Title page

Chromatin-associated protein complexes link DNA base J and transcription termination in *Leishmania*

Bryan C Jensen[1], Isabelle Q. Phan[1,2], Jacquelyn R. McDonald[1], Aakash Sur[1,3], Mark A. Gillespie[4#], Jeffrey A. Ranish[4], Marilyn Parsons[1,5,6], Peter J Myler[1,2,3,5,6*]

[1] Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, WA, USA

[2] Seattle Structural Genomics Center for Infectious Disease, Seattle, WA, USA

[3] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA USA

[4] Institute for Systems Biology, Seattle, WA, USA

[5] Department of Pediatrics, University of Washington, Seattle, WA USA

[6] Department of Global Health, University of Washington, Seattle, WA USA

current address: Allen Institute, Seattle, WA, USA

* Corresponding author: peter.myler@seattlechildrens.org

Abstract

Unlike most other eukaryotes, *Leishmania* and other trypanosomatid protozoa have largely eschewed transcriptional control of gene expression; relying instead on post-transcriptional regulation of mRNAs derived from polycistronic transcription units (PTUs). In these parasites, a novel modified nucleotide base (β-D-glucopyranosyloxymethyluracil) known as J plays a critical role in ensuring that transcription termination occurs only at the end of each PTU, rather than at the polyadenylation sites of individual genes. To further understand the biology of J-associated processes, we used tandem affinity purification (TAP-tagging) and mass spectrometry to reveal proteins that interact with the glucosyltransferase performing the final step in J synthesis. These studies identified four proteins reminiscent of subunits in the PTW/PP1 complex that controls transcription termination in higher eukaryotes. Moreover, bioinformatic analyses identified the DNA-binding subunit of *Leishmania* PTW/PP1 as a novel J-binding protein (JBP3). Down-regulation of JBP3 expression levels in *Leishmania* resulted in a substantial increase in transcriptional read-through at the 3' end of most PTUs. Additional TAP-tagging experiments showed that JBP3 also associates with two other protein complexes. One consists of subunits with domains suggestive of a role in chromatin modification/remodeling; while the other contains subunits with similarity to those found in the PAF1 complex involved in regulation of transcription in other eukaryotes. Thus, trypanosomatids utilize protein complexes similar to those used to control transcription termination in other eukaryotes and JBP3 appears to function as a hub linking these modules to base J, thereby enabling the parasites' unique reliance on polycistronic transcription and post-transcriptional regulation of gene expression.

Introduction

The genus *Leishmania* includes several species of protozoan parasites that cause a spectrum of human diseases, ranging from cutaneous lesions to disfiguring mucocutaneous and lethal visceral leishmaniasis, depending primarily on the species involved. *Leishmania* is transmitted through the bite of the sand fly and belongs to the family Trypanosomatidae, which also includes the vector-borne human pathogens *Trypanosoma brucei spp.*, causative agent of human African trypanosomiasis (African sleeping sickness), and *Trypanosoma cruzi*, causative agent of Chagas' disease. Reflecting the ancient divergence of these organisms, the Trypanosomatidae exhibit a myriad of biological differences from "higher" eukaryotes. One major difference is that each chromosome is organized into a small number of polycistronic transcription units (PTUs), which consist of tens-to-hundreds of protein-coding genes co-transcribed from a single initiation site at the 5' end of the PTU to a termination site at the 3' end. Interestingly, unlike the operons of prokaryotes, genes within each PTU are not confined to a single pathway or function. Individual genes within the primary transcript are *trans*-spliced by addition of a 39-nucleotide spliced-leader (SL) mini-exon to provide the 5' cap-4 structure and polyadenylated to form the mature individual mRNAs. As a result of this unique genomic organization, all genes within a PTU are transcribed at the same rate (as are genes across different PTUs). Hence, gene expression must be controlled by post transcriptional processes, such as splicing/polyadenylation rate, RNA stability and translational regulation.

A second distinct feature of the Trypanosomatidae (and other Euglenazoa) is that ~1% of the thymidine bases in the nuclear genome are glucosylated to form the novel nucleotide β-D-glucopyranosyloxymethyluracil, usually referred to as J (*1, 2*). While the majority of J is localized within telomeric repeat sequences (*1, 3-6*), chromosome-internal J is found at almost all transcription termination sites (*7, 8*) and centromeres, which also correspond to the major replication origin on *Leishmania* chromosomes (*9, 10*). In some trypanosomatids (although not *Leishmania*), J is also found in transcriptionally silent regions containing retrotransposons and/or other repetitive sequences (*6*). J biosynthesis occurs in two steps, whereby one of two proteins (JBP1 or JBP2) hydroxylates the methyl group of thymidine to form hydroxymethyldeoxyuracil (HmdU), which is subsequently further modified by a glucosyltransferase (HmdUGT) to form J (*7, 11*). Both JBP1 and JBP2 have an N-terminal oxygenase (Tet_JBP) catalytic domain; but JBP1 contains a central J-binding domain (*12*), while JBP2 instead contains SNF2_N, ATP-binding and helicase C-terminal domains that are suggestive of a role in chromatin binding and/or re-modeling (*13, 14*). Null mutants of *JBP1* have not been isolated (despite multiple attempts) in *Leishmania*, suggesting that it is an essential gene (and that J is required for viability) (*15*). In contrast, *JBP2* null mutants have been isolated and shown to have 40% less J than wild-

type parasites (*8*). Chromosome-internal J was gradually lost during continuous growth of *Leishmania tarentolae JBP2* null mutants, with a concomitant increase in read-through transcription at termination sites, suggesting a critical role for J (and JBP2) in transcription termination. Coupled with the identification of a JBP1 recognition motif (*16*), these data led to the model that JBP1 is responsible for J maintenance after DNA replication by binding to pre-existing J on the parental strand and modifying the thymidine twelve nucleotides downstream on the newly synthesized strand. Conversely, JBP2 is proposed to be largely responsible for *de novo* synthesis of J only when JBP1 is not able to fully restore J on both strands.

Despite this knowledge of the enzymes involved in J biosynthesis, we currently know little about how J mediates transcription termination (and/or repression of initiation), since the machinery underlying this process has not yet been identified. Here, we have used Tandem Affinity Purification (TAP)-tagging and tandem mass-spectrometry to identify a PTW/PP1-like protein complex that interacts with HmdUGT. This complex includes a novel J-binding protein (JBP3) that appears to be essential in *Leishmania.* RNA-seq analysis following conditional down-regulation of JBP3 expression shows substantially higher levels of transcriptional read-through at the 3' end of most PTUs, suggesting that it plays an important role in transcription termination. While this manuscript was being prepared, Kieft et al (*17*) reported similar results in *Leishmania* and *T. brucei*. Here, we have extended these findings by demonstrating that JBP3 also interacts with another protein complex likely involved in chromatin modification/remodeling, as well as a PAF1C-like complex that likely interacts with RNA polymerase II (RNAP II). Therefore, despite the differences in gene regulation from other eukaryotes, *Leishmania* appears to utilize proteins related to those used for chromatin remodeling and transcriptional regulation in other eukaryotes to provide the molecular machinery that links J to termination of RNAP II-mediated transcription.

Results

Identification of a protein complex containing a novel J-binding protein

To date only three proteins have been shown to be involved in J biosynthesis: JBP1, JBP2 and HmdUGT (hereafter referred to as GT). To expand the network of proteins important in J biosynthesis and/or function, we used mass spectrometry to identify proteins that co-purified with a TAP-tagged GT bait expressed in *L. tarentolae*. Two separate experiments were performed: the first using extracts from WT parasites constitutively expressing the tagged protein and the second using extracts from the tetracycline (Tet) induced T7-TR cell line over-expressing the tagged protein integrated at the *ODC* locus (Supplementary Figure S1A). After affinity purification, SDS-PAGE and silver staining confirmed successful enrichment of the bait protein in the pooled eluates (**Figure 1A)**. Proteins were identified by liquid chromatography-tandem mass spectrometry (LC-MS/MS) and their ($\log_2$-fold) enrichment calculated by comparison to a control (untransfected) parental cell line (see Supplementary Data File). The combined data from both experiments revealed seven proteins that were enriched by >150-fold (compared to the control) and were therefore considered to be part of the TAP-tagged complex. Another two proteins showed >75-fold enrichment in the first experiment, although they were not detected in the second experiment (Supplementary Table S1). Five subunits of the prefoldin complex were enriched in both experiments, while another (PFDN2) was enriched ~50-fold in the first experiment only. We suspect that the prefoldins act as a chaperone role for one or more proteins of the GT-associated complex, so they will not be further considered here.

One of the highly enriched proteins (LtaP15.0230) was annotated in TriTrypDB as a putative protein phosphatase 1 catalytic subunit, while the other three (LtaP36.0380, LtaP33.1440 and LtaP32.3990) were all annotated as “Hypothetical protein, conserved”. Constructs were made for expression of TAP-tagged proteins and transfected into the *L. tarentolae* T7-TR strain. Affinity purification was performed on two (independently-generated) cell lines for each version of the TAP-tagged proteins (the results from one replicate of each are shown in Supplementary Figure S1B) and analyzed by LC-MS/MS (Supplementary Data file). The results from these pulldowns (**Table 1** and Supplementary Tables S2-S5) show the four proteins identified above were all highly enriched (as was GT, albeit to a lesser degree than other four components). Thus, we conclude that these proteins form a stable complex, with GT perhaps more transiently associated than the other four components.

LtaP15.0230 encodes one of eight isoforms of the catalytic subunit of protein phosphatase 1 (PP1C) found in the *L. tarentolae* genome. Phylogenetic analysis (see Supplementary Figure S2) indicates that there are five different clades of PP1C paralogues in trypanosomatids and LtaP15.0230 belongs to a

clade (e) that lacks any mammalian orthologue. Interestingly, Salivarian trypanosomes (which includes *T. brucei*), also lack PP1Ce, although it is present in Stercorarian trypanosomes and the more distantly related *Blechomonas ayalai*, *Paratrypanosoma confusum* and *Bodo saltans* (data not shown). TAP-tagging of PP1Ce resulted in co-purification of seven proteins with >150-fold enrichment (Supplementary Table S2). These include GT and the other three components of the complex described above, as well as three proteins (encoded by *LtaP05.1290*, *LtaP07.0770* and *LtaP29.0170*) annotated as protein phosphatase regulatory subunits (PPP1R7/Sds22, PPP1R11/inhibitor 3, and PPP1R2/inhibitor 2, respectively). In mammalian systems, a subset of PP1C proteins are present in heterotrimeric inactive complexes with PP1R7 and PPP1R11 (*18*). Therefore, we suggest that PP1Ce forms at least three separate complexes; one being the GT-associated complex, a second with PPP1R7 and PPP1R11, and another with PPP1R2.

We performed a series of bioinformatic analyses to identify domains and/or motifs that might provide hints as to the function of the three proteins of GT-associated complex that lack a functional description. BLASTP and INTERPROSCAN searches showed high confidence matches only to orthologues in other trypanosomatids with no informative domains identified. However, HHPRED analysis, which detects remote protein homology by using hidden Markov modelling and structure prediction (*19*), revealed a structural match (Supplementary Figure S3A) between the central portion (residues 88-105) of LtaP33.1440 and the human serine/threonine-protein phosphatase 1 regulatory subunit 10 (PPP1R10), which is also known as PNUTS (*20-23*). The trypanosomatid protein is much smaller (264 *vs* 940 amino acids) than mammalian PNUTS, with the sequence and structural similarity restricted to the central region above. However, this region contains the sequence motif that is essential for phosphatase inhibitor function and interaction with PP1C (*24*). Therefore, in deference to precedent in the field (*17*), we will refer to this protein as PNUTS, despite the lack of an obvious nuclear localization signal. TAP-tagging of PNUTS showed significant enrichment of the four other components of the GT-associated protein complex and no other proteins (Supplementary Table S3).

BLASTP searches of LtaP32.3990 returned matches to orthologues in other trypanosomatids, as well as WD40 repeats in proteins from several other organisms, while HHPRED analysis (Supplementary Figure S4A) identified at least three WD40 repeats. Therefore, we will refer to this protein as WD-GT, to distinguish it from numerous other WD40 repeat-containing proteins in *Leishmania*. TAP-tagged WD-GT pulled down the four other components of the GT-associated protein complex, as well as a number of chaperone-associated proteins; including prefoldins, T-complex and heat shock proteins (Supplementary Table S4).

HHPRED analysis of LtaP36.0380 revealed three separate domains with structural similarity to different proteins (Supplementary Figure S4B). The N-terminal domain (residues 2-86) is similar to the central portion of the SWI1 subunit of the yeast SWI/SNF chromatin remodeling complex; while the C-terminal domain (residues 384-485) is related to the N-terminal TFIIS domain of PNUTS. Most importantly, the central portion (residues 137 and 269) is predicted to have substantial structural similarity to the DNA-binding domain (DBD) of JBP1 and *in silico* folding of this region revealed conservation of the functional signature D-(W/F/Y)-x-x-GGTRY motif present in all trypanosomatid JBP1 proteins (**Figure 2A**). In addition, a structural model of the DBD from LtaP36.0380 contains a binding pocket large enough to accommodate the glucose ring of J (**Figure 2B and C**) and preliminary experiments indicate that it binds preferentially to J (personal communication from Anastasis Perrakis at NKI, Amsterdam). While this manuscript was in preparation, this J-binding function of LtaP36.0380 was experimentally confirmed by others (*17*), so the protein was re-named J Binding Protein 3 (JBP3).

The molecular characteristics of the four proteins identified in the GT pulldowns – a PP1 catalytic subunit (PP1Ce), a predicted PP1 regulatory protein (PNUTS), a WD40 repeat protein (WD-GT), and a DNA binding protein (JBP3) – are highly reminiscent of the components of the mammalian PTW/PP1 complex. This complex, which contains PP1C, PNUTS, WDR82, and the DNA-binding protein TOX4, has a role in controlling chromatin structure (*23*, *25*). Importantly, the PTW/PP1 complex was recently found to be a negative regulator of RNAP II elongation rate and that dephosphorylation of the transcription elongation factor Spt5 by PP1C is necessary for transcription termination at polyadenylation sites in mammalian cells (*26*). Thus, our results indicate that GT associates with a PTW/PP1-like complex in *Leishmania* (which we will refer to as PJW/PP1), wherein JBP3 replaces the DNA-binding function of TOX4. We propose that this complex provides a direct molecular link between J and transcription termination in *Leishmania*.

JBP3 is part of another chromatin remodeling complex

While tandem affinity purification of TAP-tagged JBP3 showed >256-fold enrichment of the PJW/PP1 complex proteins (PP1Ce, PNUTS, WD-GT, and GT), another four proteins (encoded by *LtaP35.2400*, *LtaP28.2640*, *LtaP12.0900* and *LtaP14.0150*) were >3000-fold enriched (Supplementary Table S5). To confirm association of these proteins with JBP3 (and each other), we constructed TAP-tagged versions and transfected them into *L. tarentolae* T7-TR cell-lines. Cloning of *LtaP12.0900* failed because of errors in the genome sequence (see below) and transfectants containing the LtaP35.2400 construct did not express the tagged protein (perhaps because over-expression was deleterious for cell growth), but the LtaP14.0150 and LtaP28.2640 transfectants expressed tagged proteins of the expected size (see Supplementary Figure S1B), enabling affinity purification (**Figure 3A**). Subsequent mass

spectrometric analysis of co-purifying proteins showed that LtaP14.0150 (Supplementary Table S6) and LtaP28.2640 (Supplementary Table S7) pull-downs enriched JBP3 and the same JBP3-associated proteins (**Table 2**), indicating that they form a separate JBP3-containing complex. Bioinformatic analyses (described below) suggest that some or all of these proteins in this complex are likely associated with chromatin modification and/or remodeling, prompting us to call it the J3C (for JBP3-associated chromatin) complex.

BLASTP analyses of the other proteins in the J3C complex failed to reveal convincing matches to anything other than orthologues in other trypanosomatids and INTERPROSCAN analysis was uninformative. However, LtaP35.2400 is annotated as a "SET domain-containing protein, putative" in TriTrypDB and HHPRED analysis revealed that the N-terminal region (amino acids 70-227) contains structural similarity to SET domain-containing proteins, while the central portion (residues 355-385) shows weaker similarity to C4-type Zinc finger domains from several unrelated proteins (**Figure 3B** and Supplementary Figure S4C). SET domains, which are usually involved in binding to and/or methylation of histones (*27*), are also present in several other *Leishmania* proteins, so we have named this protein SET-J3C to distinguish it from the others. HHPRED analysis of LtaP14.0150 showed structural similarity to Chromatin organization modifier (Chromo) domains (*28, 29*) from numerous eukaryotic proteins at its N-terminus (amino acids 1-55) and weak similarity to Chromo shadow domains at the C-terminus (*30*) (Supplementary Figure S4D). Therefore, we have dubbed this protein Chromo-J3C and predict that it may be involved in recognition of methylated lysine residues on histone tails. LtaP28.2640 is annotated as a "Hypothetical protein, conserved", but residues 187-227 also show structural similarity to the Chromo shadow domain (Supplementary Figure S4E), and so we have called it CS-J3C. The *LtaP12.0900* gene is misassembled in the *L. tarentolae* reference genome, so we used full-length orthologues from other *Leishmania* genomes for subsequent analyses. However, BLASTP, INTERPROSCAN and HHPRED analyses were uninformative, so we have called this protein HPC-J3C (for hypothetical protein conserved in J3C). Phylogenetic analysis showed that HPC-J3C has poor sequence conservation, even in other trypanosomatids, with orthologues in other genera being shorter than in *Leishmania*. HHPRED analysis of the *T. brucei* orthologue (Tb927.1.4250) showed structural similarity at the C-terminus to subunits from several large protein complexes involved in a variety of processes, including histone remodeling. (Supplementary Figure S4F).

JBP3 interacts with PAF1C

In addition to the components of the PJW/PP1 and J3C complexes, two other proteins (encoded by *LtaP35.2870* and *LtaP29.1270*) were substantially enriched in both JBP3 TAP-tag experiments (Supplementary Table S5). LtaP35.2870 is annotated in TriTrypDB as "RNA polymerase-associated

protein LEO1, putative" and this homology was confirmed by HHPRED analyses (Supplementary Figure S4G). LEO1 is a subunit of the RNAP II-associated factor 1 complex (PAF1C), which facilitates transcription elongation by regulating chromatin modification (*31-33*). Mass spectrometric analysis of proteins that co-purified with TAP-tagged LEO1 (**Figure 4A**) identified three proteins (encoded by *LtaP29.1270*, *LtaP36.4090* and *LtaP29.2750*) that were enriched by >630-fold in both experiments (Supplementary Table S8). The last two are obvious homologues of the PAF1C subunits CDC73 and CTR9, respectively, and were also substantially enriched in one of the two JBP3 TAP-tag experiments (see Supplementary Table S5). LtaP36.4090 contains the Ras-like fold characteristic of C-terminal domain of CDC73 (Supplementary Figure S4H) and is annotated as such on TriTrypDB. LtaP29.2750 contains several tetratricopeptide repeat (TPR) domains implicated in protein-protein interactions and shows considerable overall similarity to CTR9 (Supplementary Figure S4I). Functional studies of the *T. brucei* orthologue (Tb927.3.3220) indicated that it is essential for parasite survival and depletion of its mRNA reduced the expression of many genes involved in regulation of mRNA levels (*34*). The third protein (LtaP29.1270) that was substantially enriched in both replicates of the TAP-tagged LEO1 and JBP3 experiments is not an obvious orthologue of any known PAF1C subunit. This protein is annotated as a "Hypothetical protein, conserved" in TriTrypDB, but HHPRED analysis (Supplementary Figure S4J) revealed a central domain (amino acids 232-358) with structural similarity to the PONY/DCUN1 domain found in DCN (Defective in Cullin Neddylation) proteins that are involved in regulation of ubiquitin ligation cascades (*35*). Our results (**Table 3**) suggest that LtaP29.1270, which we will refer to as DCNL (DCN-like) hereafter, forms an integral part (along with LEO1, CDC73 and CTRL) of a PAF1C-like (PAF1C-L) complex in *Leishmania*).

TAP-tagging of *T. brucei* CTR9 by others (*34*) revealed the same constellation of PAF1C-L subunits (LEO1, CDC73 and DCNL), as well an additional protein (Tb927.7.4030). Close examination of our results revealed that the *Leishmania* orthologue (LtaP14.0860) of Tb927.7.4030 is also enriched in both LEO1 TAP-tag experiments (**Table 3**). While this protein is annotated "Hypothetical protein, conserved" in TriTrypDB, HHPRED analysis revealed an N-terminal (amino acids 2-152) structural similarity to the Plus-3 domain of human RTF1 (Supplementary Figure S4K), a component of human and yeast PAF1C. Thus, LtaP14.0860 (which we have dubbed RTF1L) is likely the functional equivalent of RTF1, but may be less tightly associated with PAF1C-L, at least in *Leishmania*.

Depletion of JBP3 causes defects in transcription termination

Since the results of the TAP-tag experiments presented above suggest that JBP3 is the hub of three separate protein complexes similar to those associated with chromatin modification/remodeling and regulation of transcription in other organisms, we postulated that it is likely to control transcription

termination in *Leishmania*. To test this hypothesis, we used CRISPR/Cas9 (*36*) to delete *JBP3* in the *L. tarentolae* cell-line bearing a tetracycline (Tet)-regulated copy of *JBP3-TAP* (see above). We were able to delete both endogenous copies of *JBP3*, but only when *JBP3-TAP* expression was induced with Tet, suggesting that *JBP3* is an essential gene in *Leishmania*. To interrogate the effect of JBP3 depletion, we grew two stable transfectants lacking both endogenous copies of *JBP3* for 8-11 days (but containing a Tet-regulated copy of *JBP3-TAP*) in the presence or absence of tetracycline **(Figure 5A)**. While cells grown in the presence of drug maintained a constant growth rate (with a generation time of ~9 hours) over the length of the experiment, JBP3-TAP protein levels decreased markedly during the first day after removal of tetracycline, dropping to ~2% of the initial level by day 2 (**Figure 5B**). The growth rate in the absence of drug slowed after day 3, with the generation time increasing to >20 hours on day 6, before returning to almost the WT rate after day 10 (**Figure 5A**).

To assess the role of JBP3 on mRNA expression and transcription termination, RNA was isolated from the cells every day or two after tetracycline withdrawal and used to generate strand-specific RNA-seq libraries. Illumina sequencing reads were mapped to the *L. tarentolae* reference genome and normalized read counts calculated for every gene. Differential expression analysis revealed that *JBP3-TAP* mRNA levels were ~20-fold lower in the absence of Tet **(Figure 5C**). Interestingly, there was a marked increase in the *JBP3-TAP* mRNA levels in the Day 11 Tet- sample, coincident with resumption of normal growth. Consequently, this sample was excluded from subsequent analyses, along with the Day 1 Tet- samples (since JBP3 protein was still ~8% of the initial level). Further analysis (using the DESeq2 module of GENEIOUS) revealed that 17 genes had significantly higher mRNA abundance (>2-fold, $p<0.001$) in the remaining Tet- samples, which all had low JBP3 protein levels (Supplementary Table S9). Interestingly, 14 of these genes are located adjacent (or close) to transcription termination sites (TTSs). Indeed, 34 of the 50 most up-regulated genes are located near TTSs, with 24 of these located at convergent strand-switch regions (cSSRs) where the 3' termini of two PTUs converge.

To further characterize these increases in RNA abundance, we analyzed the read coverage for 5 kb on either side of all 192 TTSs in the *L. tarentolae* genome. As expected, when JBP3 is expressed (Tet+ samples) the median-normalized coverage on the top (coding) strand decreased sharply downstream of the TTS (Supplementary Figure S5A). However, in samples with very low JBP3 protein levels (days 2-9 Tet-), the read coverage downstream of the TTS was significantly higher, suggesting that reduction of JBP3 resulted in substantial transcriptional read-through. Importantly, this increase in read-through transcription did not occur to the same extent at different types of TTS (**Figure 5D**). It was most pronounced at the 23 non-centromeric cSSRs without RNA genes (**Figure 5E** and Supplementary Figure S5B), where transcript abundance was almost as high downstream of the TTS as it was upstream. There

was also a significant increase in read-though transcription downstream of the TTS between unidirectionally (head-to-tail) oriented PTUs (**Figure 5F** and Supplementary Figure S5C); although it was more subtle, since the gap between each PTU is small. Conversely, there was only a small increase in read-through at TTSs upstream of RNA genes transcribed by RNAP III (Supplementary Figure S5D), and essentially no read-through at centromeric (Supplementary Figure S5E) or telomeric TTSs (Supplementary Figure S5F) TTSs. Analysis of bottom (non-coding) strand transcripts revealed no significant differences between Tet+ and Tet- samples (Supplementary Figures S5A-F left-hand panels), except at cSSRs, where read-through from the second PTU results in a substantial increase in antisense transcripts upstream of the TTS (Supplementary Figure S5D). Similar analysis of transcript abundance surrounding transcription start sites (TSSs), revealed no significant changes due to JBP3 depletion, except for a small increase in top (coding) strand coverage when PTUs were oriented unidirectionally (Supplementary Figure S6C and G), presumably due to read-through from the preceding PTU. Importantly, there was little or no increase in bottom (non-coding) strand coverage upstream of most TSSs.

Discussion

Using the trypanosomatid-specific GT, which carries out the second step of J biosynthesis, as an entrée to search for the molecular machinery associated with regulation of transcription in *Leishmania*, we have identified a network of three protein complexes that contain conserved building blocks often used to assemble molecular machinery regulating transcription in other eukaryotes. A novel J-binding protein (JBP3) lies at the nexus of these complexes (**Figure 6**) and provides new insight into the molecular mechanism(s) used to mediate transcription termination at the end of the polycistronic transcription units emblematic of these (and related trypanosomatid) parasites. We have shown that JBP3 plays a central role in controlling termination of RNAP II transcription, since depletion of JBP3 leads to defects in transcriptional termination at the 3' end of PTUs in *Leishmania* (**Figure 5**), just as it does in *T. brucei (17).* However, read-through transcription is not seen to the same extent at all TTSs. The presence of RNAP III-transcribed RNA genes downstream of the TTS appears to effectively block RNAP II, as we have seen previously for JBP2 null mutants (*8*), and there is little or no read-through at TTSs immediately upstream of centromeres and telomeres. This suggests that factors other than JBP3 may play a role in reducing transcriptional read-through at these loci. Alternatively, it is possible that the higher J content at centromeres and telomeres may more effectively "capture" what little JBP3 remains in the Tet-cells. In contrast to the recent results from *T. brucei (17)*, we find little evidence for antisense transcription at the 5' end of PTUs in *Leishmania*, even after depletion of JBP3 levels. Divergent strand-switch regions (dSSRs), where adjacent PTUs are on opposite strands, tend to be smaller in *Leishmania* than those in *T. brucei* and lack the J-containing DNA found in the latter, suggesting that there may be little inappropriate antisense transcription in these regions. However, we do find evidence that depletion of JBP3 in *Leishmania* results in up-regulation of mRNA levels for protein-coding genes at the 3' end of PTUs (Supplementary Table S9). It is possible that this phenomenon is due to more efficient polyadenylation of transcripts caused by uncovering of cryptic trans-splicing sites downstream of the normal TTS. Toxic effects of a gradual accumulation of proteins from these mRNAs may be one explanation for the lag between appearance of defects in transcription termination (day 2) and decrease in growth rate (day 4). This lag period also indicates that read-through transcription if not an artifact of reduced growth rate.

Our initial TAP-tagging experiments showed that GT associates (directly or indirectly) with four other proteins that resemble components of the metazoan PTW/PP1 complex, which contains PNUTS, TOX4, WDR82 and PP1C has been implicated in numerous different cellular processes; including control of chromatin structure during cell cycle progression (*23*), repair of DNA damage by non-homologous end-joining (*37*), maintenance of telomere length (*38*), and developmental regulation of transcription (*39*).

The *Leishmania* PJW/PP1 complex shows obvious parallels to the metazoan PTW/PP1 complex by incorporating analogous (although not necessarily homologous) proteins, with JBP3 substituting for the DNA binding function of TOX4. However, there are some interesting differences between *Leishmania* PJW/PP1 and the homologous *T. brucei* complex, most notably the absence of PP1C in the latter, where the complex is therefore called PJW (*17*). This absence is intriguing in the light of a recent publication that showed the mammalian PTW/PP1 complex dephosphorylates the transcription elongation factor Spt5, thereby causing the RNAP II transcription complex to decelerate within the termination zone downstream of poly(A) sites and allowing the Xrn2 exonuclease to "track down and dislodge" the polymerase from the DNA template (*26*). This suggests that *Leishmania* and American trypanosomes may use PJW/PP1 to dephosphorylate Spt5 and mediate transcription termination, but that African trypanosomes do not use this mechanism. In addition, although *T. brucei* encodes an orthologue of GT, it was not found to be associated with the PJW complex, pointing to another potential biological difference from *Leishmania* (or merely reflecting our use of a more rapid and sensitive purification protocol). In mammalian PTW/PP1C, PNUTS not only contains the phosphatase inhibitor motif, but also contains a nuclear localization signal (NLS) and provides a "scaffold" for recruiting the other proteins. However, *Leishmania* PNUTS is much smaller and lacks an obvious NLS, so it is possible that other proteins in the complex provide these functions. For example, JBP3 contains a domain with structural similarity to the N-terminal TFIIS protein interaction domain found in mammalian PNUTS and WD proteins can act as a scaffold in other complexes (*40, 41*).

JBP3 is also part of a complex (J3C) that contains proteins with domains suggesting a role in chromatin modification and/or remodeling. One protein (SET-J3C) contains a SET domain, which is often present in histone methyltransferases (HMTs) that modify lysine and arginine residues in histone proteins. HMTs usually also contain cysteine-rich pre-SET and post-SET domains that play a crucial role in substrate recognition and enzymatic activity by coordinating zinc ions. SET-J3C contains a Zn finger domain downstream of the SET domain (Supplementary Figure S4C) that may fulfill similar functions, suggesting it is likely to have HMT activity. Two other proteins (Chromo-J3C and CS-J3C) contain Chromo and/or Chromo shadow domains typically involved in recognition of the methylated lysine residues on histone tails and may be functional homologues of the metazoan heterochromatin protein 1 (HP1) and/or fission yeast Swi6, which contain similar domains and are involved in repression of gene expression by heterochromatin (*30*). Thus, it is possible that J3C facilitates re-packing of chromatin at J-containing regions of the genome, rendering them heterochromatin-like and inaccessible to RNA polymerases. The function of the HPC-J3C subunit is unknown at this time, but the *T. brucei* orthologue (encoded by *Tb927.1.4250*) shows structural similarity to subunits from different protein complexes implicated in a several processes, including histone modification and/or chromatin remodeling, and it

localizes to the nucleus (*42*). This suggest that the primary function of HPC-J3C may be to serve as a scaffold for recruitment and/or stabilization of the J3C complex. .

JBP3 also assembles with third protein complex (PAF1C-L) that provides another potential connection between J and regulation of transcription. PAF1C-L contains proteins with functional domains similar to those in yeast and mammalian PAF1C, which associates with the large subunit of RNAP II (*31, 43*) and plays a critical role in transcription elongation and termination (*44, 45*). Four proteins (PAF1, CDC73, LEO1 and CTR9) are consistently found as components of PAF1C in all other eukaryotes, while RTF1 (*46*) and the WD40-containing protein Ski8/WDR61 (*47*) show a less ubiquitous association. The parallels between the mammalian and trypanosomatid complexes are obvious, since homologues of the CDC73, LEO1 and CTR9 subunits copurify with JBP3 and experiments performed by others in *T. brucei* showed that CTR9 has a tight association with LEO and CDC73 (*34*). Our TAP-tag experiments using LEO1 as the bait protein showed enrichment of CDC73 and CTR9, along with a protein containing a Plus-3 domain similar to that in RTF1. Interestingly, the Plus-3 domain of RTF1 has been implicated in binding to Spt5 (*48*), which is tantalizing in light of a potential role for the PJW/PP1 complex in dephosphorylating this transcription factor (*26*). We (and others) failed to identify a convincing homologue of Ski8/WDR61 (or any other WD40 protein) in the PAF1C-L complex, but this protein is not tightly associated with mammalian PAF1C either. Remarkably, the namesake of the complex, PAF1, is absent from pulldowns of Leo and CTR9 in *L. tarentolae* and *T. brucei* respectively, even though it (along with CTR9) is essential for assembly of the complex in both yeast and humans (*49*). Moreover, extensive bioinformatic analysis of the trypanosomatid genomes failed to identify a homolog of PAF1, suggesting that is truly absent from *Leishmania* and *Trypanosoma* PAF1C-L complex. However, PAF1C-L contains an additional, trypanosomatid-specific, component (DCNL) that has a putative protein-binding domain with structural similarity to the PONY/DCUN1 domain found in the eukaryotic DCN protein family. In other eukaryotes, DCN1 is required for neddylation of cullin in SCF-type E3 ubiquitin ligase complexes that mark cellular proteins for proteosomal degradation (*50*). It is interesting to speculate that DCNL may be involved in PAF1C-L recruitment/function by interaction with N-terminal acetylated residues on histones and/or other chromatin-associated proteins. Whether DCNL functionally replaces PAF1 will remain an open question until its molecular function is dissected in more detail. The lack of significant reciprocal enrichment of JBP3 in the LEO1-TAP pulldowns (this paper) and its absence from the CTR9-TAP pulldowns in *T. brucei* (*34*) suggests a more transient association of JBP3 with PAF1C-L.

We have previously postulated that J might terminate transcription by directly preventing progression of polymerase (*8*). The data presented here create an alternative hypothesis: namely that JBP3 binding to J recruits PJW/PP1 and J3C complexes to chromatin, stalling RNAP II. Extending this model,

we suggest that the interaction between JBP3 and PAF1C-L may also tether the RNAP II to the termination zone, where the DCNL subunit ubiquitinates the polymerase complex, promoting its degradation by the proteasome (*51, 52*). While our findings provide several novel insights into the role of base J in transcription termination, they also raise several interesting questions. For example, why is the GT part of the PJW/PP1 complex? One could envisage that recruitment of the PJW/PP1 complex to regions of the genome containing J recruits may allow more efficient glucosylation of nearby HOMedU. What histone modification/remodeling is mediated by J3C and what role do they play a role in transcription termination? The availability of modern genome-wide approaches will no doubt provide the appropriate tools to answer these questions.

Materials and methods

Plasmid construction

To create an expression vector that expresses epitope-tagged transgenes in *Leishmania*, the MHTAP tag (which bears a Myc epitope, six histidines, a protein A domain, and calmodulin binding peptide) was amplified from the plasmid pLEW-MHTAP (*53*) with the primers MHTAP-BamHI-S and MHTAP-Not1-AS (all primers use in this study are described in Supplementary Table S10). Following cleavage with BamHI and NotI, the PCR fragment was inserted into BglII+Not1-digested pLEXSY-I-bleCherry3 (Jena Biosciences). The resulting plasmid (pLEXSY-MHTAP) allows the TAP-tagging of introduced coding regions under the control of a Tet-regulated T7 promoter, and insertion into the *ODC* locus on chromosome 12 of *L. tarentolae*. We used a combination of published datasets to identify the 5' end of *L. tarentolae* mRNAs as marked by the 39 base pair splice leader sequence (*8*) and ribosome profiling data from *L. donovani* (unpublished data), to identify the correct CDS for bait proteins. CDSs were PCR amplified and digested with the restriction enzymes indicated in Supplementary Table S10 prior to cloning.

Parasite strains and tissue culture

The *Leishmania tarentolae* Parrot-TarII wild-type (WT) and T7-TR strains (Jena Bioscience) were grown in SDM-79 medium supplemented with 10% fetal bovine serum. Strain T7-TR has constitutively expressed T7 RNA polymerase and Tet repressor genes integrated into the rDNA locus, allowing for Tet-induced expression of integrated (or ectopically expressed) genes. Nourseothricin and hygromycin B were added to the medium at 100 µM to maintain expression of T7 RNA polymerase and Tet repressor (TetR).

Tandem affinity purification of tagged protein complexes

Ten µg of *Swa*I-digested pLEXSY-MHTAP plasmid encoding a TAP-tagged protein was electroporated into the *L. tarentolae* WT and T7-TR cell-lines as described (*54*) and transfectants selected with 100 µg/ml bleomycin and maintained in 20 µg/ml bleomycin (plus nourseothricin and hygromycin B as described above). Proteins associated with the TAP-tagged "bait" were purified from 500 ml of cells following overnight culture (in the presence of 2 µg/ml tetracycline for T7-TR transfectants) as described (*53*), except that NP-40 was omitted from the final four washes of the proteins on the calmodulin beads and from the calmodulin elution buffer. The protocol went from lysis of cells to purified samples within 6 hours. A sample from each pulldown (5% of the total eluate) was separated by 4-20% SDS-PAGE and proteins were visualized using SilverQuest Stain (Thermo Fisher Scientific Life Technologies). Fractions containing a protein with the predicted molecular weight of the bait (usually fractions 2 and 3) were

pooled. The same fractions were pooled from mock TAP purifications of the control parental line not expressing any bait protein.

Western blotting

Proteins from transfected cells were separated by SDS-PAGE on 4-20% gradient gels, transferred onto nitrocellulose and detected with either mouse anti-6×His (Clontech) at 0.25 µg/ml or rabbit antibody against calmodulin binding peptide Calmodulin Binding Peptide (GenScript) at 0.1 µg/ml, with rabbit antibody against *T. brucei* phosphoglycerate kinase serving as a control (*55*). Primary antibodies were detected with goat anti-rabbit Ig conjugated with AlexaFluor 680 (50 ng/ml) or goat anti-mouse Ig conjugated with IRDye 800 (25 ng/ml) and imaged on the LI-COR Odyssey CLX.

Proteomic analysis

Pooled protein fractions were denatured with 6 M urea, reduced with 5 mM dithiothreitol, alkylated with 25mM iodoacetamide, and digested at 37°C for three hours using 1:200 w:w endoproteinase Lys-C (Thermo Fisher Scientific). The urea was then diluted to 1.5 M and samples further digested at 37°C overnight with 1:25 w:w trypsin (Thermo Fisher Scientific). Proteinase activity was stopped with formic acid, and peptides purified using C18 reversed-phase chromatography (Waters), followed by hydrophilic interaction chromatography (HILIC; Nest Group). Purified peptides were separated by online nanoscale HPLC (EASY-nLC II; Proxeon) with a C18 reversed-phase column (Magic C18 AQ 5um 100A) over an increasing 90 minute gradient of 5-35% Buffer B (100% acetonitrile, 0.1% formic acid) at a flow rate of 300 nl/min. Eluted peptides were analyzed with an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) operated in data-dependent mode, with the 15 most intense ions per MS1 survey scan selected for MS2 fragmentation by rapid collision-induced dissociation (rCID) (*56*). MS1 survey scans were performed in the Orbitrap at a resolution of 240,000 at m/z 400 with charge state rejection enabled, while rCID MS2 was performed in the dual linear ion trap with a minimum signal of 1000. Dynamic exclusion was set to 15 seconds.

Raw output data files were analyzed using Maxquant (v1.5.3.30) (*57*) to search against a proteome predicted after resequencing and annotation of the *L. tarentolae* Parrot (LtaP) genome (Sur *et al*, manuscript in preparation). A reverse sequence decoy database was used to impose a strict 1% FDR cutoff. Label-free quantification was performed using the MaxLFQ algorithm (*58*) and further data processing was performed in PERSEUS (v1.5.3.1) (*59*) and Microsoft EXCEL. To avoid zero-value denominators, null values in the remaining data were replaced by imputation using background signal within one experiment using PERSEUS. Non-parasite contaminants, decoys, and single peptide identifications among all samples in an experiment were removed. Proteins were deemed to be part of a

complex associated with the bait protein if at least two peptides were detected and the protein showed more than 32-fold (log_2 >5) enrichment (compared to the control not expressing the bait) in both replicates. In a few cases, proteins showing >100-fold enrichment in a single replicate only, were also considered as potential subunits of the complex. Proteins with 1024-fold (log2 <10) less enrichment than the bait protein were assumed to be co-purifying contaminants and (usually) ignored. Mass spectrometry data was deposited in the MassIVE database (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) and can be accessed from ProteomeXchange using the identifier PXD020779.

Bioinformatic analysis of protein function

Structure-based similarity searches for known domains were performed with HHPRED (*19*). Domain boundaries for JBP3 DNA binding domain (DBD) were refined by aligning trypanosomatid sequences that clustered in the same OrthoMCL group as JBP3 with T-Coffee (*60*). Homology models of the JBP3-DBD domain using RosettaCM (*61*) were built with the *L. tarentolae* JBP1 structure (PDB: 2XSE) as a template. The top-scoring model covered residues 111 to 312 of JBP3-DBD with a confidence score of 0.67.

Deletion of JBP3 using SaCas9

JBP3 was deleted using *Staphylococcus aureus* Cas9 (SaCas9)-directed cleavage of sites flanking the endogenous locus as described (*36*). Briefly, guide RNAs directed at sites for SaCas9 cleavage were generated *in vitro* using T7 Megashort from ThermoFisher from PCR-generated templates. The 5' and 3' gRNA sequences used were GATGTGAAACGCTAAGCAGTCCCGAGT and AGGAACGAAAGCACACAGCAGAGGAGT, where the PAM sites are underlined. Repair fragments containing a drug resistance gene were generated as described (*62*) from pTNeo or pTPuro templates using primers LtJBP3-up and LtJBP3-down (Supplementary Table S10). Heat-denatured guide RNA was complexed with 20 µg SaCas9 recombinant protein at equimolar ratio and incubated for 15 minutes at room temperature before mixing with 2 ug of each repair fragment that had been ethanol-precipitated and resuspend in Tb-BSF (*63*). *L. tarentolae* T7/TR cells were grown overnight (in the presence of 2 ug/ml tetracycline for the latter), pelleted, washed with PBS, and resuspended in 100 µl Tb-BSF. For each transfection, 10^6 cells were mixed with the SaCas9/guide RNA complexes and repair fragments and electroporated in an Amaxa Nucleofector using program X-001. Immediately following transfection cells were split into three flasks. After allowing the cells to recover overnight, antibiotics were added to the flasks; one of which was grown in 10 µg/ml G418, one in 4 µg/ml puromycin, and one with both drugs. Deletion of the endogenous *JBP3* gene(s), was confirmed by PCR amplification of genomic DNA using primers JBP3-M84P and JBP3-P2147M (Supplementary Table S10) that flank the region being deleted. Clones cell-lines were obtained by limiting dilution of the transfectants and clones were retested for JBP3

deletion by PCR. While we were able to obtain clones where both endogenous copies of *JBP3* were deleted by selection with either puromycin or neomycin, we were unable to obtains lines where both drugs were used.

RNA-seq analysis

RNA was isolated using TRIzol (Thermo Fisher Scientific)) and resuspended in 10mM Tris, pH 7and RNA quality assessed using the Bioanalyzer 6000 Pico Chip (Agilent). mRNA was isolated from 1 µg total RNA using the NEB Poly(A) mRNA Magnetic Isolation Module (NEB) and prepared using the Stranded RNA-seq protocol (*64*), modified for *Leishmania* as described (*65*). Libraries were sequenced on an Illumina HiSeq, obtaining Paired End 150-bp reads. Reads were aligned against our in-house *L. tarentolae* genome with BOWTIE2 (*66*) using the "very high sensitivity" parameter or the GENEIOUS assembler (Geneious Prime 11.05, https://www.geneious.com) using the "Low Sensitivity/Fastest" option. Differential expression analysis was performed on the GENEIOUS assemblies using the DESeq2 module. Strand-specific read coverage was calculated directly from BAM files of the BOWTIE2 alignments using customized pysam scripts (https://github.com/pysam-developers/pysam). RNA-seq data was deposited in the Sequence Read Archive (SRA) at NCBI and can be accessed using the BioProject Identifier PRJNA657890.

Acknowledgements

We thank Lisa Jones of the Fred Hutchison Cancer Research Center Proteomics Resources facility for assistance with performing LC-MS/MS analysis. This work was supported in part by PHS grant R01 AI103858 (to P.J.M.) and contract HHSN272201700059C (to P.J.M.) from the National Institute of Allergy and Infectious Diseases.

Author contributions

Concept and experimental design were done by B.C.J and P.J.M. Experiments were performed by B.C.J and J.R.M. Proteomic analysis was done by M.A.G. and J.A.R. Protein modeling and sequence analysis was done by I.Q.P. Analysis of RNA-seq data was done by A.S and P.J.M. Manuscript was written by B.C.J., M.P. and P.J.M. M.P. cracked the whip. All authors have reviewed the manuscript.

Conflicts of interest

The authors declare no competing interest.

References

1. J. H. Gommers-Ampt *et al.*, beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. *Cell* **75**, 1129-1136 (1993).
2. F. van Leeuwen, R. Kieft, M. Cross, P. Borst, Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Mol. Cell. Biol.* **18**, 5643-5651 (1998).
3. F. van Leeuwen *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.* **95**, 2366-2371 (1998).
4. F. Van Leeuwen *et al.*, The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands. *Nucleic Acids Res.* **24**, 2476-2482 (1996).
5. F. van Leeuwen *et al.*, Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes Dev.* **11**, 3232-3241 (1997).
6. F. van Leeuwen, R. Kieft, M. Cross, P. Borst, Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **109**, 133-145 (2000).
7. L. J. Cliffe, T. N. Siegel, M. Marshall, G. A. Cross, R. Sabatini, 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 (2010).
8. H. G. van Luenen *et al.*, Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell* **150**, 909-921 (2012).
9. C. A. Marques, N. J. Dickens, D. Paape, S. J. Campbell, R. McCulloch, Genome-wide mapping reveals single-origin chromosome replication in Leishmania, a eukaryotic microbe. *Genome Biol* **16**, 230 (2015).
10. M. R. Garcia-Silva *et al.*, Identification of the centromeres of *Leishmania major:* revealing the hidden pieces. *EMBO Rep* **18**, 1968-1977 (2017).
11. A. Sekar, C. Merritt, L. Baugh, K. Stuart, P. J. Myler, Tb927.10.6900 encodes the glucosyltransferase involved in synthesis of base J in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **196**, 9-11 (2014).
12. T. Heidebrecht *et al.*, The structural basis for recognition of base J containing DNA by a novel DNA binding domain in JBP1. *Nucleic Acids Res* **39**, 5715-5728 (2011).
13. C. DiPaolo, R. Kieft, M. Cross, R. Sabatini, Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Mol. Cell* **17**, 441-451 (2005).
14. L. J. Cliffe *et al.*, JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res* **37**, 1452-1462 (2009).
15. P. A. Genest *et al.*, Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. *Nucleic Acids Res* **33**, 1699-1709 (2005).
16. P. A. Genest *et al.*, Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic Acids Res* **43**, 2102-2115 (2015).
17. R. Kieft *et al.*, Identification of a novel base J binding protein complex involved in RNA polymerase II transcription termination in trypanosomes. *PLoS Genet* **16**, e1008390 (2020).
18. B. Lesage *et al.*, A complex of catalytically inactive protein phosphatase-1 sandwiched between Sds22 and inhibitor-3. *Biochemistry (Mosc).* **46**, 8909-8919 (2007).
19. L. Zimmermann *et al.*, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237-2243 (2018).

589 20. Y. M. Kim *et al.*, PNUTS, a protein phosphatase 1 (PP1) nuclear targeting subunit.
590 Characterization of its PP1- and RNA-binding domains and regulation by phosphorylation. *J. Biol.*
591 *Chem.* **278**, 13819-13828 (2003).
592 21. J. P. Kreivi *et al.*, Purification and characterisation of p99, a nuclear modulator of protein
593 phosphatase 1 activity. *FEBS Lett.* **420**, 57-62 (1997).
594 22. P. B. Allen, Y. G. Kwon, A. C. Nairn, P. Greengard, Isolation and characterization of PNUTS, a
595 putative protein phosphatase 1 nuclear targeting subunit. *J. Biol. Chem.* **273**, 4089-4095 (1998).
596 23. J. H. Lee, J. You, E. Dobrota, D. G. Skalnik, Identification and characterization of a novel human
597 PP1 phosphatase complex. *J. Biol. Chem.* **285**, 24466-24476 (2010).
598 24. M. S. Choy *et al.*, Understanding the antagonism of retinoblastoma protein dephosphorylation
599 by PNUTS provides insights into the PP1 regulatory code. *Proc. Natl. Acad. Sci. U. S. A.* **111**,
600 4097-4102 (2014).
601 25. H. B. Landsverk, M. Kirkhus, M. Bollen, T. Kuntziger, P. Collas, PNUTS enhances in vitro
602 chromosome decondensation in a PP1-dependent manner. *Biochem. J.* **390**, 709-717 (2005).
603 26. M. A. Cortazar *et al.*, Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation
604 Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Mol. Cell* **76**, 896-908 e894
605 (2019).
606 27. S. C. Dillon, X. Zhang, R. C. Trievel, X. Cheng, The SET-domain protein superfamily: protein lysine
607 methyltransferases. *Genome Biol* **6**, 227 (2005).
608 28. R. Paro, Imprinting a determined state into the chromatin of Drosophila. *Trends Genet.* **6**, 416-
609 421 (1990).
610 29. R. Paro, D. S. Hogness, The Polycomb protein shares a homologous domain with a
611 heterochromatin-associated protein of Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 263-267
612 (1991).
613 30. R. Aasland, A. F. Stewart, The chromo shadow domain, a second chromo domain in
614 heterochromatin-binding protein 1, HP1. *Nucleic Acids Res* **23**, 3168-3173 (1995).
615 31. O. Rozenblatt-Rosen *et al.*, The parafibromin tumor suppressor protein is part of a human Paf1
616 complex. *Mol. Cell. Biol.* **25**, 612-620 (2005).
617 32. R. Pavri *et al.*, Histone H2B monoubiquitination functions cooperatively with FACT to regulate
618 elongation by RNA polymerase II. *Cell* **125**, 703-717 (2006).
619 33. N. J. Krogan *et al.*, RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted
620 proteomics approach. *Mol. Cell. Biol.* **22**, 6979-6992 (2002).
621 34. B. A. Ouna *et al.*, Depletion of trypanosome CTR9 leads to gene expression defects. *PLoS One* **7**,
622 e34256 (2012).
623 35. D. C. Scott *et al.*, Blocking an N-terminal acetylation-dependent protein interaction inhibits an E3
624 ligase. *Nat Chem Biol* **13**, 850-857 (2017).
625 36. L. C. Soares Medeiros *et al.*, Rapid, Selection-Free, High-Efficiency Genome Editing in Protozoan
626 Parasites Using CRISPR-Cas9 Ribonucleoproteins. *MBio* **8**, (2017).
627 37. S. Zhu, L. A. Fisher, T. Bessho, A. Peng, Protein phosphatase 1 and phosphatase 1 nuclear
628 targeting subunit-dependent regulation of DNA-dependent protein kinase and non-homologous
629 end joining. *Nucleic Acids Res* **45**, 10583-10594 (2017).
630 38. H. Kim *et al.*, TRF2 functions as a protein hub and regulates telomere maintenance by
631 recognizing specific peptide motifs. *Nat Struct Mol Biol* **16**, 372-379 (2009).
632 39. A. Ciurciu *et al.*, PNUTS/PP1 Regulates RNAPII-Mediated Gene Expression and Is Necessary for
633 Developmental Growth. *PLOS Genetics* **9**, e1003885 (2013).
634 40. C. U. Stirnimann, E. Petsalaki, R. B. Russell, C. W. Muller, WD40 proteins propel cellular
635 networks. *Trends Biochem. Sci.* **35**, 565-574 (2010).

636 41. M. Schapira, M. Tyers, M. Torrent, C. H. Arrowsmith, WD40 repeat domain proteins: a novel
637 target class? *Nat Rev Drug Discov* **16**, 773-786 (2017).
638 42. S. Dean, J. D. Sunter, R. J. Wheeler, TrypTag.org: A Trypanosome Genome-wide Protein
639 Localisation Resource. *Trends Parasitol* **33**, 80-82 (2017).
640 43. A. Yart *et al.*, The HRPT2 tumor suppressor gene product parafibromin associates with human
641 PAF1 and RNA polymerase II. *Mol. Cell. Biol.* **25**, 5052-5060 (2005).
642 44. B. N. Tomson, K. M. Arndt, The many roles of the conserved eukaryotic Paf1 complex in
643 regulating transcription, histone modifications, and disease states. *Biochim Biophys Acta* **1829**,
644 116-126 (2013).
645 45. S. B. Van Oss, C. E. Cucinotta, K. M. Arndt, Emerging Insights into the Roles of the Paf1 Complex
646 in Gene Regulation. *Trends Biochem. Sci.* **42**, 788-798 (2017).
647 46. C. L. Mueller, J. A. Jaehning, Ctr9, Rtf1, and Leo1 are components of the Paf1/RNA polymerase II
648 complex. *Mol. Cell. Biol.* **22**, 1971-1980 (2002).
649 47. B. Zhu *et al.*, The human PAF complex coordinates transcription with events downstream of RNA
650 synthesis. *Genes Dev.* **19**, 1668-1673 (2005).
651 48. M. K. Mayekar, R. G. Gardner, K. M. Arndt, The recruitment of the Saccharomyces cerevisiae
652 Paf1 complex to active genes requires a domain of Rtf1 that directly interacts with the Spt4-Spt5
653 complex. *Mol. Cell. Biol.* **33**, 3259-3273 (2013).
654 49. Y. Xie *et al.*, Paf1 and Ctr9 subcomplex formation is essential for Paf1 complex assembly and
655 functional regulation. *Nat Commun* **9**, 3795 (2018).
656 50. T. Kurz *et al.*, The conserved protein DCN-1/Dcn1p is required for cullin neddylation in C. elegans
657 and S. cerevisiae. *Nature* **435**, 1257-1261 (2005).
658 51. Y. Nakazawa *et al.*, Ubiquitination of DNA Damage-Stalled RNAPII Promotes Transcription-
659 Coupled Repair. *Cell* **180**, 1228-1244 e1224 (2020).
660 52. A. Tufegdzic Vidakovic *et al.*, Regulation of the RNAPII Pool Is Integral to the DNA Damage
661 Response. *Cell* **180**, 1245-1261 e1221 (2020).
662 53. B. C. Jensen *et al.*, Characterization of protein kinase CK2 from Trypanosoma brucei. *Mol.
663 Biochem. Parasitol.* **151**, 28-40 (2007).
664 54. G. M. Kapler, K. Zhang, S. M. Beverley, Nuclease mapping and DNA sequence analysis of
665 transcripts from the dihydrofolate reductase-thymidylate synthase (R) region of Leishmania
666 major *Nucleic Acids Res.* **18**, 6399-6408 (1990).
667 55. H. L. Parker *et al.*, Three genes and two isozymes: Gene conversion and the
668 compartmentalization and expression of the phosphoglycerate kinases of Trypanosoma
669 (Nannomonas) congolense *Mol. Biochem. Parasitol.* **69**, 269-279 (1995).
670 56. A. Michalski *et al.*, Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap
671 Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular &
672 cellular proteomics : MCP* **11**, O111 013698 (2012).
673 57. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range
674 mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372
675 (2008).
676 58. J. Cox *et al.*, Accurate proteome-wide label-free quantification by delayed normalization and
677 maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**,
678 2513-2526 (2014).
679 59. S. Tyanova *et al.*, The Perseus computational platform for comprehensive analysis of
680 (prote)omics data. *Nat Methods* **13**, 731-740 (2016).
681 60. P. Di Tommaso *et al.*, T-Coffee: a web server for the multiple sequence alignment of protein and
682 RNA sequences using structural information and homology extension. *Nucleic Acids Res* **39**,
683 W13-17 (2011).

684 61. Y. Song *et al.*, High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735-1742
685 (2013).
686 62. T. Beneke *et al.*, A CRISPR Cas9 high-throughput genome editing toolkit for kinetoplastids. *R Soc*
687 *Open Sci* **4**, 170095 (2017).
688 63. G. Schumann Burkard, P. Jutzi, I. Roditi, Genome-wide RNAi screens in bloodstream form
689 trypanosomes identify drug transporters. *Mol. Biochem. Parasitol.* **175**, 91-94 (2011).
690 64. A. G. Hunt, A rapid, simple, and inexpensive method for the preparation of strand-specific RNA-
691 Seq libraries. *Methods Mol. Biol.* **1255**, 195-207 (2015).
692 65. P. J. Myler, J. A. McDonald, P. J. Alcolea, A. Sur, Quantitative RNA Analysis Using RNA-Seq.
693 *Methods Mol. Biol.* **1971**, 95-108 (2019).
694 66. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359
695 (2012).

696

Tables

Table 1. Enrichment of proteins in the PJW/PP1 complex

Gene ID	Name	HHpred/InterPro domains	Tagged protein				
			HmdUGT	PP1Ce	PNUTS	JBP3	WD-GT
LtaP36.2450	HmdUGT	None found	14.1	7.2	8.2	10.6	9.2
LtaP15.0230	PP1Ce	Serine/threonine-protein phosphatase	7.4	17.1	12.2	10.9	13.9
LtaP33.1440	PNUTS	PPP1R10/PNUTS	(9.8)	12.5	17.7	10.3	13.2
LtaP36.0380	JBP3	SWI1, J-binding & N-terminal TFIIS	11.1	12.3	10.3	15.9	12.5
LtaP32.3990	WD-GT	WD40 repeats	9.3	14.8	12.7	12.3	16.3

Enrichment is expressed as the mean $\log_2$ fold-change compared to control pulldowns.
Parentheses indicate that enrichment was observed in only one replicate.

Table 2. Enrichment of proteins in the JBP3-associated chromatin remodeling complex

Gene ID	Name	HHpred/InterPro domains	Tagged protein		
			JBP3	J3C-Chromo	J3C-CS
LtaP36.0380	JBP3	SWI1, J-binding & N-terminal TFIIS	15.9	7.9	7.8
LtaP35.2400	SET-J3C	SET & C4-type Zn finger	12.8	8.2	9.4
LtaP14.0150	Chromo-J3C	Chromo & chromo shadow	11.7	11.6	8.8
LtaP28.2640	CS-J3C	Chromo shadow	12.5	9.8	11.8
LtaP12.0900	HPC-J3C	None found	12.1	9.8	6.3

Table 3. Enrichment of proteins in the PAF1C-L complex

Gene ID	Name	HHpred/InterPro domains	Tagged protein	
			JBP3	LEO1
LtaP35.2870	LEO1	Leo1-like protein	7.4	16.8
LtaP29.1270	DCNL	PONY/DCUN1 domain	5.4	13.9
LtaP36.4090	CDC73	Cell division control protein 73 (C-terminal)	(5.7)	10.4
LtaP29.2750	CTR9	RNA polymerase-associated protein Ctr9	(6.2)	9.5
LtaP14.0860	RTF1L	Plus-3 domain	0.6	3.8

Parentheses indicate that enrichment was observed in only one replicate.

Figure legends

Figure 1. The PJW/PP1 complex

A. Proteins in the peak fractions from TAP purifications of HmdUGT and other components of the PJW/PP1 complex (WD-GT, JBP3, PP1Ce and PNUTS) were separated by 4-20% SDS-PAGE and silver stained. Each lane represents 5% of the total fraction. The TAP-tagged "bait" protein is indicated by an asterisk. The first lane shows the equivalent fraction from a mock purification of control (T7-TR) cells.

B. Schematic representation of key domains in each of the associated proteins. The PNUTS domain is involved in interaction with PP1C; the SNF2/SWI6 domain is involved in protein interaction and is usually found in chromatin binding proteins; the JBP1 DNA binding domain (DBD) is involved in binding to base J; the N-terminal TIIFS domain is involved in protein interaction; WD40 domains are also involved in protein-protein interaction; and the PP1C domain shows the region of sequence homology to the PP1 catalytic subunit.

Figure 2. Modeling of the JBP3 DNA-binding domain

A. Sequence alignment of the putative JBP3 J-binding domain from *T. brucei* EATRO927 strain, *T. cruzi* Silvio strain, and *L. tarentolae* Parrot strain. Residues that are identical or conservatively replaced in all three species are shaded black, while those that are identical or conserved in two species are shaded grey.

B. The structure of the DNA binding domain from JBP3 (light blue) was modeled using RosettaCM against the J-binding domain of JBP1 (tan) from PDB entry 2XSE. The interaction between the conserved aspartic acid residue (Asp_{525} in JBP1 and Asp_{241} in JBP3-) with the glucose of base J is shown.

C. Closeup of the of the interaction between conserved aspartate of both proteins and base J.

Figure 3. The JBP3-associated Chromatin complex

A. Proteins that co-purify with TAP-tagged Chromo-J3C and CS-J3C analyzed by SDS-PAGE and silver staining as described in Fig. 1.

B. Schematic representation showing the key domains of the four proteins that co-purified with TAP-tagged JBP3. The SET domain is most often found protein methyltransferases; the Chromo domain is usually found chromatin binding proteins; the Chromo Shadow is distantly related to the Chromo domain and frequently found in chromatin binding proteins with the Chromo domains.

Figure 4. The PAF1C-like complex

A. Proteins that co-purified with TAP-tagged LEO1 were analyzed by SDS-PAGE and silver-staining as described in Fig. 1.

B. Schematic representation showing the key domains of five components of the PAF1C-L complex. LEO1 shows the region of sequence homology with mammalian LEO1; the CUE domain is involved in ubiquitin binding; CDC73 shows the region of sequence homology with mammalian CDC73; the CTR9/TPR domain contains tetratricopeptide repeats involved in protein-protein interaction; the Pony/DCUN1 domain is involved in binding and neddylation of the cullin subunit of E1-type ubiquitin ligases; and the Plus-3 is implicated in binding to single stranded DNA.

Figure 5. Depletion of JPB3 results in read through at transcription termination sites

A. Growth analysis. Two independently generated *L. tarentolae* clones lacking endogenous *JBP3* but containing a Tet-regulated TAP-tagged *JBP3* gene was grown in the presence (blue lines) or absence (orange lines) of Tet. The numbers on y-axis are corrected for dilution during sub-culturing. The solid lines show a clone where the *JBP3* genes were replaced by *pac* (and grown in the presence of puromycin), while the dotted lines show a clone where with the *JBP3* genes replaced by *neo* (and grown in G418).

B. The level of TAP-tagged JBP3 expressed by the T7-TR/JBP3-MHTAP/Δ*jbp3*::*neo* clone grown in the absence of Tet was monitored by Western blot analysis using antibodies against the calmodulin binding peptide (CBP) of the TAP-tag. Antibodies against phosphoglycerate kinase (PGK) served as a loading control. The percent JBP3 levels in comparison to day 0 is shown below the anti-CBD blot.

C. *JBP3-TAP* mRNA levels for the T7-TR/JB3-MHTAP/Δ*jbp3*::*neo* clone grown in the presence (blue) or absence (orange) or Tet for the number of days indicated. mRNA levels are expressed transcripts per million (TPM) as determined by RNA-seq analysis using GENEIOUS.

D. Box-and-whiskers plots showing the median top strand coverage in the 5 kb downstream of all 192 TTSs (All). Separate plots are shown for the 46 TTS at cSSRs (Conv), 30 TTSs between head-to-tail PTUS (Uni), 39 TTS immediately upstream of one or more RNA genes (RNA), 21 TTS adjacent to a centromere (Cent) and 56 TTSs at telomeres (Telo).

E. Median of top strand coverage at each nucleotide position in the 10 kb surrounding the 46 TTS at cSSRs. The schematic represents the protein-coding genes associated with each strand at an "average" cTTS. The second PTU at each cSSR is re-oriented so that the genes are represented on the top strand.

F. Median of top strand coverage at each nucleotide position in the 10 kb surrounding the 30 TTS at between unidirectional (head-to-tail) PTUs. The schematic represents the protein-coding genes associated with each strand at an "average" uTTS.

Figure 6. Network of interactions between JBP3-assocated protein complexes.

Solid lines denote proteins enriched in both replicates of the TAP-tag pull-downs and dashed lines indicate proteins enriched in only one sample. Double headed lines represent reciprocal enrichment with both proteins used as bait, while lines with a single arrowhead indicate where reciprocal enrichment not attempted (except in the case of the LEO1 pull-down, where JBP3 was not detected). Grey arrows represent interactions identified by co-purification of proteins with CTR9 in *T. brucei* (*34*). Subunits within the three distinct JBP3-associated protein complexes are denoted by different colors, while the individual components are not shown for within three additional protein complexes represented by boxes. The dotted line connecting the PAFC1-L complex to RNAP II reflects the interactions observed in other organisms.

List of supplementary information

Supplementary tables

Tables S1-S8. Proteins that co-purify with TAP-tagged bait proteins

Table S9. Genes with increased mRNA levels after depletion of JBP3

Table S10. Oligonucleotide primers used for construct creation

Supplementary figures

Figure S1. Expression of proteins tagged with MHTAP.

Figure S2. Phylogenetic tree for PP1C proteins from selected kinetoplastid species in comparison to proteins from human and *saccharomyces cerevisiae*.

Figures S3 and S4. HHPRED identification of domains in JBP3-associated proteins

Figure S5. Effect of depletion of JBP3 on transcription termination

Figure S6. Effect of depletion of JBP3 on transcription initiation

Figure S7. Map of plasmid pLEXSY-MHTAP showing key features and restriction sites.

Supplementary data file

Excel spreadsheet showing proteins identified by mass spectrometry of TAP-tag pull-downs.

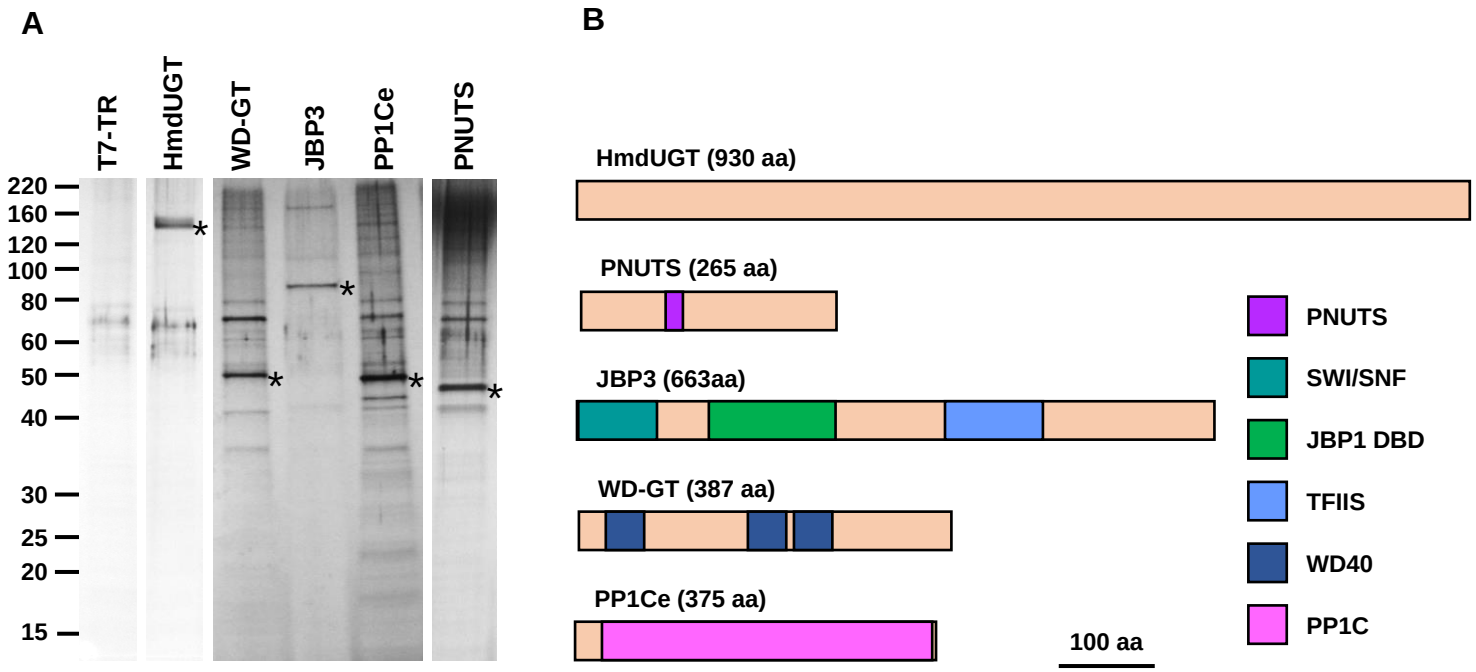

Figure 1. The PJW/PP1 complex

A. Proteins in the peak fractions from TAP purifications of HmdUGT and other components of the PJW/PP1 complex (WD-GT, JBP3, PP1Ce and PNUTS) were separated by 4-20% SDS-PAGE and silver stained. Each lane represents 5% of the total fraction. The TAP-tagged "bait" protein is indicated by an asterisk. The first lane shows the equivalent fraction from a mock purification of control (T7-TR) cells.

B. Schematic representation of key domains in each of the associated proteins. The PNUTS domain is involved in interaction with PP1C; the SNF2/SWI6 domain is involved in protein interaction and is usually found in chromatin binding proteins; the JBP1 DNA binding domain (DBD) is involved in binding to base J; the N-terminal TIIFS domain is involved in protein interaction; WD40 domains are also involved in protein-protein interaction; and the PP1C domain shows the region of sequence homology to the PP1 catalytic subunit.

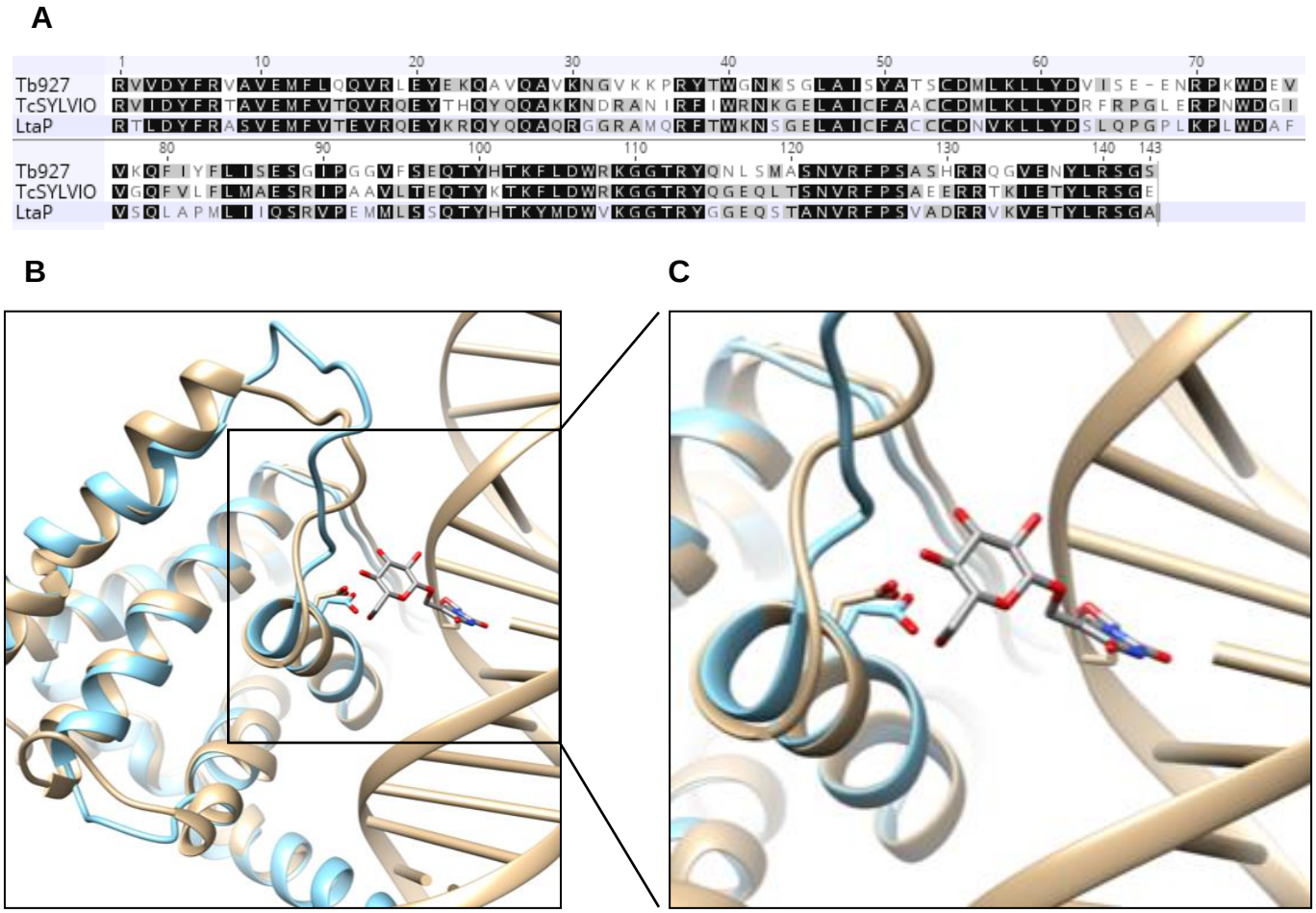

Figure 2. Modeling of the JBP3 DNA-binding domain

A. Sequence alignment of the putative JBP3 J-binding domain from *T. brucei* EATRO927 strain, *T. cruzi* Silvio strain, and *L. tarentolae* Parrot strain. Residues that are identical or conservatively replaced in all three species are shaded black, while those that are identical or conserved in two species are shaded grey.
B. The structure of the DNA binding domain from JBP3 (light blue) was modeled using RosettaCM against the J-binding domain of JBP1 (tan) from PDB entry 2XSE. The interaction between the conserved aspartic acid residue (Asp_{525} in JBP1 and Asp_{241} in JBP3-) with the glucose of base J is shown.
C. Closeup of the of the interaction between conserved aspartate of both proteins and base J.

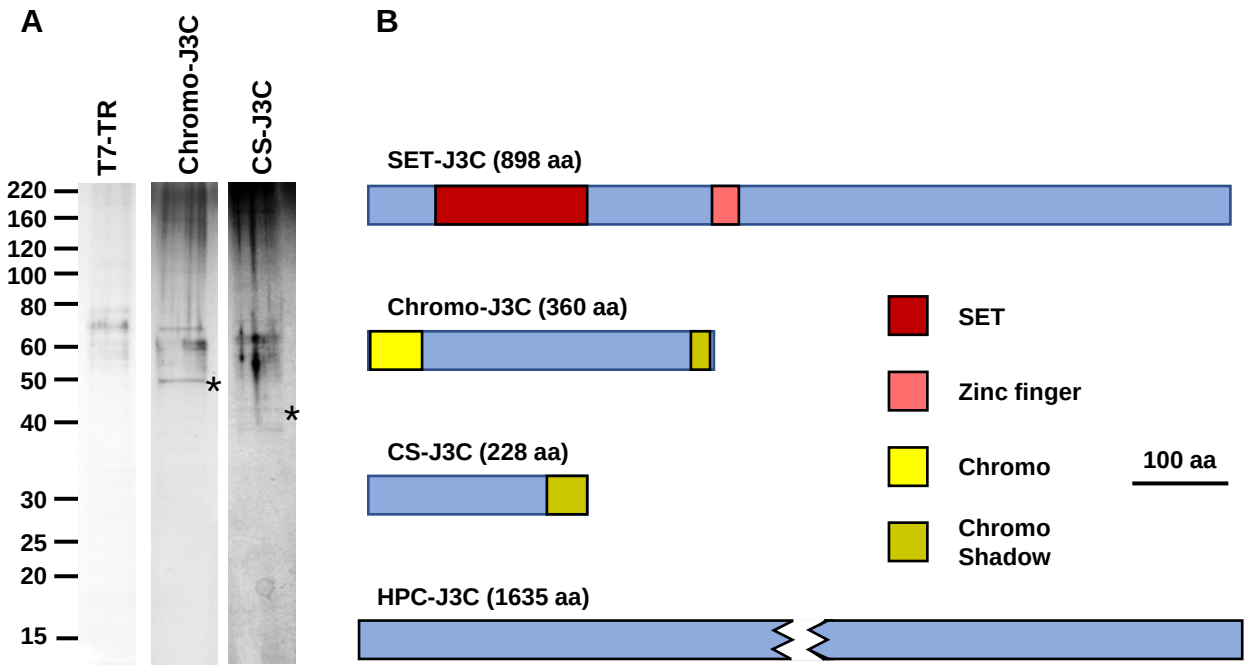

Figure 3. The JBP3-associated Chromatin complex

A. Proteins that co-purify with TAP-tagged Chromo-J3C and CS-J3C analyzed by SDS-PAGE and silver staining as described in Fig. 1.

B. Schematic representation showing the key domains of the four proteins that co-purified with TAP-tagged JBP3. The SET domain is most often found protein methyltransferases; the Chromo domain is usually found chromatin binding proteins; the Chromo Shadow is distantly related to the Chromo domain and frequently found in chromatin binding proteins with the Chromo domains.

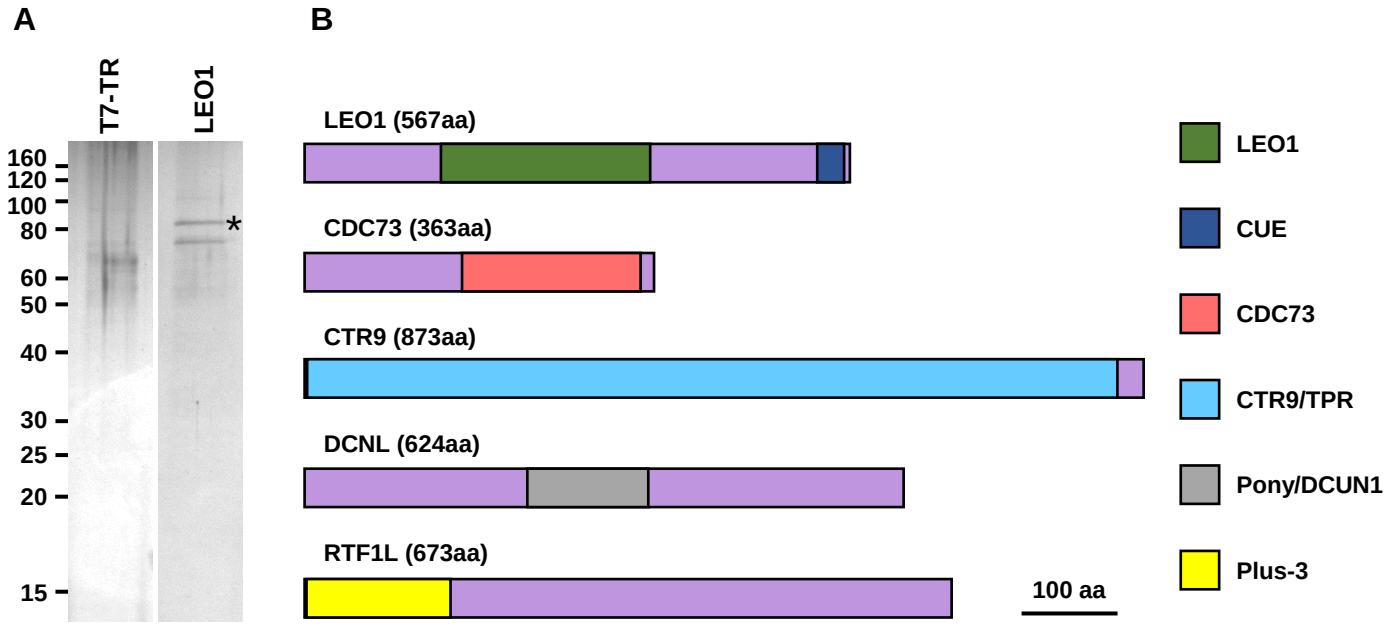

Figure 4. The PAF1C-like complex

A. Proteins that co-purified with TAP-tagged LEO1 were analyzed by SDS-PAGE and silver-staining as described in Fig. 1.

B. Schematic representation showing the key domains of five components of the PAF1C-L complex. LEO1 shows the region of sequence homology with mammalian LEO1; the CUE domain is involved in ubiquitin binding; CDC73 shows the region of sequence homology with mammalian CDC73; the CTR9/TPR domain contains tetratricopeptide repeats involved in protein-protein interaction; the Pony/DCUN1 domain is involved in binding and neddylation of the cullin subunit of E1-type ubiquitin ligases; and the Plus-3 is implicated in binding to single stranded DNA.

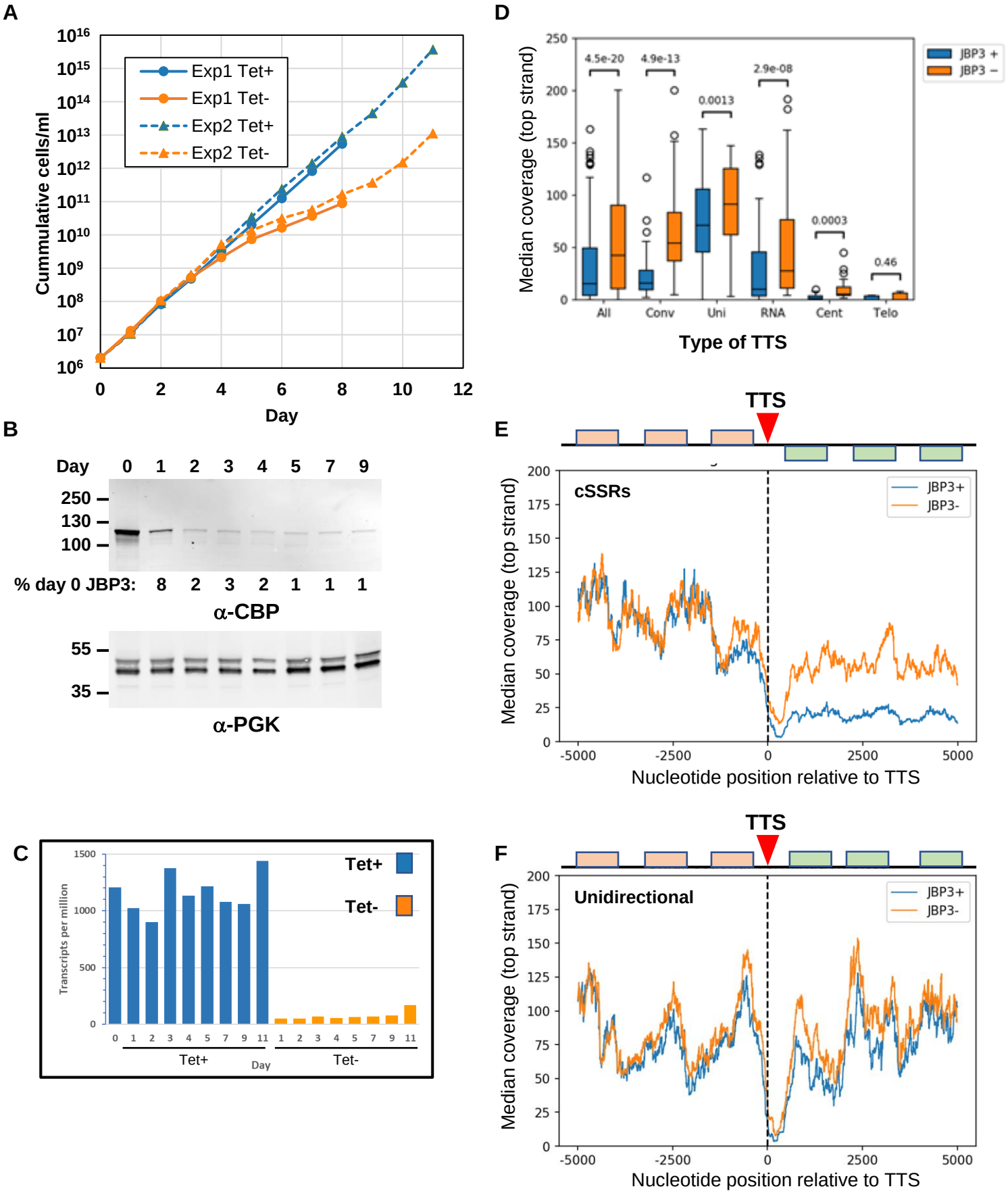

Figure 5. Depletion of JPB3 results in read through at transcription termination sites

Figure 5. Depletion of JPB3 results in read through at transcription termination sites

A. Growth analysis. Two independently generated *L. tarentolae* clones lacking endogenous *JBP3* but containing a Tet-regulated TAP-tagged *JBP3* gene was grown in the presence (blue lines) or absence (orange lines) of Tet. The numbers on y-axis are corrected for dilution during sub-culturing. The solid lines show a clone where the *JBP3* genes were replaced by *pac* (and grown in the presence of puromycin), while the dotted lines show a clone where with the *JBP3* genes replaced by *neo* (and grown in G418).

B. The level of TAP-tagged JBP3 expressed by the T7-TR/JBP3-MHTAP/Δ*jbp3*::*neo* clone grown in the absence of Tet was monitored by Western blot analysis using antibodies against the calmodulin binding peptide (CBP) of the TAP-tag. Antibodies against phosphoglycerate kinase (PGK) served as a loading control. The percent JBP3 levels in comparison to day 0 is shown below the anti-CBD blot.

C. *JBP3-TAP* mRNA levels for the T7-TR/JB3-MHTAP/Δ*jbp3*::*neo* clone grown in the presence (blue) or absence (orange) or Tet for the number of days indicated. mRNA levels are expressed transcripts per million (TPM) as determined by RNA-seq analysis using GENEIOUS.

D. Box-and-whiskers plots showing the median top strand coverage in the 5 kb downstream of all 192 TTSs (All). Separate plots are shown for the 46 TTS at cSSRs (Conv), 30 TTSs between head-to-tail PTUS (Uni), 39 TTS immediately upstream of one or more RNA genes (RNA), 21 TTS adjacent to a centromere (Cent) and 56 TTSs at telomeres (Telo).

E. Median of top strand coverage at each nucleotide position in the 10 kb surrounding the 46 TTS at cSSRs. The schematic represents the protein-coding genes associated with each strand at an “average” cTTS. The second PTU at each cSSR is re-oriented so that the genes are represented on the top strand.

F. Median of top strand coverage at each nucleotide position in the 10 kb surrounding the 30 TTS at between unidirectional (head-to-tail) PTUs. The schematic represents the protein-coding genes associated with each strand at an “average” uTTS.

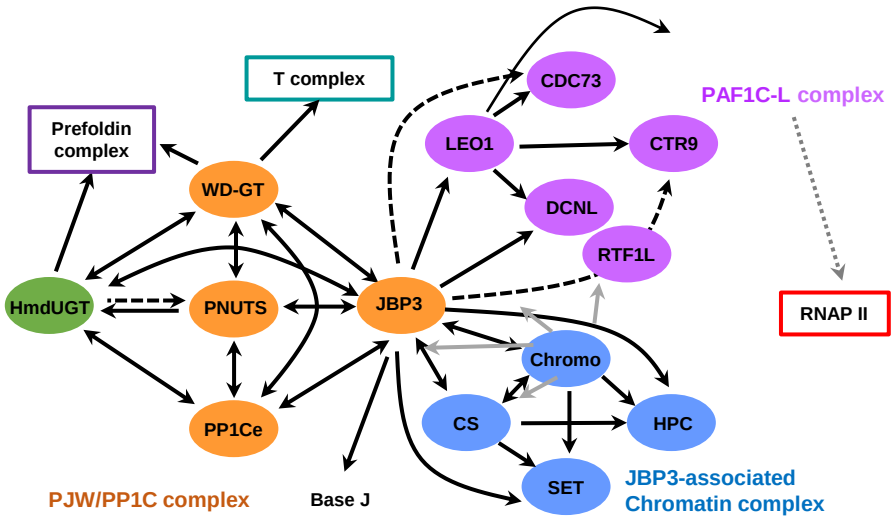

Figure 6. Network of interactions between JBP3-assocated protein complexes.
Solid lines denote proteins enriched in both replicates of the TAP-tag pull-downs and dashed lines indicate proteins enriched in only one sample. Double headed lines represent reciprocal enrichment with both proteins used as bait, while lines with a single arrowhead indicate where reciprocal enrichment not attempted (except in the case of the LEO1 pull-down, where JBP3 was not detected). Grey arrows represent interactions identified by co-purification of proteins with CTR9 in *T. brucei* (*34*). Subunits within the three distinct JBP3-associated protein complexes are denoted by different colors, while the individual components are not shown for within three additional protein complexes represented by boxes. The dotted line connecting the PAFC1-L complex to RNAP II reflects the interactions observed in other organisms.