"Comprehensive Analysis of Nascent Transcriptome Reveals Diverse Transcriptional Profiles Across the Trypanosoma cruzi Genome Underlining the Regulatory Role of Genome Organization, Chromatin Status, and Cis-Acting Elements"

Pedro Leonardo Carvalho de Lima[1,2,3], Leticia de Sousa Lopes[1,2], Juliana Nunes Rosón[1,2,3], Alyssa Borges[1,2], Natalia Karla Bellini[1,2], Ana Tahira[1], Marcelo Santos da Silva[1,2,4], David Pires[1,2], Maria Carolina Elias[1,2], Julia Pinheiro Chagas da Cunha[1,2#].

[1]Laboratory of Cell Cycle, Butantan Institute, São Paulo, Brazil.

[2]Center of Toxins, Immune Response and Cell Signaling (CeTICS), Butantan Institute, São Paulo, Brazil.

[3]Department of Microbiology, Immunology and Parasitology, Escola Paulista de Medicina – UNIFESP, São Paulo, Brazil.

[4]Department of Biochemistry, Chemistry Institute, University of São Paulo, São Paulo, Brazil.

Short title: Unraveling Transcription in *Trypanosoma cruzi*: Genome-wide Insights

#Address correspondence to Julia Pinheiro Chagas da Cunha, julia.cunha@butantan.gov.br.

Abstract

Trypanosomatids are eukaryotic parasites exhibiting polycistronic transcription and trans-splicing. Post-transcriptional mechanisms are acknowledged as pivotal in gene expression regulation of their protein-coding genes. To comprehensively investigate the impact of transcription on gene expression in *Trypanosoma cruzi* and the association with the epigenetic landscape, we conducted a genome-wide nascent transcriptomic analysis. Our findings reveal significant asymmetrical transcriptional abundance across the genome, notably between polycistronic transcription units (PTUs) enriched in conserved genes (core PTUs) and those containing virulence genes (disruptive PTUs). We found that trypanosomes exploit linear genome organization to regulate transcription abundance by embedding virulence genes into highly transcribed core-enriched PTUs, by positioning PTUs near non-coding regions of small non-coding RNAs (e.g., tRNAs, snoRNAs), and by placing core CDSs in PTUs of various sizes. Additionally, we found correlations between open chromatin status and nascent transcript levels, both globally and particularly at transcription starting regions (divergent strand switch regions - dSSRs), indicating a crucial role for chromatin architecture in transcriptional regulation. While both core and disruptive dSSRs exhibit similar levels of some epigenetic marks (H2B.V deposition and 5mC), disruptive dSSRs display significantly higher 5hmC content and nucleosome occupancy compared to core dSSRs. Furthermore, we identified distinct conserved motifs within dSSRs of core and disruptive PTUs. These findings challenge the notion of constitutive and uniform transcription in *T. cruzi*, underscoring the paramount importance of linear genome organization, cis-acting motifs, and chromatin landscape in transcriptional regulation.

Introduction

Gene expression is finely regulated by multiple mechanisms and layers, with transcription initiation standing out as one of the most tightly controlled processes. Chromatin-based mechanisms constitute the initial step, achieved by increasing its accessibility, followed by transcription itself, which involves multiple regulated steps such as the formation of the transcription preinitiation complex (PIC), initiation, promoter-proximal pausing, and pausing release (1). Both prokaryotic and eukaryotic genes typically contain promoter regions consisting of specific DNA sequence motifs that serve as platforms for the association of the transcription factors and transcription machinery, guiding them accurately to the transcription start sites (TSS). These promoters possess specialized sequences that have evolved to regulate gene transcription (2). Subsequent layers of gene expression regulation involve transcription elongation, termination, processing, and, for protein-coding genes, translation. All these layers can be finely regulated (1) and are crucial in determining the final levels of transcripts or proteins.

In trypanosomatids, protein-coding genes are organized in polycistronic transcription units (PTUs). It has long been accepted that genome and protein-coding genes are transcribed constitutively, with no regulation at transcription initiation sites (3-7). Evidence supporting the absence of transcriptional regulation in trypanosomatids existed even before the completion of genome sequencing in these organisms (3, 8-10). As a result, variations in mRNA abundances are mostly attributed to posttranscriptional mechanisms, of which differences in mRNA half-life play a significant role (7). Genome sequencing of trypanosomatids highlighted the presence of a few transcription factors, contrasting with the abundance of proteins containing RNA-binding motifs (11), likely

76 associated with posttranscriptional regulation. Additionally, efforts to identify RNA Pol
77 II promoters were elusive leading to the proposition that the "open" chromatin status
78 could drive transcription initiation in the absence of promoters in any genomic region (9).
79 Mathematical modeling of gene expression in procyclic forms of *Trypanosoma brucei*
80 did not yield evidence for the regulation of RNA polymerase II transcription initiation.
81 However, additional levels of regulation in the nucleus may play a key role in gene
82 expression in other life forms (12). Collectively, these results indicate the lack of
83 transcriptional regulation in these parasites. Consequently, significant effort was
84 dedicated to elucidating the molecular components involved in posttranscriptional
85 mechanisms in these organisms, which were recognized as excellent models for studying
86 posttranscriptional regulation.

87 The genome sequencing of trypanosomatids also revealed the presence of
88 divergent canonical histones, histone variants, and machinery associated with histone
89 modification (11), which are known to be associated with gene expression regulation in
90 many organisms (13). Thus, our group and others characterized an extensive atlas of
91 histone post-translational modifications (PTMs) (14-20) and detected the presence of
92 histone variants and PTMs, with base J marking transcription initiation and termination
93 sites in trypanosomes (6, 21-24). In many organisms, including trypanosomes, genome
94 organization and chromatin structure considerably impact gene expression, affecting
95 transcription stability and localization. In this regard, we previously showed that open
96 chromatin regions and a nucleosome-depleted region at 5' UTR are associated with higher
97 levels of mature mRNA (25, 26), indicating that chromatin reflects gene expression even
98 in the context of polycistronic expression as in *T. cruzi*. Recently, three-dimensional
99 genomic structures were also shown to be related to gene expression in trypanosomes (27,
100 28).

Although there is ample evidence that posttranscriptional mechanisms are critical for regulating mRNA abundances in trypanosomes (7), the current belief that trypanosomes do not regulate transcription initiation of RNA Pol II genes is being questioned. It has been known for decades that both *T. brucei* and *T. cruzi* exhibit a global decrease in transcription activity when replicative forms are compared to non-replicative forms, indicating that transcription regulation is evident during differentiation (8, 29, 30). Additionally, transcription of high-copy genes is lower than expected, suggesting that these genes may be subjected to transcription regulation in *T. cruzi* (29). More recently, GT-rich sequences and sequence-specific promoters were found to be necessary and sufficient to induce transcription initiation in dSSRs of *T. brucei* (31).

The *T. cruzi* genome was proposed to be compartmentalized into two main linear genomic regions: the "core compartment", composed of conserved and hypothetical genes that are syntenic in other trypanosomatid species, and the "disruptive compartment", composed of non-syntenic and repetitive multigene family (MF) genes, including the virulence factors trans-sialidase, mucins, and MASPs (32, 33). Recently, it was observed that these regions may fold differentially in the genome, with disruptive regions of *T. cruzi* showing fewer unprocessed transcripts than conserved regions (28). Extensive efforts to uncover the regulation of virulence factors (mainly the VSG family) expression in *T. brucei* reveal multifactorial players based on a combination of transcription regulation, DNA recombination, epigenetic mechanisms, and 3D organization (27, 34-36). In contrast to *T. brucei*, which has their virulence factors transcribed by RNA Pol I (37), the surface-coding virulence genes in *T. cruzi* are transcribed by RNA Pol II (8, 29). Thus, *T. cruzi* may use different strategies to ensure expression of their virulence factors. In fact, evidence of posttranscriptional regulation for individual sets of virulence factors

(trans-sialidase, mucin, amastin) indicates that both infective and non-infective forms share similar transcription rates *in T. cruzi* (8, 38).

Aiming to comprehensively investigate whether all genomic regions in *T. cruzi* are consistently and evenly transcribed, we conducted a comparative analysis of PTUs containing conserved and virulence genes. This involved a thorough evaluation of nascent transcription expression alongside extensive data integration with epigenetic datasets and an assessment of conserved sequence motifs to elucidate our findings. Our results underscore significant differences in transcriptional expression primarily between core and disruptive regions likely driven by both cis (sequence and linear genome organization) and trans (chromatin/epigenetic modifications) factors, in addition to established post-transcriptional mechanisms.

Materials and Methods

Cell Culture: Epimastigote forms of *T. cruzi* strain Dm28c were cultured at 28°C in liver infusion tryptose (LIT) medium supplemented with 10% fetal bovine serum (FBS-Vitrocell), 0.4% glucose, 0.1 µM hemin, and 59 mg/L penicillin-G. Epimastigotes were maintained in the exponential growth phase by diluting them in fresh LIT medium to 2-4×10^6 parasites/mL every 2-3 days.

GRO-seq Assay: Parasite permeabilization followed established protocols (10, 29, 39). Briefly, 2 x 10^8 epimastigotes were collected from cultures in the exponential growth phase (3200 rpm, 10 min, room temperature) in biological triplicates. The parasites were washed twice in transcription buffer (20 mM potassium glutamate, 3 mM $MgCl_2$, 150 mM sucrose, 11.7 µg/mL leupeptin, 3 mM dithiothreitol). The pellet containing the parasites was resuspended in transcription buffer and chilled on ice for 5 min. Cell

permeabilization was performed with 500 µg/mL Palmitoyl-L-lysophosphatidylcholine (Sigma-Aldrich) for 1 min at room temperature (RT), followed by washing with transcription buffer. Epimastigotes were then incubated (28 ºC, 30 min) in run-on buffer (20 mM potassium glutamate, 3 mM MgCl2, 150 mM sucrose, 11.7 µg/mL leupeptin, 3 mM dithiothreitol, 0.6 mg/mL creatine kinase, 25 mM creatine phosphate, 400 U/mL RNAse Out, 1 mM ATP, 1 mM GTP, 1 mM CTP, 1mM UTP, 1 mM BrUTP). All experiments were conducted in biological triplicates (GRO_rep1, GRO_rep2, and GRO_rep3), and epimastigotes incubated in a run-on buffer without BrUTP were used as a background control sample (GRObckg). To confirm BrUTP incorporation into nascent RNA, a direct immunofluorescence assay was performed using Alexa Fluor anti-BrdU 488 antibody (Santa Cruz Biotechnology) and analyzed under an Olympus BX51 fluorescent microscope (100x oil-immersion objective) attached to an EXFO Xcite series 120Q lamp and a digital Olympus XM10 camera with Olympus Cell F camera controller software. Total RNA was extracted using Trizol® (Invitrogen) and treated with DNase I (Thermo Scientific) according to the manufacturer's instructions. Immunoprecipitation of Br-UTP labeled RNA was performed using Dynabeads Protein G (Invitrogen) linked to Anti-BrdU antibody (Abcam). The magnetic beads preparation and RNA immunoprecipitation were performed according to the manufacturer's instructions. Nascent RNA purification was conducted via ethanol precipitation (adapted from (40)). For this, 1 mL of cold 100% ethanol with 1 µg/µL glycogen and 300 mM NaCl were added to the samples followed by overnight incubation at -20 ºC. Tubes were centrifuged (12000 x g, 30 min, 4 ºC) and washed three times with 75% ethanol at RT (with 10 min incubation interval following each washing step) to dissolve the salts. Finally, the purified RNA pellet was air-dried, resuspended in 5 µL of nuclease-free water, and incubated for 15 minutes at RT. RNA quantification was performed with NanoDrop 2000

Spectrophotometer, and RNA integrity was evaluated through electrophoresis in a 1.5% agarose gel. RNA fragments (biological triplicates and a background sample) were deep sequenced using smRNA-Seq Kit for Illumina® (1 x 75 bp) at Polyomics Facility (Glasgow University), with no poly A selection or rRNA depletion. The GRO-seq dataset are deposited at PRJNA1073942.

Bioinformatic Analyses: All experiments were run on a server configured with Ubuntu Linux Server 20.04, 224 cores, 1 TB of RAM, and 2 x Nvidia P40 GPU. The methods employed to analyze the GRO-seq data and public RNA-seq datasets are illustrated in S2 Fig and described below.

Sequence Read Quality Check, Mapping, and Normalization: Raw sequencing files (*.fastq) were checked with FastQC version 0.11.9 (41). A quality filter was applied using Trimmomatic (42), which involved trimming adaptors, limiting read size to 75 bp, and removing low-quality bases (HEADCROP:3, LEADING:10, TRAILING:0, MAXINFO:65:0.05, MINLEN:35). Reads were then mapped against the *T. cruzi* Dm28c genome using Bowtie 2 v. 2.4.5 (43) (--verysensitive (-D 25 -R 4 -N 0 -L 19 -i S,1,0.40). When indicated, alignment files were filtered using a MAPQ (Mapping Quality) and sorted and indexed using samtools version 1.12 (2021) (Li, Handsaker et al. 2009). Data coverage was assessed using deeptools version 3.3.2 (Ramírez, Dündar et al. 2014), employing the BPM (bins per million) method within the bamCoverage function. Alternatively, Kallisto version 0.46.1 (44) was used to quantify each genomic feature in transcript per million (TPM) values using default parameters. An indexed transcriptome considering the entire genome separately by PTUs or CDSs was used (see below, S2B Fig). TPM values from experimental samples were subtracted by the corresponding feature found in the GRO background sample. RNA-seq datasets by Smircich et al.,

(2015) (45) (PRJNA260933) and Berna et al. (2019) (46) (PRJNA315252) were processed as described above.

MNAse-seq and FAIRE-seq Analysis: Raw files were obtained from SRA under project numbers PRJNA763084 for FAIRE-seq (25) and PRJNA665060 for MNase-seq (26). Filtering and mapping were performed as described in the corresponding original papers.

Extra Genomic Feature Annotation: Polycistronic regions, dSSRs, and cSSRs were annotated from the *T. cruzi* Dm28c GFF file downloaded from https://tritrypdb.org/ - DB53, Dm28c 2018) as previously described (25). To evaluate transcription along the entire genome, all coordinates with no annotation in the GFF file were further annotated as "others" using a custom Python v. 3.6.7 script available at https://github.com/PedroLeoardo/annotateOthers. If "others" were between CDSs (or between CDS and any ncDNA) or at the contigs edges, they were classified respectively as "other-IR (for intergenic regions)" or "other-end." Core (conserved proteins and conserved hypothetical proteins), disruptive (mucin, MASP, and trans-sialidase), and both (GP63, RHS, DGF-1) CDSs were further annotated and visualized in a bed file. PTUs harboring 80% or more of CDSs from each genomic compartment were classified as "PTU core," "PTU disruptive," or "PTU both." "PTU mix" contained all the remaining PTUs.

PTU Abundance and DGE Analysis: Two strategies were used to evaluate if different PTUs would have similar transcription rates. TPM counts from PTUs considering the genomic coordinates of the entire PTU and the median of the TPM values from different CDSs of a given PTU were used. In all cases, TPM values were obtained from Kallisto version 0.46.1 and subtracted from the TPM counts from the background sample (GRO_bckg). DGE between infective (trypomastigote cells derived tissue cultures,

TCTs) and non-infective (epimastigotes) forms was obtained from TriTrypDB (release 66) website considering DESeq2 analysis, fold changes of 2, p-adj<0.01 from a transcriptomic analysis (28).

Data Visualization: Data coverage was normalized using the deepTools toolkit version 3.3.2 (47) by Bins Per Million (BPM) for GRO-seq and RNA-seq datasets, and Reads Per Genome Content (RPGC) for FAIRE-seq and MNase-seq datasets using the bamCoverage function to generate normalized coverage files (bigWig). Following normalization, landscape profiles and heatmaps from specific genomic features were systematically generated using the computeMatrix tool followed by plotHeatmap or plotProfile function from deeptools version 3.3.2. The Integrative Genomics Viewer (IGV) version 2.11.9 (2021) (48) was used for additional inspection and visualization.

dSSRs Analysis:

Motifs analysis

MEME platform version 5.5.5 (49) was used to identify conserved motifs in dSSR sequences. Only dSSRs between PTUs from the same genomic compartment (either core or disruptive) were considered. The motifs identified in both core and disruptive dSSRs were compared with known motifs in the TFBSshape database (50) using the TomTom platform version 5.5.5 (51). Transcription factors identified belonged to different phylogenetic groups.

Modified Bases Analysis: Modified base analysis was performed using public Oxford Nanopore sequencing datasets available at (28). Files were converted from FAST5 format to POD5 format using the pod5 library version 0.3.6 (pod5 convert fast5) (https://github.com/nanoporetech/pod5-file-format), which were used to call methylation

(m) and hydroxymethylation (h) cytosines with the Dorado program, version 0.5.3 (dorado basecaller hac,5mCG_5hmCG) (https://github.com/nanoporetech/dorado). The readings were mapped using the minimap2 program (52) through the interface offered by the Dorado program itself against the genome of the Dm28c 2018 strain (version 60). The analysis of methylated bases was carried out using the modkit program, version 0.2.5 (https://github.com/nanoporetech/modkit). The modkit pileup subcommand generated files of the type bedMethyl (modkit pileup --interval-size 1000000) and BEDGRAPH (modkit pileup --bedgraph) on the positive or negative strands.

Epigenetic Enrichment Analysis: Using core and disruptive dSSRs coordinates, nucleosome enrichment was extracted by BAMscale version v1.0 (53) using BAM files from MNase-seq dataset (26) (FPKM values); H2B.V enrichment by H2B.V Chip-seq (ChIP/input) (24) using BigWigSummary from deeptools (47)) (average Raw Counts); 5mC and 5hmC enrichment obtained as described above were extracted from bedgraph files (column Nmod / Nvalid_cov).

Statistical and downsampling analysis

Statistical were conducted using GraphPrism 8.4.3 or R 4.3.0. Downsampling involved randomly selecting 100 CDSs without replacement, repeated 50 times. Spearman correlation analyses were performed using the subset() function in R, followed by Spearman correlation tests using the cor.test() function for each genomic compartment, evaluating combinations such as "FAIRE vs RNA" and "FAIRE vs GRO." The cor.test() function was iteratively executed 50 times to ensure robustness, generating stored p-values. To enhance accuracy, false discovery rate (FDR) correction was applied using the p.adjust() function. Spearman's correlation was selected for its rank-based nature, making it appropriate for non-parametric data and providing robustness against outlier measures. Additionally, a customized R script was utilized to assess the significance of the data

using the chi-square test of independence and Fisher's test for calculating p-values. The distribution of features in the genome served as a control. Fisher's exact test was employed for frequencies less than 5.

Results

Optimization of GRO-seq Assays in *T. cruzi* and Assessment of Data Quality

To evaluate transcriptional regulation and address whether the *T. cruzi* genome undergoes widespread and constant transcription, we conducted genome-wide sequencing of nascent transcripts in biological triplicates (GRO 1-3). A control group consisting of unlabeled parasites was utilized as a background sample (GRObckg). Figs S1A-B demonstrate efficient permeabilization of parasites, labeling with Br-UTP, and RNA extraction without degradation. Over 30 Mb reads were sequenced in all samples (S1 Fig, Table S1). To validate that our GRO-seq assay results were enriched in nascent transcripts, we established a bioinformatic analysis pipeline (S2 Fig), considering both coverage (in BPM) and TPM counts. To achieve this, counts from all genomic features, either classified or not by polycistronic regions (PTUs) (S2B Fig, Tables S2-3), were utilized. GRO-seq data was compared with public RNA-seq datasets (45, 54), serving as controls for nascent transcription expression as these datasets capture the abundance of mature transcripts resulting from a balance between their synthesis, processing (trans-splicing and polyadenylation), and degradation. S1C Fig illustrates that the distribution profile of read coverage in coding regions (CDS) from nascent (GRO-seq) and mature (RNA-seq) transcripts differs significantly. The presence of transcript processing (indicated by arrows) is evident when evaluating mature transcripts due to the depletion

of reads upstream of the beginning of the CDS (ATG - S) and downstream of the end of the CDS (Stop codon – E). This pattern indicates mRNA processing and the formation of mature transcripts with the removal of upstream and downstream regions at the 5' and 3' UTR, respectively. In line with the absence of processing in nascent transcripts, this depletion is not observed in these regions in the GRO-seq data. The lack of transcript processing was detected regardless of the genomic compartments (as will be discussed below) (S1D Fig). The presence and absence of RNA processing were also evident by evaluating the TPM counts in CDSs and intergenic regions. We detected a $\log_2$ fold change of -0.16 between CDSs and their intergenic regions in nascent transcripts. In contrast, intergenic regions from mature transcripts exhibited a 1.7 $\log_2$ fold decrease compared to CDS regions (S1E Fig). The absence of processing was also evident at the spliced leader (SL) locus (S1F Fig), where the entire SL transcript of approximately 140 bp was detected in GRO-seq samples, while a processed 39 bp-SL transcript was detected in the mature transcriptome. Taken together, these results indicate that GRO-seq datasets are enriched in unprocessed RNA transcripts and, therefore, enriched in nascent transcripts.

Distribution and Abundance of Nascent Transcripts Varies Across the *T. cruzi* Genome.

We initially evaluated the transcriptional activity throughout the trypanosome genome, investigating whether specific genomic regions exhibited no preferential transcription, indicated by TPM values (GRO-GRO bckg) <= 0 (S3 Table). To achieve this, we considered all *T. cruzi* genomic regions (37,959 regions – S2C Fig) using a modified version of the annotation GFF file (see Star Methods section). We observed that

9% (4.7 Mb) of the genome (in bp) exhibited undetectable levels of nascent transcripts (Fig 1A). Among these regions, transcripts from non-coding DNA (ncDNAs), mainly snRNAs and snoRNAs, as well as from the contig ends, were overrepresented (chi-square test, p-value 10^{-12}) (Fig 1B).

Fig 1. The Distribution and Abundance of Nascent Transcripts Across the *T. cruzi* Genome. A. Percentage of *T. cruzi* Genomic Features Exhibiting No Nascent Transcription Expression (GROexp – GRO bckg ≤ 0). **B**. Bar plots representing the distribution of genomic features in GROexp – GRO bckg ≤ 0 (top) compared to the expected genome distribution (bottom). **C**. Distribution of Transcripts Per Million (TPM) Counts (log2(TPM +1)) across all Genomic Features (37,959 regions) from GRO and RNA-seq Assays. Coefficients of Variations are depicted above each violin plot. Outliers were removed (Q=1%).

We also detected an asymmetrical distribution among TPM counts of nascent transcripts (ranging from 0 to 7.8 in log2 (TPM+1)) (Fig 1C), indicating variability in nascent transcript abundance among different genomic features. The coefficient of variation (CV) was higher for mature transcripts (average 97%) than for nascent transcripts (average 77%), as expected due to intense post-transcriptional mechanisms in trypanosomes (7). However, the higher variation among transcript abundances in the nascent transcriptome indicates non-homogeneous transcription across the trypanosome genome, contrasting with previous thoughts.

Assessment of nascent expression in PTUs revealed that core PTUs are more abundant than disruptive PTUs.

The protein-coding genes of trypanosomes are organized into PTUs flanked by transcription start sites (TSS) and transcription termination sites (TTS) located at divergent and convergent strand-switch regions (dSSRs and cSSRs, respectively) (55). In *T. cruzi* (Dm28c strain), protein-coding genes are distributed among 1,286 PTUs. Polycistronic transcription precludes transcriptional control at the individual coding sequence (CDS) within the same PTU but does not preclude transcriptional differences among PTUs. To evaluate whether different PTUs share the same transcription rate, we compared their transcription abundance. As expected, TPM values from PTUs differ mostly within mature transcripts rather than nascent transcripts (Fig 2A). Nascent and mature transcript abundances from PTUs have coefficient of variation (CV) values of 66% and 72% (on average), respectively, implying that nascent transcript abundance is also not homogeneous among PTUs.

Fig 2. Assessment of Nascent Expression in PTUs Reveals Greater Abundance in Core PTUs Compared to Disruptive PTUs. A. Distribution of TPM Counts (log2(GRO-GRObckg +1)) across all PTUs (1218) from GRO and RNA-seq assays. **B.** Boxplots of TPM Counts from PTUs on the first three major contigs (PRFA:01000001-01000003). Individual TPM Values (averaged among biological replicates) for each CDS are plotted for each PTUs. **C.** Violin plots of TPM Distribution from PTUs Containing 80% of Core, Disruptive, or Both Genes, respectively. **D.** IGV snapshot showing read coverage (BPM) from three replicates of GRO along two *T. cruzi* contigs (PRF01000009 and PRF010000011). Polycistrons are depicted in blue rectangles, while green and red rectangles indicate CDSs from Core and Disruptive compartments, respectively.

For a more detailed assessment, we evaluated the transcript abundance profile of the first hundred PTUs (Fig 2B and S4A Fig) distributed in the three major contigs (PRFA01000001, PRFA01000002, and PRFA01000003), highlighting that nascent expression profiles differ among PTUs. Interestingly, PRFA01000002 contains PTUs that are more abundant than those from the other contigs. This contig is enriched in CDSs of syntenic and conserved genes, which are part of the core compartment, whereas the lower abundant PTUs are mainly composed of non-syntenic/virulence genes that form the disruptive compartment.

To further evaluate whether these two compartments have different nascent expression profiles, PTUs composed of at least 80% of CDSs either from the core or disruptive or both compartments (composed of DGF-1, GP63, and RHS) were respectively classified as "core," "disruptive," or "both," and their nascent expression profile was measured. Fig 2C-D and S3A Fig show that PTUs from the disruptive compartment indeed have lower nascent expression than those from the core and both compartments (One-way Ordinary ANOVA, p.adj<0.0001). Additionally, the nascent expression profile of individual CDSs from the core and disruptive compartments confirms that the former is more abundant than the latter (S3B Fig). To account for the repetitive nature of the disruptive compartment (33) and potential mapping issues, we used highly-quality mapped reads (MAPQ1 $\geq$ 30) (S3C Fig) and an RNA-seq dataset enriched in reads from disruptive regions (S3D Fig) to confirm the differential expression of these compartments. Taken together, these results indicate that virulence factor genes (disruptive compartment) in epimastigote forms have lower nascent transcription rates when compared to conserved genes (core compartment).

Nascent Transcription of Core PTUs Depends on the Number of CDSs and Proximity to ncDNA Loci

To delve deeper into the differential transcriptional expression between core and disruptive compartments, we classified core and disruptive PTUs based on the number of CDSs per PTU. We observed that core PTUs containing more than 4 CDSs (369 PTUs) exhibited higher nascent expression compared to core PTUs with one (162 PTUs) or 2-4 CDSs (128 PTUs) (Fig 3A, S4A, and S4B, One-way ANOVA, p-adjusted <0.0001). Intriguingly, no correlation was observed between the number of CDSs per PTU within the disruptive compartment (Fig 3B, One-way ANOVA). Despite the median size (in bp) of core PTUs being larger than disruptive PTUs (S4C Fig), the length of PTUs increased with the number of CDSs, both for core and disruptive PTUs (S4C Fig). Hence, the lack of association between the number of CDSs in disruptive PTUs and nascent expression is unclear but aligns with the distinct nascent expression patterns observed in these two compartments.

Fig 3. Nascent Transcription of Core PTUs Depends on the Number of CDSs and the Proximity of ncDNA Loci. A. Violin plots illustrating the TPM distribution from PTUs containing 80% of core (left) or disruptive (right) genes, categorized by the number of CDSs per PTU. Core PTUs are divided into those with one (162 PTUs), 2-4 (128 PTUs), or 4 or more CDSs (369 PTUs). Disruptive PTUs are categorized similarly with one (40 PTUs), 2-4 (68 PTUs), or 4 or more CDSs (65 PTUs). Statistical analysis was performed using one-way ANOVA with Tukey's range test at 95% confidence level (**** p<0.0001**). B.** Percentage of pairwise PTU comparisons that are statistically significant, classified by genomic compartments and CDS number per PTU using one-

way ANOVA with Tukey's range test at 95% confidence level. **C.** TPM counts from four core PTUs with 4 or more CDSs located on the same contig. **D.** Abundance values from PTUs not adjacent to any ncDNA (none) or adjacent to a tDNA or snoDNA locus. Statistical analysis was conducted using the Kruskal-Wallis test (* $p<0.01$). E. TPM counts of seven canonical PTUs (#59 - contig 2 - 100 CDSs, #81 - contig 2 - 56 CDSs, #110 - contig 4 - 129 CDSs, #111 - contig 4 - 74 CDSs, #112 - contig 4 - 59 CDSs, #173 - contig 9 - 19 CDSs, #181 - contig 10 - 57 CDSs). TPM values for each PTU were obtained considering the entire PTU coordinates. No statistical significance was observed.

Subsequently, we explored whether nascent expressions varied within PTUs previously classified based on their genomic compartment and number of CDSs. Between 5.4% to 8.1% and 4.2% to 9.6% of the pairwise comparisons were significantly different within disruptive and core PTUs, respectively (Fig 3C) (ANOVA test using Tukey's range test as post-hoc, $p<0.01$). Among the differentially abundant core PTUs with four or more CDSs located on the same contig, we observed that differential expression was associated with PTUs positioned adjacent to a tRNA locus or at contig ends (Fig 3D). Among the top 20 differentially expressed PTUs, 35% contained ncDNA (mainly tDNAs and snoDNAs) adjacent either upstream or downstream. To assess whether the presence of ncDNA could globally interfere with transcriptional expression, we measured nascent expression from PTUs adjacent to tDNA and snoDNA compared to those without any associated ncDNA (Fig 3E). PTUs adjacent to tDNA and snoDNA exhibited, on average, higher transcriptional rates than those without any nearby ncDNA. This suggests that the presence of ncDNA may regulate *T. cruzi* PTU transcriptional expression.

Finally, we investigated whether nascent transcription levels indeed differed among "canonical core PTUs." We manually selected seven PTUs that met specific criteria: i. contained more than 4 CDSs; ii. were located between intergenic regions delimited by dSSRs and cSSRs; iii. were not at the chromosome/contig end; iv. were not interrupted by tRNA, snoRNA, snRNA, or rRNA loci; and v. contained only genes from the core compartment (Fig 3F). No statistical differences were found among these PTUs, indicating that PTUs with similar features (related to genomic compartment, number of CDSs, and genomic location) exhibited the same transcription rate.

Virulence factor genes located within core PTUs exhibit higher transcriptional rates than those located within disruptive PTUs.

Since disruptive PTUs exhibit lower nascent expression than core PTUs, we investigated whether a disruptive CDS within a core PTU (classified as >80% of core CDSs) would show higher nascent transcription levels compared to those disruptive CDSs within disruptive PTUs (classified as >80% of disruptive CDSs), and vice versa for core CDSs within disruptive PTUs (Fig 4A). We found that core CDSs within disruptive PTUs had lower expression compared to core CDSs from core PTUs. Similarly, disruptive CDSs within core PTUs exhibited higher expression compared to those disruptive CDSs from disruptive PTUs (Fig 4B). Virulence factors such as trans-sialidases, mucins, and MASPs are part of the disruptive compartment (46). Therefore, we evaluated the nascent expression of these genes separately (Fig 4C). Consistently, we observed higher nascent expression levels for trans-sialidases, mucins, and MASPs when they were located within core PTUs compared to disruptive or mixed PTUs. Interestingly,

472 genes for trans-sialidase type I, II, and VIII are overrepresented in core PTUs, while
473 MASP and mucin are underrepresented (Fig 4D).
474
475 **Fig 4. Nascent Transcription Expression Depends on the PTU Context. A.** Schematic
476 representation of four PTU categories based on the content (at least 80%) of core (PTU
477 core), disruptive (PTU disruptive), both (PTU both) genes, and mixed PTUs composed
478 of core, disruptive, or both CDSs with less than 80% representation. The number of PTUs
479 in each category is shown. Arrows indicate the transcription direction of PTUs. **B.** Bar
480 plots of TPM counts from individual CDSs classified as core, disruptive, or both, from
481 PTUs with more than 80% composition of core, disruptive, and both CDSs, respectively.
482 **C.** Similar analysis shown in B for trans-sialidases, mucin, and MASP CDSs. Statistical
483 analysis was conducted using one-way ANOVA with Tukey's range test as post-hoc
484 ($p<0.0001$). **D.** Distribution of trans-sialidase, mucin, MASP, and pseudogenes in core
485 and disruptive PTUs (chi-square test with a p-adjusted < 0.0001). ** for adjusted p-value
486 < 0.001, *** for adjusted p-value < 0.0001 (chi-square test). **E.** Percentage of
487 differentially expressed disruptive genes (Epimastigotes versus TCTs, p-adjusted <0.01,
488 fold change of 2, based on (28)) located in core or disruptive PTUs.
489
490 It is known that *T. cruzi* infective forms contain higher levels of mature transcripts
491 from virulence factors (54); however, these levels are regulated post-transcriptionally
492 since infective forms do not increase transcription levels of virulence factors compared
493 to non-infective forms (8, 10, 38). Here, we demonstrate that virulence-factor genes are
494 more transcribed when located within core PTUs. Thus, we hypothesize that genes
495 preferentially enriched in infective forms (cellular trypomastigotes) originate from those
496 located preferentially in core PTUs. Accordingly, we found that 82% of virulence-factor

497 genes located in core PTUs are upregulated in TCT forms compared to 63% of those
498 located in disruptive PTUs (p-adj <0.01, fold change of 2, Fig 4E). These findings suggest
499 that genome organization, particularly linear gene order, plays a critical role in regulating
500 transcription to ensure the expression of essential virulence-factor genes in *T. cruzi*.
501
502 **Open chromatin status, both globally and locally at dSSRs, correlates with**
503 **nascent transcript levels.**
504
505 Regions with more open chromatin are generally associated with higher
506 transcriptional activity in several eukaryotes (56). Previously, we demonstrated a positive
507 correlation between open chromatin (based on FAIRE-seq data) and nucleosome
508 occupancy (measured by MNAse-seq data - depth of the nucleosome-depleted region at
509 5'UTR) with mature transcript levels (25, 26). Here, we show a stronger correlation
510 between nascent transcriptome and open chromatin status (ρ = 0.65) than mature
511 transcriptome and open chromatin (ρ = 0.43) (Fig 5A). Furthermore, we classified
512 genomic regions into three groups (high, medium, and low) based on levels of open
513 chromatin (measured by RPGC) and compared them with nascent expression levels. As
514 expected, higher levels of open chromatin corresponded to higher nascent expression (Fig
515 5B).
516
517 **Fig 5. Global Open Chromatin Status at Core Region Correlates with Nascent**
518 **Transcript Levels. A.** Scatter plots of open chromatin (FAIRE-seq dataset, RPGC levels)
519 profiles versus nascent (GRO-seq) or mature (RNA-seq) transcriptomes for all *T. cruzi*
520 genomic features. R values are indicated. **B.** Bar plots illustrating TPM counts of genomic
521 regions based on their open chromatin levels (high - 25% highest RPGC levels, low -

522 25% lowest RPGC levels, medium – the 50% remaining). Outliers were removed using
523 the ROUT method, Q=1% at PRISMA. **C.** Scatter plots of open chromatin (FAIRE-seq
524 dataset, RPGC levels) profiles versus nascent (GRO-seq) transcripts for coding DNA
525 regions classified as core, disruptive, and both regions. R values are displayed. **D.** R
526 values of open chromatin profile and nascent transcription obtained through
527 downsampling analysis of a randomly selected 100 genes (for the indicated compartment)
528 repeated 50 times. Statistical analysis was performed using one-way ANOVA with
529 Tukey's range test as post-hoc (****p<0.0001).
530
531 Remarkably, we detected a moderate positive correlation (ρ = 0.52) between
532 "open" chromatin and nascent transcription exclusively for genes from the core
533 compartment, confirmed by down-sampling analysis (100 genes representing each
534 compartment randomly selected 50 times) (Fig 5C). In contrast, for global analysis, no
535 correlation was observed between nucleosome occupancy and nascent transcription,
536 regardless of the genomic compartment (S5A-B Fig). These findings suggest that
537 expression in core and disruptive compartments is subject to different transcriptional
538 regulatory effects based on chromatin landscape.
539 Although the observation of a global association between chromatin landscape
540 and nascent expression is intriguing for these organisms, establishing whether chromatin
541 genuinely could influence transcription necessitates the examination of local variations,
542 especially at transcription start regions. In trypanosomes, transcription initiates mainly at
543 dSSRs (6, 57), which are known to harbor enrichment of open chromatin, H2A.Z/H2B.V,
544 and active histone marks (6, 21, 24, 58, 59). Here, we demonstrate that in epimastigote
545 forms, dSSRs can be clustered into three classes based on their levels of chromatin
546 opening (Fig 6A). For simplicity, we termed dSSRs associated with cluster 1 as "open"

and those with cluster 3 as "closed." We observed that PTUs associated with open dSSRs have higher abundance of nascent transcripts than those associated with medium and closed dSSRs (Fig 6A).

Fig 6. Core and Disruptive dSSRs Differ in Their Epigenetic Landscape and Sequence Motifs. A. Left: Hierarchical cluster analysis of the distribution of the RPGC log2 ratio in PTUs from epimastigotes considering 1 Kb upstream or downstream, highlighting the dSSRs. Right: Bar plots depicting TPM levels from PTUs associated with dSSRs that harbor different levels of open chromatin as shown on the left. PTU abundance was based on median TPM values from CDSs of each PTU. Wilcoxon test, p.adj <0.05**.** **B.** Distribution of core, disruptive, both, and mixed PTUs associated with open and closed dSSRs as depicted in A. (chi-square test under a p-adjusted < 0.01). * for adjusted p-value < 0.0125, ** for adjusted p-value < 0.00125, *** for adjusted p-value < 0.000125 (chi-square test). **C.** Sequence motifs determined by the MEME platform using dSSRs from either core or disruptive PTUs. Only dSSRs between PTUs from the same compartment were analysed. The top 3 sequence motifs are shown. E-values were higher than $1.5e^{-076}$ **D.** Distribution of abundance levels for 5hmC and 5mC (Nmod/Nvalid_cov, bedgraph score of coverage), H2.BV (ChIP/input) deposition (average raw counts), and nucleosome levels (MNase-seq) (FPKM) for core and disruptive dSSRs. Mann-Whitney test, ** p.adj <0.01. Only dSSRs between PTUs from the same compartment were used in C and D analysis.

To gain further insights into these dSSRs clusters, we evaluated their size in base pairs, PTU compartment classification, and number of CDSs per PTU. Open dSSRs (cluster 1) are shorter (median size ~1000 bp) than those with closed chromatin states

572 (clusters 2 and 3) (S6 Fig). Comparing open (cluster 1) and closed (cluster 3) chromatin
573 dSSRs with the expected genome distribution of core and disruptive PTUs, we found that
574 the former has no disruptive PTUs while the latter has 17% of disruptive PTUs.
575 Additionally, core PTUs are enriched (86%, chi-square $p < 0.05$) in open dSSRs and
576 impoverished (38% of their PTUs, chi-square $p < 0.05$) in closed dSSRs (Fig 6B). Hence,
577 it is evident that the disruptive compartment exhibits not only reduced open chromatin
578 levels but also a higher prevalence of closed dSSRs. These findings strongly indicate that
579 chromatin accessibility at transcription start regions correlates with the abundance of
580 nascent transcripts, potentially influencing transcription rates. This offers a compelling
581 molecular rationale for the regulation of transcription initiation.
582
583 **Transcription start regions from core and disruptive PTUs share similar**
584 **levels of some epigenetic marks but differ in their nucleosome content, 5hmC levels**
585 **and sequence motifs.**
586
587 Here, we demonstrated that PTUs composed of genes from core and disruptive
588 genomic compartments exhibit distinct levels of transcriptional expression, which
589 correlate with varying degrees of open chromatin at their dSSRs. Motivated by these
590 findings, we explored whether core and disruptive dSSRs also exhibit differential levels
591 of other epigenetic marks, including H2B.V deposition (a hallmark of dSSRs in
592 trypanosomes), 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and
593 nucleosome content. We found that while both core and disruptive dSSRs harbor similar
594 levels of H2B.V and 5mC; disruptive dSSRs exhibit a higher 5hmC content and
595 nucleosome density compared to core dSSRs (Table S5-S6).

Finally, we investigated whether dSSRs situated between core or disruptive PTUs would display differential conserved motifs that resemble putative promoter sequences, as recently identified in trypanosomes. Remarkably, we observed substantial differences in the enriched motifs within dSSRs from core versus disruptive PTUs. While the former exhibited enrichment of poly-T and poly-A repeats, the latter displayed a diverse array of motifs with varying nucleotide compositions, albeit with a predominance of G and T nucleotides (Fig 6C). Collectively, these findings underscore the divergent regulatory mechanisms based on cis (sequence) and trans (chromatin) features of core and disruptive dSSRs, suggesting a potential regulatory mechanism governing transcription initiation in *T. cruzi*.

Discussion

Traditional transcriptome analyses are typically based on detecting the total set of mature transcripts, encompassing both unprocessed and processed RNA forms. To gain a deeper understanding of transcriptional activity, strategies such as "run-on" assays are employed. Such assays utilize labeled nucleotides during *in vitro* transcription coupled with large-scale sequencing techniques, such as Global Run-On sequencing (GRO-seq) (1, 60). This approach enables the detection of nascent RNA levels, providing a more precise understanding of global transcriptional regulation. By conducting a genome-wide analysis of nascent transcripts in *T. cruzi*, we identified an asymmetrical distribution of transcription abundance, indicating that genomic regions are not uniformly transcribed. Specifically, we observed different transcriptional levels between core and disruptive regions, with PTUs from disruptive regions being less transcribed than those from core

621 regions. Furthermore, within core PTUs, variations in nascent expression levels were
622 detected corresponding to the number of coding sequences (CDSs). Although the
623 underlying reasons for these observations remain unclear, it has been previously
624 demonstrated that longer PTUs exhibit anomalous expression in trypanosomes (61). It is
625 important to note that among core PTUs with the same number of CDSs, no
626 transcriptional differences were observed, indicating conserved levels of transcription
627 among core PTUs of similar sizes.

628 In PTUs from the same genomic compartment and with a similar number of CDSs,
629 few statistical differences were detected (less than 10%). These differences were mainly
630 associated with the presence of *loci* containing snoRNAs and tRNAs in their vicinity.
631 This finding suggests that the presence of non-coding RNA *loci* may interfere with
632 adjacent transcriptional expression. Although the underlying mechanisms for that are
633 unknown, we envisage that it may be associated with the fact that tRNAs and some
634 snoRNAs are transcribed by a different RNA Polymerase, the RNA polymerase III (62).
635 Basal transcripts factors of RNA polymerase III and II can be shared resulting in
636 competition, which could result in transcription interference (63, 64). Additionally, some
637 tDNA *loci* function as insulators, contributing to the 3D architectural organization of the
638 genome, potentially impacting gene expression (65). However, this is yet to be proven in
639 trypanosomes. Together, these data suggest that the trypanosome genome is organized to
640 promote transcriptional gene regulation by positioning tDNAs and snoRNAs close to
641 some protein-coding genes.

642 In *T. brucei*, virulence factors (VSGs) are transcribed unusually by RNA Pol I
643 (37), and considerable effort has been dedicated to understanding the underlying
644 mechanism of their gene expression (35, 66). In contrast, *T. cruzi* surface-coding virulent
645 genes are transcribed by RNA Pol II, and efforts to uncover their regulation have found

646 solid evidence of posttranscriptional regulation in a particular set of genes (8, 10, 29, 38).
647 In this way, understanding how parasites regulate RNA polymerase II transcription of
648 protein-coding genes may provide important insights into how *T. cruzi* regulates the
649 expression of their repertoire of virulence genes. By evaluating the nascent transcriptome
650 and linear organization of the *T. cruzi* genome, we observed a group of virulence genes
651 embedded in core PTUs, exhibiting higher transcription rates. Notably, a significant
652 proportion of these highly transcribed virulence genes are enriched in infective forms,
653 suggesting a strategic genomic positioning to ensure robust expression. This finding
654 reinforces the notion that the trypanosome genome is strategically organized to regulate
655 gene expression. Linear genome organization has been shown to be critical in gene
656 expression regulation in *T. brucei* (67). In these organisms, genes differentially expressed
657 during the heat shock response are located either proximal or distal to transcription start
658 sites.

659 Different organisms regulate gene expression by modulating the chromatin
660 landscape. Consistent with a role of chromatin architecture in transcription regulation
661 (68), we detected a positive correlation between open chromatin levels and nascent
662 transcription. The coding regions of virulence factors (disruptive regions) are enriched in
663 nucleosomes, depleted in open chromatin (25, 26), and form more inter-chromosomal
664 contacts (28), suggesting that they are enriched in condensed and less accessible
665 chromatin. In contrast, core regions are more open with fewer nucleosomes and exhibit
666 higher transcription levels. Therefore, this chromatin landscape may be involved in
667 modulating the binding of RNA polymerase II machinery and basal transcription factors,
668 justifying the differential transcription rates in these regions.

669 Moreover, while a clear correlation between core nascent transcription and
670 chromatin openness was observed, this correlation is not evident when analyzing

671 disruptive regions. The reasons for that are unclear but suggest that virulence and
672 conserved regions are under different regulatory players. While a clear correlation
673 between core nascent transcription and chromatin openness was observed, such a
674 relationship is not evident for the disruptive regions. Interestingly, levels of open
675 chromatin and nucleosome abundance are not altered in disruptive virulence regions
676 (compared to core regions) in infective forms when compared to non-infective ones (25),
677 suggesting that the chromatin landscape of disruptive regions may remain silent and
678 impervious to transcriptional regulation. Run-on experiments indicated a similar
679 transcription rate of virulence genes in infective and non-infective forms, highlighting
680 that trypanosomes employ tight post-transcriptional control to ensure high expression of
681 virulence factor genes in infective forms (8, 10, 38). Thus, we hypothesize that disruptive
682 genomic regions are silenced both in infective and non-infective forms, and parasites may
683 take advantage of the presence of virulence factors located within core PTUs. Future
684 experimental analyses should be conducted to verify the validity of this hypothesis.

685 While observing a global association between chromatin landscape and gene
686 expression is intriguing, determining whether chromatin indeed impacts transcription
687 requires examining local differential alterations, particularly at transcription start regions.
688 In this regard, our previous observations revealed varying levels of open chromatin and
689 nucleosomes in dSSRs between infective and non-infective forms (25). Focusing on non-
690 infective forms in this study, we found that PTUs associated with open dSSRs exhibit
691 higher transcription rates compared to those with closed dSSRs. This data indicates that
692 open chromatin at regions specifically associated with transcription initiation may indeed
693 play a critical role in gene expression, suggesting a mechanism of regulation at
694 transcription initiation in *T. cruzi* that contrasts with previous proposals in the field.

Until recently, the only promoter sequence described for RNA polymerase II was found at spliced leader RNA genes in trypanosomatids (69). However, the discovery of GT-rich elements and a 75 bp promoter in *T. brucei* dSSRs challenged the previous belief that there were no promoter sequences for mediating RNA polymerase II transcription of protein-coding genes (6, 31). Here, we determined that transcription start regions from core and disruptive PTUs differ greatly in their nucleotide content. Disruptive dSSRs have a higher content of G and T, with some repeats of GT, as seen in *T. brucei* (6). Guanine-rich motifs are common features of promoter regions (70). On the contrary, dSSRs from core PTUs are enriched in poly-A and polyT sequences. Poly dA-dT elements, ranging from 15 to 30 base pairs, are commonly found in the genomes of all eukaryotes, particularly in promoter regions. Studies have demonstrated the importance of these tracts in transcriptional regulation and recombination (71, 72). Regions containing homopolymeric elements (dA-dT) form more rigid and less flexible structures than conventional DNA, enabling them to resist nucleosome entanglement and providing access sites for transcription factors (73, 74). Consistent with this, dSSRs from core PTUs harbor fewer nucleosomes than those from disruptive PTUs. Collectively, these data reinforce the idea that sequence-specific elements within core and disruptive dSSRs may determine whether a region becomes enriched in nucleosomes, suggesting a putative mechanism of transcription initiation regulation. If nucleosome content is primarily determined by these cis-elements, variations among life forms should not be expected. However, we found that dSSRs from infective forms are enriched in nucleosomes compared to non-infective (26). Thus, other factors may guide the deposition of nucleosomes in these regions, contributing to regulation. In this context, we found no association between the deposition of the epigenetic marks H2B.V and 5mC in core versus disruptive regions, but an increase in 5hmC in disruptive dSSRs. 5mC and 5hmC

720 exhibit distinct distribution patterns in the genome of eukaryotes; while the former is more
721 closely linked to gene repression, the latter is predominantly found in active enhancers
722 and surrounding expressed genes (75). This may appear contradictory to our findings, as
723 we detected lower expression associated with disruptive regions; however, it is still
724 necessary to investigate the role of these cytosine modifications in trypanosomes to
725 determine if they play the same role described in other eukaryotes.

726 In summary, this study challenges the paradigm that the genome of *T. cruzi* is
727 constitutively and uniformly transcribed. Moreover, we demonstrate that even in genomes
728 organized into PTUs it is possible to implement strategies to regulate transcription, such
729 as positioning of virulent factor genes within higher transcriptional core PTUs or flanking
730 PTUs with ncDNA *loci*. The wealth of biological resources is vast, and studying
731 eukaryotes from different domains of life continues to provide surprises.

732

733 **DATA AVAILABILITY**

734 GRO-seq, MNAse-seq, FAIRE-seq data can be accessed at the Sequence Read Archive
735 (SRA) (https://www.ncbi.nlm.nih.gov/sra) with the following accession number:
736 PRJNA1073942, PRJNA665060 and PRJNA763084. Oxford Nanopore sequencing
737 datasets available at (28).

738

739 **AUTHOR CONTRIBUTIONS**

740 Conceptualization: MCE and JPCC

741 Data curation: PLCL, DP, JPCC

742 Formal Analysis PCL, LL, JNR, AB, NKB, AT, MSS, DP, JPCC

743 Funding acquisition MCE and JPCC

744 Investigation PCL, LL, JNR, AB, NKB, AT, MSS, DP, JPCC

Methodology PCL, LL, JNR, AB, NKB, AT, MSS, DP, JPCC

Project administration MCE and JPCC

Resources: MCE and JPCC

Software: PCL, DP, AT

Supervision: JPCC, DP, AT

Writing – original draft: PCL, JPCC

Writing – review & editing: PCL, LL, JNR, AB, NKB, AT, MSS, DP, MCE, JPCC

ACKNOWLEDGEMENTS

We thank Ivan Novaski Avino for technical assistance and Dr. Alex Ranieri Lima for important inputs on the data and bioinformatic analysis. We thank Herbert Guimarães de Sousa Silva, Dr. Ana Paula de Jesus Menezes and Dr. Simone Calderano for reading this manuscript and providing critical comments. We are also grateful to Dr. Sergio Verjovski de Almeida and Dr. Ariel Silber for critical suggestions.

FUNDING

This work was supported by fellowships from FAPESP and by grants (#13/07467-1, #18/15553-9, #21/03219-0, #22/15610-8, #22/15610-8) from the Sao Paulo Research Foundation (FAPESP). MCE has a fellowship from the National Council for Scientific and Technological Development (CNPq). JPCC has a fellowship from the Serrapilheira Institute (grant number Serra-1709-16865).

CONFLICT OF INTEREST

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used https://chat.openai.com/ to correct grammar, spelling, typos and improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

1. Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nature Reviews Genetics. 2019;20(12):705-23.
2. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19(10):621-37.
3. Clayton CE. Life without transcriptional control? From fly to man and back again. Embo j. 2002;21(8):1881-8.
4. Teixeira SM, daRocha WD. Control of gene expression and genetic manipulation in the Trypanosomatidae. Genet Mol Res. 2003;2(1):148-58.
5. Martínez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martínez LE, Manning-Cela RG, Figueroa-Angulo EE. Gene expression in trypanosomatid parasites. J Biomed Biotechnol. 2010;2010:525241.
6. Wedel C, Förstner KU, Derr R, Siegel TN. GT-rich promoters can drive RNA pol II transcription and deposition of H2A.Z in African trypanosomes. Embo j. 2017;36(17):2581-94.

793 7. Clayton C. Regulation of gene expression in trypanosomatids: living with
794 polycistronic transcription. Open Biol. 2019;9(6):190072.
795 8. Teixeira SM, Russell DG, Kirchhoff LV, Donelson JE. A differentially
796 expressed gene family encoding "amastin," a surface protein of Trypanosoma cruzi
797 amastigotes. J Biol Chem. 1994;269(32):20509-16.
798 9. McAndrew M, Graham S, Hartmann C, Clayton C. Testing Promoter Activity in
799 the Trypanosome Genome: Isolation of a Metacyclic-Type VSG Promoter, and
800 Unexpected Insights into RNA Polymerase II Transcription. Experimental Parasitology.
801 1998;90(1):65-76.
802 10. Abuin G, Freitas-Junior LH, Colli W, Alves MJ, Schenkman S. Expression of
803 trans-sialidase and 85-kDa glycoprotein genes in Trypanosoma cruzi is differentially
804 regulated at the post-transcriptional level by labile protein factors. J Biol Chem.
805 1999;274(19):13041-7.
806 11. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et
807 al. The genome of the kinetoplastid parasite, Leishmania major. Science.
808 2005;309(5733):436-42.
809 12. Antwi EB, Haanstra JR, Ramasamy G, Jensen B, Droll D, Rojas F, et al.
810 Integrative analysis of the Trypanosoma brucei gene expression cascade predicts
811 differential regulation of mRNA processing and unusual control of ribosomal protein
812 expression. BMC Genomics. 2016;17:306.
813 13. Millán-Zambrano G, Burton A, Bannister AJ, Schneider R. Histone post-
814 translational modifications — cause and consequence of genome function. Nature
815 Reviews Genetics. 2022;23(9):563-80.

816 14. da Cunha JP, Nakayasu ES, Elias MC, Pimenta DC, Tellez-Inon MT, Rojas F, et
817 al. *Trypanosoma cruzi* histone H1 is phosphorylated in a typical cyclin dependent
818 kinase site accordingly to the cell cycle. Mol Biochem Parasitol. 2005;140(1):75-86.
819 15. da Cunha JP, Nakayasu ES, de Almeida IC, Schenkman S. Post-translational
820 modifications of *Trypanosoma cruzi* histone H4. Mol Biochem Parasitol.
821 2006;150(2):268-77.
822 16. Siegel TN, Kawahara T, Degrasse JA, Janzen CJ, Horn D, Cross GA.
823 Acetylation of histone H4K4 is cell cycle regulated and mediated by HAT3 in
824 Trypanosoma brucei. Mol Microbiol. 2008;67(4):762-71.
825 17. Nardelli SC, da Cunha JP, Motta MC, Schenkman S. Distinct acetylation of
826 *Trypanosoma cruzi* histone H4 during cell cycle, parasite differentiation, and after DNA
827 damage. Chromosoma. 2009;118(4):487-99.
828 18. de Lima LP, Poubel SB, Yuan ZF, Rosón JN, Vitorino FNL, Holetz FB, et al.
829 Improvements on the quantitative analysis of Trypanosoma cruzi histone post
830 translational modifications: Study of changes in epigenetic marks through the parasite's
831 metacyclogenesis and life cycle. J Proteomics. 2020;225:103847.
832 19. Kraus AJ, Vanselow JT, Lamer S, Brink BG, Schlosser A, Siegel TN. Distinct
833 roles for H4 and H2A.Z acetylation in RNA transcription in African trypanosomes. Nat
834 Commun. 2020;11(1):1498.
835 20. de Almeida RF, Fernandes M, de Godoy LMF. An updated map of
836 Trypanosoma cruzi histone post-translational modifications. Sci Data. 2021;8(1):93.
837 21. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, et al.
838 Four histone variants mark the boundaries of polycistronic transcription units in
839 *Trypanosoma brucei*. Genes Dev. 2009;23(9):1063-76.

840 22. Wright JR, Siegel TN, Cross GA. Histone H3 trimethylated at lysine 4 is
841 enriched at probable transcription start sites in Trypanosoma brucei. Mol Biochem
842 Parasitol. 2010;172(2):141-4.
843 23. Schulz D, Zaringhalam M, Papavasiliou FN, Kim HS. Base J and H3.V Regulate
844 Transcriptional Termination in Trypanosoma brucei. PLoS Genet.
845 2016;12(1):e1005762.
846 24. Rosón JN, Vitarelli MO, Costa-Silva HM, Pereira KS, Pires DDS, Lopes LS, et
847 al. H2B.V demarcates divergent strand-switch regions, some tDNA loci, and genome
848 compartments in Trypanosoma cruzi and affects parasite differentiation and host cell
849 invasion. PLoS Pathog. 2022;18(2):e1009694.
850 25. Lima ARJ, Silva HGS, Poubel S, Rosón JN, de Lima LPO, Costa-Silva HM, et
851 al. Open chromatin analysis in Trypanosoma cruzi life forms highlights critical
852 differences in genomic compartments and developmental regulation at tDNA loci.
853 Epigenetics Chromatin. 2022;15(1):22.
854 26. Lima ARJ, de Araujo CB, Bispo S, Patané J, Silber AM, Elias MC, da Cunha
855 JPC. Nucleosome landscape reflects phenotypic differences in Trypanosoma cruzi life
856 forms. PLoS Pathog. 2021;17(1):e1009272.
857 27. Müller LSM, Cosentino RO, Förstner KU, Guizetti J, Wedel C, Kaplan N, et al.
858 Genome organization and DNA accessibility control antigenic variation in
859 trypanosomes. Nature. 2018;563(7729):121-5.
860 28. Díaz-Viraqué F, Chiribao ML, Libisch MG, Robello C. Genome-wide chromatin
861 interaction map for Trypanosoma cruzi. Nat Microbiol. 2023;8(11):2103-14.
862 29. Elias MC, Marques-Porto R, Freymuller E, Schenkman S. Transcription rate
863 modulation through the *Trypanosoma cruzi* life cycle occurs in parallel with changes in
864 nuclear organisation. Mol Biochem Parasitol. 2001;112(1):79-90.

865 30. Ferreira LR, Dossin Fde M, Ramos TC, Freymüller E, Schenkman S. Active
866 transcription and ultrastructural changes during Trypanosoma cruzi metacyclogenesis.
867 An Acad Bras Cienc. 2008;80(1):157-66.
868 31. Cordon-Obras C, Gomez-Liñan C, Torres-Rusillo S, Vidal-Cobo I, Lopez-
869 Farfan D, Barroso-Del Jesus A, et al. Identification of sequence-specific promoters
870 driving polycistronic transcription initiation by RNA polymerase II in trypanosomes.
871 Cell Rep. 2022;38(2):110221.
872 32. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et
873 al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.
874 Science. 2005;309(5733):409-15.
875 33. Pita S, Díaz-Viraqué F, Iraola G, Robello C. The Tritryps Comparative
876 Repeatome: Insights on Repetitive Element Evolution in Trypanosomatid Pathogens.
877 Genome Biol Evol. 2019;11(2):546-51.
878 34. Horn D. The molecular control of antigenic variation in Trypanosoma brucei.
879 Curr Mol Med. 2004;4(6):563-76.
880 35. Sima N, McLaughlin EJ, Hutchinson S, Glover L. Escaping the immune system
881 by DNA repair and recombination in African trypanosomes. Open Biol.
882 2019;9(11):190182.
883 36. Faria J, Luzak V, Müller LSM, Brink BG, Hutchinson S, Glover L, et al. Spatial
884 integration of transcription and splicing in a dedicated compartment sustains monogenic
885 antigen expression in African trypanosomes. Nat Microbiol. 2021;6(3):289-300.
886 37. Kooter JM, Borst P. Alpha-amanitin-insensitive transcription of variant surface
887 glycoprotein genes provides further evidence for discontinuous transcription in
888 trypanosomes. Nucleic Acids Res. 1984;12(24):9457-72.

889 38. Gentil LG, Cordero EM, do Carmo MS, dos Santos MR, da Silveira JF.
890 Posttranscriptional mechanisms involved in the control of expression of the stage-
891 specific GP82 surface glycoprotein in Trypanosoma cruzi. Acta Trop. 2009;109(2):152-
892 8.
893 39. Ekanayake DK, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R,
894 et al. Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-
895 linked thymine glucosylation of DNA. Mol Cell Biol. 2011;31(8):1690-700.
896 40. Gardini A. Global Run-On Sequencing (GRO-Seq). Methods Mol Biol.
897 2017;1468:111-20.
898 41. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and
899 quality control. F1000Res. 2018;7:1338.
900 42. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
901 sequence data. Bioinformatics. 2014;30(15):2114-20.
902 43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat
903 Methods. 2012;9(4):357-9.
904 44. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-
905 seq quantification. Nat Biotechnol. 2016;34(5):525-7.
906 45. Smircich P, Eastman G, Bispo S, Duhagon MA, Guerra-Slompo EP, Garat B, et
907 al. Ribosome profiling reveals translation control as a key mechanism generating
908 differential gene expression in Trypanosoma cruzi. BMC Genomics. 2015;16(1):443.
909 46. Berná L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, et al.
910 Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. Microb
911 Genom. 2018;4(5).

912 47. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible
913 platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(Web Server
914 issue):W187-91.
915 48. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer
916 (IGV): high-performance genomics data visualization and exploration. Brief Bioinform.
917 2013;14(2):178-92.
918 49. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to
919 discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36.
920 50. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, Rohs R.
921 TFBSshape: a motif database for DNA shape features of transcription factor binding
922 sites. Nucleic Acids Res. 2014;42(Database issue):D148-55.
923 51. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity
924 between motifs. Genome Biol. 2007;8(2):R24.
925 52. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
926 2018;34(18):3094-100.
927 53. Pongor LS, Gross JM, Vera Alvarez R, Murai J, Jang S-M, Zhang H, et al.
928 BAMscale: quantification of next-generation sequencing peaks and generation of scaled
929 coverage tracks. Epigenetics & Chromatin. 2020;13(1):21.
930 54. Berná L, Chiribao ML, Greif G, Rodriguez M, Alvarez-Valin F, Robello C.
931 Transcriptomic analysis reveals metabolic switches and surface remodeling as key
932 processes for stage transition in Trypanosoma cruzi. PeerJ. 2017;5:e3017.
933 55. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al.
934 Comparative genomics of trypanosomatid parasitic protozoa. Science.
935 2005;309(5733):404-9.

936 56. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome.
937 Epigenetics & Chromatin. 2014;7(1):33.
938 57. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ.
939 Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions
940 within a single region. Mol Cell. 2003;11(5):1291-9.
941 58. Respuela P, Ferella M, Rada-Iglesias A, Aslund L. Histone acetylation and
942 methylation at sites initiating divergent polycistronic transcription in *Trypanosoma*
943 *cruzi*. J Biol Chem. 2008;283(23):15884-92.
944 59. Maree JP, Tvardovskiy A, Ravnsborg T, Jensen ON, Rudenko G, Patterton H-G.
945 Trypanosoma brucei histones are heavily modified with combinatorial post-translational
946 modifications and mark Pol II transcription start regions with hyperacetylated H2A.
947 Nucleic Acids Research. 2022;50(17):9705-23.
948 60. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread
949 pausing and divergent initiation at human promoters. Science. 2008;322(5909):1845-8.
950 61. Fadda A, Ryten M, Droll D, Rojas F, Färber V, Haanstra JR, et al.
951 Transcriptome-wide analysis of trypanosome mRNA decay reveals complex
952 degradation kinetics and suggests a role for co-transcriptional degradation in
953 determining mRNA levels. Mol Microbiol. 2014;94(2):307-26.
954 62. Maraia RJ, Arimbasseri AG. It's Sno'ing on Pol III at nuclear pores. Genome
955 Biol. 2013;14(10):137.
956 63. Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, et al. Close
957 association of RNA polymerase II and many transcription factors with Pol III genes.
958 Proc Natl Acad Sci U S A. 2010;107(8):3639-44.
959 64. Vannini A, Cramer P. Conservation between the RNA polymerase I, II, and III
960 transcription initiation machineries. Mol Cell. 2012;45(4):439-46.

961 65. Van Bortle K, Corces VG. tDNA insulators and the emerging role of TFIIIC in
962 genome organization. Transcription. 2012;3(6):277-84.
963 66. Horn D. Antigenic variation in African trypanosomes. Mol Biochem Parasitol.
964 2014;195(2):123-9.
965 67. Kelly JM, Taylor MC, Horn D, Loza E, Kalvinsh I, Bjorkling F. Inhibitors of
966 human histone deacetylase with potent activity against the African trypanosome
967 Trypanosoma brucei. Bioorg Med Chem Lett. 2012;22(5):1886-90.
968 68. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the
969 regulatory epigenome. Nature Reviews Genetics. 2019;20(4):207-20.
970 69. Gilinger G, Bellofatto V. Trypanosome spliced leader RNA genes contain the
971 first identified RNA polymerase II gene promoter in these organisms. Nucleic Acids
972 Res. 2001;29(7):1556-64.
973 70. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the
974 human genome. Nucleic Acids Res. 2007;35(2):406-13.
975 71. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates
976 transcription via its intrinsic DNA structure. Embo j. 1995;14(11):2570-9.
977 72. Shimizu M, Mori T, Sakurai T, Shindo H. Destabilization of nucleosomes by an
978 unusual DNA conformation adopted by poly(dA) small middle dotpoly(dT) tracts in
979 vivo. Embo j. 2000;19(13):3358-65.
980 73. Koch KA, Thiele DJ. Functional analysis of a homopolymeric (dA-dT) element
981 that provides nucleosomal access to yeast and mammalian transcription factors. J Biol
982 Chem. 1999;274(34):23752-60.
983 74. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ.
984 Genome-scale identification of nucleosome positions in S. cerevisiae. Science.
985 2005;309(5734):626-30.

75. Song CX, Szulwach KE, Fu Y, Dai Q, Yi C, Li X, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol. 2011;29(1):68-72.

Supporting Information

Supporting Legends

S1 Fig. GRO-seq Assays: Comparison of Nascent and Mature Transcriptomes. A. Fluorescence microscopy images showing nascent RNA labeled with Br-UTP in the nuclear space of *T. cruzi* epimastigote forms. **B.** Agarose gel displaying RNA integrity extracted from parasites subjected to transcription assays with or without Br-UTP labeling. **C.** Summary plots of read coverage (median values) from coding DNA sequence (CDS) regions of three biological replicates (rep1, rep2, and rep3) of GRO-seq (top) (GRO/GRO bckg) and RNA-seq (bottom) [45], assays of *T. cruzi* epimastigote forms. Median coverage values were scaled individually for each replicate. **D.** Read coverage of three biological replicates of GRO-seq (left) (GRO/GRO bckg) and RNA-seq (right) assays of core and disruptive genomic compartments. **E.** TPM counts from CDSs and intergenic regions (IR) of GRO and RNA-seq assays. Median values are depicted above each boxplot. Median coverage values were scaled individually for each replicate. **F.** Summary plots of read coverage (median values) from the spliced leader locus of one representative GRO and RNA-seq sample. Yellow lines and boxes represent results from GRO-seq assays; blue lines and boxes from RNA-seq assays. The start (S – first ATG) and end (E – stop codon) of CDS are shown, along with 1 Kb upstream and downstream. **S2 Fig.A.** Scheme of the bioinformatic pipeline used to retrieve coverage and TPM counts from GRO and RNA-seq datasets. **B**. Scheme of two GFF files obtained to cover all

genomic features, considering either the genome classified by PTUs or by all CDSs. **C.** Distribution of genomic features considering GFF files from all features (37,959 regions) in the *T. cruzi* genome (Dm28c strain). IR stands for Intergenic Region; IPR - Intergenic polycistronic region is located between a polycistron and a locus of non-coding RNA.

S3 Fig. A. Top: Boxplots of TPM counts from PTUs of the first three major contigs (PRFA:01000001-01000003). Individual CDS TPM values for each biological replicate (GRO-seq) are plotted for each PTU. **B**. Left: Violin plot of the TPM distribution from individual CDSs of the core (conserved proteins and conserved hypothetical proteins), disruptive (mucin, MASP, and trans-sialidase), and both (GP63, RHS, DGF-1) compartments. Right: Confidence interval graph from the one-way ANOVA with Tukey's range test as post-hoc (95% confidence). **** $p<0.0001$. **C**. Distribution of TPM values in core, disruptive, and both compartments based on MAPQ filters. D. IGV snapshot of GRO and RNA-seq coverages showing that disruptive regions (ref rectangules) are covered by transcripts in trypomastigote forms (TRYPO).

S4 Fig. A. Confidence interval graph of the one-way ANOVA with Tukey post-hoc (95% confidence) from Figure 3A. **B.** Violin plots of TPM distribution from PTUs containing 80% of core genes classified by the number of CDS per PTU. PTUs TPM values were based on the median TPM values from individual CDSs from each core PTUs with one (162 PTUs), 2-4 (128 PTUs), or 4 or more CDSs (369 PTUs). PTU length in bp classified by genomic compartment (**C**) and by compartment and number of CDS (**D**).

S5 Fig. A. Scatter plots illustrating the nucleosome occupancy (MNase-seq dataset, RPGC levels) profile versus the nascent (GRO-seq) transcriptome for coding DNA regions classified as core, disruptive, and both regions. R values are presented. **B**. R values of the open chromatin profile and nascent transcription obtained through

downsampling analysis of a randomly selected 100 genes (for the indicated compartment) repeated 50 times.

S6 Fig. dSSRs size distribution according to chromatin openness. The first 20 contigs of *T. cruzi* strain Dm28c were examined for length of dSSRs located between PTUs within the same cluster. We identified 11 dSSRs between PTUs belonging to cluster 1 (representing regions with more open chromatin), 51 dSSRs between PTUs in cluster 2, and 11 dSSRs between PTUs in cluster 3 (indicating regions with more closed chromatin). Statistical analysis was performed using one-way ANOVA with Tukey's range test as post-hoc (adjusted p-value: 0.0034).

Supporting Tables

S1 Table. Total and mapped reads

S2 Table. Raw and processed data of all genomic features classified by their genomic compartment and CDSs per PTU

S3 Table.TPM values all PTUs classified by their genomic compartment, number of CDSs per PTU, presence of adjacent ncDNA

S4 Table. Genome features with undectable levels of nascent expression (GROexp-GRObck<=0)

S5 Table. FPKM values of epigenetic marks and nucleosome occupancy in epimastigote forms from all dSSRs classified by genomic compartment and number of CDSs per PTU

S6 Table. Abundance values from 5mC (m) and 5hmC (h) in dSSRs from epimastigote forms classified by genomic compartment and strand

FIGURES LEGENDS

Fig 1. The Distribution and Abundance of Nascent Transcripts Across the *T. cruzi* Genome. A. Percentage of *T. cruzi* Genomic Features Exhibiting No Nascent Transcription Expression (GROexp – GRO bckg ≤ 0). **B**. Bar plots representing the distribution of genomic features in GROexp – GRO bckg ≤ 0 (top) compared to the expected genome distribution (bottom). **C**. Distribution of Transcripts Per Million (TPM) Counts (log2(TPM +1)) across all Genomic Features (37,959 regions) from GRO and RNA-seq Assays. Coefficients of Variations are depicted above each violin plot. Outliers were removed (Q=1%).

Fig 2. Assessment of Nascent Expression in PTUs Reveals Greater Abundance in Core PTUs Compared to Disruptive PTUs. A. Distribution of TPM Counts (log2(GRO-GRObckg +1)) across all PTUs (1218) from GRO and RNA-seq assays. **B.** Boxplots of TPM Counts from PTUs on the first three major contigs (PRFA:01000001-01000003). Individual TPM Values (averaged among biological replicates) for each CDS are plotted for each PTUs. **C.** Violin plots of TPM Distribution from PTUs Containing 80% of Core, Disruptive, or Both Genes, respectively. **D.** IGV snapshot showing read coverage (BPM) from three replicates of GRO along two *T. cruzi* contigs (PRF01000009 and PRF010000011). Polycistrons are depicted in blue rectangles, while green and red rectangles indicate CDSs from Core and Disruptive compartments, respectively.

Fig 3. Nascent Transcription of Core PTUs Depends on the Number of CDSs and the Proximity of ncDNA Loci. A. Violin plots illustrating the TPM distribution from PTUs containing 80% of core (left) or disruptive (right) genes, categorized by the number

of CDSs per PTU. Core PTUs are divided into those with one (162 PTUs), 2-4 (128 PTUs), or 4 or more CDSs (369 PTUs). Disruptive PTUs are categorized similarly with one (40 PTUs), 2-4 (68 PTUs), or 4 or more CDSs (65 PTUs). Statistical analysis was performed using one-way ANOVA with Tukey's range test at 95% confidence level (**** $p<0.0001$**). B.** Percentage of pairwise PTU comparisons that are statistically significant, classified by genomic compartments and CDS number per PTU using one-way ANOVA with Tukey's range test at 95% confidence level. **C.** TPM counts from four core PTUs with 4 or more CDSs located on the same contig. **D.** Abundance values from PTUs not adjacent to any ncDNA (none) or adjacent to a tDNA or snoDNA locus. Statistical analysis was conducted using the Kruskal-Wallis test (* $p<0.01$). E. TPM counts of seven canonical PTUs (#59 - contig 2 - 100 CDSs, #81 - contig 2 - 56 CDSs, #110 - contig 4 - 129 CDSs, #111 - contig 4 - 74 CDSs, #112 - contig 4 - 59 CDSs, #173 - contig 9 - 19 CDSs, #181 - contig 10 - 57 CDSs). TPM values for each PTU were obtained considering the entire PTU coordinates. No statistical significance was observed.

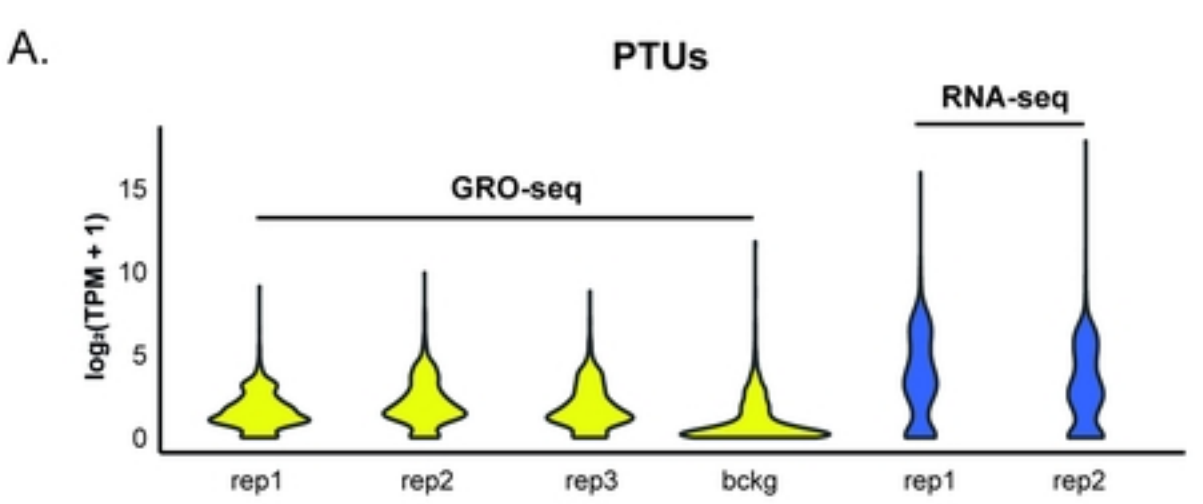
Fig 4. Nascent Transcription Expression Depends on the PTU Context. A. Schematic representation of four PTU categories based on the content (at least 80%) of core (PTU core), disruptive (PTU disruptive), both (PTU both) genes, and mixed PTUs composed of core, disruptive, or both CDSs with less than 80% representation. The number of PTUs in each category is shown. Arrows indicate the transcription direction of PTUs. **B.** Bar plots of TPM counts from individual CDSs classified as core, disruptive, or both, from PTUs with more than 80% composition of core, disruptive, and both CDSs, respectively. **C.** Similar analysis shown in B for trans-sialidases, mucin, and MASP CDSs. Statistical analysis was conducted using one-way ANOVA with Tukey's range test as post-hoc

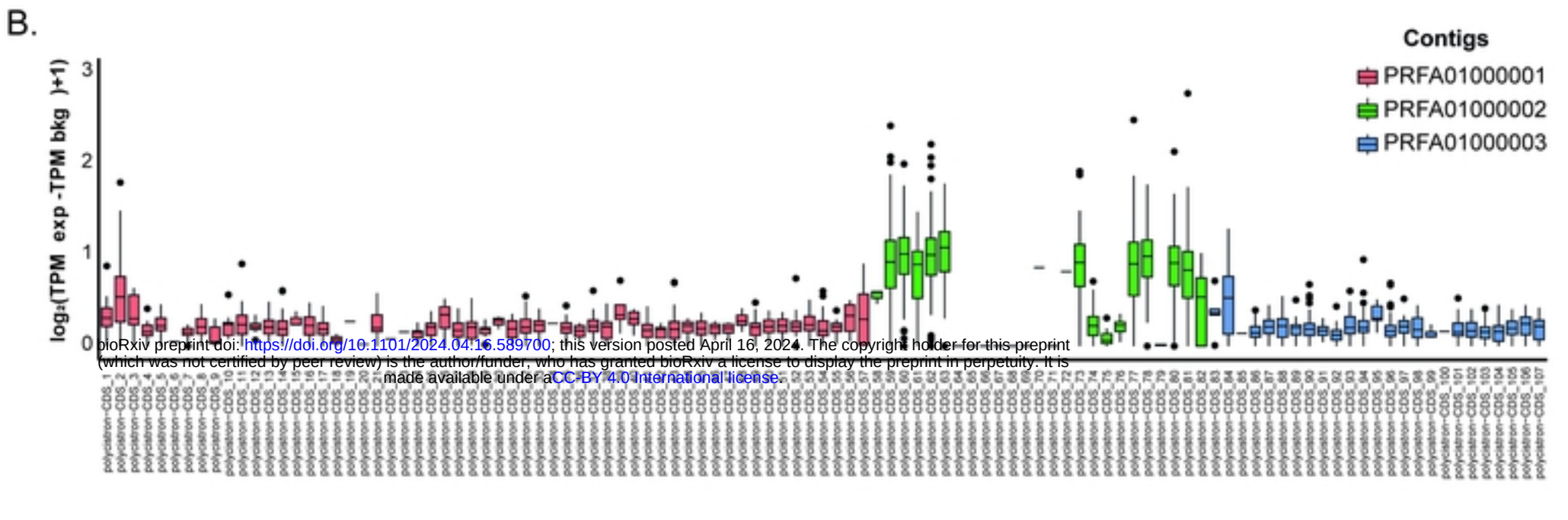
(p<0.0001). **D.** Distribution of trans-sialidase, mucin, MASP, and pseudogenes in core and disruptive PTUs (chi-square test with a p-adjusted < 0.0001). ** for adjusted p-value < 0.001, *** for adjusted p-value < 0.0001 (chi-square test). **E.** Percentage of differentially expressed disruptive genes (Epimastigotes versus TCTs, p-adjusted <0.01, fold change of 2, based on (28)) located in core or disruptive PTUs.

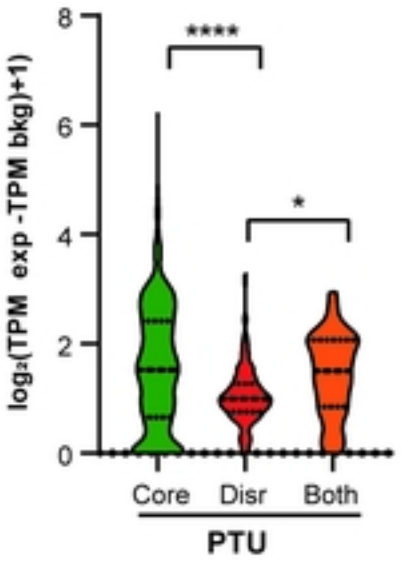
Fig 5. Global Open Chromatin Status at Core Region Correlates with Nascent Transcript Levels. A. Scatter plots of open chromatin (FAIRE-seq dataset, RPGC levels) profiles versus nascent (GRO-seq) or mature (RNA-seq) transcriptomes for all *T. cruzi* genomic features. R values are indicated. **B.** Bar plots illustrating TPM counts of genomic regions based on their open chromatin levels (high - 25% highest RPGC levels, low - 25% lowest RPGC levels, medium – the 50% remaining). Outliers were removed using the ROUT method, Q=1% at PRISMA. **C.** Scatter plots of open chromatin (FAIRE-seq dataset, RPGC levels) profiles versus nascent (GRO-seq) transcripts for coding DNA regions classified as core, disruptive, and both regions. R values are displayed. **D.** R values of open chromatin profile and nascent transcription obtained through downsampling analysis of a randomly selected 100 genes (for the indicated compartment) repeated 50 times. Statistical analysis was performed using one-way ANOVA with Tukey's range test as post-hoc (****p<0.0001).

Figure 6. Core and Disruptive dSSRs Differ in Their Epigenetic Landscape and Sequence Motifs. A. Left: Hierarchical cluster analysis of the distribution of the RPGC log2 ratio in PTUs from epimastigotes considering 1 Kb upstream or downstream, highlighting the dSSRs. Right: Bar plots depicting TPM levels from PTUs associated with dSSRs that harbor different levels of open chromatin as shown on the left. PTU abundance was based on median TPM values from CDSs of each PTU. Wilcoxon test, p.adj <0.05**. B.** Distribution of core, disruptive, both, and

mixed PTUs associated with open and closed dSSRs as depicted in A. (chi-square test under a p-adjusted < 0.01). * for adjusted p-value < 0.0125, ** for adjusted p-value < 0.00125, *** for adjusted p-value < 0.000125 (chi-square test). **C.** Sequence motifs determined by the MEME platform using dSSRs from either core or disruptive PTUs. Only dSSRs between PTUs from the same compartment were analysed. The top 3 sequence motifs are shown. E-values were higher than $1.5e^{-076}$ **D.** Distribution of abundance levels for 5hmC and 5mC (Nmod/Nvalid_cov, bedgraph score of coverage), H2.BV (ChIP/input) deposition (average raw counts), and nucleosome levels (MNase-seq) (FPKM) for core and disruptive dSSRs. Mann-Whitney test, ** p.adj <0.01. Only dSSRs between PTUs from the same compartment were used in C and D analysis.







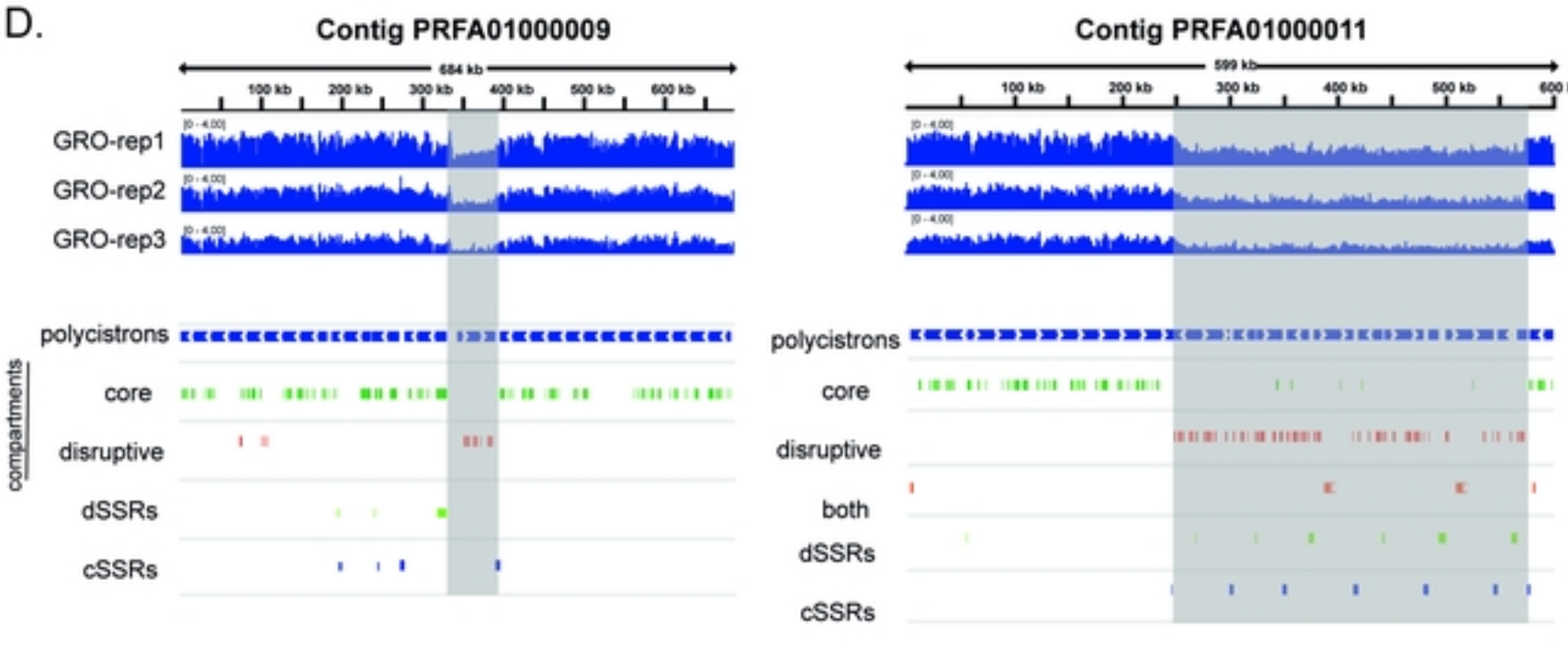


Fig2

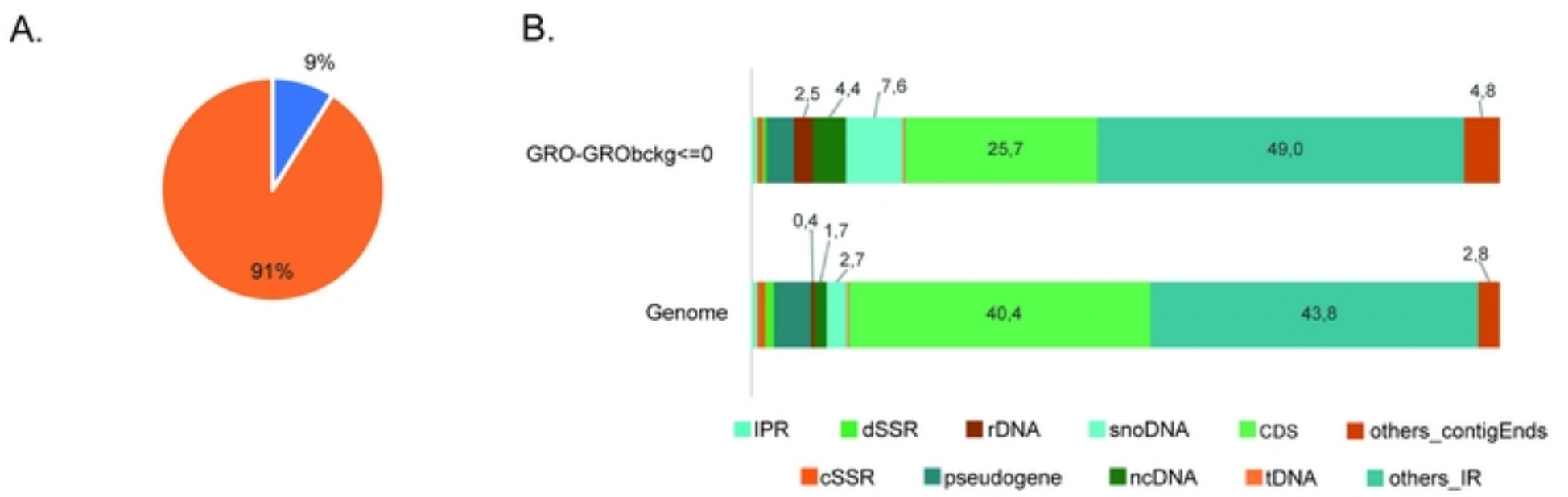
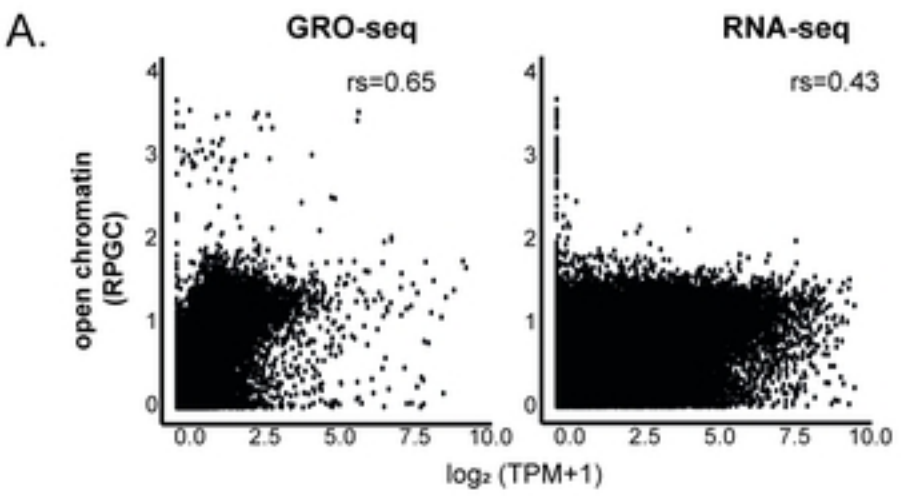
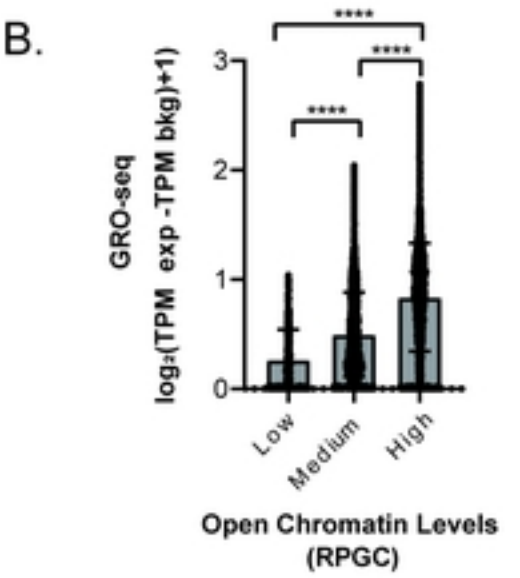
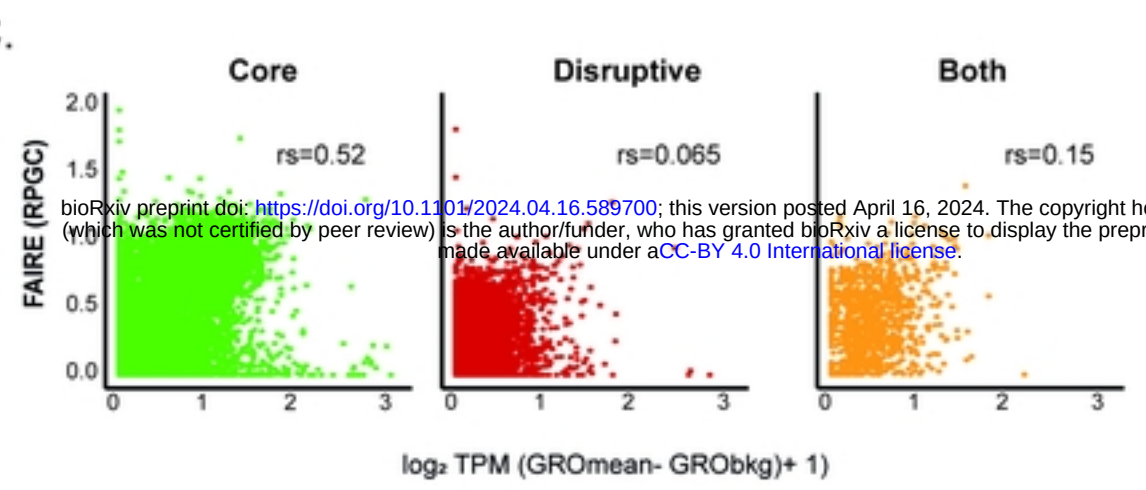


Fig1







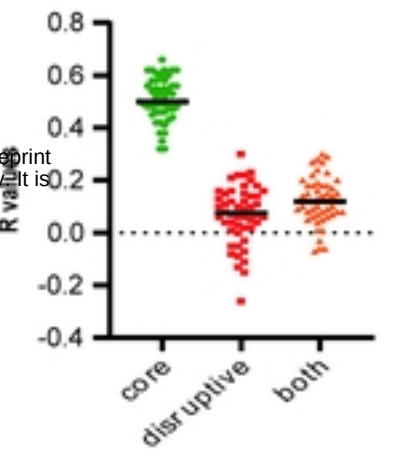


Fig5

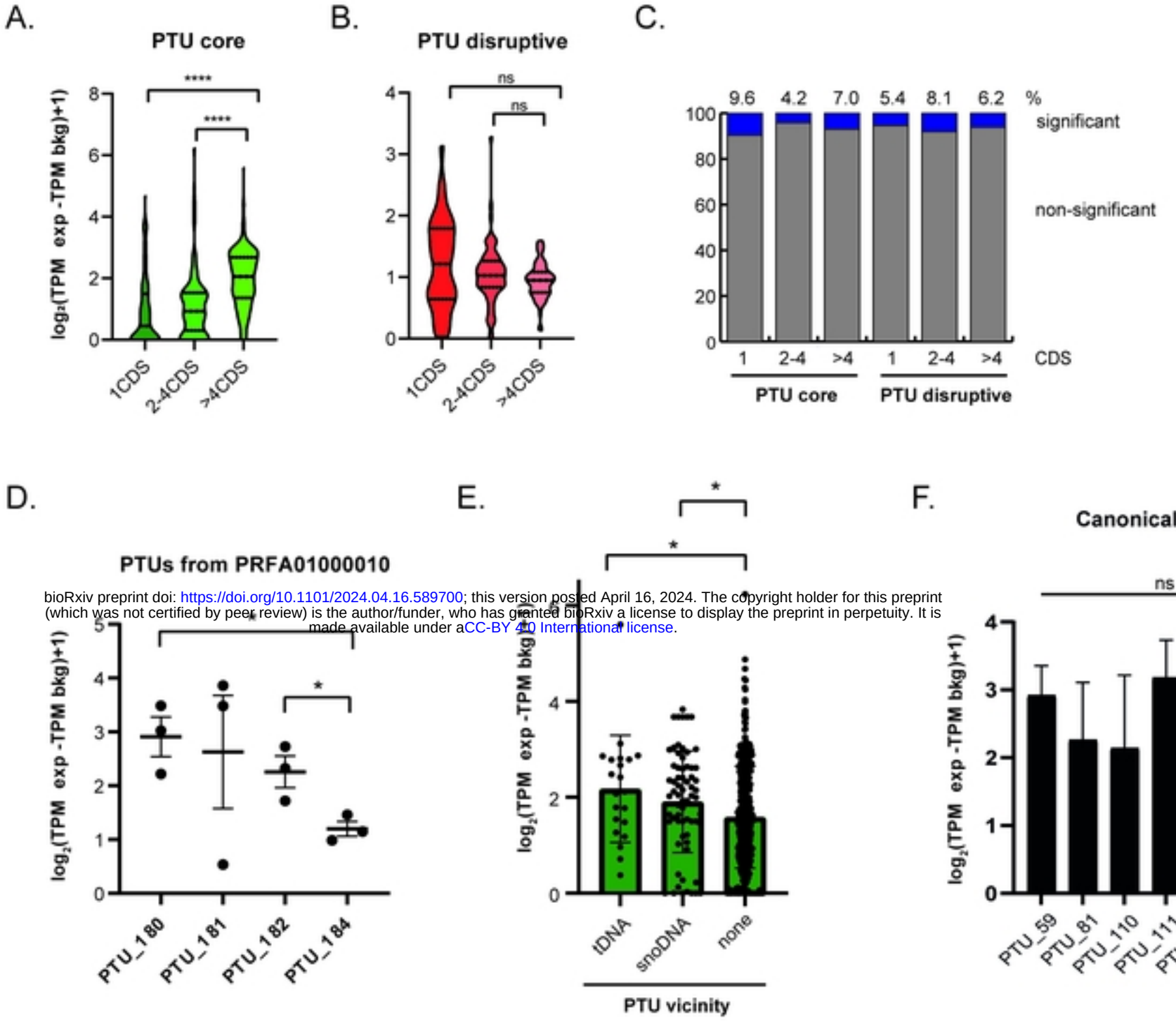


Fig3

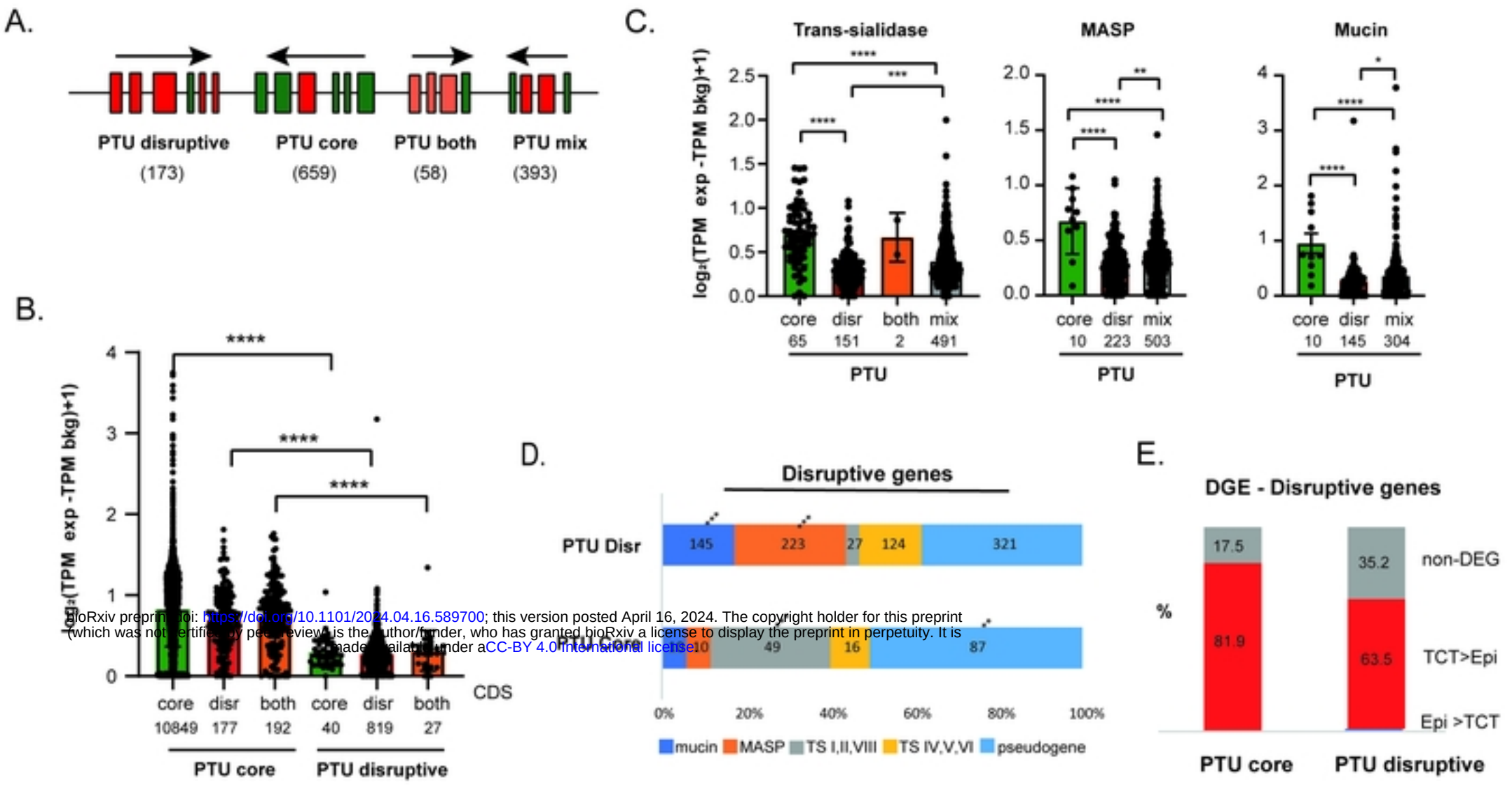
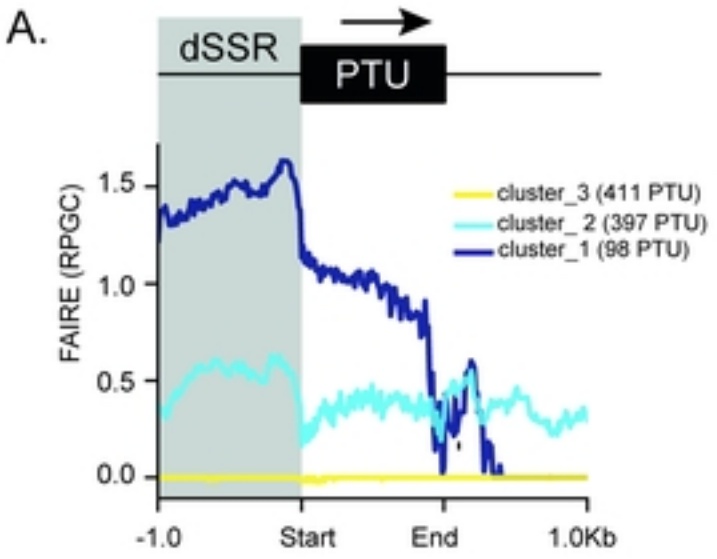
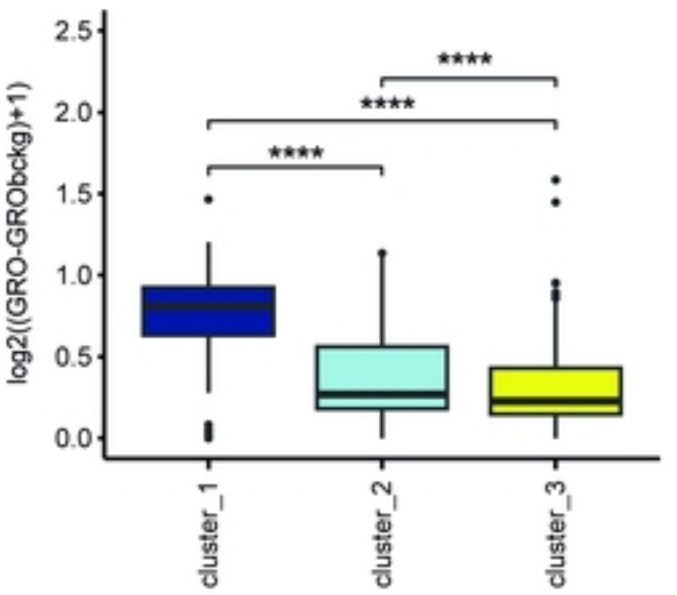
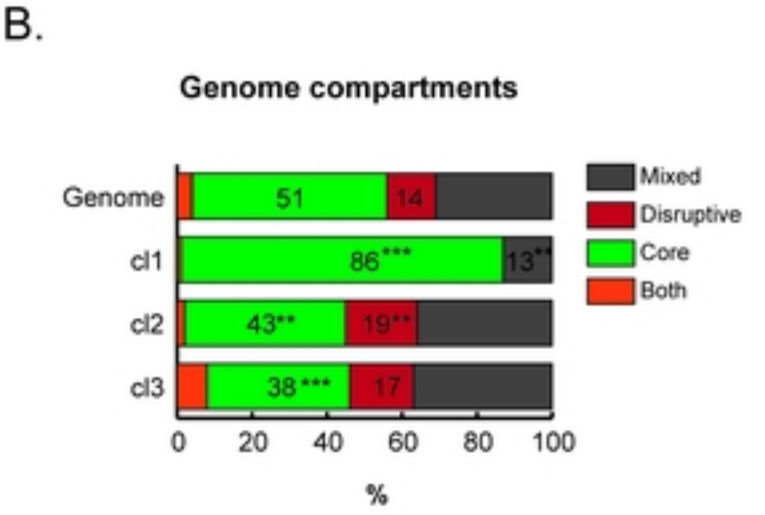
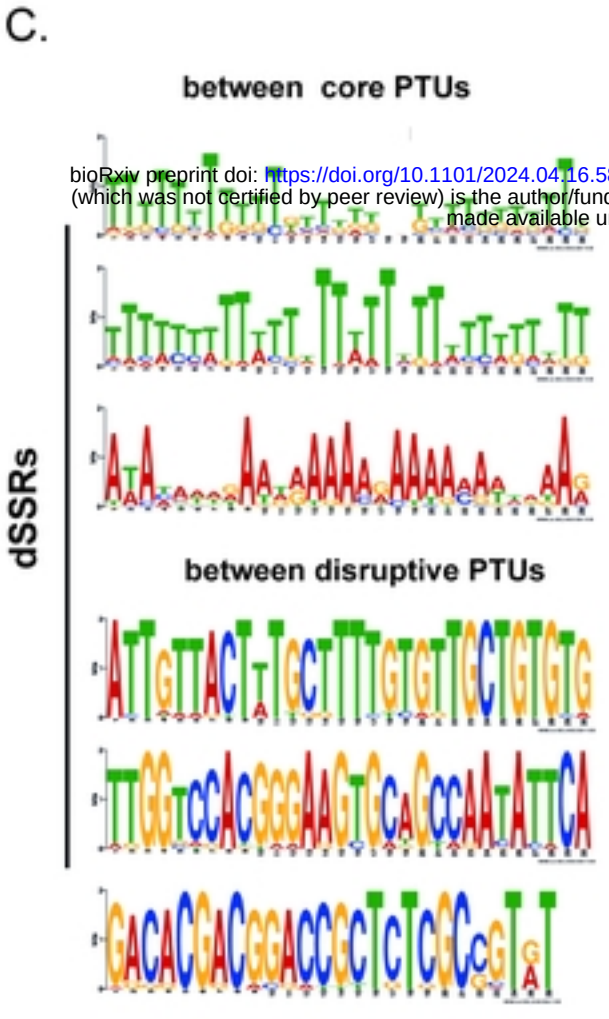


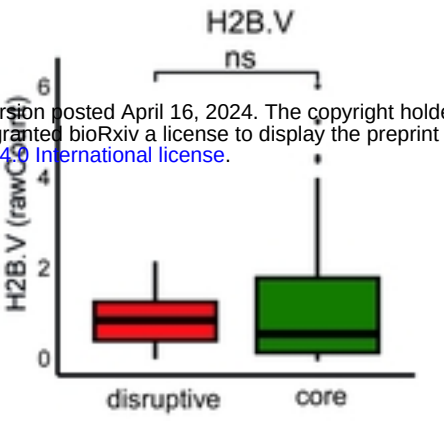
Fig4

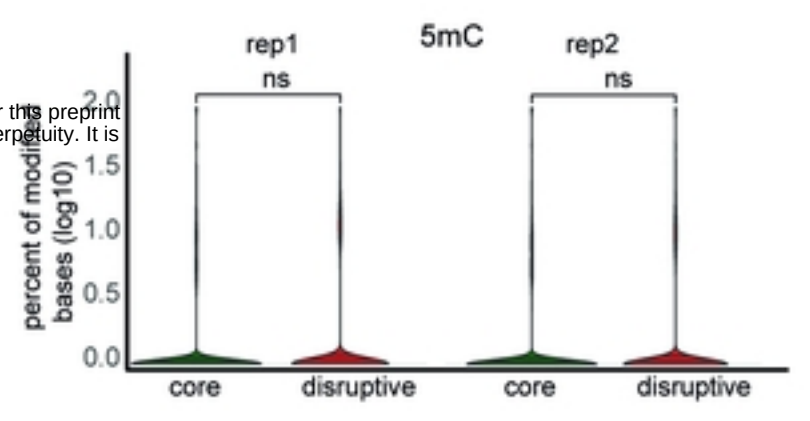


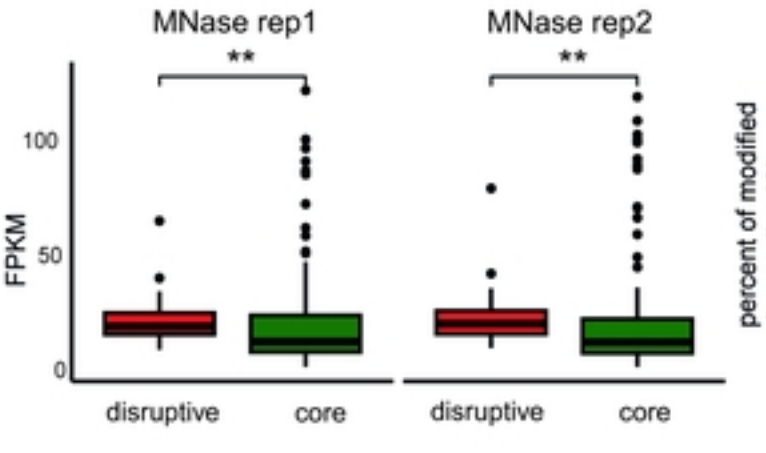












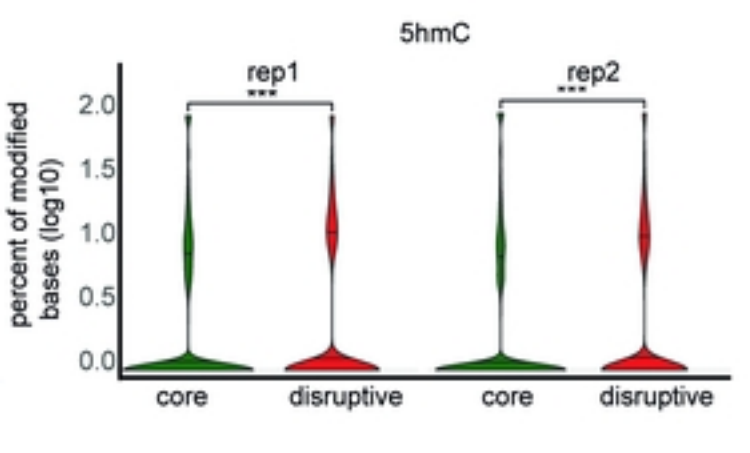


Fig6