J loss affects *L. major* differentiation, future studies should aim to assess the significance of J loss on mammalian host cell invasion and egress, which are altered by J loss in *T. cruzi* (9), and the consequence of J loss in the context of a mammalian infection.

In addition to increased expression of genes at the end of gene clusters, loss of J from cSSRs in *L. major* leads to read through transcription and the production of antisense RNAs. Although these antisense RNAs do not negatively affect the abundance of sense mRNAs (Chapter 5), they do appear to be processed similarly to sense mRNAs in that they contain a spliced leader sequence at the 5' end and are polyadenylated at the 3' end ((10) and Chapter 5). It remains possible that these antisense RNAs alter processes other than the steady state abundance of sense mRNAs, which we would have missed in our mRNA-seq analyses. For example, although it is unclear whether any read through antisense RNAs contain an open reading frame, the presence of similar 5' and 3' end processing could result in their disruption of sense mRNA translation if they are present in the cytoplasm. It will therefore be important to investigate the effect of J reduction on the *L. major* proteome. Such analysis will not only confirm that genes with increased mRNA abundance following J loss, i.e. those at the end of gene clusters, are also increased at the protein level, but will also reveal any changes in gene

expression that are not reflected in the steady state level of mRNAs. Proteomic analyses will therefore shed additional light on the essential nature of base J in *L. major*.

Why termination?

Given that the function of most chromatin modifications in kinetoplastids are only now being investigated, it is likely that additional mechanisms of transcriptional regulation will be identified. It is possible though that the organization of genes into clusters greatly limits the ways in which transcriptional regulation can effectively achieve changes in gene expression in kinetoplastids, particularly at the level of transcription initiation, which would potentially involve transcription through multiple genes irrelevant for a given cellular response. Placement of genes at the end of gene clusters and regulating their expression by inhibiting RNAP II elongation could represent a way to transcriptionally regulate gene expression in the absence of sophisticated regulation of RNAP II initiation that occurs in most other eukaryotes.

SUMMARY

Although much work remains to more fully assess the role of chromatin modifications in kinetoplastid parasites, not only in regulating gene expression, but also more broadly in affecting processes such as DNA replication, DNA repair, recombination, nuclear organization, etc., the findings presented here strongly demonstrate that chromatin modifications impact gene expression by promoting transcription termination. This work further opens the door to additional investigations of transcriptional regulation in kinetoplastids, which have been largely overlooked because of the dogma of solely post-transcriptional gene expression regulation. Given the evolutionary divergence of the currently characterized transcriptional regulatory

mechanisms utilized by kinetoplastid parasites, this work may reveal novel drug targets to combat the diseases caused by these organisms.

REFERENCES

1. Clayton, C.E. (2002) Life without transcriptional control? From fly to man and back again. *EMBO J*, **21**, 1881-1888.
2. Campbell, D.A., Thomas, S. and Sturm, N.R. (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.*, **5**, 1231-1240.
3. Rudenko, G. (2010) Epigenetics and transcriptional control in African trypanosomes. *Essays in biochemistry*, **48**, 201-219.
4. Respuela, P., Ferella, M., Rada-Iglesias, A. and Aslund, L. (2008) Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. *J. Biol. Chem.*, **283**, 15884-15892.
5. Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S. and Cross, G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev.*, **23**, 1063-1076.
6. Thomas, S., Green, A., Sturm, N.R., Campbell, D.A. and Myler, P.J. (2009) Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*, **10**, 152.
7. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A. and Sabatini, R. (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res*, **38**, 3923-3935.

8. Ekanayake, D. and Sabatini, R. (2011) Epigenetic regulation of Pol II transcription initiation in Trypanosoma cruzi: Modulation of nucleosome abundance, histone modification and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryotic cell*, **10**, 1465-1472.

9. Ekanayake, D.K., Minning, T., Weatherly, B., Gunasekera, K., Nilsson, D., Tarleton, R., Ochsenreiter, T. and Sabatini, R. (2011) Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. *Mol. Cell. Biol.*, **31**, 1690-1700.

10. van Luenen, H.G., Farris, C., Jan, S., Genest, P.A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G. *et al.* (2012) Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, **150**, 909-921.

11. Schulz, D., Zaringhalam, M., Papavasiliou, F.N. and Kim, H.-S. (2016) Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005762.

12. Reynolds, D., Hofmeister, B.T., Cliffe, L., Alabady, M., Siegel, T.N., Schmitz, R.J. and Sabatini, R. (2016) Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLoS Genet.*, **12**, e1005758.

13. Reynolds, D., Cliffe, L., Forstner, K.U., Hon, C.C., Siegel, T.N. and Sabatini, R. (2014) Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res*, **42**, 9717-9729.

14. Toaldo, C.B., Kieft, R., Dirks-Mulder, A., Sabatini, R., van Luenen, H.G. and Borst, P. (2005) A minor fraction of base J in kinetoplastid nuclear DNA is bound by the J-binding protein 1. *Molecular and biochemical parasitology*, **143**, 111-115.

15. Kellinger, M.W., Song, C.X., Chong, J., Lu, X.Y., He, C. and Wang, D. (2012) 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nature structural & molecular biology*, **19**, 831-833.

16. Vanhamme, L., Poelvoorde, P., Pays, A., Tebabi, P., Van Xong, H. and Pays, E. (2000) Differential RNA elongation controls the variant surface glycoprotein gene expression sites of Trypanosoma brucei. *Molecular microbiology*, **36**, 328-340.

17. van Leeuwen, F., Wijsman, E.R., Kieft, R., van der Marel, G.A., van Boom, J.H. and Borst, P. (1997) Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes & development*, **11**, 3232-3241.

18. Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C. and Zhang, Y. (2011) Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science*, **333**, 1300-1303.

19. Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y. and Sabatini, R. (2014) Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. *J. Biol. Chem.*, **289**, 20273-20282.

20. Marina, R.J., Sturgill, D., Bailly, M.A., Thenoz, M., Varma, G., Prigge, M.F., Nanan, K.K., Shukla, S., Haque, N. and Oberdoerffer, S. (2016) TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. *EMBO J*, **35**, 335-355.

21. van Leeuwen, F., Kieft, R., Cross, M. and Borst, P. (1998) Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Molecular & Cellular Biology*, **18**, 5643-5651.

22. Bullard, W., Cliffe, L., Wang, P., Wang, Y. and Sabatini, R. (2015) Base J glucosyltransferase does not regulate the sequence specificity of J synthesis in trypanosomatid telomeric DNA. *Molecular and biochemical parasitology*, **204**, 77-80.

23. Talbert, P.B. and Henikoff, S. (2010) Histone variants--ancient wrap artists of the epigenome. *Nature reviews. Molecular cell biology*, **11**, 264-275.

24. Schlimme, W., Burri, M., Bender, K., Betschart, B. and Hecker, H. (1993) Trypanosoma brucei brucei: differences in the nuclear chromatin of bloodstream forms and procyclic culture forms. *Parasitology*, **107**, 237-247.

25. van Leeuwen, F., Dirks-Mulder, A., Dirks, R.W., Borst, P. and Gibson, W. (1998) The modified DNA base beta-D-glucosyl-hydroxymethyluracil is not found in the tsetse fly stages of Trypanosoma brucei. *Molecular & Biochemical Parasitology*, **94**, 127-130.

26. Siegel, T.N., Hekstra, D.R., Wang, X., Dewell, S. and Cross, G.A. (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res*, **38**, 4946-4957.

27. Dillon, Laura A.L., Okrah, K., Hughitt, V.K., Suresh, R., Li, Y., Fernandes, M.C., Belew, A.T., Corrada Bravo, H., Mosser, D.M. and El-Sayed, N.M. (2015) Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. *Nucleic Acids Res*, **43**, 6799-6813.